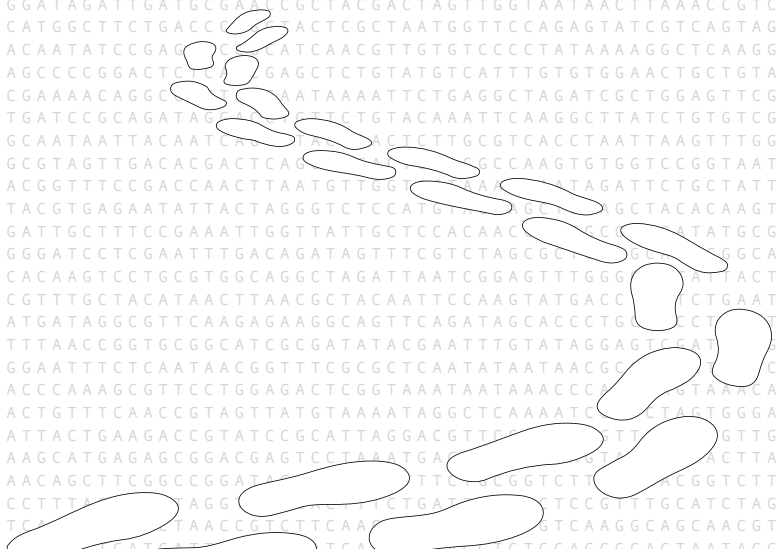
Understanding and diagnosing cancer through its mutational history

Luan Nguyen



Understanding and diagnosing cancer through its mutational history

Luan Ngoc Nguyen

ISBN: 978-94-6423-953-9

Design and layout by: Luan Nguyen
Printed by: ProefschriftMaken // www.proefschriftmaken.nl

Understanding and diagnosing cancer through its mutational history

Kanker begrijpen en diagnosticeren door zijn mutatiegeschiedenis
(met een samenvatting in het Nederlands)

Proefschrift

ter verkrijging van de graad van doctor aan de
Universiteit Utrecht
op gezag van de
rector magnificus, prof. dr. H.R.B.M. Kummeling,
ingevolge het besluit van het college voor promoties
in het openbaar te verdedigen op

woensdag 30 november 2022 des middags te 12:15 uur

door

Luan Ngoc Nguyen
geboren op 5 januari 1992
te Ho Chi Minh Stad, Vietnam

Promotiecommissie

Promoter
Prof. dr. E.P.J.G. Cuppen

Co-promoter
Dr. A. Van Hoeck

Overige leden
Dr. R. van Boxtel
Dr. J. de Ridder

Contents

Chapter 1

General introduction

Studying cancer genomes

The human genome contains the instructions to build, maintain and reproduce a human organism, and is encoded in deoxyribonucleic acid (DNA). DNA is a polymer of nucleotides (adenine (A), cytosine (C), thymine (T) and guanine (G)) that coils together to form a double helix. The average human genome consists of roughly 6 billion nucleotides that are organized over 23 pairs of chromosomes. About 1-2% of the genome codes for the ~20,000 genes, most of which are eventually transcribed into mRNA and subsequently translated into proteins. The remainder of the genome (at least in part) plays an important role in regulating the expression of genes [1].

In normal cells, growth and division (together called cell proliferation) is tightly controlled by hundreds of genes. However, mutations in these genes (or its regulatory elements) can lead to the outgrowth of cells that are able to proliferate uncontrollably and avoid cell death; in other words, cancer. Cancerous cells may eventually escape their original environment and colonize other tissues (i.e. metastasize), which can be driven by mutations but also other factors such as environmental pressures [2]. The mutations that contribute to cancer are primarily those acquired over the lifetime of an individual (somatic mutations), though inherited germline mutations can also increase cancer risk [3].

DNA sequencing is commonly used to study mutations in cancer. The first generation of sequencing technology (Sanger sequencing) was used to determine the full sequence of the human genome, but was limited to sequencing one DNA fragment of up to 1,000 bases at a time. The next-generation sequencing (NGS) technologies that emerged shortly after allowed for massively parallel sequencing (thousands to millions) of multiple short DNA fragments (tens to hundreds of bases). These next generation 'short-read' sequencing techniques typically involve breaking the DNA of a sample into random short fragments that are amplified and then sequenced to produce 'reads'. Mapping assembly is then performed, whereby reads are compared ('mapped') to a reference genome and pieced together to form a continuous genomic sequence of the sample which allows for the detection of genetic differences (mutations) [4]. NGS technology enabled fast and cost effective whole genome sequencing (WGS), which has recently led to large scale pan-cancer studies involving thousands of cancer patients [5,6].

Typically, to study the full spectrum of mutations in cancer patients, a tumor and normal sample (commonly blood or nearby healthy tissue) are both sequenced and mapped to a human reference genome. Important to note is that the human reference genome is an aggregate of the full genome sequences across multiple donors and thus does not represent the genome of one individual [7]. Changes in nucleotide sequences between the sample and reference genome are considered 'variants'. Variants found in the normal sample are considered germline mutations, and these are subtracted from those found in the tumor to obtain the somatic mutations specific to the tumor. Tumor genomes typically have thousands to tens of thousands of somatic mutations. On average only 4-5 of these are 'driver' mutations, which provide a growth advantage to cells and drive cancer development [5]. Driver mutations define the characteristics of a tumor, and are often used to guide treatment decisions [8]. The remainder of mutations are considered 'passenger' mutations that do not contribute to cancer development. Passenger mutations do however reflect the mutations history in tumors, and as such can be used to understand cancer development, but also as cancer diagnostic markers [9]. The following sections will provide a detailed overview of how mutations have been utilized to study cancer as well as to develop tools for cancer diagnostics.

Functional impact of mutations on cancer driver genes

Cancer driver genes are genes in which mutations can confer a growth advantage to cancer cells, and broadly fall under two categories. Oncogenes normally stimulate cell proliferation, and can be activated when a mutation prevents the gene from being 'turned off'. Tumor suppressors genes (TSG)

normally regulate cell proliferation, and are inactivated typically by two mutations that render both gene copies (in a diploid genome) defective [10]. A very small number of genes, such as TP53, can act both as oncogenes and TSGs [11]. The various mutational mechanisms by which oncogene activation and TSG inactivation can occur will be described below.

Single and multi-base substitutions

The most common mutations in the cancer genomes are point mutations (also known as single base substitutions; SBS), followed by insertions/deletions of multiple nucleotides (indels), and multi-base substitutions (MBS) which are defined as clusters of two or more nearby variants on the same haplotype [5,12]. These 'small mutations' (as opposed to structural variants, which will be discussed in a later section) are caused by diverse mechanisms, from environmental causes such as ultraviolet light exposure [13], to endogenous causes such as errors during DNA replication [14] or repair [15].

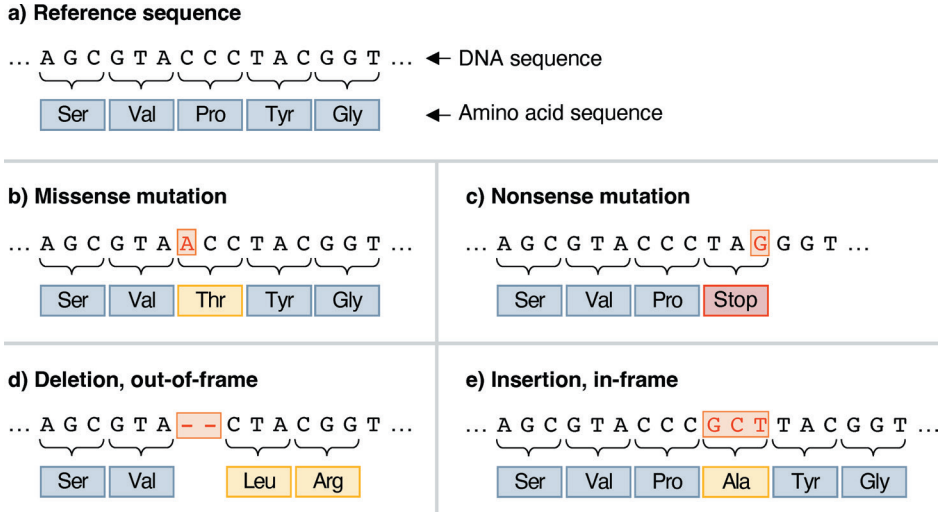


Figure 1: Examples of the effects of small mutations on protein coding sequences. **a)** Alterations to the DNA sequence with respect to a reference genome are considered mutations. **b)** Missense mutations result in the change of one amino acid to another. **c)** Nonsense mutations introduce a premature stop codon. **d)** An out-of-frame deletion alters the downstream amino acid sequence. **e)** An in-frame insertion does not alter the triplet reading frame, but does introduce additional amino acids.

SBSs and MBSs can lead to 'non-synonymous' base substitutions which alter the amino acid sequence of a protein, with missense and nonsense mutations being the main types [12,16]. Missense mutations lead to a change of a single amino acid into another (**Figure 1b**), which can lead to gene activation (e.g. *BRAF* V600E [17]) but also inactivation (e.g. *KRAS* G12D [18]). Nonsense mutations introduce a premature stop codon which leads to early termination of transcription and translation (**Figure 1c**), resulting in a truncated protein. Nonsense mutations located earlier in the protein coding sequence result in a more truncated protein and are thus more likely to lead to a defective protein [16]. Indels can lead to protein loss of function via out-of-frame frameshift mutations. Indels with lengths not divisible by 3 alter the triplet reading frame of the protein coding sequence and result in incorrect incorporation of amino acids into the eventual protein product after the location of the indel (**Figure 1d**). On the other hand, indels with lengths divisible by 3 lead to an in-frame mutation (i.e. reading frame not disrupted) generally do not impact protein function unless they affect important amino acids of a protein [19] (**Figure 1e**).

Non-coding mutations (i.e. those occurring outside of protein coding sequences) can also result in altered gene expression or protein function via numerous mechanisms [20]. Mutations in regulatory regions, such as promoters and enhancers, can disrupt or create transcription factor (TF) binding sites. For example, the C228T and C250T *TERT* promoter mutations create a binding site for the ETS family of proteins, resulting in promoter activation and increased *TERT* gene transcription [21]. Likewise, mutations at the 5′ and 3′ ends of introns can disrupt splice sites or create new sites, resulting in altered mRNA splicing (i.e. mRNA post-processing). With the MET oncogene, intronic mutations can for instance exon 14 being skipped (in the mature mRNA of the gene) which codes for a binding site for the ubiquitin ligase *CBL*. The oncoprotein escapes ubiquitination and degradation, thus resulting in aberrant activation *MET* [22]. Mutations in the 5' or 3' untranslated regions (UTR) can have various effects, such as altered translation efficiency or mRNA stability. For example, the *RB1* G17C and G18U mutations stabilize the 5'-UTR secondary structure, reducing the accessibility of translation machinery and hence protein expression [23].

Structural variants

Structural variants (SV) are rearrangements of large genomic segments (typically >50bp in size) and are caused by errors during DNA replication, repair, or recombination, but also by insertion of foreign (e.g. viral) DNA [24,25]. Simple SVs, which only involve a single genomic rearrangement, can be divided into two classes. Balanced SVs do not change the amount of DNA in the genome, and include inversions and balanced translocations. Unbalanced SVs lead to gains or losses of DNA, also known as copy number alterations (CNA), and include deletions, insertions, duplications, and unbalanced translocations [25,26]. Aneuploidy is a special case of unbalanced SVs which affect whole chromosomes or chromosome arms, and are caused by chromosome missegregation and unbalanced translocations respectively [27]. It should be noted that the distinction between indels (variants <50bp in size) and structural insertions, duplications and deletions (essentially indels >50bp in size) is not biological, but more the result of the technical limitations of algorithms that detect ('call') indels and SVs [28].

In contrast to simple SVs, complex SVs involve multiple inter-connected rearrangements and can occur through various mechanisms. Breakage-fusion-bridge (BFB) results from cycles of telomere breakage, sister chromatid telomeric fusion which produces a dicentric chromosome, and breakage of this chromosome during anaphase [29]. BFB is characterized by multiple foldback inversions and a 'staircase' amplification CNA profile due to the stepwise accumulation of DNA [30]. Chromothripsis is a single catastrophic chromosome shattering and rearrangement event that leads to a pattern of oscillating copy-number changes and localized clustering of tens to hundreds of breakpoints [31]. Double minutes (DM) are circular fragments of DNA that may form as a result of chromothripsis [32] or aberrant DNA repair [29].

One way SVs can impact cancer driver genes is by amplifying (i.e. increasing the copy number of) oncogenes, such as *EGFR, MYC, and MDM2* [33]. This can occur by linear amplification such as via structural duplications (**Figure 2c**), inverted duplications (possibly due to BFB [34]), or whole chromosome or chromosome arm gains [35]. Alternatively, circular amplification of oncogenes in DMs can result in extremely high copy numbers, possibly due to unequal segregation of DMs due to their lack of centromeres coupled with positive selection [33,36]. SVs can also inactivate TSGs if one or more of the breakpoints (i.e. the location(s) of the double strand break) occurs within the gene region. This may be for example due a full or partial structural deletion of the gene (e.g. of *VHL* in kidney cancer [37] or *BRCA2* in prostate cancer [38]), or an inversion that overlaps with the gene (**Figure 2b,d**). Gene inactivation can also occur from insertions of mobile elements such as from long interspersed nuclear elements (LINE) which are prevalent in gastrointestinal cancers [39], or from viral

sequence insertions from viruses that can integrate into host genomes such as human papillomavirus (HPV), common in cervical and head and neck cancers [40] (**Figure 2f**).

SVs may also result in gene fusions, which can occur via a deletion or inversion between two genes, or a translocation (**Figure 2b,d,e**) [41]. This can lead to oncogene overexpression, such as with the prostate cancer specific *TMPRSS2-ERG* fusion that is caused by an interstitial deletion (i.e. a deletion between the gene partners) and leads to overexpression of *ERG* via the constitutively highly active *TMPRSS2* promoter [42]. Oncogene overexpression can also result from aberrant interaction of the functional domains between two genes, such as with the lung cancer specific *EML4-ALK* fusion, caused by an interstitial inversion [43]. Fusions can also lead to decreased TSG expression (e.g. *CBFB–MYH11* fusions downregulating *CBFB [44]*) or TSG inactivation via gene truncations (e.g. *TP53* fusions [45]).

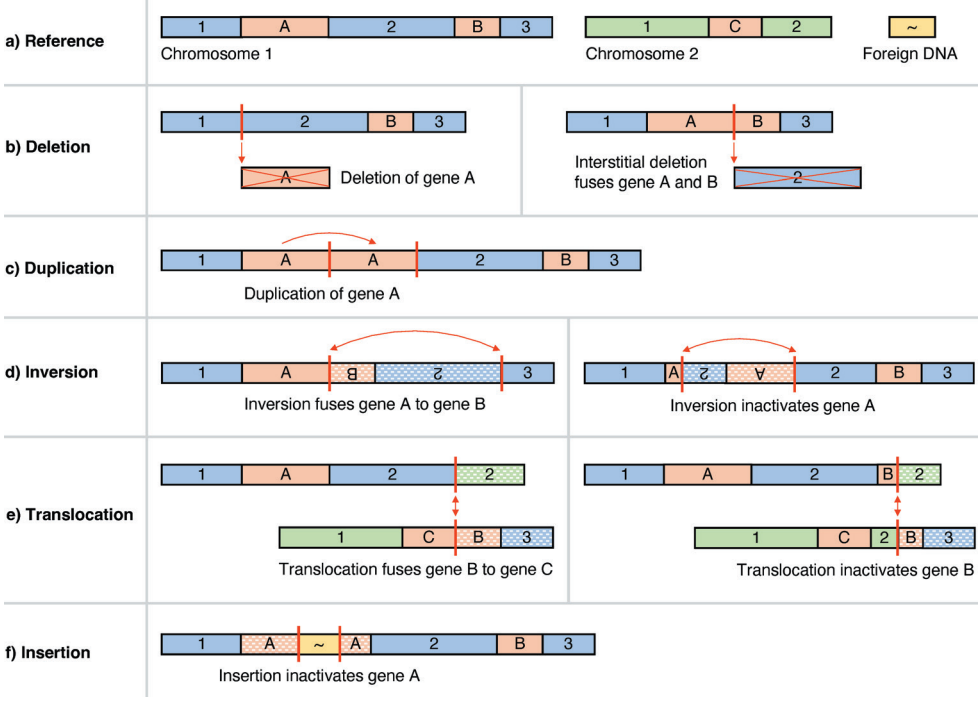


Figure 2: Examples of the effects of structural variants (SV) on genes. **a)** SVs are large rearrangements of DNA with respect to a reference genome. **b)** Deletions can result in loss of genes, or fusions of two genes. **c)** Duplications can create additional copies of genes. **d,e)** Inversions and translocations can result in the fusion of two genes, or inactivation of genes if the breakpoints occur within genes. **f)** Insertions of for example foreign DNA can also similarly inactivate genes.

Identifying driver mutations and genes

The sequencing of whole exomes and genomes has enabled the detection of millions of somatic mutations in the genomes of large tumor cohorts [5,6,46,47]. A major challenge is to distinguish the small number of driver mutations from the vast abundance of passenger mutations, as well as to identify the genes that are affected by driver mutations (i.e. cancer driver genes).

Numerous computational approaches have been developed to infer which small mutations (SBSs and indels) are drivers [48]. These are based on identifying recurrent mutations and/or mutations that likely

impact protein function. Mutations that occur significantly more frequently compared to the background mutation rate in a tumor cohort are likely to have been positively selected for and are thus likely driver mutations. This has been exploited by methods such as dNdScv [49], CBaSE [50], and mutSigCV [51]. Driver mutations can also be identified based on whether they cluster more frequently at specific (likely functional) regions on the DNA. OncodriveCLUSTL [52] detects linear clustering of mutations due to increased mutation recurrence at protein functional domains. Likewise, HotMAPS3D [53] detects mutation recurrence at regions important for protein structure leading to 3D clustering of mutations. Enrichment of mutations with high functional impact scores (e.g. those computed using CADD [54]) at specific regions may also indicate that those enriched mutations are drivers, and is the basis for OncodriveFML [55]. The above methods can also be combined into an ensemble approach such as was done with Intogen [56].

Similar as for small mutations, identifying driver SVs is based on recurrence (i.e. detecting SVs that occur more often than the background rate). The main approach has been to identify recurrent CNAs, which are gains and losses of DNA due to unbalanced SVs [57,58]. GISTIC has been the most widely used algorithm for this purpose [59,60], and addresses two major challenges with identifying recurrent CNAs. First, determining which CNAs are considered the same in different samples is not trivial, as a chromosome arm gain in one sample and a focal gain in another sample can both have the same consequence of amplifying an oncogene. Second, because the aggregate CNA profile across a cohort of samples represents overlapping underlying CNA events, it is difficult to determine the background CNA rate. Aside from CNAs, recurrent (i.e. clustered) SV breakpoints have also been used to identify driver SVs [61], following a similar rationale used by OncodriveCLUSTL [52]. Likewise, recurrent juxtapositions of genomic regions (i.e. those brought close together by SVs) were used to identify driver SVs leading to gene fusions [61].

Patterns of passenger somatic mutations

Mutational processes leave characteristic patterns of mutations in the genome, such as ultraviolet light exposure primarily resulting in C>T mutations at C<u>C</u> and T<u>C</u> nucleotides. Passenger mutations constitute the majority of mutations and thus capture such mutational patterns. Genome-wide patterns of passenger mutations thus can be used to provide insight into the mutational processes that have been active in cancer cells [13].

The first mutational patterns that were characterized were based on the 6 possible strand-agnostic base substitutions (C>A, C>G, C>T, T>A, T>C, T>G) together with the 16 possible combinations of 5' and 3' flanking bases, yielding a total of 96 possible 'trinucleotide contexts'. Using a computational framework based on non-negative matrix factorization (NMF), the mutation counts of these 96 trinucleotide contexts from whole-genome and whole-exome sequenced tumors across numerous cancer types were clustered into 30 trinucleotide context profiles, or 'mutational signatures' [62,63] (**Figure 3b**). These initial 30 single base substitution (SBS) signatures have been cataloged in the COSMIC database [64], with more SBS signatures having been added since. Signatures based on mutation contexts of other mutation types have also been added, including double base substitution (DBS) signatures based on the 78 possible strand-agnostic DBS mutation types, and indel (ID) signatures based on 83 different indels types that considered indel size, which nucleotides were affected, and presence of repetitive and/or microhomology regions [13].

Some of these signatures have been associated with the activity of specific mutational processes, including but not limited to: DNA repair deficiencies, such as homologous recombination deficiency (HRD; SBS3, ID6) or mismatch repair deficiency (MMRD; SBS6/14/15/20/21/26/44); environmental mutagens, such as ultraviolet radiation (SBS7/38, DBS1, ID13) or smoking (SBS4, DBS2, ID3); treatments such as platinum chemotherapy (SBS31/35, DBS5); and DNA editing activity of APOBEC proteins (SBS2/13) which normally serve to defend against viruses [13]. However, at the time of writing,

about a quarter of signatures in the COSMIC database [64] have yet to be assigned a potential cause or origin ('etiology').

Since its conception, mutational signature analysis has been widely adopted to study cancer. Many studies sought to discover additional sources of mutagenesis in cancer, such as the *Escherichia coli* metabolite colibactin in a subset of colorectal cancers (SBS88) [65], or the chemotherapeutic drug 5-fluorouracil (SBS17) [66]. A study by Kucab *et al. [67]* also linked ~50 environmental carcinogens to distinct mutational patterns. Other studies focused on the mechanisms that shape mutational signatures, such as DNA replication timing [68], as well as the interplay between DNA damage and repair [69]. Mutational signatures have also been frequently used for cancer subtyping, such as for esophageal [70], gastric [71], liver [72], and pancreatic [73] cancer. Lastly, mutational signatures have been used to develop classifiers of DNA repair deficiencies such as for HRD [74] and MMRD [75], which can be used to guide treatment decisions.

Machine learning applications for cancer diagnostics

The wealth of available WGS data [46,76] now enables the development of machine learning (ML) models for data driven decision making in cancer diagnostics. ML is the application of statistics to train computers to identify patterns or make predictions from existing data. In the context of cancer genomics, this data would for example consist of a (preferably large) set of whole genome sequenced tumor samples, with each sample having a set of measurable features (e.g. genome ploidy, number of mutations, presence of a gene mutation, etc). ML can be split into supervised and unsupervised learning (**Figure 3**).

Unsupervised learning broadly includes clustering and dimension reduction, and aims to find underlying patterns in the data without the need for labeled samples [77]. Clustering aims to group similar samples based on their features, and can be for example used to define genomic subtypes of cancer [78–80] (**Figure 3a**). Dimension reduction tries to represent a set of features using a smaller set of 'meta-features' which retain the underlying patterns in the original feature set. This includes methods such as principal component analysis (PCA) [81], as well as NMF which was used for signature extraction as described in the previous section [13,64] (**Figure 3b**). Unsupervised learning is mainly used for data analysis and is therefore typically not what is meant by ML, though dimension reduction can be used with 'classical' ML to reduce computational load and/or improve model performance [77].

Supervised learning uses labeled samples to train models that can make predictions on future unlabeled samples. 'Classical' ML usually refers to supervised learning which is divided into classification and regression [77]. Classification aims to define a decision boundary between sample groups based on their features so that future unseen samples can be categorized (**Figure 3c**). Common classification algorithms often used in genomics include (in increasing complexity): logistic regression (a type of regression for classification), support vector machines (SVM), random forest (RF), and neural networks (NN), with deep learning referring to multilayered NNs of varying architectures [77,82,83]. One application of classification is to determine the pathogenicity of coding [54,84,85] and non-coding [86] mutations, and by extension to predict which mutations are cancer drivers [56,87,88]. Classification has also been used for patient stratification, such as using somatic mutations to detect DNA repair defects such as HRD [74] and MMRD [75], detect Ras pathway activation [89], and to determine tumor tissue of origin [90,91]. Regression on the other hand aims to predict a numeric value, such as using simple linear regression to estimate mutation load based on age [92] (**Figure 3d**). Like with unsupervised learning, regression is mainly used for data analysis, though regression has been used to predict cancer drug response [93–96] based on genomic features from cell lines. Despite the numerous clinical applications of machine learning, it has yet to be widely adopted in clinical practice [97].

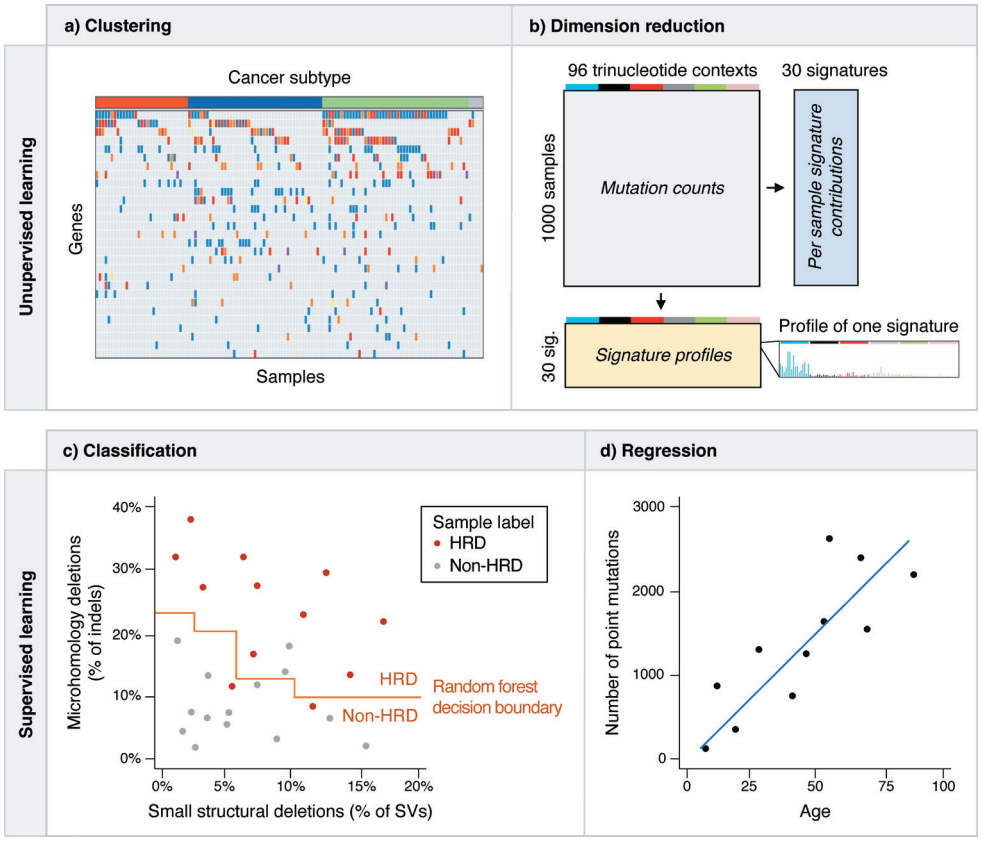


Figure 3: Overview of unsupervised and supervised machine learning. **a)** Clustering aims to group similar samples based on their features (e.g. gene mutations). **b)** Dimension reduction, using an algorithm such as non-negative matrix factorization (NMF; depicted in the graphic), tries to represent a set of features (e.g. the 96 trinucleotide contexts) using a smaller set of 'meta-features' (e.g. mutational signatures). The graphic depicts NMF decomposing a 1000x96 mutation count matrix into two matrices, a 30x96 signature profile matrix and a 1000x30 signature contribution matrix. The two latter matrices multiplied together produce an approximation of the original mutation count matrix. **c)** Classification aims to determine a decision boundary that can best separate two (or more) labeled sample groups (e.g. those with and without homologous recombination deficiency (HRD)) based on their features (e.g. the number of microhomology deletions and the number of structural deletions, as depicted in the graphic). **d)** Regression is used to correlate a numerical label (e.g. number of point mutations) with one or more features (e.g. age, as depicted in the graphic).

Thesis outline

In this thesis, we focus on using driver and/or passenger mutations detected from WGS to develop machine learning classifiers for cancer diagnostics (**Chapter 2** and **Chapter 3**), as well as to study cancer development (**Chapter 4** and **Chapter 5**). **Chapter 2** describes a Classifier of HOmologous Recombination Deficiency (CHORD) that detects HRD using various mutation types including microhomology deletions, small structural deletions and large structural duplications. Using CHORD, we found that HRD was most common in ovarian, breast, prostate and pancreatic cancer. We also found that HRD was often associated with loss-of-heterozygosity in all cancer types, with increased contribution of deep deletions in prostate cancer. **Chapter 3** describes Cancer of Unknown Primary Location Resolver (CUPLR), a machine learning model that predicts tumor tissue of origin using a wide range of genomic features, including RMD, mutational signatures, driver gene mutations, aneuploidy, viral insertions, and various SV types. We found that SVs were important for and improved the performance of tumor tissue of origin classification for cancer types where SV related features were important, such as in pilocytic astrocytomas (characterized by *KIAA1549-BRAF* fusions) or cervical cancer (characterized by human papillomavirus DNA insertions). In **Chapter 4**, we use liver organoids to study mutation accumulation and its contribution to cancer development in three liver diseases, alcoholic cirrhosis, non-alcoholic steatohepatitis (NASH), and primary sclerosing cholangitis (PSC). We find surprisingly that these liver diseases do not contribute to detectable alterations in the mutation landscape. In **Chapter 5**, we perform a pan-cancer comparison of primary versus metastatic cancer and find that metastatic tumors differ from primary tumors mainly in their increased burden of small mutations and SVs as a result of treatment exposure and cancer type specific endogenous mutational processes. Lastly, in **Chapter 6** we discuss the limitations and challenges faced in using WGS data to understand and diagnose cancer, and provide directions for future research.

Chapter 2

Pan-cancer landscape of homologous recombination deficiency

Luan Nguyen[1], John Martens[2,3], Arne Van Hoeck[1,$], Edwin Cuppen[1,4,$,*]

[1] University Medical Center Utrecht, Utrecht, The Netherlands
[2] Erasmus Medical Center, Rotterdam, The Netherlands
[3] Center for Personalized Cancer Treatment, Rotterdam, The Netherlands
[4] Hartwig Medical Foundation, Amsterdam, The Netherlands

[$] These authors jointly supervised this work
* Corresponding author

Adapted from: Nat Commun 11, 5584 (2020)
URL: https://doi.org/10.1038/s41467-020-19406-4
QR code to URL:

Abstract

Homologous recombination deficiency (HRD) results in impaired double strand break repair and is a frequent driver of tumorigenesis. Here, we develop a genome-wide mutational scar-based pan-cancer **C**lassifier of **HO**mologous **R**ecombination **D**eficiency (CHORD) that can discriminate between BRCA1- and *BRCA2*-subtypes. Analysis of a metastatic (n=3,504) and primary (n=1,854) pan-cancer cohort reveals that HRD is most frequent in ovarian and breast cancer, followed by pancreatic and prostate cancer. We identify biallelic inactivation of *BRCA1*, *BRCA2*, *RAD51C* or *PALB2* as the most common genetic cause of HRD, with *RAD51C* and *PALB2* inactivation resulting in *BRCA2*-type HRD. We find that while the specific genetic cause of HRD is cancer type specific, biallelic inactivation is predominantly associated with loss-of-heterozygosity (LOH), with increased contribution of deep deletions in prostate cancer. Our results demonstrate the value of pan-cancer genomics-based HRD testing and its potential diagnostic value for patient stratification towards treatment with e.g. poly ADP-ribose polymerase inhibitors (PARPi).

Introduction

The homologous recombination (HR) pathway is essential for high-fidelity DNA double strand break (DSB) repair and involves numerous genes including *BRCA1* and *BRCA2*. HR deficiency (HRD) due to inactivation of such genes leads to increased levels of genomic alterations [98]. HRD is a common characteristic of many tumors and is frequently observed in breast and ovarian cancer [99]. Accurate detection of HR deficiency (HRD) is of clinical relevance as it is indicative of sensitivity to targeted therapy with poly ADP-ribose polymerase inhibitors (PARPi) [100,101] as well as to DNA damaging reagents [98].

In the clinic, germline *BRCA1/2* mutation status is currently the main genetic biomarker of HRD [102]. However, germline testing has its drawbacks: i) it is dependent on the completeness and accuracy of clinical variant annotation databases (e.g. ClinVar); ii) epigenetic silencing is overlooked; iii) partial/complete deletions of the *BRCA1/2* loci are missed by current clinical genetic testing, resulting in *BRCA1/2* status reporting based on the wild type allele from contaminating normal tissue; and iv) HRD can be driven purely by somatic events. Furthermore, the focus on *BRCA1/2* overlooks inactivation of other HR pathway genes. Consequently, patients may receive incorrect treatment or miss out on treatment opportunities, thus necessitating the development of better biomarkers for HRD.

It was recently shown that somatic passenger mutations, which are identified efficiently by whole genome sequencing (WGS), can provide insights into the mutational processes that occurred before and during tumorigenesis, paving the way for novel opportunities for clinical tumor diagnostics [9]. For the repair of DSBs, HRD tumors are dependent on alternative more error-prone pathways including microhomology mediated end-joining (MMEJ) [15], resulting in a characteristic mutational footprint across the genome that can be used to detect HRD regardless of the underlying cause (whether genetic or epigenetic). Indeed, some mutational footprints were found to be associated with *BRCA1/2* deficiency, namely deletions with flanking microhomology, as well as several 'mutational signatures' including two COSMIC single nucleotide variant (SNV) signatures and two structural variant (SV) signatures [103]. These features were used to develop a breast cancer-specific predictor of HRD known as HRDetect [74]. Application of this tool in primary tumors revealed that the prevalence of HRD extends beyond *BRCA1/2*-deficient breast cancer tumors, and occurs at varying frequencies in different cancer types [104]. However, HRD rates in advanced metastatic cancer remain unclear, although these are the patients that are increasingly targeted with personalized treatments including PARP inhibitors for *BRCA*-deficiency [9].

Here, we describe the development of a random forest-based **C**lassifier of **HO**mologous **R**ecombination **D**eficiency (CHORD) for pan-cancer HRD detection. With this model, we demonstrate that accurate prediction of HRD is possible across cancer types using specific SNV, indel and SV types. We identify inactivation of *BRCA1*, *BRCA2*, *RAD51C* and *PALB2* as the most frequent genetic cause of HRD pan-cancer in both primary and metastatic cancer, with the latter two genes resulting in the same mutational footprints as *BRCA2* (consistent with the findings of recent studies in breast cancer [105,106]). In addition, we find that the underlying genetic inactivation of these genes is cancer type specific, but independent of tumor progression state.

Results

Random forest classifier training

For the development of CHORD, we used WGS data of 3,824 solid tumors from 3,584 patients from the pan-cancer metastatic cohort of the Hartwig Medical Foundation (HMF) [6]. From these, we selected tumor samples with biallelic loss of *BRCA1* or *BRCA2*, and non-mutated *BRCA1/2*, to obtain a high confidence set of samples belonging to 3 classes for classifier training (*BRCA1*-deficient, *BRCA2*-deficient, and *BRCA1/2*-proficient). To this end, we screened each sample to identify those samples with one of the following events in *BRCA1/2*: (i) complete copy number loss (i.e. deep deletion), (ii) loss-of-heterozygosity (LOH) in combination with a pathogenic germline or somatic SNV/indel (as annotated in ClinVar, or a frameshift), or (iii) 2 pathogenic SNV/indels. This unbiased approach revealed 35 and 89 samples with *BRCA1* or *BRCA2* biallelic loss of function, respectively, which were labeled as HRD for the training. Conversely, 1,902 samples were labeled as HR proficient (HRP) as these samples were observed to carry at least one functional allele of *BRCA1/2*. In total, 2,026 out of 3,824 samples (53% of the HMF dataset) were used to train the classifier (**Supplementary figure 1**).

The occurrence of three main somatic mutation categories were used as features for training (**Figure 1a**), which included (i) single nucleotide variants (SNVs) subdivided by base substitution (SBS) type; (ii) indels stratified by the presence of sequence homology, tandem repeats, or the absence of either; and (iii) structural variants (SV), stratified by type and length. An initial feature analysis revealed that small deletions with ≥2bp flanking homology were together more predictive of *BRCA1/2* deficiency versus deletions with 1bp flanking homology (**Supplementary figure 3**). Thus, deletions with flanking homology were further split into these two homology length bins. The occurrence of the 29 features together formed a contribution profile for each sample. From this, relative contributions per mutation category were calculated to account for differences in mutational load across samples (**Figure 1a**). These features are henceforth collectively referred to as 'mutation contexts'.

A random forest was then trained to predict the probability of *BRCA1* or *BRCA2* deficiency (**Figure 1b**). Briefly, a core training procedure performed feature selection and class resampling (to alleviate the imbalance between the 3 classes). This core procedure was subjected to 10-fold cross-validation (CV) which was repeated 100 times to filter samples from the training set that were not consistently HRD or HRP. A sample was considered HRD if the sum of the *BRCA1* and *BRCA2* deficiency probabilities (henceforth referred to as the HRD probability) was ≥0.5. This core procedure was reapplied to the filtered training set to yield the final random forest model which we refer to as 'CHORD' (**Supplementary figure 2a,b**; **Supplementary figure 4**).

The presence of deletions with ≥2bp flanking homology (del.mh.bimh.2.5) was found to be the most important predictor of HRD. Additionally, CHORD uses 1-10kb and to a lesser extent 10-100kb duplications (DUP_1e03_1e04_bp and DUP_1e04_1e05_bp, respectively) for distinguishing *BRCA1* from *BRCA2* deficiency. Given that deficiencies in other HR genes may lead to similar phenotypes, we have coined the terms '*BRCA1*-type HRD' and '*BRCA2*-type HRD' to describe these HRD subtypes (**Figure 1c**). Together, the features that are predictive of HRD are in line with those of a previously described HRD classifier HRDetect [74]. However, the feature weights differ markedly likely due to differences in the background mutational landscape between the pan-cancer cohort used here compared to the breast cancer cohort used for training HRDetect.

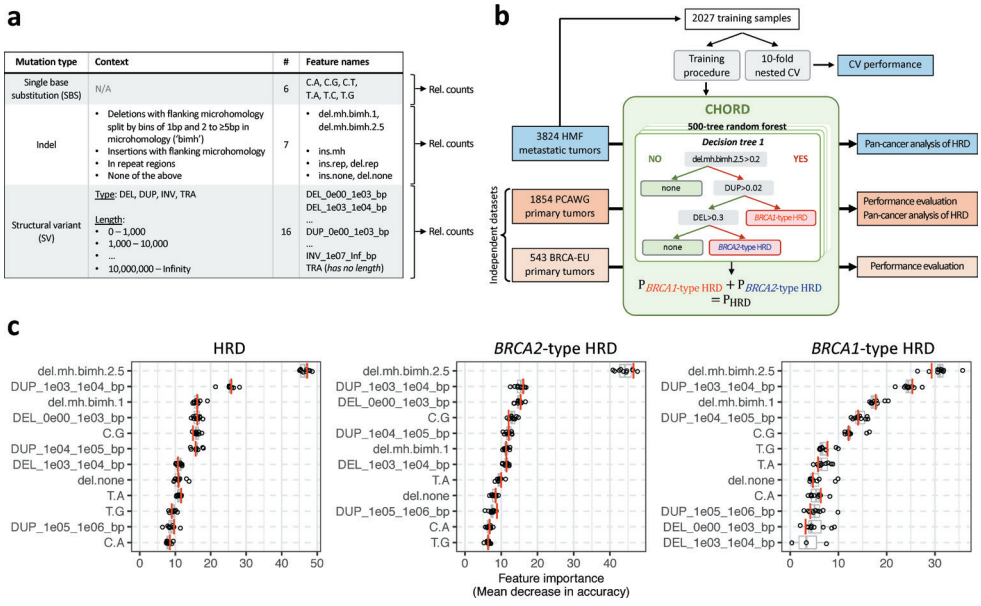


Figure 1: CHORD is a random forest Classifier of HOmologous Recombination Deficiency able to distinguish between *BRCA1*- and *BRCA2*-type HRD phenotypes in a pan-cancer context. (**a**) The features used for training CHORD are relative counts of different mutation contexts which fall into one of three groups based on mutation type. (i) Single nucleotide variants (SNV): 6 possible base substitutions (C>A, C>G, C>T, T>A, T>C, T>G). (ii) Indels: indels with flanking microhomology (del.mh, ins.mh), within repeat regions (del.rep, del.none), or not falling into either of these 2 categories (del.none, ins.none). (iii) Structural variants (SV): SVs stratified by type and length. Relative counts were calculated separately for each of the 3 mutation types. (**b**) Training and application of CHORD. From a total of 3,824 metastatic tumor samples, 2,026 samples were selected for training CHORD. The model outputs the probability of *BRCA1*-type HRD and *BRCA2*-type HRD, with the probability of HRD being the sum of these 2 probabilities. The performance of CHORD was assessed via a 10-fold nested cross-validation (CV) procedure on the training samples, as well as by applying the model to the BRCA-EU dataset (543 primary breast tumors) and PCAWG dataset (1,854 primary tumors). Lastly, CHORD was applied to all samples in the HMF and PCAWG dataset in order to gain insights into the pan-cancer landscape of HRD. (**c**) The features used by CHORD to predict HRD as well as *BRCA1*-type HRD and *BRCA2*-type HRD, with their importance indicated by mean decrease in accuracy. Deletions with 2 to ≥5bp (i.e. ≥2bp) of flanking microhomology (del.mh.bimh.2.5) was the most important feature for predicting HRD as a whole, with 1-100kb structural duplications (DUP_1e03_1e04_bp, DUP_1e04_1e05_bp) differentiating *BRCA1*-type HRD from *BRCA2*-type HRD. Boxplot and dots (n=10) show the feature importance over 10-folds of nested CV on the training set, with the red line showing the feature importance in the final CHORD model. Boxes show the interquartile range (IQR) and whiskers show the largest/smallest values within 1.5 times the IQR.

Performance of CHORD

Two approaches were used to assess the performance of CHORD. In the first approach, 10-fold CV was performed on the training data which allows every sample to be excluded from the training set after which unbiased HRD probabilities can be determined (**Supplementary figure 2c**). The probabilities of all prediction classes (i.e. HRD, *BRCA1*-type HRD, *BRCA2*-type HRD) were highly concordant with the genetic annotations (**Figure 2a**). The concordance between predictions and annotations was quantified by calculating the area under the curve of receiver operating characteristic (AUROC) and precision-recall (AUPRC) curves (**Figure 2b,c**). CHORD achieved excellent performance as shown by the high AUROC and AUPRC for all prediction classes (0.98 and 0.87 respectively). Additionally, CHORD achieved a maximum F1-score (~0.88) for predicting HRD at a cutoff of 0.5 which was thus set to be the classification threshold (**Supplementary figure 6**).

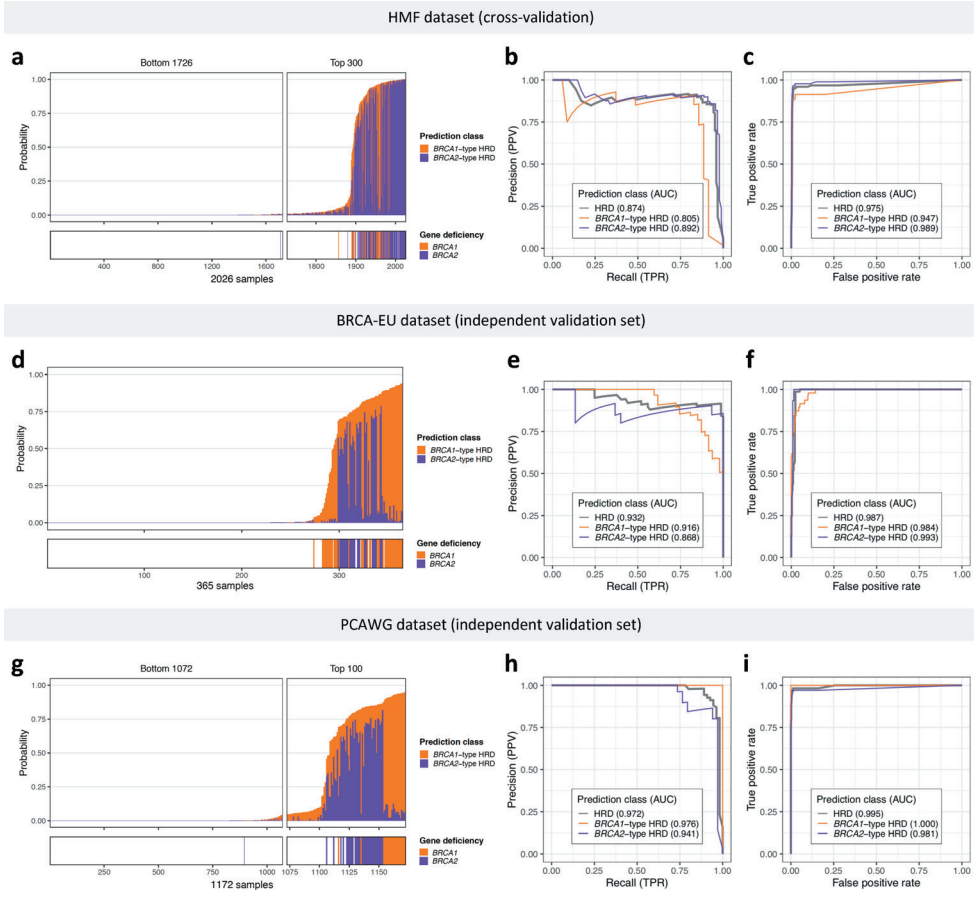


Figure 2: Performance of CHORD. Performance was determined by 10-fold cross-validation (CV) on the HMF training data or prediction on two independent datasets: BRCA-EU (primary breast cancer dataset) and PCAWG (primary pan-cancer dataset). BRCA-EU and PCAWG samples shown here all passed CHORD's QC criteria (i.e. MSI absent, ≥50 indels, ≥30 SVs if a sample was predicted HRD). (**a, d, g**) The probability of HRD for each sample (total bar height) with each bar being divided into segments indicating the probability of *BRCA1*- (orange) and *BRCA2*-type HRD (purple). Stripes below the bar plot indicate biallelic loss of *BRCA1* or *BRCA2*. In (**a**), probabilities have been aggregated from the 10 CV folds. (**b, e, h**) Receiver operating characteristic (ROC) and (**c,**

f, i) precision-recall curves (PR) and respective area under the curve (AUC) values showing the performance of CHORD when predicting HRD as a whole (grey), *BRCA1*-type HRD (orange), or *BRCA2*-type HRD (purple).

In the second approach, performance was evaluated on two independent datasets: the BRCA-EU dataset [103] (543 primary breast tumors) and the PCAWG dataset [5] (1,854 primary tumors, pan-cancer). For both datasets, samples that (i) passed CHORD's QC criteria (i.e. MSI absent, ≥50 indels, ≥30 SVs if a sample was predicted HRD; **Supplementary notes; Supplementary figure 26**, **Supplementary figure 27**) and (ii) for which the biallelic status of *BRCA1/2* could confidently be determined were selected for validation of CHORD. For the BRCA-EU dataset, this included the 365 samples that were used to train and evaluate the performance of HRDetect [74]. For the PCAWG dataset, this included 1,172 samples for which the same genetic criteria used for selecting samples from the HMF dataset for training CHORD applied. Applying CHORD on these samples revealed that the HRD probabilities were concordant to their *BRCA1/2* genetic status for both the BRCA-EU and PCAWG datasets (**Figure 2d,g**). The AUROC (>0.98) and AUPRC (>0.93) values were comparable to those obtained by CV on the HMF training data for all prediction classes for both datasets (**Figure 2e,f,h,i**). In the BRCA-EU dataset, we still observed some *BRCA1* deficient samples classified as HRP by CHORD (while HRDetect classified these as HRD) and tested whether this was due to differences in somatic calling algorithms. Indeed, using the variants obtained from the native pipeline of the HMF dataset (HMF pipeline [6]) for HRD prediction resulted in overall higher HRD probabilities compared to using the variants downloaded from ICGC, especially for *BRCA1*-deficient samples. This was apparent for sample PD4017 which became HRD using HMF pipeline called mutation profiles, with PD24186, PD11750 and PD23578 having greatly increased HRD probabilities (**Supplementary figure 7**). Our results thus demonstrate that CHORD is robust when applied to other datasets. However, differences in variant calling pipelines can affect CHORD's ability to predict HRD (especially considering the still existing challenges of indel and SV calling from WGS data, and CHORD's dependency on these features). Additional validation and threshold optimisation is thus recommended when applying CHORD on data from other variant calling pipelines.

We note that CHORD performs similarly to HRDetect based on predictions on the BRCA-EU dataset (AUROC=0.98 for both models) [74]. Additionally, the predictions of CHORD and HRDetect on the PCAWG dataset [104] were concordant for the vast majority of samples (1506/1526; 99%) (**Supplementary figure 10**). Of the 8 HRD samples only detected by CHORD, 3 showed biallelic loss of *BRCA1/2*, while none of the HRDetect-only samples could be explained by genetic biallelic loss. Given that CHORD, unlike HRDetect, does not rely on COSMIC SBS signatures [9] and SV signatures [103], the similar performance between the two models suggests that accurate detection of HRD is possible without using an intermediate mutational signature analysis step [107]. To further validate this, we trained a random forest model (CHORD-signature) that uses the SBS and SV signatures as input instead of mutation contexts. CHORD-signature performed similarly to CHORD (**Supplementary figure 12**), which can be explained by the reliance on similar features (**Supplementary figure 11**), namely microhomology deletions and SV signature 3 (analogous to 1-100kb duplications). We thus conclude that accurate detection of HRD does not require mutational signatures, thereby simplifying HRD calling and avoiding potential complications associated with the fitting step required for computing signature contributions in individual samples (for which there is currently no consensus approach) [107].

Effect of treatment on HRD predictions

The HMF dataset comprises tumors from patients with metastatic cancer who have been exposed (some heavily) to treatment which could potentially affect CHORD's predictions. Two recent studies showed that common cancer treatments in general do not induce mutations that may interfere with CHORD predictions [67,108]. However, these two studies (as well as one by Behjati *et al.* [109]) did

show that radiotherapy had the potential to induce deletions with flanking microhomology, which could potentially lead to false positive HRD classifications. To investigate this, we used random forests to identify and compare the mutational features associated with radiotherapy and *BRCA1/2* deficiency when using clonal variants versus subclonal variants (which are enriched for treatment induced mutations [66,108]) as input features. This revealed that small deletions with 1bp of flanking homology (del.mh.bimh.1) are highly associated with radiotherapy (**Supplementary figure 14**) and less with *BRCA1/2* deficiency. When we retrain CHORD with all microhomology deletions merged into a single feature (CHORD-del.mh.merged; **Supplementary figure 15**), there were only few discrepant predictions (9 CHORD-specific and 5 CHORD-del.mh.merged-specific out of 3715; **Supplementary figure 16**). All 5 samples that were CHORD-del.mh.merged-specifc did have radiotherapy as a previous treatment, while of the 9 samples predicted HRD only by CHORD, 5 had radiotherapy although 2 had evidence of *BRCA1/2* biallelic loss. These data suggest that splitting microhomology deletions into two microhomology length bins may slightly reduce false positive predictions resulting from radiotherapy treatment, although the low number of discrepant samples between CHORD and CHORD-del.mh.merged also indicates that the impact of radiotherapy on HRD prediction is minimal, at least when using all somatic variants (clonal plus subclonal) as input (which is likely the default setting for routine application).

On the other hand, we observed more samples being predicted as HRD based on subclonal variants but HRP based on clonal variants for CHORD-del.mh.merged (97 samples) compared to CHORD (64 samples) (**Supplementary figure 17a,b**). This indicates that having microhomology deletions split by these two homology length bins may mitigate false positive predictions when CHORD is applied to subclonal variants, whether due to mutations induced by radiotherapy, other treatments, or noise from variant calling algorithms. Alternatively, some samples that are scored HRP based on clonal variants but HRD on subclonal variants could truly be HRD, especially since 4 of these samples had evidence of *BRCA1/2* biallelic loss (deep deletion: n=1; LOH and a pathogenic variant: n=1; 2 pathogenic variants: n=2). For these samples, it is likely that *BRCA1/2* biallelic loss occurred relatively late in the tumor progression stage which results in an insufficient number of HRD-associated mutations for clear HRD classification by CHORD. Furthermore, subclonal-only HRD could potentially also be explained by transient inactivation of HR e.g. through epigenetic silencing of key components. Thus, CHORD predictions on subclonal variants must be interpreted with caution, especially given the extra challenges associated with accurately detecting subclonal variants with low variant allele frequency (VAF).

BRCA2, RAD51C and PALB2 are associated with BRCA2-type HRD while only BRCA1 is associated with BRCA1-type HRD

To gain insights into the genetic causes of HRD, we applied CHORD to both the HMF and PCAWG datasets and selected the samples that passed CHORD's QC criteria (i.e. MSI absent, ≥50 indels, ≥30 SVs if a sample was predicted HRD; **Supplementary notes**). For the HMF dataset, we also selected a single tumor per patient (based on highest tumor purity) for those with multiple biopsies, though all patients had consistent HRD probabilities across all biopsies (**Supplementary data 1**). This yielded a total of 5,122 patients (3,504 from HMF and 1,618 from PCAWG), with 310 (6%) being classified as being homologous recombination deficient (CHORD-HRD). Of these, 118 were classified as having *BRCA1*-type HRD and 192 as having *BRCA2*-type HRD. The remaining 4,812 patients were classified as homologous recombination proficient (CHORD-HRP) (**Figure 3a, Supplementary data 1**).

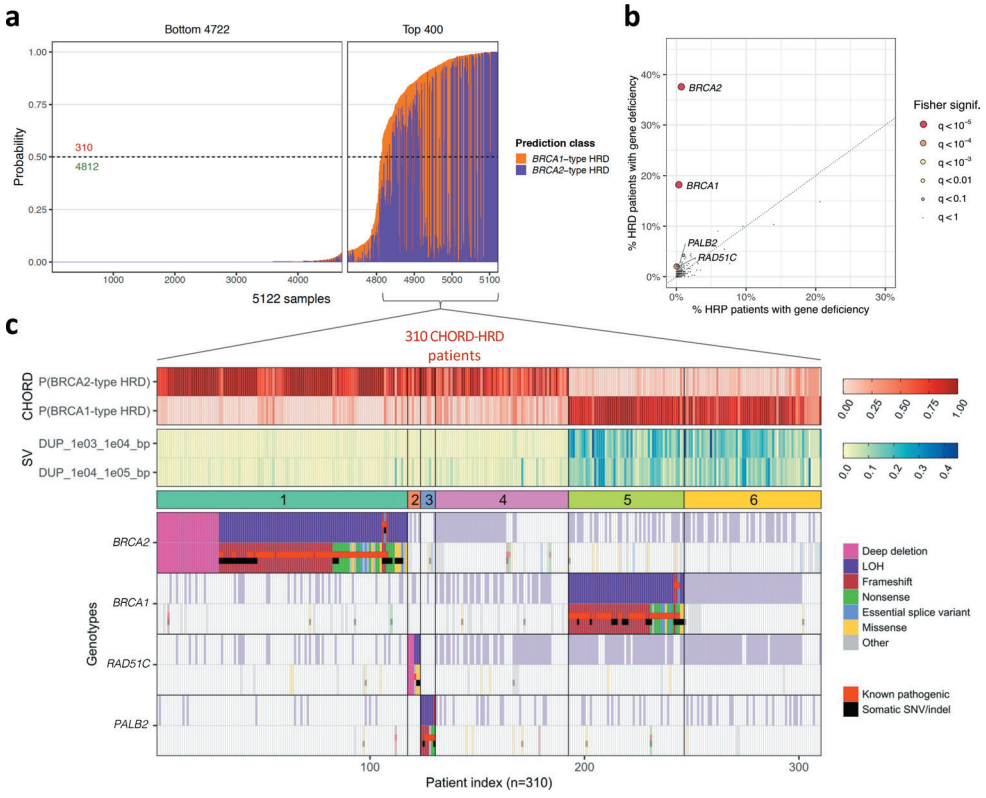


Figure 3: The genetic causes of HRD in patients from the HMF and PCAWG datasets. (**a**) The bar plot shows the probability of HRD for each patient (total bar height) with each bar being divided into segments indicating the probability of *BRCA1*-type HRD (orange) and *BRCA2*-type HRD (purple). 310 patients were predicted HRD while 4,812 were predicted HRP by CHORD. (**b**) A one-tailed Fisher's exact test identified enrichment of *BRCA1* (q=9.4e-51), *BRCA2* (q=4.8e-101), *RAD51C* (q=5.6e-5) and *PALB2* (q=0.02) biallelic inactivation in CHORD-HRD vs. CHORD-HRP patients (from a list of 781 cancer and HR related genes). Each point represents a gene with its size/color corresponding to the statistical significance as determined by the Fisher's exact test, with axes indicating the percentage of patients (within either the CHORD-HRD or CHORD-HRP group) in which biallelic inactivation was detected. Multiple testing correction was performed using the Hochberg procedure. (**c**) Biallelic inactivation of *BRCA2*, *RAD51C* and *PALB2* was associated with *BRCA2*-type HRD, whereas only *BRCA1* inactivation was associated with *BRCA1*-type HRD. Top: *BRCA1*- and *BRCA2*-type HRD probabilities from CHORD. Middle: SV contexts (duplications 1-10kb and 10-100kb) used by CHORD to distinguish *BRCA1*- from *BRCA2*-type HRD. Bottom: The biallelic status of each gene. Samples were clustered according to HRD subtype, and by the impact of a biallelic/monoallelic event (based on 'P-scores' as detailed in the methods). Clusters 1, 2, 3, and 5 correspond to patients with identified inactivation of *BRCA2*, *RAD51C*, *PALB2* and *BRCA1*, while clusters 4 and 6 correspond to patients without clear biallelic inactivation of these 4 genes. Tiles marked as 'Known pathogenic' refer to variants having a 'pathogenic' or 'likely pathogenic' annotation in ClinVar. 'Other' variants include various low impact variants such as splice region variants or intron variants (these are fully specified in **Supplementary data 4**). LOH: loss-of-heterozygosity. Only data from samples that passed CHORD's QC criteria are shown in this figure (MSI absent, ≥50 indels, and ≥30 SVs if a sample was predicted HRD).

We then sought to identify the key mutated genes underlying the HRD phenotype by performing an enrichment analysis of biallelically inactivated genes in CHORD-HRD vs. CHORD-HRP patients. For this analysis, we started from a list of 781 genes that are cancer related (based on the catalog of genes from Cancer Genome Interpreter) and/or HR related (manually curated based on the KEGG HR pathway, as well as via literature search) (**Supplementary data 3**). For these genes, we considered likely pathogenic variants (according to ClinVar) as well as predicted impactful variants such as nonsense mutations to contribute to gene inactivation (see *Methods*). This revealed that, in addition to *BRCA1* and *BRCA2* ($q<10^{-5}$ for both genes, one sided Fisher's exact test), *RAD51C* and *PALB2* ($q<0.001$ and $q<0.05$ respectively) were also significantly enriched amongst HRD patients using a q-value threshold of 0.05 (**Figure 3b**).

Of all CHORD-HRD HMF patients, ~60% (184/310) could be explained by biallelic inactivation of either *BRCA2* (cluster 1; n=117), *BRCA1* (cluster 5; n=54), *RAD51C* (cluster 2; n=6), or *PALB2* (cluster 3; n=7), which was most often caused by LOH in combination with a pathogenic variant or frameshift, or a deep deletion (**Figure 3c, Supplementary data 4**). *RAD51C* and *PALB2* were recently linked to HRD as incidental cases using mutational signature based approaches [105,110] and our results now confirm that biallelic inactivation of these two genes results in HRD and is actually a common cause of HRD (albeit to a lesser extent than for *BRCA1/2*). *RAD51C* and *PALB2* deficient patients shared the *BRCA2*-type HRD phenotype (absence of duplications) with *BRCA2* deficient patients (clusters 1-3; **Figure 3c**), consistent with previous studies [105,106]. On the other hand, only *BRCA1* deficient patients (cluster 5) harbored the *BRCA1*-type HRD phenotype (1-100kb duplications).

Of note, we observed one patient (**Figure 3c**; patient #6) bearing a known pathogenic frameshift mutation in *BRCA1* (**Supplementary data 4**; patient HMF001925, c.1961dupA), which based on current practices for detecting HRD in the clinic (testing for pathogenic SNVs/indels) [102] would be considered the driver mutation. However, our genetic analysis indicates that the deep deletion in *BRCA2* (which would be missed by testing for SNVs/indels) was the cause of HRD, which is supported by the lack of LOH in *BRCA1*, as well as the *BRCA2*-type HRD phenotype of this patient.

In ~40% of CHORD-HRD patients (126/310; clusters 4 and 6, **Figure 3c**), there was no clear indication of biallelic loss of *BRCA1/2*, *RAD51C* or *PALB2* (henceforth referred to as the 'HRD associated genes'). However, 109 of these patients had a deleterious event in a single allele of one of the HRD associated genes (the majority due to LOH (including copy number neutral LOH)), with a similar cancer type distribution in these patients as in the biallelically affected patients (**Supplementary figure 21**). Some samples had, as a second hit, variants not known to be pathogenic, but could potentially be novel pathogenic variants (**Supplementary Notes**, **Supplementary figure 28**, **Supplementary data 7**). We also found enrichment of LOH in *BRCA1*, *BRCA2*, as well as *RAD51C* in HRD samples (**Supplementary figure 22**), which implies the involvement of LOH in the inactivation of these genes for the patients in clusters 4 and 6. This is consistent with the finding by Jonsson *et al.* [111] that LOH is enriched in tumors with *BRCA1/2* germline pathogenic variants or somatic loss-of-function variants. Davies *et al.* [74] showed that promoter methylation of *BRCA1* was present in 22% of ovarian and 16% of breast primary cancers with HRD (**Table 1**). *BRCA1* and *RAD51C* promoter methylation with loss of the other allele was also reported in HRD tumors in other studies [105,106,110]. Thus, *BRCA1* and *RAD51C* promoter methylation, likely in combination with LOH, may have led to the HRD phenotype for a sizable portion of the ovarian and of breast cancer patients with no clear biallelic loss of the HRD associated genes, and potentially for patients with other cancer types as well (**Supplementary figure 23**). Unfortunately, we could not directly assess this as methylation data was not available for the HMF nor the PCAWG dataset.

Cancer type	HRD patients with *BRCA1* promoter methylation (data from Davies et al. 2017)	CHORD-HRD patients without biallelic loss of HRD associated genes
Ovarian	22% (10/46)	47% (36/76)
Breast	16% (14/86)	49% (50/103)

Table 1: Comparing prevalence of *BRCA1* promoter methylation in HRDetect-HRD patients with CHORD-HRD patients without biallelic loss of *BRCA1/2*, *RAD51C* or *PALB2.*

We also cannot rule out the possibility that deficiencies in other HR genes that did not reach significance in our enrichment analysis, underlie the HRD phenotype for a small number of patients in clusters 4 and 6. We indeed identified 17 patients with biallelic inactivation of a HR gene other than *BRCA1/2*, *RAD51C* or *PALB2*, and 1 patient with a likely inactivating biallelic event (LOH in combination with a nonsense variant in *CHEK1*) (**Supplementary figure 24**). Notably, the 4 patients with *RAD51B* (n=2) and *XRCC2* (n=2) deficiency were all predicted to have *BRCA2*-type HRD, a phenotype shared with *RAD51C* deficient patients [112]. Given that these 3 genes all belong to the *RAD51* paralog complex *BCDX2* [113], the *BRCA2*-type HRD suggests that *RAD51B* and *XRCC2* deficiency could have led to HRD in these patients. Likewise, the 4 patients with deficiencies in the *BRCA1*-binding proteins, *BARD1* [114] (n=1), *BRIP1* [115] (n=1), *FAM175A* [116] (n=1) and *FANCA* [117] (n=1), were all predicted as having *BRCA1*-type HRD. Thus, while we could not conclusively determine the cause of HRD for patients in clusters 4 and 6, we postulate that HRD in these patients may have been a result of epigenetic silencing of *BRCA1/2* or *RAD51C*, deficiencies in other HR genes (not associated to HRD in our analysis), or possibly a result of other unknown regulatory mechanisms.

The incidence and genetic cause of HRD varies in different tissue types and cancer stage

We next investigated the differences in the incidence and genetic causes of HRD based on primary tumor location in both primary (PCAWG) and metastatic (HMF) cancer datasets (**Figure 4**). HRD was most prevalent in ovarian, breast, prostate and pancreatic cancer (85% combined), and only occurred sporadically in other cancer types (15%) (**Supplementary data 5**). Compared to metastatic cancer, HRD is found much more often in primary ovarian (52% vs 30%) and breast (24% vs 12%) cancers, and less often in primary prostate (5.6% vs 13%) and pancreatic (7.3% vs 13%) cancer (**Figure 4a**). Notably, in metastatic cancer, prostate and pancreatic cancer has a similar incidence of HRD to breast cancer (all ~13%). However, the observed differences in HRD rates between the primary and metastatic cohorts may not necessarily be conclusive as we cannot rule out confounding factors such as patient inclusion criteria.

Across different cancer types, we observed pronounced diversity in HR function loss (**Figure 4b**). *BRCA2*-type HRD deficiencies (including *BRCA2*, *RAD51C*, *PALB2* deficiencies) were more frequent in pancreatic and prostate cancer. On the other hand, *BRCA1*-type HRD deficiencies were found more often in ovarian and breast cancer. Interestingly, for ovarian and prostate cancer, *BRCA1*-type HRD deficiencies were more prominent in primary cancer compared to metastatic cancer. Whether these differences in gene deficiencies in different cancer types can be linked to a biological cause or have prognostic value remains to be determined.

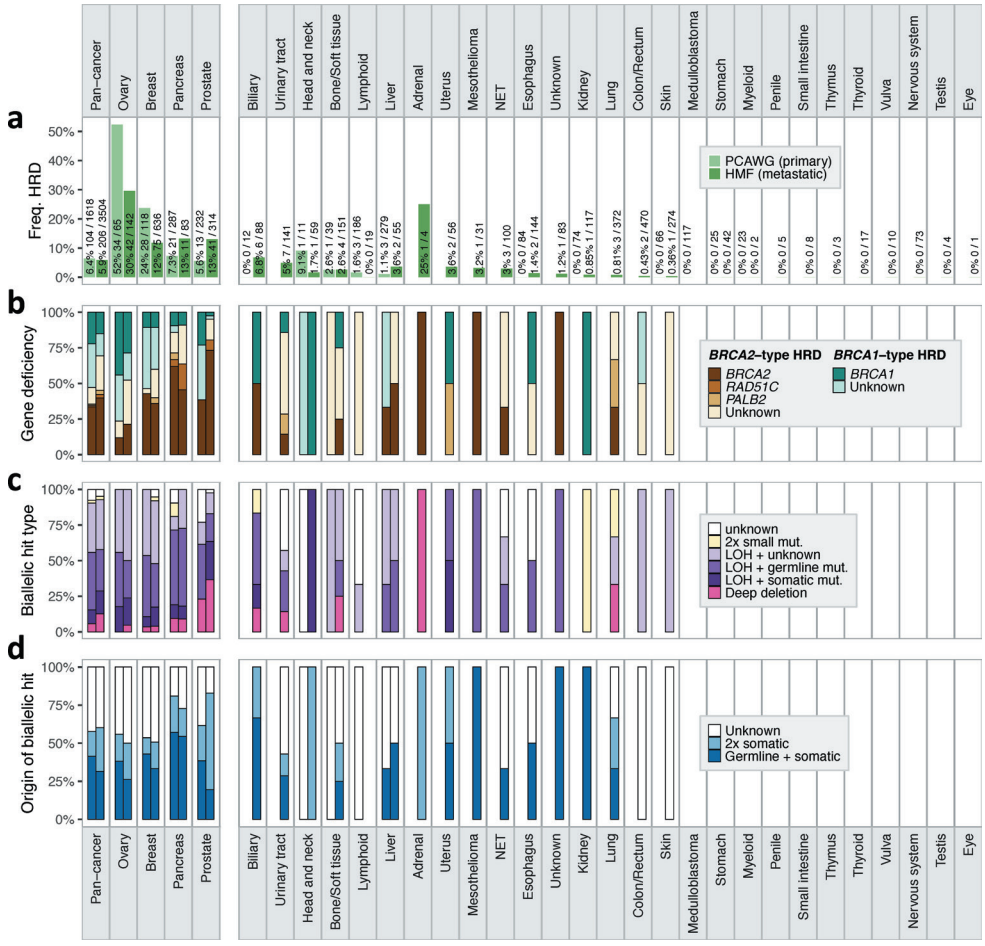


Figure 4: Percentage breakdown of the incidence and genetic causes of HRD in CHORD-HRD patients pan-cancer and by cancer type. Left and right bars represent the HMF and PCAWG datasets respectively. The vertical split in the figure separates cancer types with (left side) and without (right side) ≥10 CHORD-HRD patients in at least one of the datasets. (**a**) Frequency of HRD. Cancer types where no frequency of HRD is displayed contain no data in either the HMF or PCAWG datasets (**b**) The gene deficiency associated with HRD. Bar segments are grouped into *BRCA2*-type HRD genes (*BRCA2*, *RAD51C*, *PALB2*) and *BRCA1*-type HRD genes (*BRCA1* only). (**c**) The likely combination of biallelic events in *BRCA1/2*, *RAD51C* or *PALB2* causing HRD. (**d**) Whether the genetic cause of HRD was purely due to somatic events, due to germline predisposition, or unknown. In (**c** and **d**), 'Unknown' and/or 'LOH + unknown' bar segments refer to patients where no clear biallelic loss of the aforementioned *BRCA1/2*, *RAD51C*, or *PALB2* was identified (i.e. clusters 4 and 6 of **Figure 3c**). LOH: loss-of-heterozygosity. Only data from samples that passed CHORD's QC criteria are shown in this figure (MSI absent, ≥50 indels, and ≥30 SVs if a sample was predicted HRD).

In 94% (292/310) of all CHORD-HRD patients, we found mono- or biallelic inactivation of at least one of the four HRD associated genes (*BRCA1*, *BRCA2*, *PALB2*, *RAD51C*; **Figure 4c**). In the case of biallelic inactivation, we observed LOH to be the dominant secondary event, occurring in combination with a germline SNV/indel (33%) or with a somatic SNV/indel (14%) of CHORD-HRD patients. LOH of *BRCA1/2*, *RAD51C* or *PALB2* was also found as a monoallelic event, mainly in ovarian (47%) and breast (49%) cancer patients (**Supplementary data 5**). As indicated earlier, the other allele may be

inactivated by epigenetic mechanisms in these patients (or alternatively HRD was caused by inactivation of another HR gene). Interestingly, we find that deep somatic deletions do frequently contribute to biallelic loss of *BRCA2* or *RAD51C*, occurring in 10% of CHORD-HRD patients pan-cancer (**Supplementary data 5**). However, deep deletions (primarily of *BRCA2*; **Supplementary figure 23**) occurred much more frequently in prostate cancer (33%) compared to other cancer types, consistent with previous observations [118]. Nevertheless, deep deletions of HRD genes did occur in every cancer type with a high frequency of CHORD-HRD patients indicating that complete somatic gene loss is a common and underestimated cause of HRD in both primary and metastatic cancer.

We find that biallelic gene loss is often associated with germline predisposition (**Figure 4d**) in ovarian (32%), breast (36%), and pancreatic (56%) cancer patients, but to a lesser extent in prostate cancer patients (24%) (**Supplementary data 5**). On the other hand, biallelic gene loss exclusively by somatic events occurs in sizable proportion of CHORD-HRD patients (35% pan-cancer), being most frequent in prostate cancer (54%) (**Supplementary data 5**) mainly due to the deep deletions (**Supplementary figure 23**). Although these frequencies may not be fully representative for each cancer type due to the proportion of patients with unknown mutation status in at least one allele (indicated as 'Unknown' in **Figure 4d**), these observations do emphasize that somatic-only events should not be overlooked as a mechanism of HR gene inactivation.

Discussion

Here we describe a classifier (CHORD) that can detect HRD (as well as HRD sub-phenotypes) across cancer types based on mutation profiles. By using this tool in a systematic pan-cancer analysis, we reveal novel insights into the mechanisms and incidence of HRD across cancer types with potentially important clinical relevance.

HRD targeted therapy with PARPi is mostly restricted to breast and ovarian cancer [102], though its use for treating pancreatic cancer was recently approved by the FDA (US Food and Drug Administration) [119]. However, we show that HRD is common not only in ovarian and breast cancer, but also in prostate, pancreatic cancer. The incidence of HRD was relatively higher in metastatic prostate and pancreatic cancer, and lower for ovarian and breast cancer as compared with primary tumors. This may reflect more intensive familial (germline) testing for *BRCA1/2* mutations in ovarian and breast cancer [120] and consequently earlier diagnosis and treatment with fewer cases of progression to metastatic cancer as a result. However, we cannot formally exclude that these observations originate from differences in cohort inclusion criteria that could skew numbers (e.g. due to more recruitment of patients with triple negative breast cancer which has higher HRD rates [106]).

We show that HRD is also found sporadically in cancer types other than breast, ovarian, prostate or pancreatic, but collectively this constitutes a sizable group of patients (15% of all patients). We do acknowledge that there may be underestimation of HRD frequency in these other cancer types due to the low prevalence of *BRCA1/2* deficient samples, which served as examples of HRD samples for training CHORD (**Supplementary figure 23**). On the other hand, we have shown that the HRD mutational footprint is not tissue type specific (**Supplementary figure 8**) suggesting that cancer type biases in the training set should not impact CHORD predictions. Our results thus indicate that a large number of patients who would potentially benefit from PARPi therapy still remain unnoticed. Since the mutational phenotype of HRD is independent of cancer type, mutational scar based HRD detection such as with CHORD would be valuable for cancer type agnostic patient stratification for future PARPi trials [121]. This is particularly important for metastatic patients (who depend on systemic treatments and benefit most from targeted treatments like PARPi), as well as for cancer types currently lacking good markers for patient stratification for such treatment (such as prostate [122] and biliary [123] cancer).

Genetic based detection of HRD in the clinic is commonly done by testing for pathogenic *BRCA1/2* germline mutations [102]. However, such hereditary mutations are only present in 30% of CHORD-HRD patients (**Supplementary notes; Supplementary figure 29**) indicating that germline testing likely misses a substantial number of HRD patients. Germline variant testing is particularly unsuitable for prostate cancer where gene inactivation is frequently caused by somatic deep deletions, which prevent the identification of any SNVs/indels at the affected locus when using panel- or PCR-based sequencing methods (exon scanning). This problem also exists for other cancer types where deep deletions also make up a non-negligible fraction of HR gene inactivation cases. While somatic mutation testing improves diagnostic yield and is indeed increasingly performed in the clinic [102], WGS based genetic testing is ultimately necessary to capture the full spectrum of genetic alterations and to accurately determine the mutational status of HR genes. However, even such broad genetic testing with focus on biallelic gene inactivation still potentially misses roughly 50% of all HRD patients (**Supplementary notes; Supplementary figure 29**).

We do acknowledge that mutational scars represent genomic history and not current on-going mutational processes that can result in false positive CHORD predictions, which could be for example due to reversion of HRD by secondary frameshifts [124,125], or recent acquisition of the HRD phenotype. False positive predictions could also arise from treatments producing similar mutational scars (in particular, microhomology deletions) to HRD. The most common cancer treatments have been shown to have little or no contribution to microhomology deletions, with the exception of radiotherapy [67,108,109]. However, we showed that radiotherapy itself likely does not lead to false positive predictions. We cannot exclude the possibility however that clonal expansion of a radiotherapy resistant tumor cell leads to sufficient enrichment of radiotherapy associated microhomology deletions in the tumor, resulting in a false positive prediction. Ultimately, the ability for CHORD to improve patient stratification and treatment outcome will need to be evaluated in direct comparisons and prospective clinical trials.

Thus, while CHORD can detect HRD independent of the underlying cause, genetic testing of HRD genes is complementary and can provide supporting information for making a final verdict on a patient's HR status. The unique advantage of using WGS, although not routine in clinical diagnostics yet, but likely in the near future [126], is that both genetic testing and mutational scar based HRD detection with CHORD can be performed simultaneously with the same assay. We envision that the findings from our analyses incentivizes improvements to current clinical practices for detecting HRD, and that the application of genomics-based approaches, like CHORD, in the clinic will support these endeavors and provide additional treatment options for patients. CHORD is freely available as an R package at https://github.com/UMCUGenetics/CHORD.

Methods

Datasets

We have used patient data for which re-use for cancer research was consented by the patients as part of two clinical studies (NCT01855477, NCT02925234) unrelated to the current work. Matched tumor/blood samples from these patients were sequenced and uniformly analyzed by the Hartwig Medical Foundation (HMF; https://www.hartwigmedicalfoundation.nl/en/appyling-for-data/). The data transfer agreement (Data Request 10 and 47) were approved by the medical ethical committees (METC) of the University Medical Center Utrecht. We received germline and somatic VCF files of the 3,824 metastatic tumor samples from 3,584 patients in May 2019. For patients with multiple biopsies that were taken at different timepoints, patient IDs were suffixed by 'A' for the first biopsy, 'B' for the

second biopsy, etc (e.g. HMF001423A, HMF001423B). A detailed description of the whole patient cohort has been described in detail in Priestley *et al.* 2019 [6].

Somatic variant TSV files of the 560 breast cancer (BRCA-EU) dataset were downloaded from the International Cancer Genome Consortium (ICGC; https://dcc.icgc.org/repositories) in August 2017. BAM files for the 44 BRCA-EU samples are available from EGA (datasets: EGAD00001000063 [https://www.ebi.ac.uk/ega/datasets/EGAD00001000063], EGAD00001001322 [https://www.ebi.ac.uk/ega/datasets/EGAD00001001322], EGAD00001001337 [https://www.ebi.ac.uk/ega/datasets/EGAD00001001337]). *BRCA1/2* status annotations for this dataset being obtained from the supplementary data in Davies *et al.* [74].

Somatic variant VCF files and somatic copy-number TSV files for the ICGC portion of the Pan-Cancer Analysis of Whole Genomes (PCAWG) dataset (consisting of 1,854 patient tumors) were downloaded from https://dcc.icgc.org/releases/PCAWG on March 3, 2020. PCAWG access for germline data has been granted via the Data Access Compliance Office (DACO) Application Number DACO-1050905 on October 6, 2017 and via https://console.cancercollaboratory.org download portal on December 4, 2017. Germline VCF files were downloaded from the cancer collaboratory download portal on March 21, 2018.

Variant calling

Variant calling in the HMF dataset was performed previously by HMF (https://github.com/hartwigmedical/pipeline) [6]. Briefly, reads were mapped to GRCh37 using BWA-MEM v0.7.5a with duplicates being marked for filtering. Indels were realigned using GATK v3.4.46 IndelRealigner. GATK Haplotype Caller v3.4.46 was used for calling germline variants in the reference sample. For somatic SNV and indel variant calling, GATK BQSR3 was first used to recalibrate base qualities, followed by Strelka v1.0.14 for the variant calling itself. Somatic SV calling was performed using GRIDSS v1.8.0. Copy-number calling was performed using PURity & PLoidy Estimator (PURPLE), that combines B-allele frequency (BAF), read depth and structural variants to estimate the purity and copy number profile of a tumor sample [127] as well as VAF and clonality (either clonal, subclonal or inaccurate) estimates of each variant.

Determining gene biallelic status

For samples in the HMF and PCAWG cohorts, biallelic status was determined for 781 genes (**Supplementary data 3**) which included genes associated with cancer, according to Cancer Genome Interpreter (https://www.cancergenomeinterpreter.org/genes), as well as a manually curated set of genes involved in HR (based on the KEGG HR pathway (https://www.genome.jp/), as well as via a literature search). This was performed using an in-house pipeline that interprets copy-number, and germline and somatic SNV/indel data from the HMF variant calling pipeline to determine biallelic gene status (https://github.com/UMCUGenetics/hmfGeneAnnotation).

First, the copy number status in the gene region was determined. If the minimum copy number was <0.3, the gene was considered to have a deep deletion (and by default biallelically inactivated). Else, the gene was screened for 2 mutation events, which included following combinations: (i) loss-of-heterozygosity (LOH) with a germline or somatic SNV/indel; (ii) a germline and somatic SNV/indel; or (iii) 2 somatic SNV/indels.

LOH was considered pathogenic and was automatically given a P-score of 5. LOH occurred if the minimum minor allele copy number within a gene region was <0.2.

Pathogenicity of SNVs/indels was assessed based on pathogenicity annotations from ClinVar (https://www.ncbi.nlm.nih.gov/clinvar/; GRCh37, database date 2020-02-24). For variants without an

entry in ClinVar, pathogenicity was assessed based on variant type as determined by SnpEff (http://snpeff.sourceforge.net/; v4.1h). Briefly, variants can be given one of the following annotations from ClinVar: pathogenic, likely pathogenic, variant of unknown significance (VUS), likely benign, and benign. A pathogenicity score (P-score) of 1-5 was also assigned to each annotation, with 1=benign and 5=pathogenic. Additionally, variant types as determined by SnpEff were assigned similar annotations and scores: out-of-frame frameshifts were considered pathogenic (P-score=5); nonsense and splice variants were considered likely pathogenic (P-score=4); missense variants, essential splice variants, and inframe frameshifts were considered VUS's (P-score=3); the remaining variant types were considered likely benign or benign (P-score ≤2). The final P-score of a variant was the ClinVar P-score if a ClinVar annotation exists for that variant, and if not, the SnpEff P-score was used. See **Supplementary data 6** for details on pathogenicity scoring.

P-scores from pairs of mutation events (i.e. SNV, indel, or LOH) were summed to yield a biallelic pathogenicity score (BP-score), giving a maximum possible score of 10. Deep deletions were automatically given a score of 10. Per gene, the biallelic event with the highest score was taken the biallelic status of the gene. If multiple events had the same score, a biallelic event was greedily selected.

Extracting mutation contexts

The counts of 3 types of mutation contexts (SNV, indel, and structural variant (SV) contexts) were determined from the somatic variant data from the HMF, PCAWG and BRCA-EU cohorts (**Supplementary data 2**). This was performed using the R package *mutSigExtractor* (https://github.com/UMCUGenetics/mutSigExtractor).

The SNV contexts comprised of 96 trinucleotide contexts, which are composed of one of six classes of base substitutions (C>A, C>G, C>T, T>A, T>C, T>G) in combination with the immediate 5' and 3' flanking nucleotides.

The indel contexts comprised of 6 types based on the presence of: short tandem repeats (ins.rep, del.rep); short stretches of identical sequence at the breakpoints, also known as microhomology (ins.mh, del.mh); or the presence of neither (ins.none, del.none). Indels in repeat regions were defined as the presence of ≥1 copy of the indel sequence downstream (i.e. in the 3' direction) from the breakpoint, where sequence length must be <50bp. Indels with flanking microhomology were defined as the presence of the following sequence features up or downstream from the breakpoint: (i) ≥1 copy of the indel sequence if the indel sequence length is ≥50bp; (ii) ≥2bp sequence identity to the indel sequence; or (iii) ≥1bp sequence identity if the indel sequence length is ≥3bp. For (ii) and (iii) the number of up or downstream bases searched was equal to the length of the indel. The 6 indel contexts types were further expanded into 30 indel contexts by stratifying ins.rep, del.rep, ins.none, and del.none by indel sequence length (1-4bp and ≥5bp); and ins.mh and del.mh by the number of bases in microhomology ('bimh'; 1-4bp and ≥5).

The 16 SV contexts were composed of the SV type (deletion, duplication, inversion, translocation) and the SV length (1-10kb, 10-100kb, 100kb-1Mb, 1Mb-10Mb, >10Mb). Note that SV length is not applicable for translocations.

Random forest training

Features

To construct the features for training the **C**lassifier of **HO**mologous **R**ecombination **D**eficiency (CHORD), the 96 trinucleotide contexts were simplified to 6 base substitution contexts by discarding the 5' and 3' flanking nucleotide information. For CHORD-del.mh.merged, the 30 indel contexts were

simplified to the 6 indel types. For CHORD and CHORD-signature, the del.mh indel type was split into 2 bins: del.mh with 1bp homology and 2 to ≥5 (i.e. equivalent to ≥5bp) homology (del.mh.bimh.1 and del.mh.bimh.2.5 respectively). Then, relative contribution was calculated for each feature per mutation context type (i.e. SNV, indel and SV contexts separately). For CHORD-signature, the 96 trinucleotide contexts were fitted to the 30 COSMIC SBS signatures [74] using the non-negative least squares algorithm (incorporated in *mutSigExtractor*). The SV contexts were fitted in the same manner to the 6 SV signatures [74]. The relative contribution of the SBS signatures, SV signatures, and indel contexts was then calculated per mutation type.

Training set

The training set consisted of samples which we could confidently consider *BRCA1/2* deficient or proficient based on the P-scores/BP-scores as described in *Determining gene biallelic status* and **Supplementary data 6**. *BRCA1/2* deficiency was defined as having a BP-score = 10. This includes samples with: (i) a deep deletion, (ii) LOH in combination with a pathogenic SNV/indel or an out-of-frame frameshift, or (iii) two pathogenic SNV/indels and/or or out-of-frame frameshifts. Within the *BRCA1/2* deficient group, samples where the absolute frequency of indels within repeat regions was >14000 were considered to have microsatellite instability (MSI) and were removed. This filtering step was done as the relative contribution of indels in repeat regions are grossly overrepresented in samples with MSI, thereby masking the contribution of microhomology deletions. This sample group ultimately consisted of 35 *BRCA1* (*'BRCA1'* class) and 89 *BRCA2* (*'BRCA2'* class) deficient samples which were both considered HRD during the training. Conversely, *BRCA1/2* proficiency required the following criteria: (i) Absence of deep deletions or LOH; (ii) all SNV/indels had a P-score ≤ 3 (VUS or lower in impact); (iii) for the highest impact pair of SNV/indels (i.e. highest BP-score), both variants had a P-score ≤ 3 (VUS or lower in impact). This *BRCA* proficient group (*'none'* class) consisted of 1,902 samples which were considered HRP during the training (**Supplementary figure 1**).

Training procedure

The training procedure for CHORD (as well as other models described in this study) is illustrated in **Supplementary figure 2**. A core training procedure, which performs feature selection and class resampling, forms the basis for the full training procedure (**Supplementary figure 2a**). Feature selection was done to retain mutation contexts which were significantly higher ($p < 0.01$, determined by one-tailed Wilcoxon tests) in *BRCA1/2* deficient versus proficient samples. Class resampling serves to reduce the difference in the number of samples between each class (i.e. class imbalances). Here, a grid search was performed to determine the optimal pair of the following parameters: (i) down-sampling of the *'none'* class: 1x (i.e. no down-sampling), 2x or 4x; (ii) up-sampling of the *'BRCA1'* class: 1x (i.e. no up-sampling), 1.5x or 2x. For each iteration of the grid search, 10-fold cross-validation (CV) was performed, after which the area under the precision-recall curve (AUPRC) was calculated. The parameter pair with the highest AUPRC was chosen. With the selected features and resampling parameters, a random forest was then trained that predicts the probability of a new sample being in one of the aforementioned 3 classes (i.e. *'BRCA1'*, *'BRCA2'* or *'none'*). We defined the HRD probability as the sum of the probability of belonging to the *'BRCA1'* and *'BRCA2'* classes, where a sample was considered HRD if the HRD probability was ≥0.5. Random forests were trained using the *randomForest* R package.

The full training procedure was split into 2 stages (**Supplementary figure 2b**). The first stage serves to filter *'BRCA1'* or *'BRCA2'* samples from the which are likely not HRD (e.g. due to reversal of biallelic inactivation via a second frameshift bringing the gene in frame), or *'none'* samples which are likely not HRP (e.g. due to deficiencies in other HR genes). Here, the core training procedure is encapsulated by a 10-fold CV loop to allow every sample to be excluded from the training set to subsequently calculate an unbiased HRD probability. This was repeated 100 times and the number of times each sample was

HRD or HRP was calculated. *'BRCA1'* or *'BRCA2'* samples that were predicted HRD < 60 times were blacklisted while *'none'* samples that were predicted HRD > 40 times were blacklisted. In the second training stage, the core training procedure was performed on a training set without the blacklisted samples. This yielded the final random forest model.

The performance of the final random forest model was assessed using 2 approaches: (i) 10-fold CV of the training set by further encapsulating the full training procedure in a 10-fold CV loop; (ii) applying the final random forest model to an external dataset (BRCA-EU dataset). An AUPRC was then calculated for both approaches. In the case of the BRCA-EU dataset, *BRCA1/2* deficiency annotations were obtained from Davies *et al.* 2017 [74]. All performance metrics were calculated using the *mltoolkit* R package (https://github.com/UMCUGenetics/mltoolkit).

Determining the genetic cause of HRD

To determine the genetic cause of HRD, tumors were first selected from the HMF cohort based on the absence of MSI, having ≥50 indels, and ≥30 SVs for HRD predicted samples (**Supplementary data 1**). Furthermore, for patients with multiple biopsies, a single tumor per patient was selected (based on the one with highest tumor purity). In total, 3504 tumors were selected (from the 3824 in total) to represent each patient. The following procedure was then employed for identifying biallelic loss in each of the 781 cancer/HR associated genes. First, high frequency germline SNV/indels (**Supplementary figure 18**) were marked as benign (P-score = 0). Then, each gene was screened for the following events: (i) a deep deletion; (ii) LOH in combination with a germline SNV/indel with a P-score ≥ 4 (likely pathogenic or higher in impact); (iii) LOH in combination with a somatic SNV/indel with a P-score ≥ 3 (VUS or higher in impact); or (iv) two SNVs/indels (germline + somatic, or 2x somatic) both with a P-score = 5 (pathogenic). See **Supplementary data 6** for details of the P-score thresholds used.

After applying CHORD to the HMF cohort, we then determined whether each of the 781 genes was significantly more frequently deficient in CHORD-HRD vs. CHORD-HRP patients using a one-tailed Fisher's exact test, with multiple testing correction performed with the Hochberg procedure using the p.adjust() function in R. This was done to determine the genes most likely to cause HRD when inactivated. Six genes were found with a q-value < 0.1 and had at least 5 patients with deficiency in the corresponding gene: *BRCA1, BRCA2, RAD51C, PALB2, NF1, and STARD13* (**Supplementary figure 19**). *NF1* and *STARD13* have not been reported to be involved in HR, and thus further analyses were performed to validate the enrichment for these 2 genes.

Since *BRCA1* and *NF1* are both located on Chr17, we reasoned that copy number alterations (CNA; in this case referring to deep deletions or LOH) that affect *BRCA1* also affect *NF1*. This leads to frequent biallelic loss of *NF1* even though the gene is likely not associated with HRD. A similar situation was suspected for *BRCA2* and *STARD13* which are both located on Chr13. Thus, one-tailed Fisher's exact tests were performed to determine whether CNAs in each of the 781 genes significantly co-occurred more often with a CNA in *BRCA1* or *BRCA2*. Multiple testing correction was performed using the Hochberg procedure. Indeed, enrichment in the co-occurrence of *BRCA1* and *NF1* CNAs was found, and was similarly the case for *BRCA2* and *STARD13* (**Supplementary figure 20**). We thus concluded that biallelic loss of *NF1* and *STARD13* are likely not associated with HRD and were therefore excluded from **Figure 3a**.

Clustering of CHORD-HRD samples

Clustering of CHORD-HRD samples based on biallelic inactivating events (as in **Figure 3c**) is illustrated in **Supplementary figure 25**. First, samples were split into 4 groups according to their HRD subtype and whether a sample had an impactful biallelic event (P-score pair of 5 and ≥3).

For each of these groups, the HRD associated gene with the max BP-score was greedily determined per sample and assigned a score of 1, with the remaining genes being assigned a score of 0. Genes were prioritized as follows *BRCA2*, *BRCA1*, *RAD51C*, *PALB2*. This was based on highest to lowest enrichment of gene deficiency in CHORD-HRD vs. CHORD-HRP as described above. With the resultant (1,0) matrix, a sorting operation was performed. A post-processing step (done purely for cosmetic purposes) ensured that samples with deep deletions, LOH + frameshift, and LOH + other SNV/indels in the corresponding gene were ranked first. The sorted (1,0) matrices from the 4 sample groups were combined, and consecutive rows of 1's were considered a cluster. For the 2 groups representing samples with no impactful biallelic event, all samples were considered to be in one cluster. These 2 groups corresponded to clusters 4 and 6 in **Figure 3c**, and samples in these clusters were considered to have an unknown cause of HRD.

For **Supplementary figure 23**, samples were first split by cancer type before performing the above procedure.

Code availability

CHORD is available as an R package at https://github.com/UMCUGenetics/CHORD (DOI: 10.5281/zenodo.4020925). The code used for data processing and generating the figures is also available in this repository.

Data availability

Metastatic WGS data and corresponding metadata have been obtained from the Hartwig Medical Foundation and provided under data request numbers DR-010 and DR-047. Both WGS data and metadata is freely available for academic use from the Hartwig Medical Foundation through standardized procedures and request forms can be found at https://www.hartwigmedicalfoundation.nl. WGS data for the 560 primary breast cancer (BRCA-EU) dataset and Pan-Cancer Analysis of Whole Genomes (PCAWG) primary cancer dataset are publically available from the International Cancer Genome Consortium (ICGC) (https://dcc.icgc.org/repositories; https://dcc.icgc.org/releases/PCAWG). For access to identifying data (e.g. germline or raw read data) for the PCAWG or BRCA-EU datasets, researchers will need to request access via the ICGC Data Access Compliance Office (DACO). All other data are available within the article, Supplementary Information or available from the authors upon request.

Acknowledgements

This publication and the underlying study have been made possible partly on the basis of the data that Hartwig Medical Foundation and the Center of Personalised Cancer Treatment (CPCT, The Netherlands) have made available to the study. We thank Neeltje Steeghs (Netherlands Cancer Institute), Martijn Lolkema (Erasmus Medical Center Rotterdam), Geert Cirkel (Meander Medical Center), Els Witteveen (UMC Utrecht), Mariette Labots (Amsterdam UMC, location VUmc) and Laurens Beerepoot (Elisabeth-TweeSteden Ziekenhuis, Tilburg) for study inclusion of a significant part of the patients that were used in this study and Peter Bouwman (Netherlands Cancer Institute, Amsterdam) for critically reading the manuscript. This work was financially supported by the gravitation program CancerGenomiCs.nl from the Netherlands Organisation for Scientific Research (NWO) and Oncode Institute to E.C.

Author contributions

L.N. performed analyses, wrote/edited the paper.
J.M. edited the paper and provided discussion.
A.v.H. conceived the study, performed analyses, wrote/edited the paper.
E.C. edited the paper and provided discussion.
E.C. and A.V.H. supervised the study. All authors proofread, made comments and approved the paper.

Competing interests

The authors declare no competing interests.

Supplementary data

Supplementary data 1: Predictions from CHORD and CHORD-signature on the HMF, BRCA-EU, and PCAWG datasets as well as metadata for each sample

Supplementary data 2: Mutation contexts and mutational signatures extracted from the HMF, BRCA-EU, and PCAWG datasets

Supplementary data 3: List of 781 cancer and HR related genes used for the pan-cancer analysis of HRD and results of the enrichment analysis to determine HRD associated genes. The enrichment analysis was performed using one-sided Fisher's exact tests with multiple testing correction using the Hochberg procedure

Supplementary data 4: Genotypes of *BRCA1/2, RAD51C, PALB2* and other HR genes for the 310 CHORD-HRD patients from the HMF and PCAWG datasets

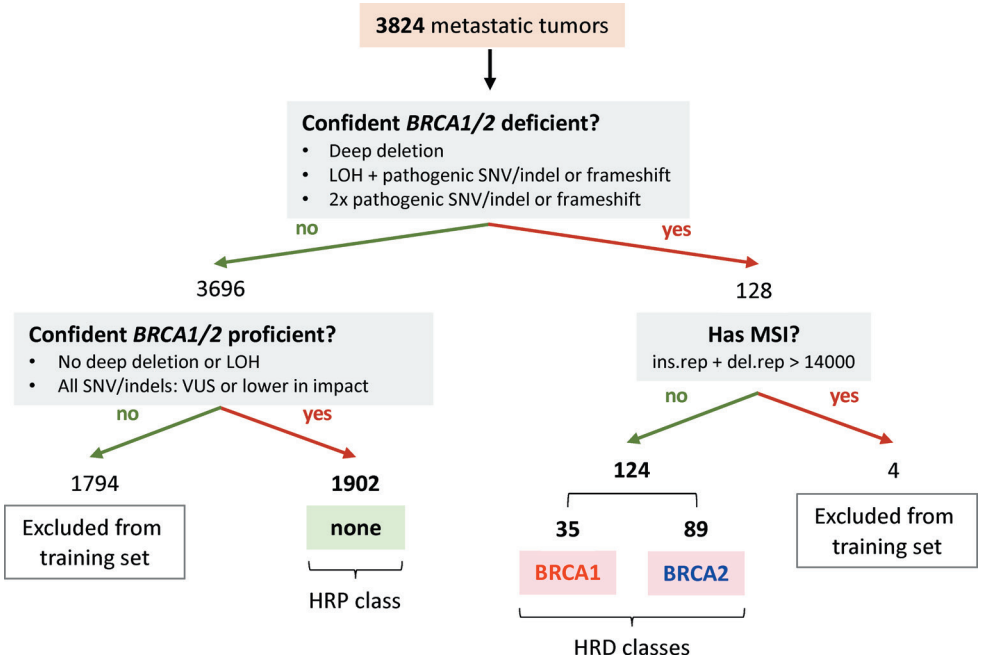
Supplementary data 5: Incidence and genetic cause of HRD by cancer type in the HMF and PCAWG dataset

Supplementary data 6: Pathogenicity scoring of variants used to determine biallelic gene status, including biallelic pathogenicity score inclusion criteria for CHORD training data

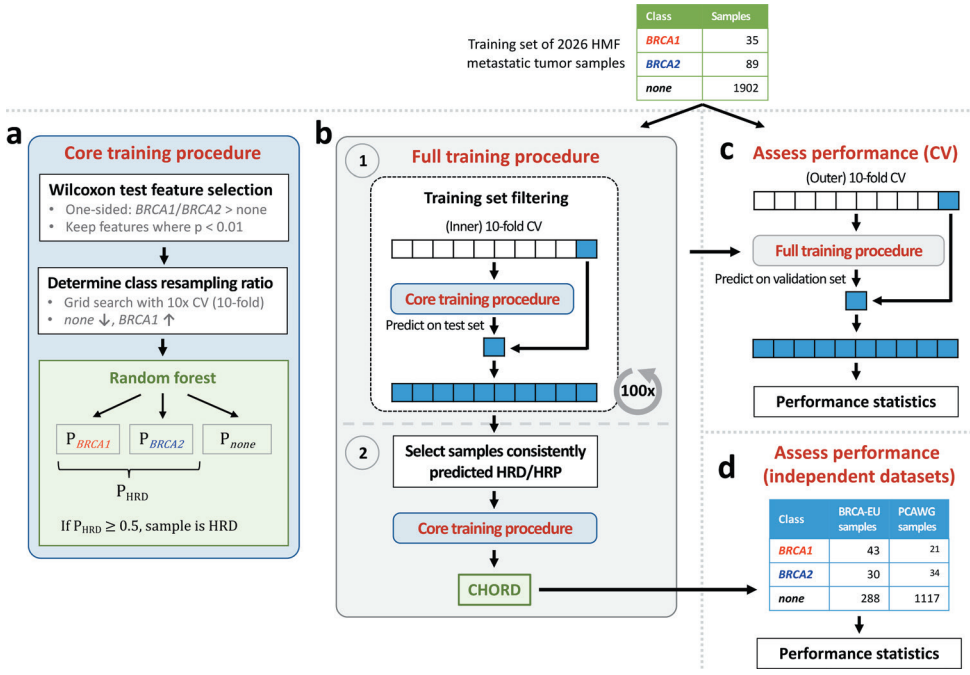
Supplementary data 7: Details for the novel, potentially pathogenic variants of unknown significance (VUS)

Supplementary data are available online at https://www.nature.com/articles/s41467-020-19406-4#Sec21, or by scanning the QR code below:

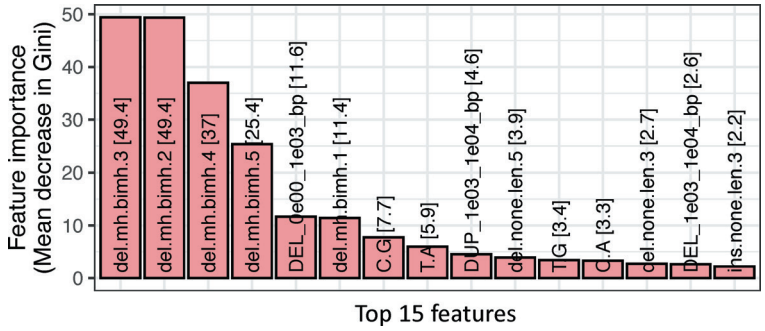
Supplementary figures



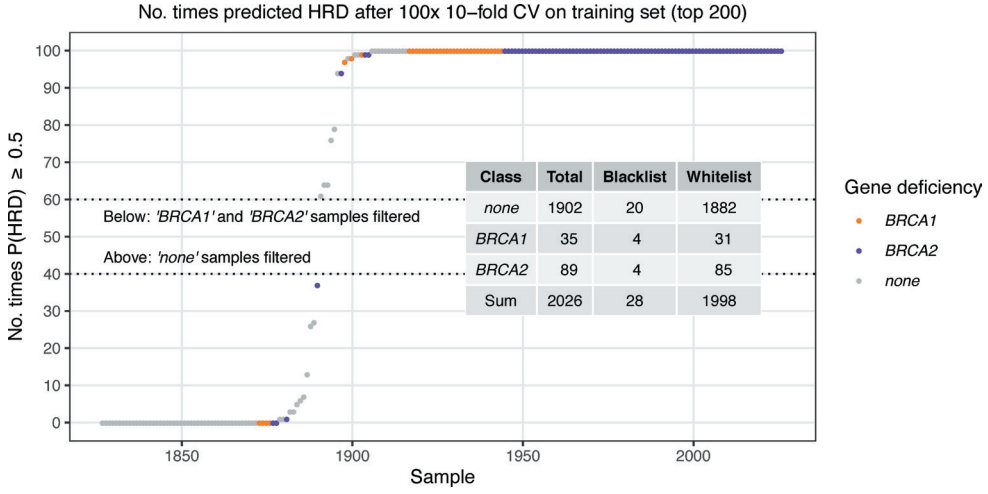
Supplementary figure 1: Sample selection for training CHORD. Only samples where biallelic inactivation of *BRCA1/2* could confidently be identified (and did not have MSI), or where no disruption of *BRCA1/2* was apparent were recruited into the training set. A total of 2,026 samples were eventually selected, consisting of 35 *BRCA1* and 89 *BRCA2* deficient which were both considered HRD during training, and 1,902 *BRCA1/2* proficient (*'none'*) which were considered HRP. LOH: loss-of-heterozygosity; VUS: variant of unknown significance.



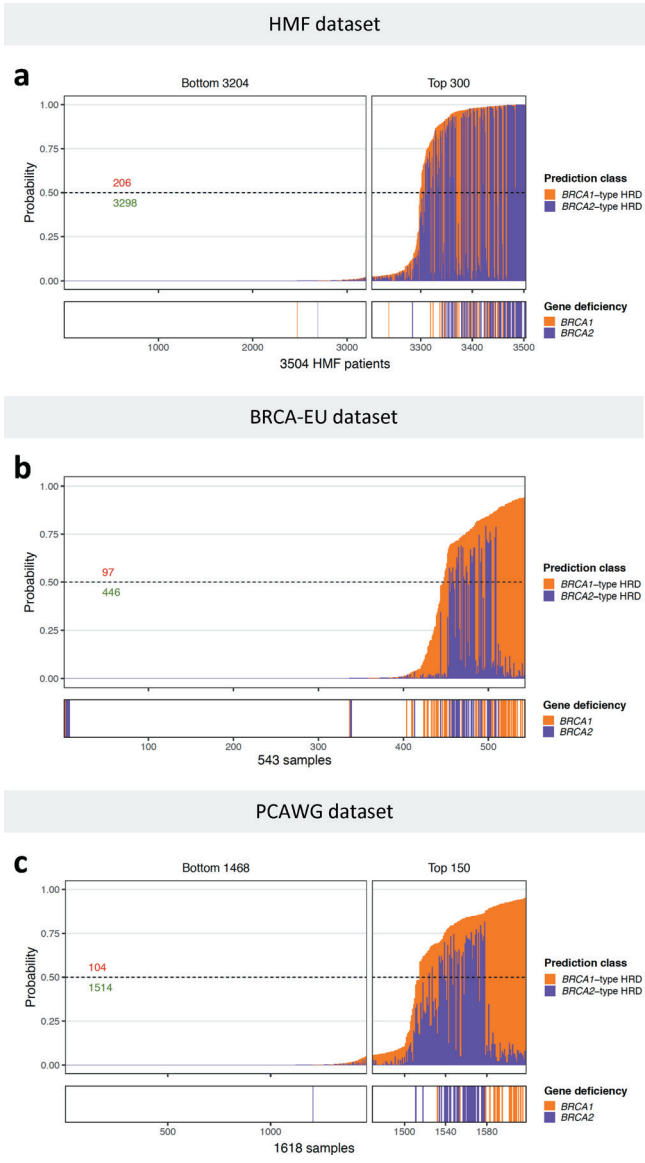
Supplementary figure 2: Workflow for training CHORD. *'BRCA1'*, *'BRCA2'* and *'none'* classes refer to *BRCA1* deficient, *BRCA2* deficient and *BRCA1/2* proficient sample groups, respectively. (**a**) The core training procedure on which the full training procedure is based on feature selection and class resampling. This returns a random forest that outputs the probability of a new sample being in one of the aforementioned 3 classes. The probability of HRD (P_{HRD}) is the sum of the probability of belonging to the *'BRCA1'* and *'BRCA2'* classes, where a sample is considered HRD if P_{HRD} is ≥ 0.5. (**b**) The full training procedure is split into 2 stages. The first stage serves to blacklist *'BRCA1'* or *'BRCA2'* class samples which are likely not HRD (e.g. due to reversal of biallelic inactivation via a secondary frameshift), or *'none'* samples which are likely not HRP (e.g. due to deficiencies in other HR genes). Here, the core training procedure is encapsulated by a 10-fold cross-validation (CV) loop to allow every sample to be excluded from the training set to subsequently calculate an unbiased P_{HRD}. This was repeated 100 times and the number of times each sample was HRD or HRP was calculated. *'BRCA1'* or *'BRCA2'* samples that were predicted HRD < 60 times were blacklisted. *'none'* samples that were predicted HRD > 40 times were blacklisted. In the second training stage, the core training procedure was performed on a training set without the blacklisted samples. This produced the final random forest model, CHORD. (**c**) The full training procedure was further encapsulated by a 10-fold CV to assess the performance of CHORD. (**d**) Performance was also assessed by applying CHORD on two independent datasets: BRCA-EU (371 primary breast tumors) and PCAWG (1176 primary tumors, pan-cancer).



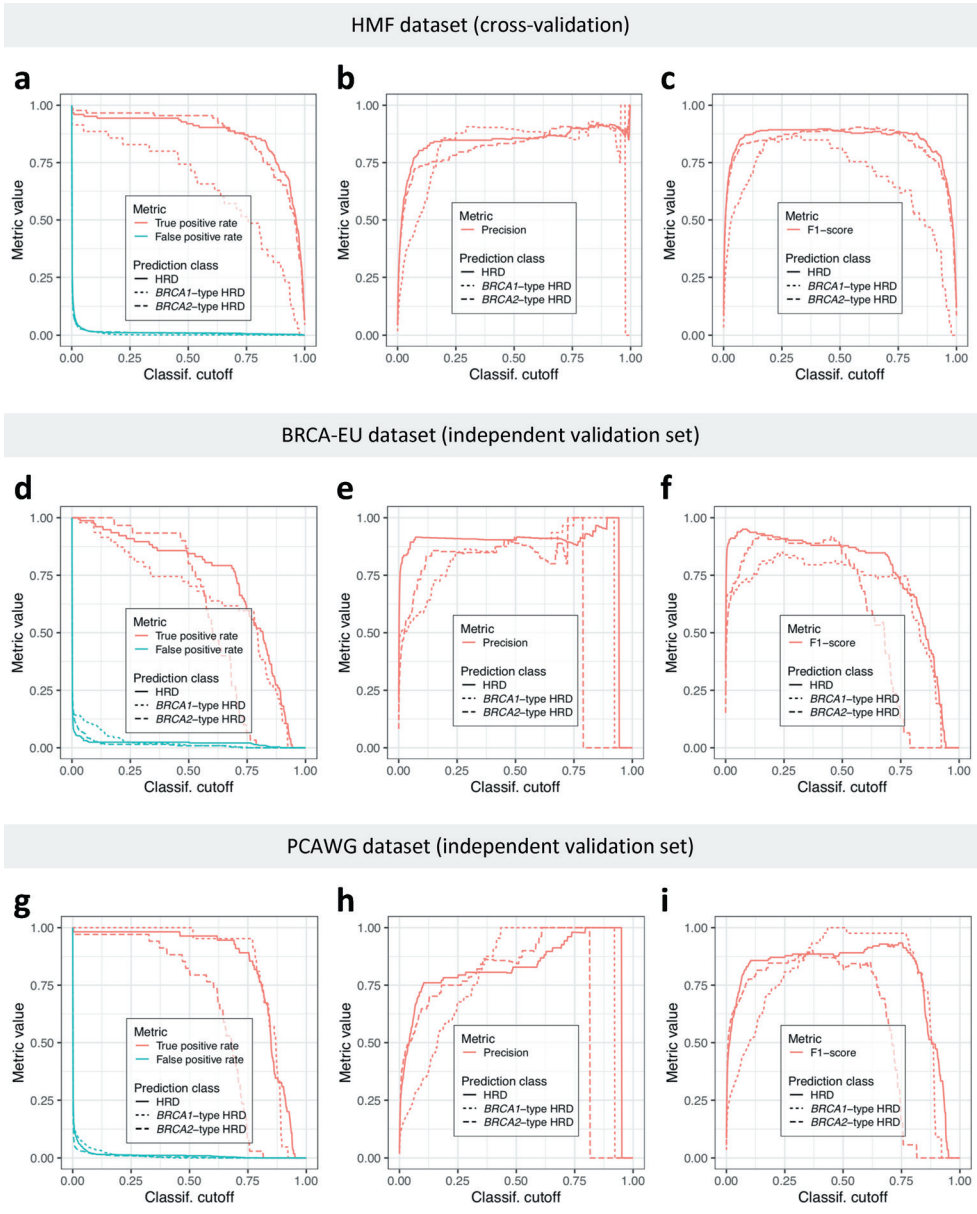
Supplementary figure 3: Deletions with ≥2bp flanking homology were most predictive of *BRCA1/2* deficiency. A random forest was applied to the training set and the importance of each feature was quantified using the mean decrease in Gini. The features used for training are relative counts of different mutation contexts which fall into one of three groups based on mutation type. (i) Single nucleotide variants (SNV): 6 possible base substitutions (C>A, C>G, C>T, T>A, T>C, T>G). (ii) Indels: indels with flanking microhomology stratified by homology length (del.mh, ins.mh), within repeat regions (del.rep, del.none), or not falling into either of these 2 categories (del.none, ins.none). (iii) Structural variants (SV): SVs stratified by type and length.



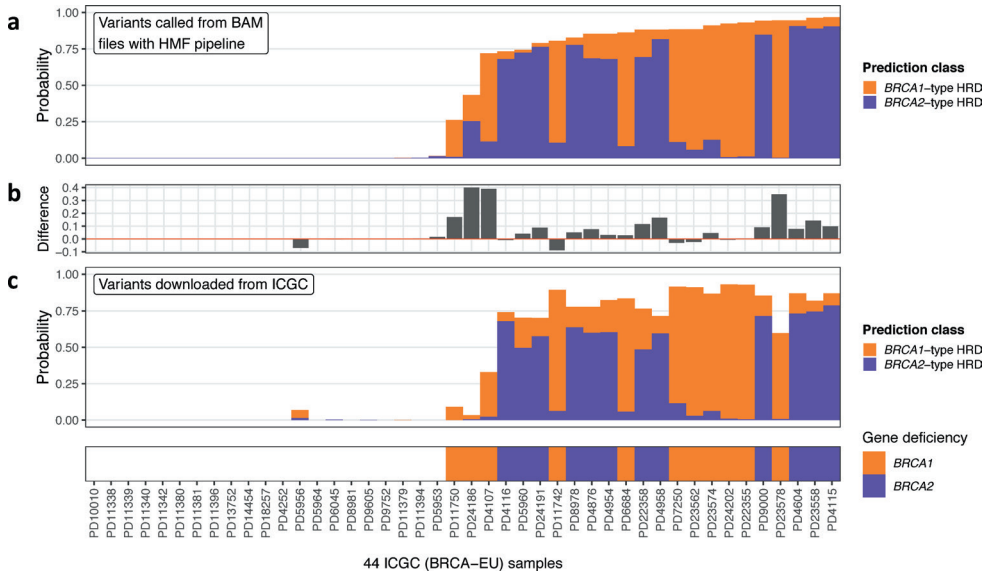
Supplementary figure 4: **Results of the 100x repeated 10-fold CV procedure for filtering CHORD training samples.** This was performed to blacklist *'BRCA1'* or *'BRCA2'* class samples that were consistently predicted HRP (no. times HRD < 60), or *'none'* class samples that were consistently predicted HRD (no. times HRD > 40). Samples with a probability of HRD ≥ 0.5 were considered HRD. The overlaid table summarizes the number of samples per class before and after removing the blacklisted samples from the training set. Note that only the top 200 samples are shown here as the remaining samples belonged to the *'none'* class and were always predicted HRP.



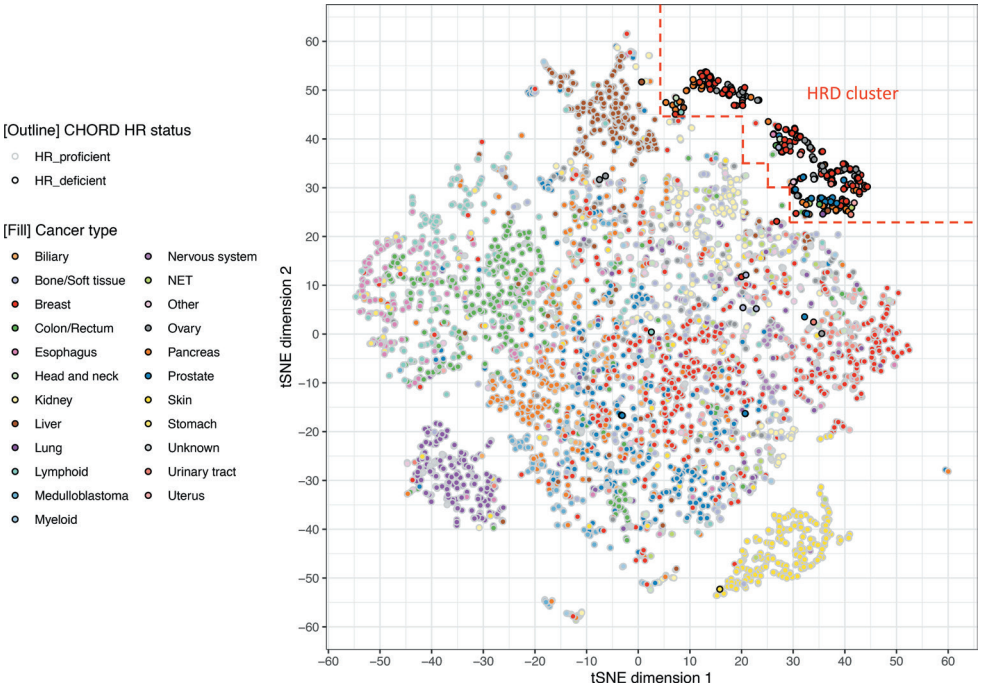
Supplementary figure 5: CHORD predictions on all HMF, BRCA-EU and PCAWG patients. Predictions on the HMF, BRCA-EU and PCAWG datasets are shown in **a**, **b** and **c**, respectively. The probability of HRD for each sample (total bar height) with each bar being divided into segments indicating the probability of *BRCA1*- (orange) and *BRCA2*-type HRD (purple). Stripes below the bar plot indicate biallelic loss of *BRCA1* or *BRCA2*. Only samples that passed CHORD's QC criteria are shown (MSI negative, ≥50 indels, and ≥30 SVs if a sample was predicted HRD).



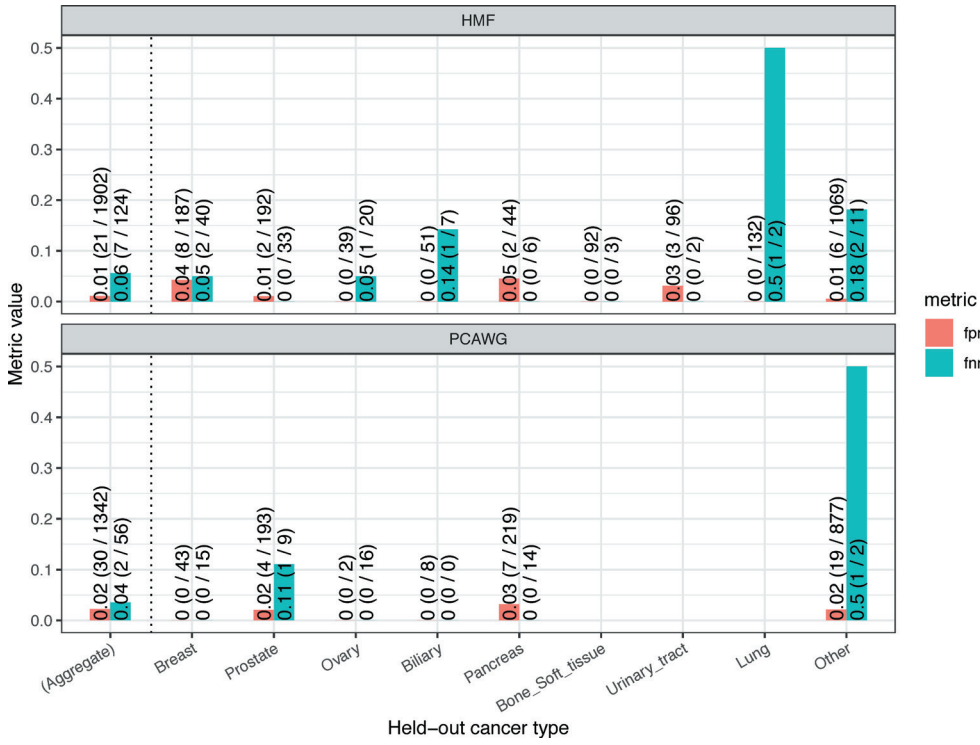
Supplementary figure 6: Additional performance metrics for CHORD. Performance was determined by 10-fold cross-validation (CV) on the HMF training data (**a-c**) or prediction on two independent datasets, BRCA-EU (primary breast cancer dataset; **d-f**) and PCAWG (primary pan-cancer dataset; **g-i**). Data from the BRCA-EU and PCAWG datasets are from samples that passed CHORD's QC criteria (i.e. MSI absent, ≥50 indels, ≥30 SVs if a sample was predicted HRD).



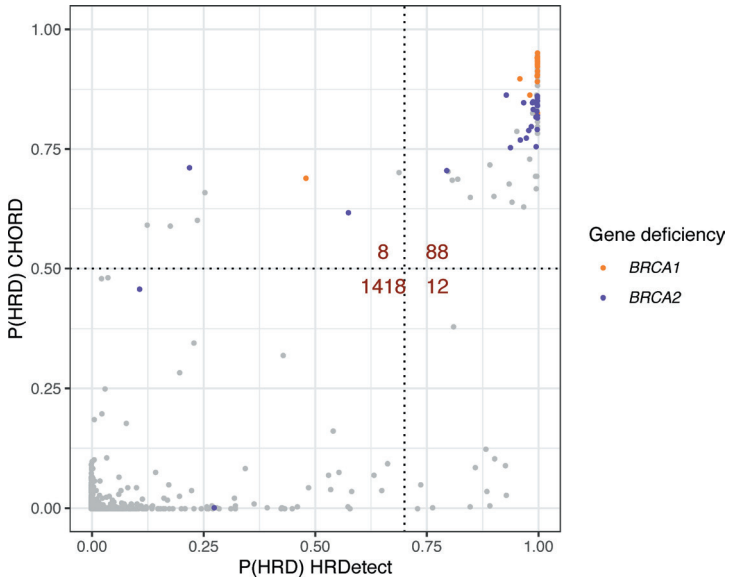
Supplementary figure 7: Variant calling pipeline differences affects CHORD performance. (**a**) Using variants called with the native pipeline of the HMF dataset (HMF pipeline) for HRD prediction with CHORD resulted in overall higher HRD probabilities in *BRCA1/2* deficient tumors when compared to (**c**) using variants downloaded from ICGC. The differences in HRD probabilities are quantified in (**b**). All samples shown here passed CHORD's QC criteria (i.e. MSI absent, ≥50 indels, ≥30 SVs if a sample was predicted HRD).



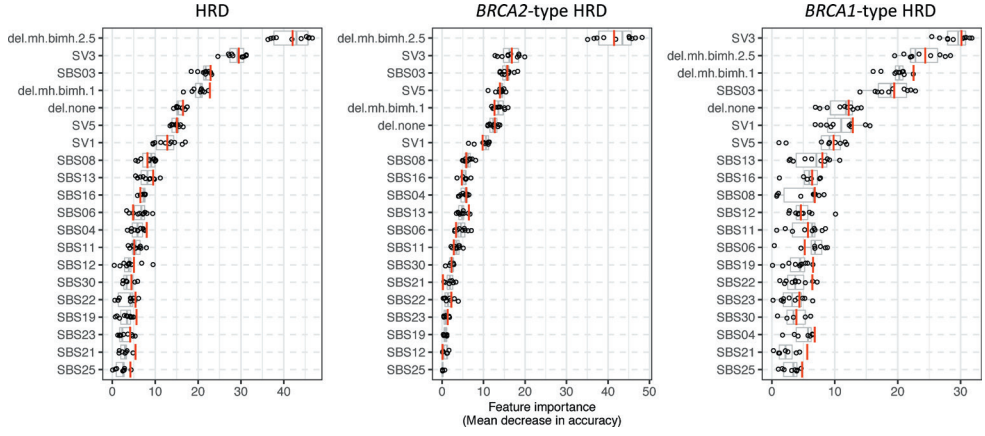
Supplementary figure 8: Clustering of samples (HMF, n=2026; PCAWG, n=1854) by t-distributed stochastic neighbor embedding (t-SNE) on the features used as input for CHORD. The dashed red lines are a manual annotation of the HRD cluster. All samples shown here passed CHORD's QC criteria (i.e. MSI absent, ≥50 indels, ≥30 SVs if a sample was predicted HRD).



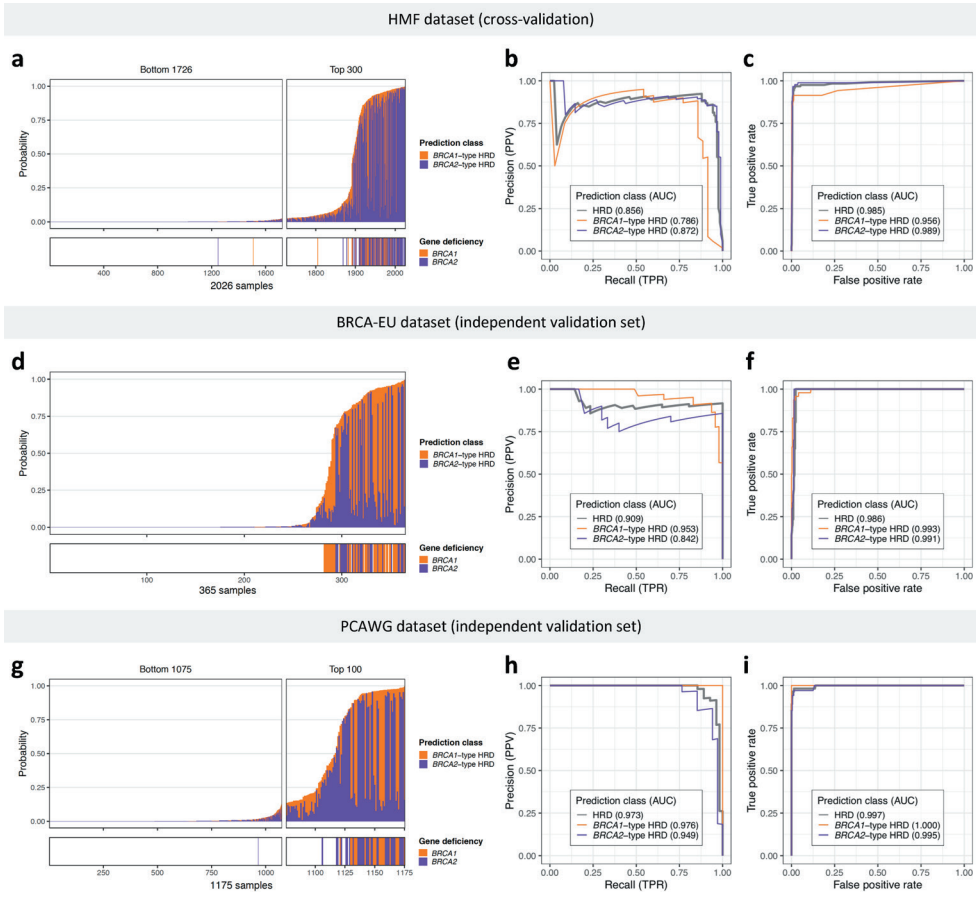
Supplementary figure 9: Performance of random forests trained when HMF samples belonging to each cancer type were held out. Training was performed in the same manner as for CHORD. These models were then applied to the held out HMF samples as well as the PCAWG samples to calculate performance metrics (at a classification threshold of 0.5). Only samples where *BRCA1/2* deficiency status could be confidently determined were included in this analysis (HMF, n=2026; PCAWG, n=1854). False positive rate (FPR) = number of *BRCA1* or *BRCA2* proficient samples misclassified as HR proficient / total number of *BRCA1* or *BRCA2* proficient samples. False negative rate (FNR) = number of *BRCA1* or *BRCA2* deficient samples misclassified as HR deficient / total number of *BRCA1* or *BRCA2* deficient samples. 'Aggregate' refers to the aggregated FPR and FNR across all cancer types. 'Other' refers to cancer types with <2 *BRCA1/2* deficient samples.



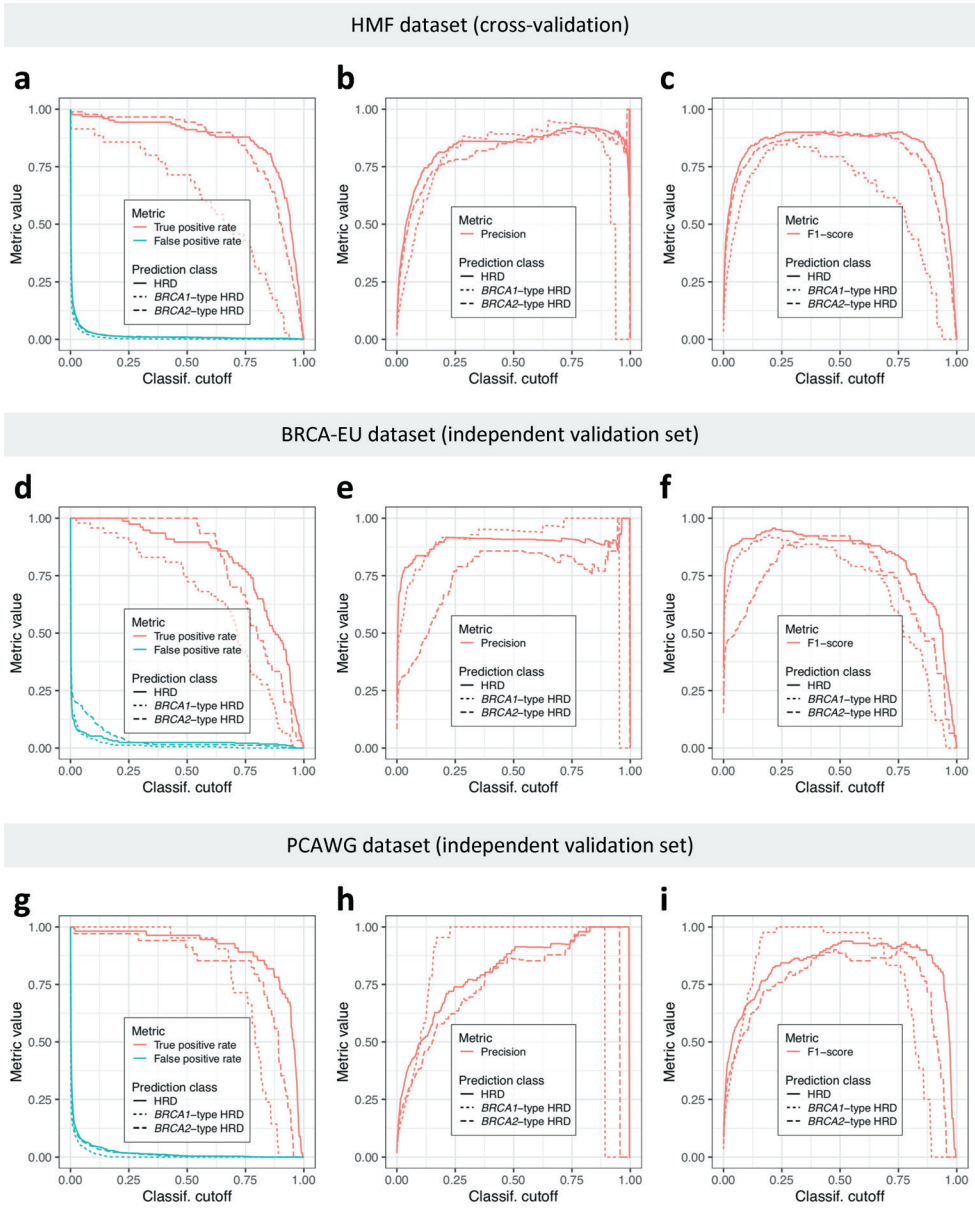
Supplementary figure 10: CHORD vs HRDetect predictions on the PCAWG dataset. The 1526 samples shown here had predictions from HRDetect and passed CHORD's QC criteria for predicting HRD (i.e. MSI absent, ≥50 indels, ≥30 SVs if a sample was predicted HRD). The dotted lines indicate the classification thresholds for the two models (CHORD: 0.5, HRDetect: 0.7). P(HRD): probability of HRD.



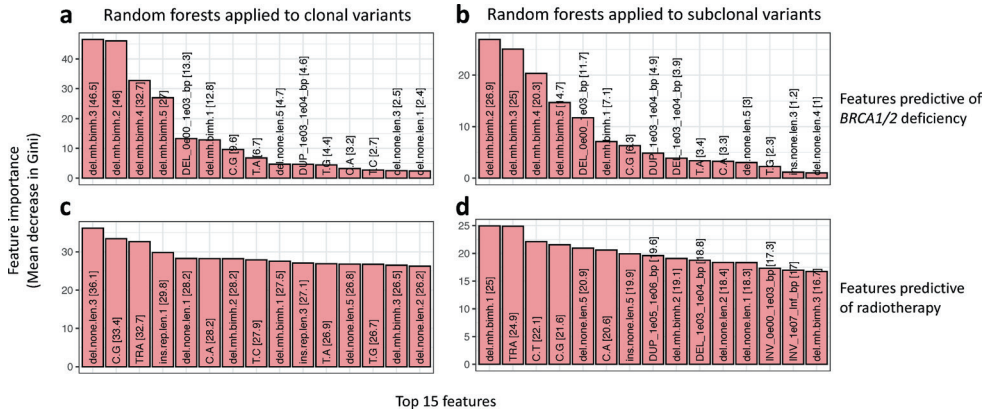
Supplementary figure 11: The features used by CHORD-signature to predict HRD as well as *BRCA1*-type HRD and *BRCA2*-type HRD. Importance is indicated by mean decrease in accuracy. SV# refers to the 6 SV mutational signatures and SBS# refers to the 30 single base substitution signatures as used by HRDetect. Indels are stratified into those with flanking microhomology further stratified by homology length (del.mh, ins.mh), within repeat regions (del.rep, del.none), or not falling into either of these 2 categories (del.none, ins.none). Boxplot and dots (n=10) show the feature importance over 10-folds of nested CV on the training set, with the red line showing the feature importance in the final CHORD model. Boxes show the interquartile range (IQR) and whiskers show the largest/smallest values within 1.5 times the IQR.



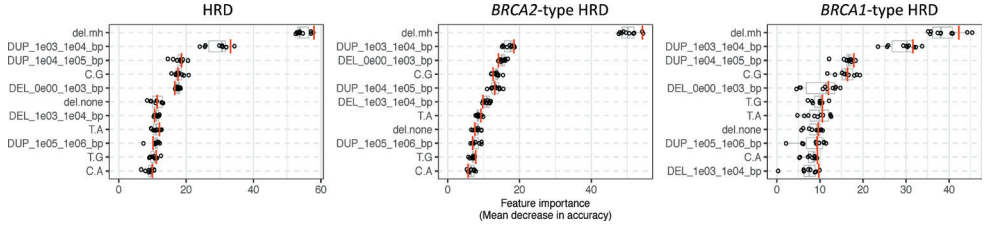
Supplementary figure 12: Performance of CHORD-signature. Performance was determined by 10-fold cross-validation (CV) on the HMF training data or prediction on two independent datasets: BRCA-EU (primary breast cancer dataset) and PCAWG (primary pan-cancer dataset). BRCA-EU and PCAWG samples shown here all passed CHORD's QC criteria (i.e. MSI absent, ≥50 indels, ≥30 SVs if a sample was predicted HRD) (**a, d, g**) The probability of HRD for each sample (total bar height) with each bar being divided into segments indicating the probability of *BRCA1*- (orange) and *BRCA2*-type HRD (purple). Stripes below the bar plot indicate biallelic loss of *BRCA1* or *BRCA2*. In (**a**), probabilities have been aggregated from the 10 CV folds. (**b, e, h**) Receiver operating characteristic (ROC) and (**c, f, i**) precision-recall curves (PR) and respective area under the curve (AUC) values showing the performance of CHORD when predicting HRD as a whole (grey), *BRCA1*-type HRD (orange), or *BRCA2*-type HRD (purple).



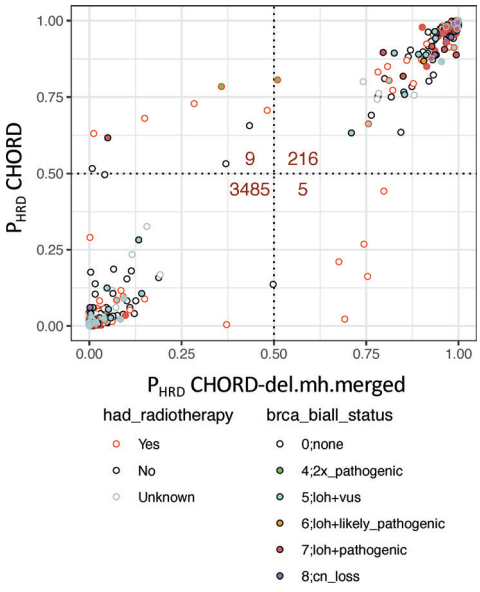
Supplementary figure 13: Additional performance metrics for CHORD-signature. Performance was determined by 10-fold cross-validation (CV) on the HMF training data (**a-c**) or prediction on two independent datasets, BRCA-EU (primary breast cancer dataset; **d-f**) and PCAWG (primary pan-cancer dataset; **g-i**). Data from the BRCA-EU and PCAWG datasets are from samples that passed CHORD's QC criteria (i.e. MSI absent, ≥50 indels, ≥30 SVs if a sample was predicted HRD).



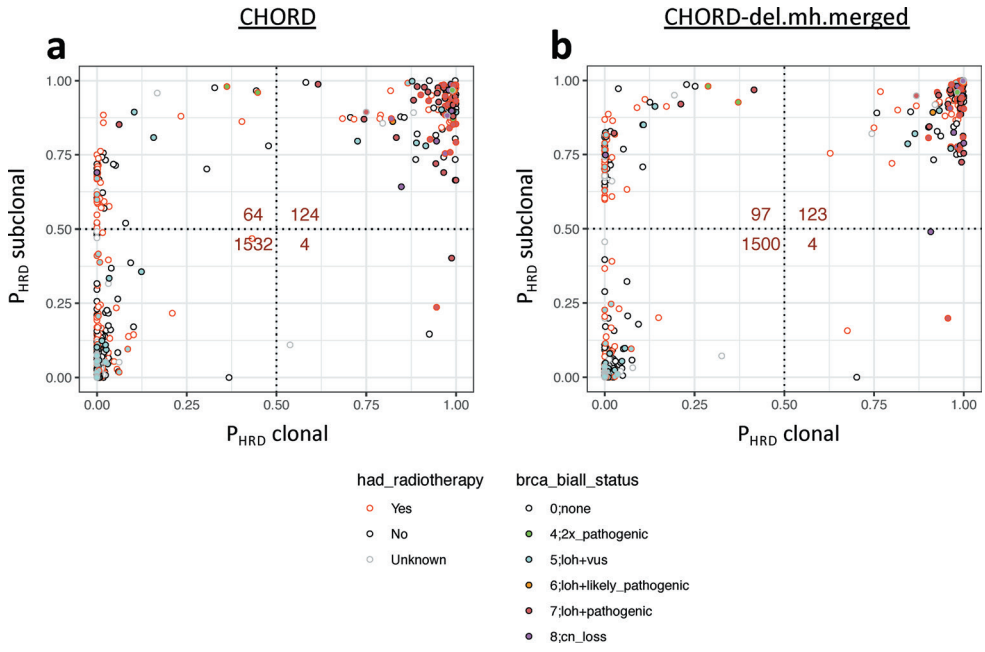
Supplementary figure 14: Random forests were used to identify the features predictive of *BRCA1/2* deficiency (a,b) compared to those predictive of radiotherapy (c,d) when using clonal versus subclonal variants as input. Feature importance is indicated by mean decrease in Gini, with only the top 15 important features being shown. The features used for training are relative counts of different mutation contexts which fall into one of three groups based on mutation type. (i) Single nucleotide variants (SNV): 6 possible base substitutions (C>A, C>G, C>T, T>A, T>C, T>G). (ii) Indels: indels with flanking microhomology stratified by homology length (del.mh, ins.mh), within repeat regions (del.rep, del.none), or not falling into either of these 2 categories (del.none, ins.none). (iii) Structural variants (SV): SVs stratified by type and length. Deletions with 1bp of flanking homology (del.mh.bimh.1) was more associated with radiotherapy especially in the subclonal fraction, while deletions with ≥2bp flanking homology (del.mh.bimh.2.5) was more associated with *BRCA1/2* deficiency.



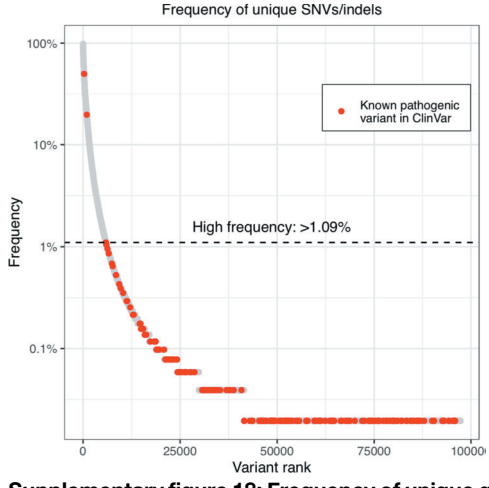
Supplementary figure 15: The features used by CHORD-del.mh.merged to predict HRD as well as *BRCA1*-type HRD and *BRCA2*-type HRD. Importance is indicated by mean decrease in accuracy. The features used for training are relative counts of different mutation contexts which fall into one of three groups based on mutation type. (i) Single nucleotide variants (SNV): 6 possible base substitutions (C>A, C>G, C>T, T>A, T>C, T>G). (ii) Indels: indels with flanking microhomology stratified by homology length (del.mh, ins.mh), within repeat regions (del.rep, del.none), or not falling into either of these 2 categories (del.none, ins.none). (iii) Structural variants (SV): SVs stratified by type and length. Deletions with flanking microhomology (del.mh) was the most important feature for predicting HRD as a whole, with 1-100kb structural duplications (DUP_1e03_1e04_bp, DUP_1e04_1e05_bp) differentiating *BRCA1*-type HRD from *BRCA2*-type HRD. Boxplot and dots (n=10) show the feature importance over 10-folds of nested CV on the training set, with the red line showing the feature importance in the final CHORD model. Boxes show the interquartile range (IQR) and whiskers show the largest/smallest values within 1.5 times the IQR.



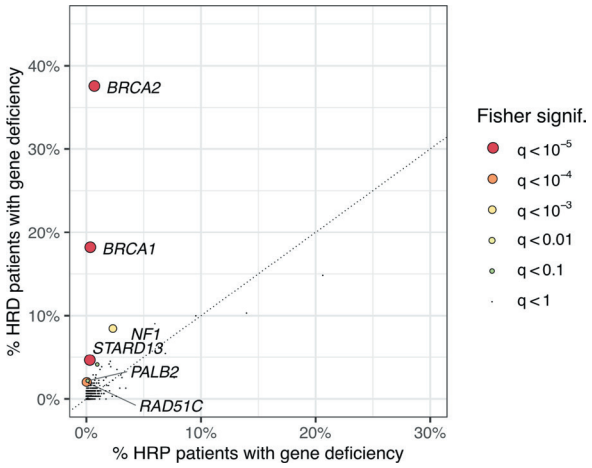
Supplementary figure 16: HRD predictions (on HMF samples) from CHORD and CHORD-del.mh.merged on all variants. Only samples passing CHORD's QC criteria were shown (MSI negative, ≥50 indels, and ≥30 SVs if a sample was predicted HRD). Dots with a red outline indicate samples which had radiotherapy prior to the tumor biopsy. Dot fill color indicates the biallelic status of *BRCA1* or *BRCA2*. LOH: loss-of-heterozygosity; VUS: variant of unknown significance; P_{HRD}: probability of HRD.



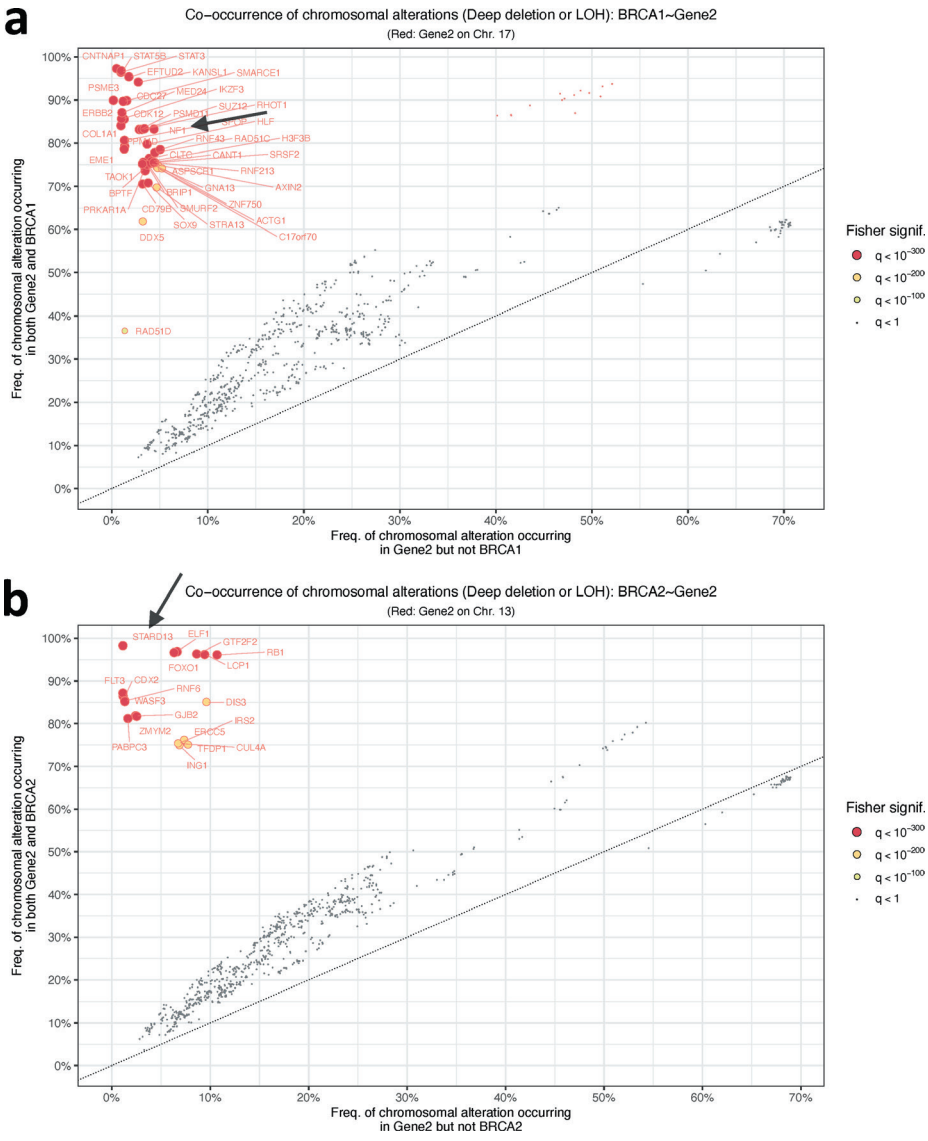
Supplementary figure 17: HRD predictions (on HMF samples) on subclonal versus clonal variants and from (a) CHORD and (b) CHORD-del.mh.merged. Only samples passing CHORD's QC criteria were shown (MSI negative, ≥50 indels in both clonal and subclonal fractions, and ≥30 SVs if a sample was predicted HRD). Dots with a red outline indicate samples which had radiotherapy prior to the tumor biopsy. Dot fill color indicates the biallelic status of *BRCA1* or *BRCA2*. LOH: loss-of-heterozygosity; VUS: variant of unknown significance; P_{HRD}: probability of HRD.



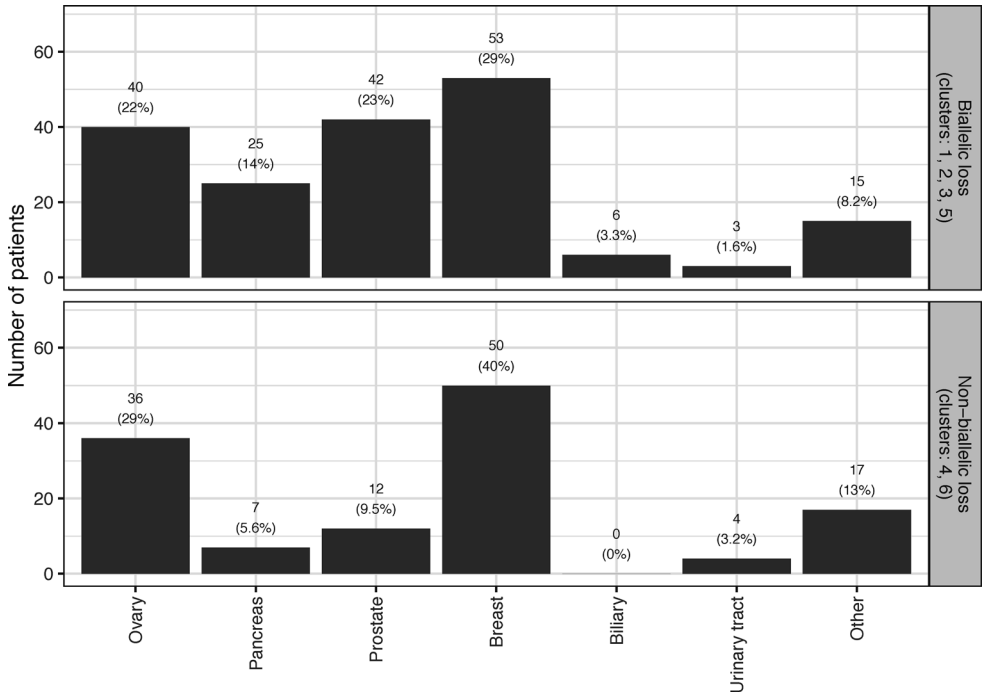
Supplementary figure 18: Frequency of unique germline SNVs/indels of 781 cancer related genes in patients of the HMF cohort. Germline variants with a frequency >1.09% were marked as benign prior to performing the pan-cancer analysis of HRD. Two known pathogenic variants above this frequency were also considered benign.



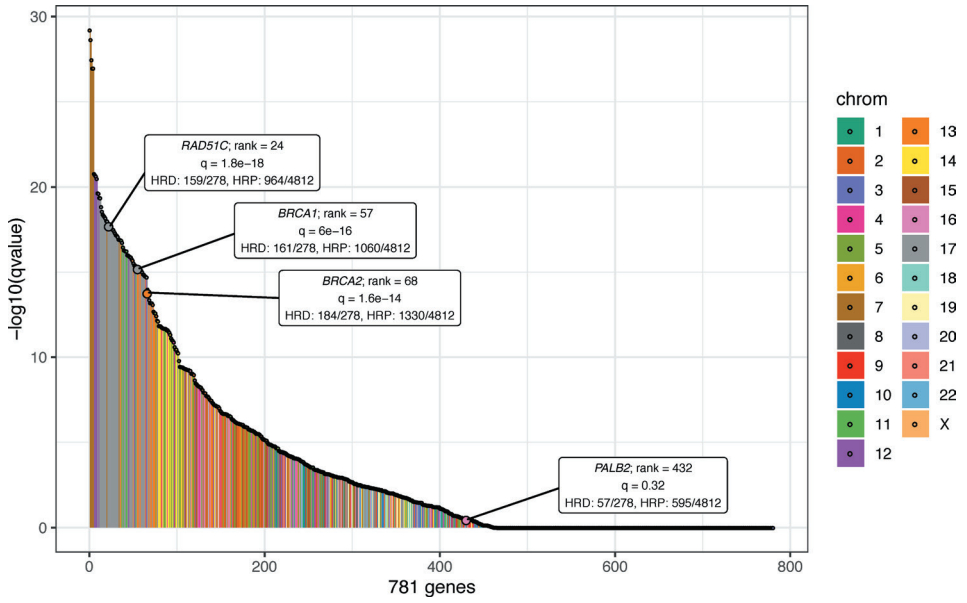
Supplementary figure 19: **A one-tailed Fisher's exact test identified enrichment of *BRCA1, BRCA2, RAD51C* and *PALB2* biallelic inactivation in CHORD-HRD vs. CHORD-HRP patients.** Each point represents a gene (from a list of 781 cancer and HR related genes) with its size/color corresponding to the statistical significance as determined by the Fisher's exact test, with axes indicating the percentage of patients (within either the CHORD-HRD or CHORD-HRP group) in which biallelic inactivation was detected. Multiple testing correction was performed using the Hochberg procedure. This figure is the same as **Figure 3b** except with *NF1* and *STARD13* included. A one-tailed Fisher's exact test determined *BRCA1* (q=9.4e-51), *BRCA2* (q=4.8e-101), *RAD51C* (q=5.6e-5), *PALB2* (q=0.02), *NF1* (q=3.5e-7) and *STARD13* (q=3.0e-8) to be significantly enriched (from a list of 781 cancer related genes) in CHORD-HRD vs. CHORD-HRP patient groups. Each point represents a gene with its size/color corresponding to the statistical significance as determined by the Fisher's exact test, with axes indicating the percentage of patients (within either the CHORD-HRD or CHORD-HRP group) in which biallelic inactivation was detected. Multiple testing correction was performed using the Hochberg procedure.



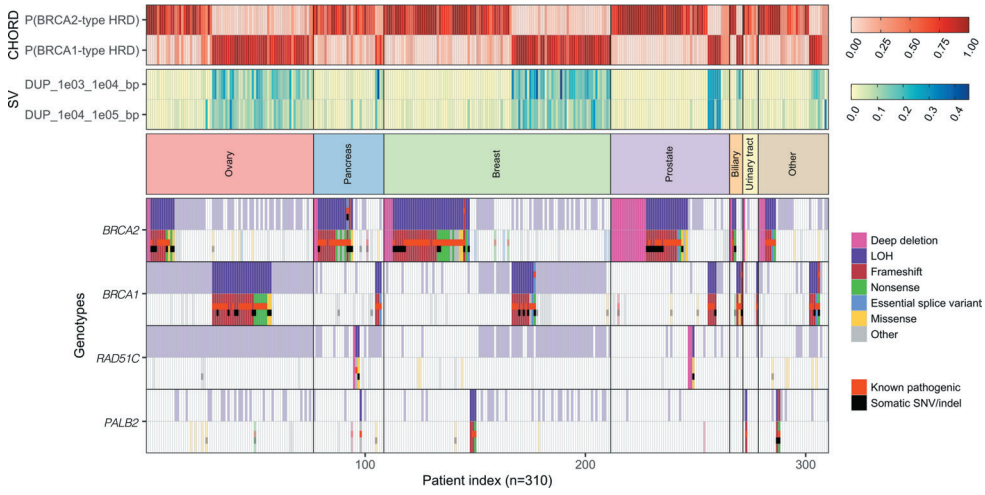
Supplementary figure 20: Chromosomal alterations (deep deletions or loss of heterozygosity (LOH)) affecting *BRCA1* and *BRCA2* also affects nearby genes including *NF1* and *STARD13* respectively. (**a**) Enrichment of *NF1* (q<1e-300) biallelic loss as shown in **Supplementary figure 19** is likely due to the gene being within proximity of *BRCA1* and not because the gene is associated with HRD, since *NF1* is not considered to be involved in HR in literature. The size/color of each point on the plot represents the significance of enrichment of CNAs occurring both in *BRCA1* and in each of the 781 genes (one vs. all comparison). This was determined by a one-tailed Fisher's exact test, where multiple testing correction was performed using the Hochberg procedure. Genes residing on the same chromosome as *BRCA1* were marked with a red outline and text. (**a**) Similarly, as with (**b**), a chromosomal alteration affecting *BRCA2* also affects the nearby gene *STARD13* (q<1e-300). Enrichment of *STARD13* biallelic loss as shown in **Supplementary figure 19** is likely due to the gene being within proximity of *BRCA2* and not because the gene is associated with HRD.



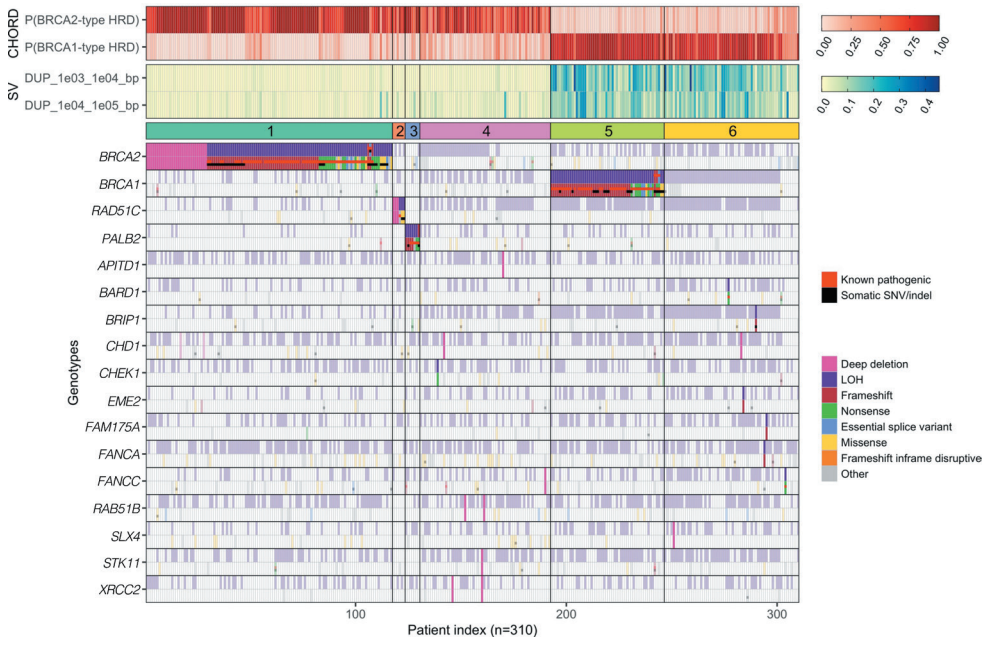
Supplementary figure 21: The number of CHORD-HRD patients of each cancer type which did (top) and did not (bottom) have biallelic loss of *BRCA1*, *BRCA2*, *RAD51C*, or *PALB2*. The top panel corresponds to patients in cluster 1,2,3 and 5 of **Figure 3c** while the bottom panel corresponds to patients in cluster 4 and 6. The percentages shown are within-group proportions.



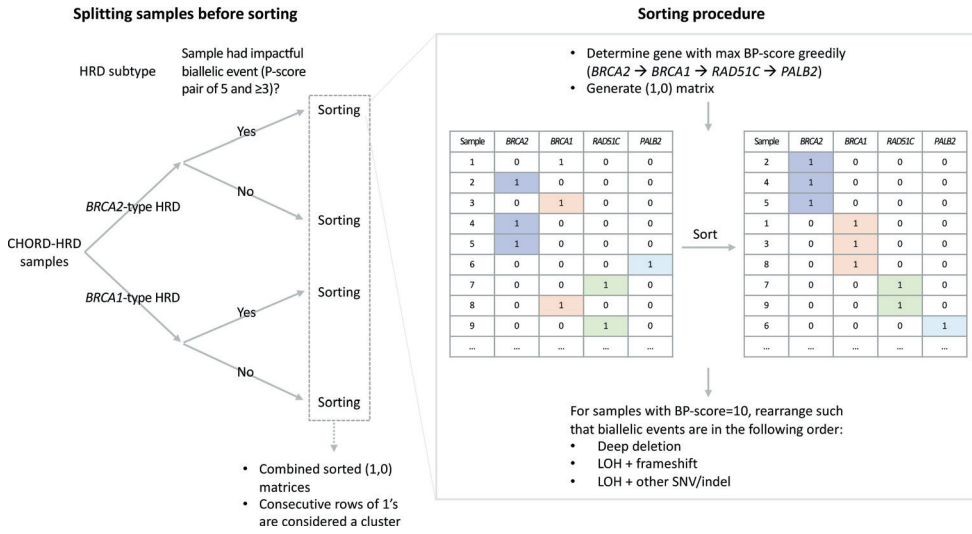
Supplementary figure 22: Enrichment of LOH in the 781 HR/cancer related genes. Enrichment for each gene was determined between CHORD-HRD samples (excluding those with deep deletions in *BRCA1/2*, *RAD51C* or *PALB2*) and CHORD-HRP samples using one-sided Fisher's exact tests.



Supplementary figure 23: Biallelic status of *BRCA2*, *BRCA1*, *RAD51C* and *PALB2* in CHORD-HRD patients from both the HMF and PCAWG datasets. Patients were clustered both by cancer type and by HRD type. Top: *BRCA1*- and *BRCA2*-type HRD probabilities from CHORD. Middle: SV contexts used by CHORD to distinguish *BRCA1*- from *BRCA2*-type HRD. Bottom: The biallelic status of each gene. Tiles marked as 'Known pathogenic' refer to variants having a 'pathogenic' or 'likely pathogenic' annotation in ClinVar. Only data from samples that passed CHORD's QC criteria are shown in this figure (MSI absent, ≥50 indels, and ≥30 SVs if a sample was predicted HRD). LOH: loss-of-heterozygosity.



Supplementary figure 24: Biallelic status of *BRCA2*, *BRCA1*, *RAD51C*, *PALB2*, as well as other HR genes in CHORD-HRD patients from both the HMF and PCAWG datasets. Top: *BRCA1*- and *BRCA2*-type HRD probabilities from CHORD. Middle: SV contexts used by CHORD to distinguish *BRCA1*- from *BRCA2*-type HRD. Bottom: The biallelic status of each gene. Tiles marked as 'Known pathogenic' refer to variants having a 'pathogenic' or 'likely pathogenic' annotation in ClinVar. Only HR genes with one of the following events in at least one patient from cluster 4 or 6 was shown here: deep deletion; or LOH in combination with a pathogenic/likely pathogenic variant or a frameshift/nonsense variant. Only data from samples that passed CHORD's QC criteria are shown in this figure (MSI absent, ≥50 indels, and ≥30 SVs if a sample was predicted HRD). Furthermore, only genes where at least one patient had an impactful biallelic event are shown. LOH: loss-of-heterozygosity.

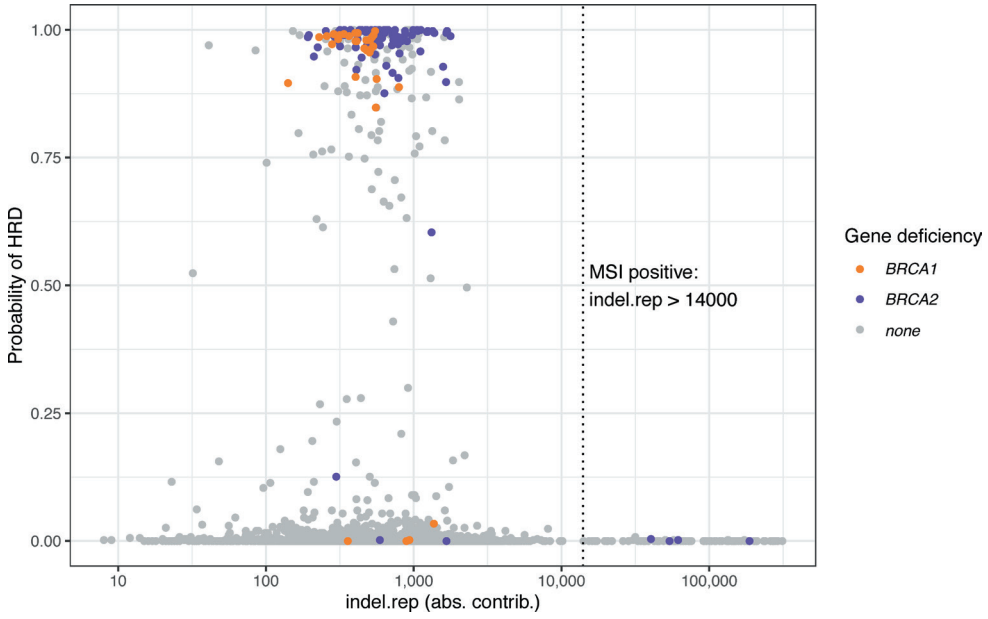


Supplementary figure 25: Overview of the procedure to sort and cluster CHORD-HRD samples. CHORD-HRD samples were first split by HRD subtype (*BRCA1*- or *BRCA2*-type HRD), and further split by whether a sample had an impactful biallelic event (defined as having a P-score pair of 5 and >=3). This produced 4 distinct groups of samples, which were sorted such that samples were ordered by deficiency of *BRCA2, BRCA1, RAD51C* and *PALB2* (in that order). This gene order corresponds to the enrichment significance of these genes as shown in **Figure 3b**. For each gene, samples were sorted such that samples with the HRD causing gene deficiency due to deep deletions were ordered first, followed by those with LOH (loss-of-heterozygosity) + frameshift and LOH + SNV/indel.

Supplementary notes

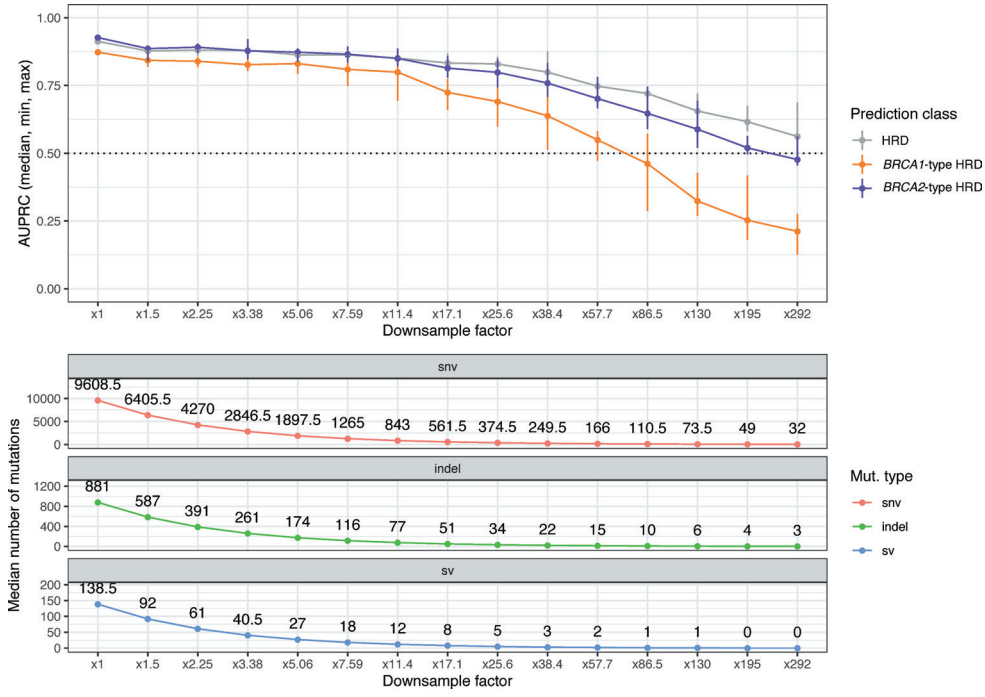
Supplementary Note 1: Prerequisites for accurate HRD prediction

It is important to note that microsatellite instability (MSI) negatively affects CHORD's ability to accurately predict HRD. MSI is a hypermutator phenotype characterized by an exceptional number of indels in regions with (tandem) repeats. As CHORD uses relative values of mutation contexts as input, MSI (defined as having more than 14,000 indels within repeat regions) results in a reduction of the relative contribution of microhomology deletions and thus an underestimated HRD probability ('false negatives'). **Supplementary figure 26** shows the CHORD being applied to all 3,824 HMF samples. All MSI samples were predicted HRP by CHORD even though 4 of these samples had biallelic loss of *BRCA2*. This could be circumvented by incorporating a MSI trained HRD classifier within CHORD. However, the number of HRD predicted samples with MSI is currently too small for training such a classifier and therefore we have built in MSI status checking as a quality control (QC) step within CHORD.



Supplementary figure 26: Microsatellite instability (MSI) impacts HRD prediction by CHORD. After applying CHORD to all 3824 tumors of the HMF dataset, all MSI positive tumors (those with more than 14,000 indels within repeat regions) had a low HRD probability, including 4 with *BRCA2* biallelic loss.

A minimum number of mutations is also required for accurate HRD prediction. To determine this, we progressively down-sampled all mutations for each sample in the training set and measured the reduction in performance (**Supplementary figure 27**). We found that at least 50 indels were required for accurately predicting HRD, and if a sample was predicted HRD, at least 30 SVs were required for distinguishing *BRCA1*-type from *BRCA2*-type HRD. These threshold levels may be particularly relevant for samples with low tumor purity and/or read coverage, and are as such also included as a QC step within CHORD.

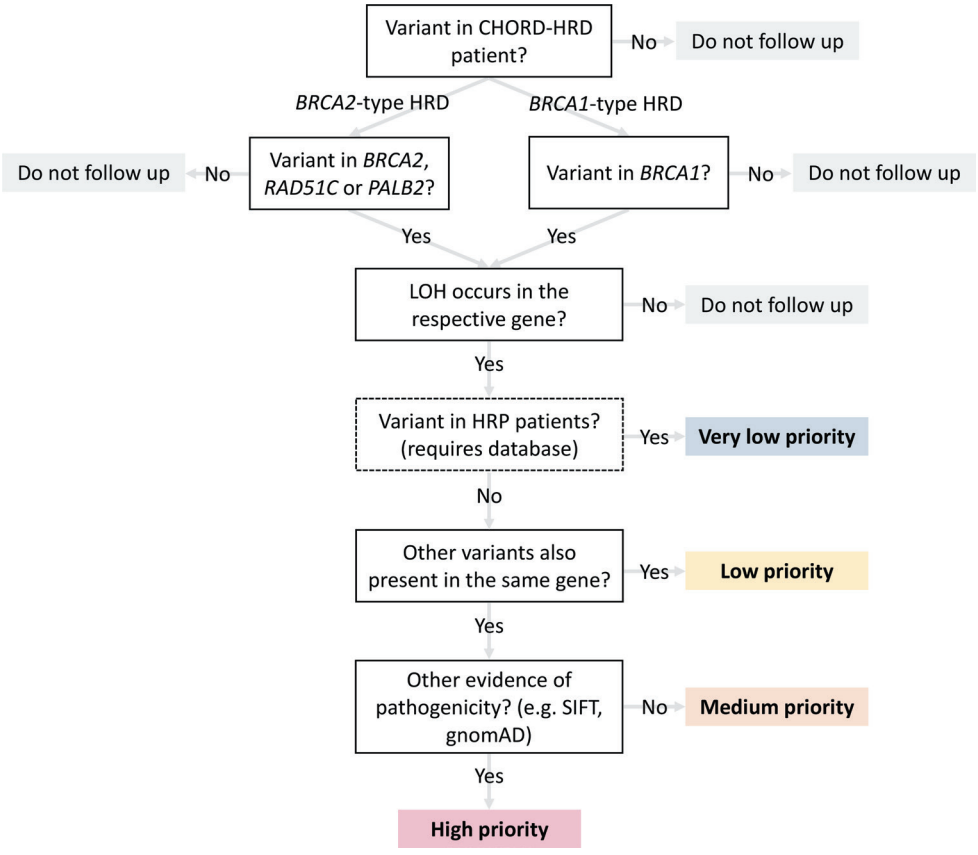


Supplementary figure 27: Performance of predicting HRD declines below ~50 indels. Similarly, performance for distinguishing *BRCA1*-type from *BRCA2*-type HRD declines below ~30 SVs. The set of mutations in each sample in the training set (used for training CHORD) were progressively downsampled. At each stage, CHORD was applied to the downsampled set of mutations, and thereafter, the area under the precision-recall curve was calculated.

Supplementary Note 2: CHORD as a tool to uncover novel pathogenic variants

The HRD predictions from CHORD can provide supportive evidence for interpreting variants of unknown significance (VUS), either germline or somatic, which can in turn be used for prioritizing rare VUS's for experimental validation as shown in **Supplementary figure 28**.

From 13 CHORD-HRD patients, we identified 14 variants not previously described to be pathogenic but which in combination with LOH could explain biallelic loss of a HR gene (**Supplementary data 7**). Moreover, biallelic loss of the respective gene corresponded to the associated HRD subtype, and all of these variants were not present in HRP samples, providing additional support that these variants are potentially pathogenic. From these variants, 2 *BRCA2* missense variants (c.9230T>C, c.9254C>T) were both found in patient HMF000429. Here, further validation is required to determine which was the driver mutation (or possibly the combination). The remaining 12 variants were the sole variant found in the respective gene and respective patient, and are thus more likely to be pathogenic driver variants. Of these, one *BRCA2* missense variant (c.8045C>T) had a 'Uncertain_significance' annotation in ClinVar but a low population frequency according to gnomAD as well as being predicted as 'deleterious' by SIFT and 'probably damaging' by PolyPhen, supporting its potential pathogenicity.

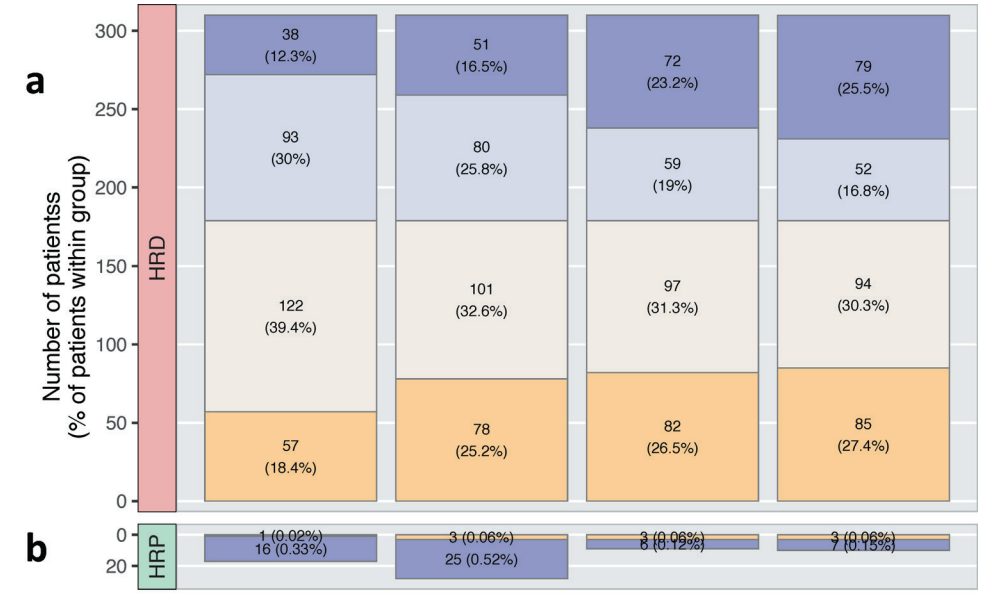


Supplementary figure 28: Scheme for prioritizing variants of unknown significance (VUS) for experimental validation starting from the HRD predictions from CHORD.

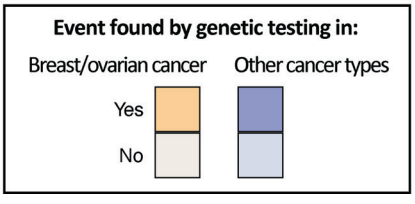
Supplementary Note 3: CHORD can detect HRD in a substantial number of cases that would be missed by genetic testing

To assess the potential value of CHORD in a clinical setting, we compared CHORD's predictions for HMF patients to the hypothetical outcomes of common genetic testing approaches. In the clinic, HRD detection is currently often done by screening for pathogenic *BRCA1/2* SNVs/indels based on annotations from curated databases (e.g. ClinVar). This is performed either on blood biopsies (blood genetic testing), analogous to screening for germline SNVs/indels in sequencing data of the HMF cohort (which was analysed by tumor-normal pair whole genome sequencing); or on tumor biopsies (tumor genetic testing), analogous to screening both germline and somatic SNVs/indels. Our genetic analyses (below figure) indicate that blood genetic testing would identify a pathogenic *BRCA1/2* SNV/indel (according to ClinVar; or an out-of-frame frameshift) in 18% of CHORD-HRD breast/ovarian cancer patients (which genetic testing is often restricted to), while tumor genetic testing would increase this proportion to 25%. If patients with other cancer types would be included, blood and tumor genetic testing would identify 31% and 42% of CHORD-HRD tumors (respectively) with a pathogenic *BRCA1/2* SNV/indel (**Supplementary figure 29a, columns 1 and 2**).

While not currently routinely performed in the clinic, WGS based genetic testing with matched blood/tumor biopsies (WGS genetic testing) would allow the detection of any event (including structural events such as LOH or deep deletions) that contributes to HR gene inactivation, and enables the determination of biallelic gene status. Our analyses show that WGS genetic testing would increase the number of patients that are considered HRD compared to SNV/indel-based blood/tumor genetic testing, with biallelic loss of *BRCA1/2* being identified in 27% of CHORD-HRD breast/ovarian cancer patients and in 50% of patients pan-cancer (**Supplementary figure 29a, column 3**). Additionally, WGS genetic testing would consider 9 CHORD-HRP patients as HRD (**Supplementary figure 29b, column 3**; 'genetic testing false positives'; these could also be CHORD false negatives), a marked decrease compared to tumor genetic testing which would identify 28 false positives (**Supplementary figure 29b, column 2**). By including the two other main HRD associated genes (*RAD51C* and *PALB2*) in WGS genetic testing, biallelic gene loss would be identified in 53% of patients (**Supplementary figure 29a, column 4**), while only increasing the number of genetic testing false positives from 9 to 10 patients (**Supplementary figure 29b, column 5**). Our findings show that while WGS genetic testing (for biallelic loss) offers improved detection of HRD patients compared to testing for pathogenic SNVs/indels, it still misses around half of HRD patients as classified by CHORD.



Genetic testing type	On blood	On tumor	WGS, matched blood/tumor	
Events tested for	*BRCA1/2* germline pathogenic mut.	*BRCA1/2* germline/somatic pathogenic mut.	*BRCA1/2* biallelic loss	*BRCA1/2, RAD51C, PALB2* biallelic loss



Supplementary figure 29: CHORD identifies a large proportion of HRD patients that would be missed by genetic testing. (**a**) and (**b**) show the percentage of CHORD-HRD or CHORD-HRP patients (respectively) from the HMF dataset in which a pathogenic event was found based on four genetic testing setups. In the first two setups ('On blood', 'On tumor'), a pathogenic event was identified if a pathogenic SNV/indel was found on one allele, which was defined as a frameshift, or a likely pathogenic or pathogenic variant according to ClinVar. In the 'WGS based testing' setups, the pathogenic SNV/indel must also have occurred in combination with loss of heterozygosity (LOH); or, a deep deletion was identified. Only data from samples that passed CHORD's QC criteria are shown in this figure (MSI absent, ≥50 indels, and ≥30 SVs if a sample was predicted HRD).

Chapter 3

Machine learning-based tissue of origin classification for cancer of unknown primary diagnostics using genome-wide mutation features

Luan Nguyen[1], Arne Van Hoeck[1], Edwin Cuppen[1,2,*]

[1] University Medical Center Utrecht, Utrecht, The Netherlands
[2] Hartwig Medical Foundation, Amsterdam, The Netherlands

* Corresponding author

Adapted from: Nat Commun 13, 4013 (2022)
URL: https://doi.org/10.1038/s41467-022-31666-w
QR code to URL:

Abstract

Cancers of unknown primary (CUP) origin account for ~3% of all cancer diagnoses, whereby the tumor tissue of origin (TOO) cannot be determined. Using a uniformly processed dataset encompassing 6756 whole-genome sequenced primary and metastatic tumors, we develop Cancer of Unknown Primary Location Resolver (CUPLR), a random forest TOO classifier that employs 511 features based on simple and complex somatic driver and passenger mutations. CUPLR distinguishes 35 cancer (sub)types with ~90% recall and ~90% precision based on cross-validation and test set predictions. We find that structural variant derived features increase the performance and utility for classifying specific cancer types. With CUPLR, we could determine the TOO for 82/141 (58%) of CUP patients. Although CUPLR is based on machine learning, it provides a human interpretable graphical report with detailed feature explanations. The comprehensive output of CUPLR complements existing histopathological procedures and can enable improved diagnostics for CUP patients.

Introduction

Cancers of unknown primary (CUPs) is an umbrella term for advanced stage metastatic tumors for which the tumor tissue of origin (TOO) cannot be conclusively determined based on routine diagnostics (typically via histopathology [128]), and there is also a significant fraction of patients with indeterminate or differential diagnoses, especially with poorly differentiated tumors [129]. Patients with uncertain TOO diagnosis suffer from a lack of therapeutic options as primary cancer type classification is a dominant factor in guiding treatment decisions [130].

Thus far, TOO classifiers have been developed on data from a wide range of molecular methods including DNA sequencing (targeted [131], whole exome [132], and whole genome [90,91]), RNA profiling (from coding RNA [133], microRNA [134–136], as well as whole transcriptome profiling [137,138]), and methylation profiling [139]. Driven by its ability to comprehensively capture actionable biomarkers that enable precision medicine [126], whole genome sequencing (WGS) is maturing rapidly as a diagnostic tool [140] and increasingly adopted in clinical systems in various countries [141–143], and could thus be an interesting basis for a diagnostic TOO classifiers. Recently developed WGS-based classifiers [90,91] were shown to outperform targeted or whole exome sequencing based approaches [131,132] due to being able to utilize mutations from all genomic regions. The main features employed by these classifiers included mutational signatures which are patterns of somatic mutations resulting from exogenous or endogenous mutational processes (e.g. C>T mutations due to ultraviolet radiation exposure in melanoma) [13], as well as regional mutational density (RMD) which represents the genomic distribution of somatic mutations that are associated with tissue type specific chromatin states, whereby late-replicating closed chromatin regions show increased mutation rates [144]. However, not all WGS based features are yet fully explored for TOO classification including complex mutagenic features such as viral DNA integrations, driver gene fusions, and other complex structural events (e.g. chromothripsis), as well as non-mutagenic features such as gender, all of which have been shown to be correlated with specific tumor type(s). Indeed, human papillomavirus (HPV) sequence insertions are specifically and frequently found in cervical and head and neck cancer [14], *KIAA1549-BRAF* fusions in pilocytic astrocytomas [13], and liposarcomas frequently harbor *FUS-DDIT3* fusions [15] as well as chromothripsis events [145].

Here we describe the development of CUPLR (Cancer of Unknown Primary Location Resolver), a TOO classifier that integrates current state-of-the-art WGS based mutation features, including complex structural variant (SV) features. CUPLR comprises an ensemble of binary random forest classifiers that each discriminate one of 35 cancer types with an overall recall of 90%. We find that while RMD and mutational signatures were highly predictive of cancer type (in line with existing classifiers [90,91]), the incorporation of SV features improves prediction performance for cancer types that currently lack highly informative features. Furthermore, we have ensured that the output of CUPLR, namely the prediction probabilities and the features supporting each prediction, are human interpretable to facilitate diagnostic use and clinical decision making with CUPLR.

Results

Extraction of genomic features

To develop CUPLR, we constructed a harmonized dataset from two large pan-cancer WGS datasets from the Hartwig Medical Foundation (Hartwig) and Pan-Cancer Analysis of Whole Genomes consortium (PCAWG) [146]. The raw sequencing reads were analyzed with the same mutation calling pipeline to construct a catalog of uniformly called simple and complex mutations. The harmonized dataset consisted of tumors from 6756 patients across 35 different cancer types (**Figure 1a, Supplementary data 1**). In contrast to many previously published papers [90,91,131,132], this dataset includes a large proportion of samples taken from metastatic lesions, which is relevant for TOO classification as CUP samples are by definition from patients with metastatic cancer.

A wide range of features (n=4131) were extracted for classifying cancer type based on driver/passenger and simple/complex mutations (**Figure 1c**). First, we determined the presence of gain of function (amplifications and activating mutations) and loss of function (deep deletions and biallelic loss) events in 203 cancer associated genes. These genes were selected based on having enrichment of gain and/or loss of function events in at least one cancer type (see methods). Second, we calculated the mutational load of single base substitutions (SBS), double base substitutions (DBS), and indels for each sample. Third, we determined the number of contributing mutations to the SBS, DBS and indel signatures from the COSMIC catalog [13]. Fourth, the number of SBSs in each 1Mb bin across the genome (n=3071) was calculated to determine the RMD [147]. Mutational signatures and RMD were normalized by the mutational load of the respective mutation type to account for differences in mutational load across samples. Fifth, copy number data was used to infer the genome ploidy, diploid proportion, whole genome duplication status, and gender for each sample [127]. Sixth, for each sample we determined the copy number change of each chromosome arm relative to the genome ploidy [35]. Lastly, we parsed the called simple and complex SVs to determine: (i) the total SV load per sample; (ii) the number of deletions, duplications stratified by length; (iii) the number of complex events stratified by size; (iv) the size of the largest complex event, (v) the number of long interspersed nuclear element (LINE) insertions and double minutes; and (vi) the presence of gene fusions and viral sequence insertions [28,127].

Classifier training

The extracted genomic features were then used to develop CUPLR, a classifier consisting of two components (**Figure 1d**). The first component is an ensemble of binary random forest classifiers that each discriminates one cancer type versus other cancer types (i.e. one-vs-rest). We chose to use an ensemble of binary classifiers as opposed to one multiclass classifier so that feature selection could be performed per cancer type, since different features are important for each cancer type. Additionally, we chose to use random forests over other algorithms (e.g. neural networks) as they can natively handle different feature types (continuous, boolean, categorical, etc) without requiring feature values to be scaled, which also improves model interpretability. The second component of CUPLR is an ensemble of isotonic regressions to calibrate the probabilities produced by each random forest. Random forests tend to be overconfident at probabilities towards 0 and underconfident at probabilities towards 1, and this bias varies between random forests [148]. The calibration we have performed here ensures that probabilities are comparable between random forests. Furthermore, calibration allows for the probabilities to have the following intuitive interpretation: a probability of e.g. 0.8 means that there is an 80% chance of a prediction being correct (this relationship does not hold for the raw ‘probabilities’ from random forests).

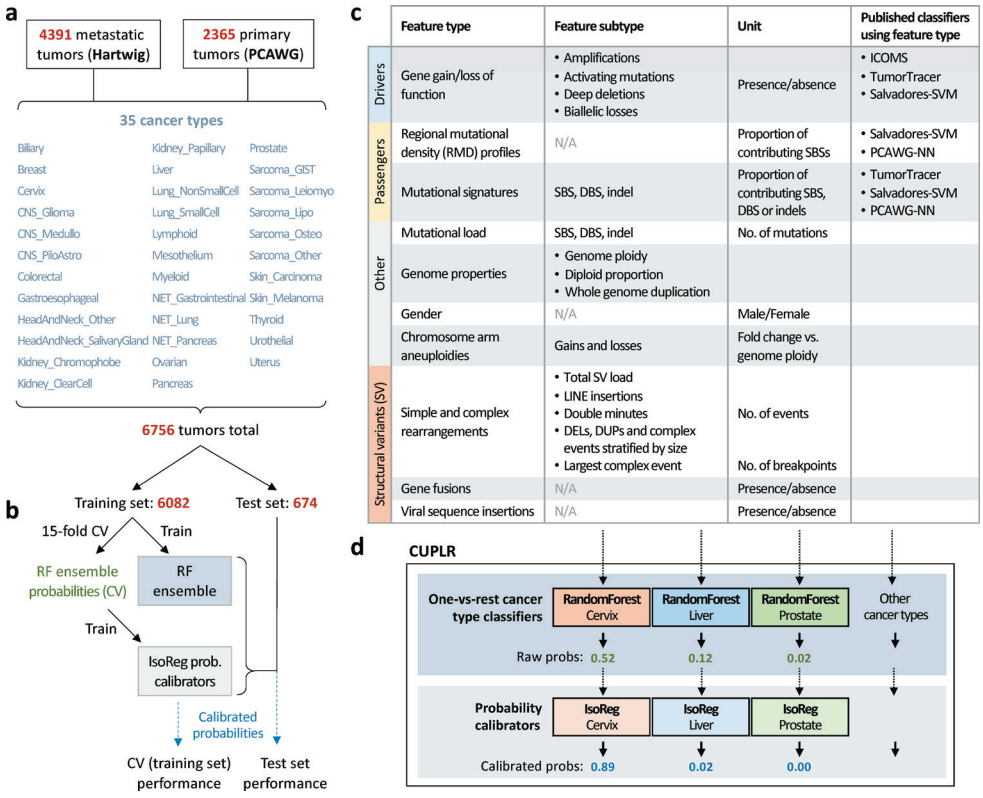


Figure 1: **Cancer of Unknown Primary Location Resolver (CUPLR) classifies 35 different cancer types using features derived from all mutation types**. **(a)** CUPLR was developed using whole-genome sequencing data 4391 metastatic tumors from the Hartwig Medical Foundation (Hartwig) and 2365 primary tumors from the Pan-Cancer Analysis of Whole Genomes (PCAWG) consortium, totalling 6756 samples across 35 different cancer types. **(b)** 6082 samples were used to train CUPLR and 674 were held out as an independent test set. The whole training set was used to train the final random forest ensemble. 15-fold cross-validation was performed to obtain the random forest cancer type probabilities on the training set, which were then used to train the ensemble of isotonic regressions (for probability calibration). CUPLR is composed of the random forest and isotonic regression ensembles as shown in **(d)**. The performance of CUPLR was assessed using the calibrated cross-validation probabilities as well as probabilities obtained by applying CUPLR to the test set. (**c**) A summary of the genomic features extracted from the whole-genome sequencing data and used by CUPLR. Detailed descriptions of each feature can be found in **Supplementary data 3**. The names of the published classifiers refer to the following studies; ICOMS: Inferring Cancer Origins from Mutation Spectra, Dietlen *et al.* 2014 [131]; TumorTracer: Marquard *et al.* 2015 [132]; Salvadores-SVM: support vector machine by Salvadores *et al.* 2019 [91]; PCAWG-NN: PCAWG neural network by Jiao *et al.* 2020 [90]. **Cancer type abbreviations**; CNS: central nervous system; CNS_Medullo: medulloblastoma; CNS_PiloAstro: pilocytic astrocytoma; NET: neuroendocrine tumor; Sarcoma_GIST: gastrointestinal stromal tumor; Sarcoma_Leiomyo: leiomyosarcoma; Sarcoma_Lipo: liposarcoma; Sarcoma_Osteo: osteosarcoma; Sarcoma_Other: sarcomas other than leiomyosarcoma, liposarcoma or gastrointestinal stromal tumors. **Other abbreviations**; RF: random forest, IsoReg: isotonic regression, CV: cross-validation, SBS: single base substitutions, DBS: double base substitutions, SV: structural variants, DEL: structural deletions, DUP: structural duplications, LINE: long interspersed nuclear element.

Performance of CUPLR

To assess the performance of CUPLR, we used the cancer type predictions based on the isotonic regression calibrated cross-validation (CV) probabilities, as well as the predictions upon applying CUPLR to the held out test set (**Figure 1b**). Both the training set (n=6082) and the held out test set (n=674) had the same cancer type and cohort distribution (**Supplementary data 2**). CUPLR could predict TOO with 90% (CV) and 89% (test set) overall recall, and 90% (CV) and 89% (test set) overall precision (**Figure 2b,c**). Differences between CV and test set recall and precision in certain cancer types were due to low sample sizes in the test set (**Figure 2a,b**, **Supplementary figure 6**).

High misclassification rates for certain cancer types could likely be explained by shared cancer type characteristics (**Figure 2c**). This could be due to a common developmental origin, such as with Uterus being misclassified as Ovarian (CV: 7%, test: 29%) due to both being gynecological cancers [149], and Biliary being misclassified as Pancreas (CV: 24%, test: 42%) and Liver (CV: 9%) due to being cancers of the foregut [150,151]. Cancer subtypes were also often misclassified as other subtypes, which was the case for Lung_SmallCell towards Lung_NonSmallCell (CV: 40%, test: 60%); Kidney_Papillary towards Kidney_ClearCell (CV: 38%, test: 67%); and Sarcoma_Leiomyo (CV: 35%, test: 43%) and Sarcoma_Osteo (CV: 17%) towards Sarcoma_Other (sarcomas other than leiomyo-/lipo-/osteosarcomas or gastrointestinal stromal tumors). Neuroendocrine tumor (NET) subtypes were occasionally misclassified as each other, such as NET_Lung towards NET_Gastrointestinal (CV: 9%) and NET_Pancreas (CV: 9%, test: 33%), and NET_Gastrointestinal towards Pancreas (CV: 6%) which may (at least partially) reflect cancer type misannotation amongst these samples due to neuroendocrine tumors having similar morphological features [152]. Likewise, HeadAndNeck_SalivaryGland samples that were misclassified as breast (CV: 23%, test: 33%) were potentially misannotated due to being adenoid cystic carcinomas (i.e. salivary gland-like cancers) of the breast [153].

Thus far, we have mainly assessed performance based on whether the highest probability cancer type is the correct cancer type (i.e. recall; **Figure 2b,c**). However, if we consider whether the correct cancer type is amongst the top-2 highest probabilities (top-2 recall; **Figure 2b**), overall recall increases from 90% to 95% (CV) and 89% to 94% (test set), with the greatest increases being for the cancer subtypes including Lung_SmallCell (CV: 50% to 83%, test: 40% to 100%), Kidney_Papillary (CV: 62% to 79%, test: 33% to 100%), Sarcoma_Leiomyo (CV: 56% to 89%, test: 57% to 86%) and Sarcoma_Other (CV: 63% to 89%, test: 54% to 77%). A large gain in recall was also observed for Biliary (CV: 52% to 73%, test: 42% to 83%) which was often misclassified as Pancreas. Similar increases in recall were seen based on predictions on the test set. The top-2 (and even top-3) probabilities of CUPLR can be particularly useful for differential diagnosis purposes to narrow down potential TOOs when routine diagnostics are not fully conclusive.

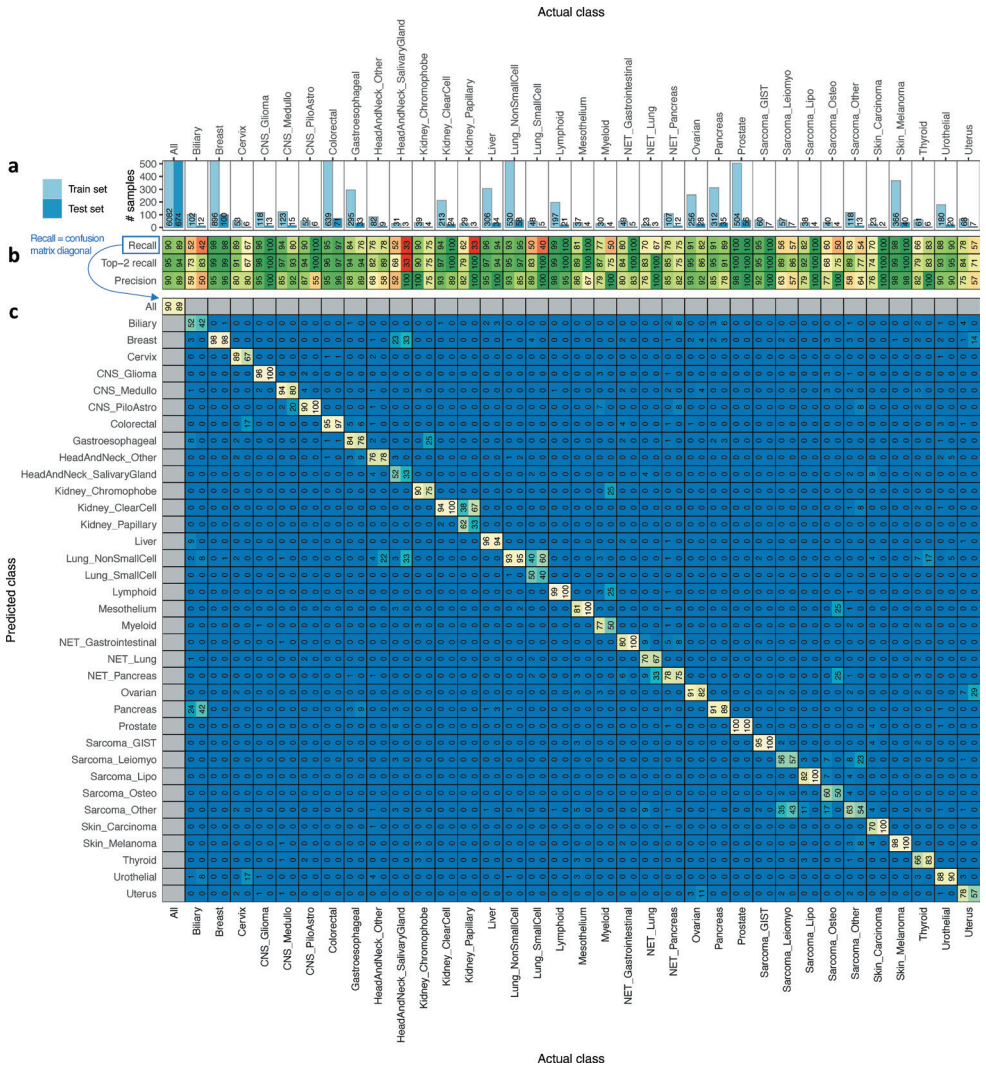


Figure 2: **Performance of CUPLR**. **(a)** The total number of samples for each cancer type for the training set (left columns) and held out test set (right columns). Note that the y-axis of **(a)** is truncated to better visualize cancer types with small sample sizes. Panels **b-c** also have the same left/right column layout corresponding to the training set cross-validation and test set performance. **(b)** Summary of performance metrics. Recall: % of correctly classified samples per cancer type, and is equivalent to the diagonal values in panel **c**. Top-2 recall: % of correctly classified samples when considering the 2 highest probability cancer types as correct. Precision: % of correctly classified samples amongst samples predicted as a particular cancer type. Overall performance metrics (i.e. under the 'All' column) are micro-averages. **(c)** Confusion matrix showing the performance of CUPLR where columns represent the % of samples in a cancer type cohort predicted as a particular cancer type. The diagonal represents the % of samples correctly predicted as a particular cancer type (equivalent to recall). Raw data for performance metrics and confusion matrix in panels **b** and **c** can be found in **Supplementary data 4**. **Cancer type abbreviations**; CNS: central nervous system; CNS_Medullo: medulloblastoma; CNS_PiloAstro: pilocytic astrocytoma; NET: neuroendocrine tumor; Sarcoma_GIST: gastrointestinal stromal tumor; Sarcoma_Leiomyo: leiomyosarcoma; Sarcoma_Lipo: liposarcoma; Sarcoma_Osteo: osteosarcoma; Sarcoma_Other: sarcomas other than leiomyosarcoma, liposarcoma or gastrointestinal stromal tumors.

Added predictive value of SV related features

When examining the most important feature types from each random forest within CUPLR (**Figure 3a**), RMD profiles ('rmd') were consistently the most predictive of cancer type (in line with the findings from Jiao *et al.* 2020 [90]), as well mutational signatures ('sigs') including those with known cancer type associations such as SBS4 (associated with smoking [13]) in lung cancer (**Figure 3b**). As these mutation features are derived from genome-wide SBSs and indels, we assessed whether the presence of certain confounding factors that affect the SBS and indel genomic landscape (including DNA repair deficiencies [38,154], chemotherapy treatment [66,108], smoking history [155]) may lead to more incorrect predictions. However, these confounding factors showed minimal impact on classification performance (**Supplementary Note 1**, **Supplementary table 1**, **Supplementary figure 13**).

In addition to RMD profiles and mutational signatures, gender (as expected) was highly important for predicting cancers of the reproductive organs including breast, cervical, ovarian, prostate, and uterine cancer (**Figure 3b, Supplementary figure 7**). Notably, SV related features were important for classifying certain cancer types (**Figure 3b**). This included known cancer type specific gene fusions such as *TMPRSS2-ERG* for Prostate [42], *EML4-ALK* for Lung_NonSmallCell [156], *KIAA1549-BRAF* for CNS_PiloAstro (pilocytic astrocytomas) [157], and *FUS-DDIT3* for Sarcoma_Lipo [158]. We also find known viral DNA integrations as important features, including from human papillomavirus (viral_ins.HPV) in Cervix and HeadAndNeck_Other (non-salivary gland head and neck cancers) [159], Epstein-Barr virus in HeadAndNeck_Other [160], hepatitis B virus (viral_ins.HBV) in Liver [161], and Merkel cell polyomavirus (viral_ins.MCPyV) in Skin_Carcinoma [162]. Lastly, the largest complex SV cluster (i.e. by number of breakpoints) (sv.COMPLEX.largest_cluster) which we use as a proxy for the presence of chromothripsis was also predictive for liposarcomas, which are known to frequently harbor chromothripsis events. However, in contrast to the features mentioned above, the presence of chromothripsis alone is not sufficient for classifying a tumor as liposarcoma as chromothripsis is also highly prevalent in other tumor types [145].

To further assess the added value of SV related features, we excluded entire feature types from the training and examined the decrease in classifier performance (**Supplementary figure 8**). Indeed, removal of the viral integration features ('viral_ins') led to decreased recall for Skin_Carcinoma (70% to 61%) and Cervix (89% to 83%), likely due to loss of the viral_ins.MCPyV and viral_ins.HPV features respectively. Removal of the simple and complex SV features ('sv') resulted in a drop in recall for Sarcoma_Leiomyo (56% to 49%), likely as the size of the largest complex SV cluster (sv.COMPLEX.largest_cluster) can discriminate Sarcoma_Lipo and Sarcoma_Osteo from Sarcoma_Leiomyo (**Supplementary figure 7**). Lastly, removal of the gene fusion features ('fusion') resulted in a large decrease in recall for Lung_SmallCell (50% to 38%) likely as *EML4-ALK* fusions are characteristic of non-small cell (but not small cell) lung cancer.

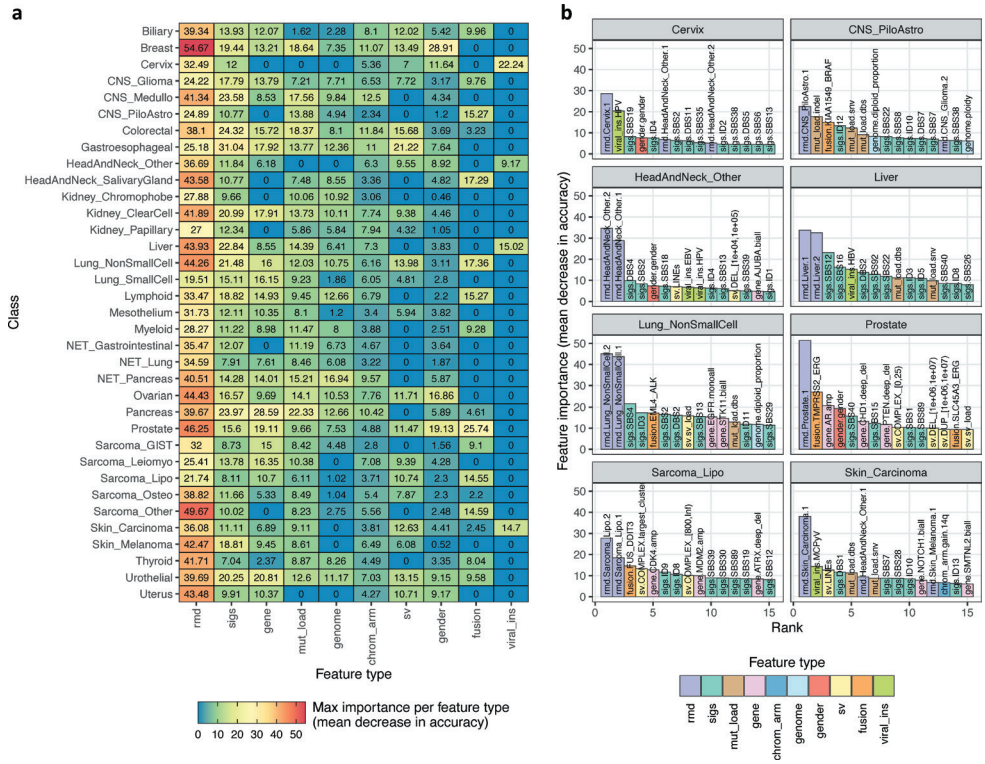


Figure 3: **Features most predictive of cancer type. (a)** Maximum feature importance per feature type for each cancer type random forest from CUPLR. See **Supplementary data 3** for the descriptions as well as feature importance values for each feature. **(b)** Feature importances from the top 15 features for selected cancer type random forests. Feature names are in the form {feature type}.{feature name}. **Feature type abbreviations**; rmd: regional mutation density profiles; sigs: mutational signatures; mut_load: total number of single base substitutions, double base substitutions or indels; gene: presence of gene gain or loss of function events; chrom_arm: chromosome arm copy number fold change versus overall genome ploidy; genome: genome properties including genome ploidy, diploid proportion, whole genome duplication status; gender: sample gender as determined by copy number data; sv: structural variants; fusion: presence of gene fusions; viral_ins: presence of viral sequence insertions. **Cancer type abbreviations**; CNS: central nervous system; CNS_Medullo: medulloblastoma; CNS_PiloAstro: pilocytic astrocytoma; NET: neuroendocrine tumor; Sarcoma_GIST: gastrointestinal stromal tumor; Sarcoma_Leiomyo: leiomyosarcoma; Sarcoma_Lipo: liposarcoma; Sarcoma_Osteo: osteosarcoma; Sarcoma_Other: sarcomas other than leiomyosarcoma, liposarcoma or gastrointestinal stromal tumors.

When compared to existing WGS-based CUP classifiers [90,91,132], CUPLR is able to classify more cancer (sub)types and showed improved recall and precision (**Supplementary figure 9**, **Supplementary figure 10**) for cancer types where SV related features were important, including for CNS_PiloAstro, Lung_NonSmallCell, and Prostate. Overall, CUPLR achieved similar recall to existing classifiers for the remaining cancer types (except for head and neck, myeloid, pancreatic neuroendocrine and thyroid cancers). Likewise, precision was also similar to other classifiers for the remaining cancer types (except for head and neck, myeloid, thyroid, and uterine cancers)

In summary, the incorporation of all feature types resulted in the best performance, with SV related features being important for specific cancer types.

Graphical prediction report

Aside from cancer type probabilities, CUPLR also outputs explanations as to which features support these probabilities. These allow one to verify the predictions based on existing knowledge, and could be included in diagnostic reports to support decision making, e.g. in molecular tumor boards. The feature explanations are based on feature contribution calculations which enable feature importance determination at the sample level (rather than at the cohort level as in **Figure 3**). Specifically, feature contributions represent the total gain (or loss) in probability upon passing a feature through a random forest [163]. To ease the interpretation of CUPLR's output, we have implemented a graphical report (**Figure 4**) which can be generated per patient that shows the cancer type probabilities (left panel), feature contributions for the top features for the top cancer types (middle panel). Also shown are the corresponding feature values in the patient in relation to the average feature value amongst patients in the training set (right panel).

We will use patient HMF004048A as an example demonstration of the graphical report (**Figure 4a**). Since the Lung_NonSmallCell probability (0.96) was much higher than the probability of other cancer types, only one cancer type (i.e. panel row) is shown. Up to 3 panel rows can be shown when more than one cancer type probability is high (such as for patient DO7304; **Figure 4b**). One of the top features for HMF004048A was the presence of an *EML4-ALK* fusion (middle panel) which is on average found in ~4% of Lung_NonSmallCell patients (pink label), but only ~0.1% in all other patients (blue label). Since this feature is of boolean type, a feature value of 1 (red label) indicates the presence of the *EML4-ALK* fusion in HMF004048A (whereas a feature value of 0 would indicate absence). Additionally, the contribution of the non-small cell lung cancer RMD profiles (rmd.Lung_NonSmallCell.2) as well as the contribution of the APOBEC associated signatures SBS2 and SBS13 [13] in HMF004048A (red labels) is higher than in Lung_NonSmallCell patients (pink labels), but also compared to all other patients (blue label). Whether a feature value in the patient of interest is higher or lower than that in all other patients is also indicated in text in the feature contribution panel (middle panel) for non-boolean features.

Patient DO7304 (**Figure 4b**) is an example of a situation where more than one cancer type probability is high, with the probability of Lymphoid being 0.78 and CNS_PiloAstro being 0.75 (**Figure 4b**). Here, two feature panels are shown for both of these cancer types to aid with resolving this uncertainty. Since *IGLL5* loss is specific to lymphomas [164], we can confirm that the likely correct prediction is indeed Lymphoid.

This graphical report presents the detailed machine learning-based classification output of CUPLR in a human readable format. We acknowledge that the output is highly detailed, which is inevitable due to the large amounts of data used by the algorithm. However, as these details may not be necessary in all circumstances, we have implemented the option in the software to hide the feature contribution and/or feature value panels in the final graphical output.

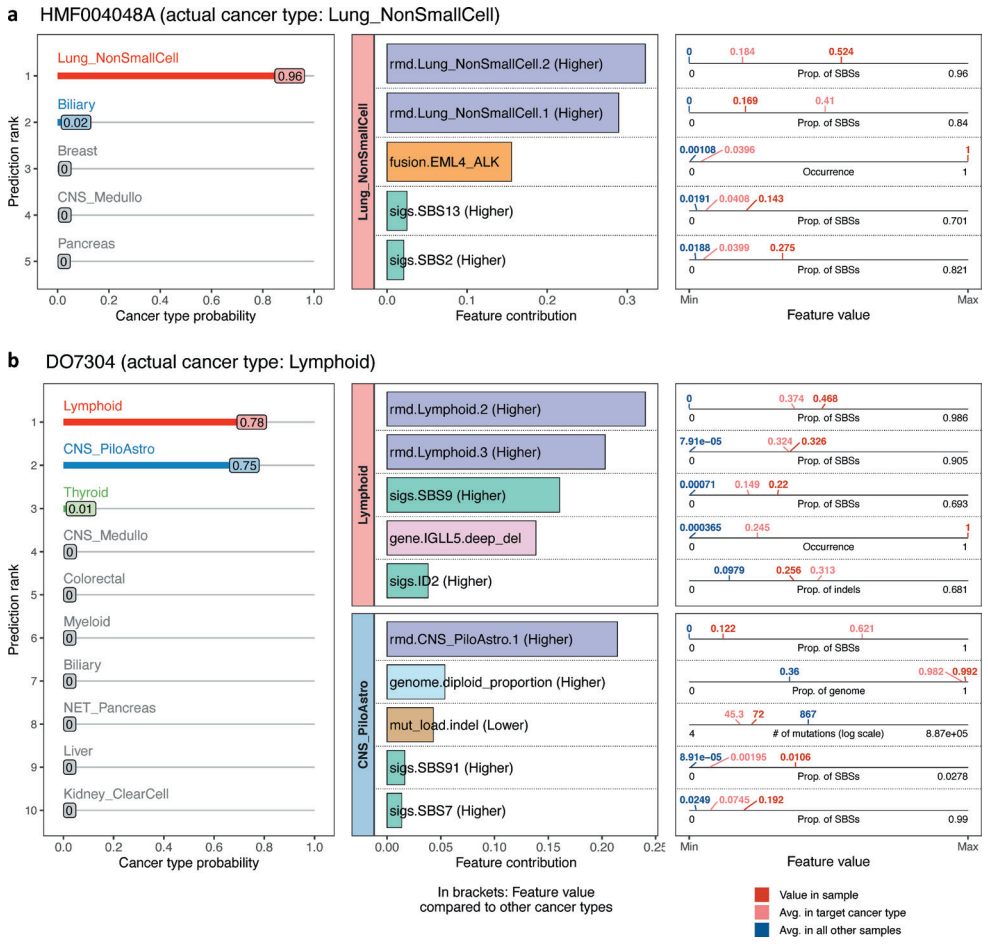


Figure 4: Graphical report for CUPLR predictions. Example reports are shown for two patients from the holdout set: **(a)** HMF004048A, annotated as having non-small cell lung cancer (Lung_NonSmallCell), and **(b)** DO7304, annotated as having lymphoma (Lymphoid). The leftmost panels show the predicted cancer type probabilities. In the middle panels, contributions of the top features are shown for the top predicted cancer types. When there is uncertainty in the cancer type probabilities such as in **(b)**, more feature contribution panel rows are shown. In the rightmost panel, the feature values in the patient (red label) are plotted in relation to the average feature value amongst patients in the training set with the target cancer type (pink label) as well as all other samples not belonging to the target cancer type (blue label). For a full description of each feature, see **Supplementary data 3**. **Feature type abbreviations**; rmd: regional mutation density profiles; sigs: mutational signatures; mut_load: total number of single base substitutions, double base substitutions or indels; gene: presence of gene gain or loss of function events; genome: genome properties including genome ploidy, diploid proportion, whole genome duplication status; fusion: presence of gene fusions.

Feature contributions aid in clarifying the primary tumor location of CUPs

To showcase how CUPLR could be used in a real life clinical setting, we applied CUPLR to 141 tumors with a CUP diagnosis from the Hartwig dataset. From these, we could confidently (n=68) or partially (n=14) resolve the cancer type for 82 (58%) patients by examining the top predicted cancer types and corresponding top contributing features for each patient (**Supplementary data 5**).

Of the 68 patients with fully resolved cancer types, 44 exhibited features with well known cancer type associations in combination with high contribution of respective cancer type RMD profile. This included: 4 patients predicted as Breast with breast cancer specific driver mutations (in *CHD1*, *GATA3*, *PIK3CA*, and/or *ZNF703*) [165]; 7 patients predicted as Colorectal with APC mutations [5] and/or the presence of the colibactin signature (SBS88) [65]; 5 patients predicted as Gastroesophageal with the ROS damage signature SBS17 [13], LINE insertions [39], and/or *FHIT* deletions [70]; 21 patients predicted as Lung_NonSmallCell with signatures of smoking (SBS4, DBS2 and/or ID3 [13]); 3 patients predicted as Pancreas with *KRAS* mutations [5]; 1 patient predicted as Prostate with a *TMPRSS2-ERG* fusion [42]; 2 patients predicted as Urothelial with *TERT* mutations [166]; and 1 patient predicted as Uterus with a *PIK3R1* mutation [167]. The remaining 24 patients with a top cancer type probability ≥0.8 were likely correctly classified as most of these patients were predicted as high recall cancer types (>0.9, **Figure 2**) with highly specific RMD profiles respective to each cancer type, including Breast Colorectum, Gastroesophageal, Kidney_ClearCell, Liver, Kidney_ClearCell, and Ovarian (**Supplementary data 5**, **Supplementary figure 11**).

For the 14 patients with partially resolved cancer types, we could only determine the cancer supertype. For example, 4 patients had >0.6 probability of both Lung_NonSmallCell and Lung_SmallCell and exhibited smoking signatures SBS4, DBS2, and/or ID3 [13], indicating that these patients most likely had lung cancer, though the subtype remains uncertain. Likewise, we could narrow down the TOO to 2 probable cancer types for 3 patients. For example, patient HMF002806A had >0.5 probability of Uterus and Breast, and had mutations in *PIK3CA* and *PTEN* which are common in both these cancer types [5]. This is indicative of gynecological cancer which often are treated in a similar manner [168,169].

We thus demonstrate that CUPLR can potentially clarify the TOO for over half of patients with CUP. It is important to note that even if the TOO is only partially resolved for a patient, such a patient could now potentially have more treatment options. However, for the CUP patients discussed above, additional evidence would be required for validation and final diagnosis, which could for example be based on histopathological examinations. These were unfortunately not available for the retrospectively analyzed samples included here, but such information would be readily available in routine diagnostics.

Discussion

Here, we have developed a tissue of origin classifier (CUPLR) for the analysis and diagnostics of CUPs using a large uniformly processed WGS dataset of both primary and metastatic tumors. Our classifier is to our knowledge the first to incorporate genomic features derived from SVs, as well as provide human interpretable explanations alongside each prediction which allows for manual resolving of CUPs, especially in cases of lower scoring samples or for samples for which multiple tumor types are suggested by CUPLR.

Current state-of-the-art WGS-based classifiers, including those by Jiao *et al.* [90] and Salvadores *et al.* [91], achieve high recall (≥80%) for over three quarters of the cancer types they predict by primarily employing RMD and mutational signatures as features, which are derived from simple mutations. CUPLR builds upon these classifiers, with the inclusion of SV and driver gene related features improving performance for certain cancer types, such as for pilocytic astrocytoma and prostate cancer. Of note, genomic-based TOO classifiers published thus far have only used data from primary tumors [90,91,131,132]. However, since CUPs are by definition metastatic tumors, the inclusion of 4391 metastatic tumors with known tissue of origin for training CUPLR may make it a more suitable tool for the purpose of clarifying CUPs. We do acknowledge however that the metastatic tumor data used here may harbor treatment effects which are absent in treatment-naive CUPs, such as driver mutations in *AR* [170] or *ESR1* [171] conferring resistance to therapy or presence of characteristic mutations induced by for example 5-fluorouracil [66] or radiotherapy [108,172]. Identification and removal of such treatment associated features could potentially improve TOO classification. Additionally, CUPLR is able to distinguish the largest number of cancer (sub)types (35 cancer types) compared to existing WGS-based classifiers, with the Jiao [4] and Salvadores [5] classifiers discriminating 24 and 18 cancer types respectively, and also more than a recently published whole histology slide image based classifier which discriminates 18 cancer types [173]. We do acknowledge that discriminating even more cancer types is warranted, but this is currently limited by the amount of available training samples that were sequenced and uniformly bioinformatically analyzed. For example, thymic cancer had too few (<15) samples to be included as a separate class for training (**Supplementary data 1**). Furthermore, certain cancer types could be divided into their subtypes, such as Myeloid (currently only with 34 samples in total; **Figure 2a**) into acute myeloid leukemia and multiple myeloma [174], and sarcomas in a broader range of subtypes [175]. The availability of more WGS data for less frequent, but also (ultra-)rare cancers would allow for training of an updated CUPLR model that classifies additional cancer types and subtypes.

While CUPLR achieved overall excellent recall (90%) which is similar to or better than other WGS-based classifiers (**Supplementary figure 9**), it should be noted that this is driven by several large sub-cohorts of common cancer types (e.g. breast and colorectal cancer) and that performance for certain cancer types is still suboptimal. For cancer subtypes that CUPLR has difficulty discriminating, such as small cell versus non-small cell lung cancer and papillary versus clear cell renal carcinoma, additional information from histopathological stainings could be used to clarify these cases. Here, the application of artificial intelligence-based histology image analysis [173] could further improve the prediction performance and reliability of resolving CUPs. Clinical metadata, such as biopsy location and metastasis grade [176], when used together with CUPLR can also aid with clarifying primary tumor location. For instance, there may be uncertainty as to whether a tumor with human papillomavirus DNA integration was a head and neck cancer or cervical cancer based on CUPLR predictions (e.g. DO51592 with probability of Cervix=0.754 and HeadAndNeck_Other=0.310; **Supplementary data 5**), since human papillomavirus DNA integration is characteristic of these two cancer types [159]. However, in clinical practice, it will always be known if the tumor biopsy was taken from the upper body and if the

lesions were local, and with this information one can conclude that this tumor can only be of head and neck origin.

Given that RMD [90,91], mutational signatures [177], and SVs [28,178] are still active areas of research, we expect improvements to these features could also lead to better TOO classification. Here, we have demonstrated that extraction of cancer type specific RMD profiles is possible from raw mutation counts, similarly as was done for mutational signatures [13]. However, CUPLR does not heavily rely on RMD profiles for classification of certain cancer types such as liposarcoma and non-melanoma skin cancer (**Figure 3b**), potentially because more training samples are required to extract more stable and informative RMD profiles which could improve classification for these cancer types. Improvements to TOO classification may also come from extraction of more comprehensive mutational signatures, for example by incorporating information of mutation timing or genome localization [179,180]. Development of more sophisticated signature extraction methods may also allow for quantification of low signal tissue type specific signatures, such as SBS88 (associated with colibactin induced DNA damage in the colon) which has only been extracted from colon healthy tissue [65,181] but not cancer tissue likely because other mutational processes in cancer tissue mask the presence of this signature (**Supplementary figure 7**, **Supplementary figure 11**). Lastly, while CUPLR uses a wide range of features derived from SVs (including gene fusions, viral DNA integrations, LINE insertions, structural deletion and duplication size, and chromothripsis), there is still an opportunity to explore other SV related features to improve TOO classification, such as SV signatures [178].

Given that WGS is rapidly maturing and is now slowly being adopted in routine diagnostics for comprehensive molecular diagnostics [126,140], CUPLR serves as a viable complementary tool to standard procedures for determining TOO (e.g. histopathological stainings). CUPLR can be run from the output of open source tools and is freely available as an R package [https://github.com/UMCUGenetics/CUPLR]. The trained model as well as the code for generating the input features is provided to enable prediction on new samples and for facilitating diagnostic use.

Methods

Datasets

We have matched tumor/normal whole genome sequencing data from cancer patients from two cohorts: the Hartwig Medical Foundation (Hartwig) cohort and the Pan-Cancer Analysis of Whole Genomes (PCAWG) cohort.

The Hartwig cohort included 4902 metastatic tumor samples from 4572 patients. The data was provided under data request DR-104 from the Hartwig Medical Foundation. All patients in this resource have given consent for reuse of their genomic and clinical data for research purposes. For patients with multiple biopsies that were taken at different timepoints, patient IDs were suffixed by ‘A’ for the first biopsy, ‘B’ for the second biopsy, etc (e.g. HMF001423A, HMF001423B). Normal samples (blood) had a mean read coverage of ~30x while tumor samples had a coverage of ~90x [6]. Only a single sample of each patient was used for this study. To do this, we selected the tumor sample with the earliest biopsy date, and if this information did not exist we selected the sample with the highest tumor purity. However, some Hartwig patients had biopsies from different primary tumor locations. In these cases, we kept at least one sample from each primary tumor location, and when there were multiple samples from the same primary tumor location, we applied the aforementioned biopsy date and tumor purity filtering criteria.

The PCAWG cohort consisted of 2835 patient tumors. Access for raw sequencing data for these tumors was granted via the Data Access Compliance Office (DACO) Application Number DACO-1050905 on 6 October 2017 and via the Cancer Genome Collaboratory download portal [https://console.cancercollaboratory.org] on 4 December 2017. Normal samples (blood, adjacent tumors or distant tumors) had a mean read coverage of 39x, while tumor samples had a bimodal coverage distribution with modes at 38x and 60x [5]. Samples with <0.2 tumor purity were excluded from this study as somatic variant calling was not reliable for these samples. PCAWG samples that were gray- or blacklisted by the PCAWG consortium were also excluded [https://dcc.icgc.org/releases/PCAWG/donors_and_biospecimens].

For both cohorts, we only kept samples with >=50 SNVs/indels, and removed an additional set of samples for several reasons including due to failed variant calling, insufficient informed consent for use of the WGS data, and one duplicate PCAWG sample (DO217844) that was already included in the Hartwig cohort. Lastly, we only selected samples from cancer types with at least 15 samples. Ultimately, we selected 4391 Hartwig samples and 2365 PCAWG samples for training, as well as 141 Hartwig CUP samples for the CUP analysis (**Supplementary data 1**).

Variant calling

Somatic mutation data of the CPCT, DRUP and WIDE projects were kindly shared by Hartwig on 6 February 2020 with an update received on 20 October 2021. To exclude technical noise from PCAWG and Hartwig somatic variant calling workflows, we have reanalyzed the PCAWG samples with the Hartwig pipeline for somatic variant calling [https://github.com/hartwigmedical/pipeline5] which was hosted on the Google Cloud Platform using Platinum [https://github.com/hartwigmedical/platinum] [146]. Details of the full pipeline are described by Priestley *et al.* [6] as well as in the Hartwig pipeline GitHub page. Briefly, reads were mapped to GRCh37 using BWA (v0.7.17). GATK (v3.8.0) Haplotype Caller was used for calling germline variants in the reference sample. SAGE (v2.2) was used to call somatic single and multi-nucleotide variants as well as indels. GRIDSS (v2.9.3) was used to call SVs. PURPLE combines B-allele frequency (BAF) from AMBER (v3.3), read depth ratios from COBALT (v1.7), and structural variants from GRIDSS to estimate copy number profiles, variant allele frequency (VAF) and variant clonality. Additionally, PURPLE also determines the gender (based on sex chromosome ploidy), the proportion of the genome that is diploid, as well as the presence of whole genome duplication in a sample. LINX interprets SVs (to identify simple and complex structural events) from PURPLE, and also detects gene fusions, viral DNA integrations, and homozygously disrupted genes. Importantly, we ensured that mutation (simple and complex) filtering and annotation tools were run with the same versioning for PCAWG and HMF cohort. For PURPLE we relied on v2.53 whereas for LINX we used v1.17.

Extraction of features

Regional mutational density

RMD was defined as the number of somatic SBSs in each 1Mb bin across the genome (n=3071), normalized by the total number of SBSs in the sample. Extraction of RMD profiles from these RMD bins was performed within the CUPLR training procedure (**Supplementary figure 1a**) using non-negative matrix factorization (NMF). This is described in detail in the “CUPLR training procedure” methods section. See **Supplementary data 6** for a visualization of each RMD profile.

Mutational signatures

The number of somatic mutations falling into the 96 SBS, 78 DBS and, 83 indel contexts (as described in COSMIC: [https://cancer.sanger.ac.uk/signatures/]) was determined using the R package mutSigExtractor [https://github.com/UMCUGenetics/mutSigExtractor], v1.23). To obtain the mutational signature contributions for each sample, the mutation context counts were fitted to the COSMIC catalog of mutational signatures using the nnlm() function from the NNLM R package.

The contributions of the child signatures SBS7a, SBS7b, SBS7c and SBS7d were summed to yield the parent signature SBS7. Similarly SBS10a-d and SBS17a-b were merged to yield SBS10 and SBS17. Lastly, the SBS, DBS and indel signature contributions were normalized by the total number of SBSs, DBSs and indels respectively.

Chromosome arm ploidy

Chromosome arm ploidies were determined in a similar method as described by Taylor et al. 2018 [35].

Somatic copy number (CN) segments (called by PURPLE) were split by their respective chromosome arms. Only chr1-22 and chrX were included. All chromosomes have p and q arms, except for chr13, chr14, chr15, chr21, and chr22 which are considered to only have the q arm. For each chromosome arm, the CN values of each segment were converted to integer values. The arm coverage of each CN integer value was then determined (e.g. 70% of the arm has a CN of 2, 20% a CN of 1 and 10% a CN of 3). The CN with the highest coverage was assigned as the preliminary arm CN. The most common CN across all arms was assigned as the genome CN.

Two filtering steps were then performed to obtain the final arm CN values. For each arm, if the CN with the highest coverage has <50% coverage, and if any of the CN values equal the genome CN, then assign that CN as the final CN of the arm. Else, assign the genome CN as the final CN of the arm.

To determine the CN gains and losses of each arm, the fold change between the arm CN and genome CN was calculated.

Features extracted from LINX output

LINX combines simple mutations (point mutations and indels), structural variants, and copy number variants to resolve simple and complex structural rearrangements, and subsequently identify gene driver events, gene fusions and detect viral DNA integrations.

The presence of 4 types of gene driver events were determined from the output of LINX: (i) amplifications, (ii) deep deletions, (ii) biallelic loss, and (iv) monoallelic hits (pathogenic mutation in one allele). Amplified genes were marked as likelihoodMethod=='AMP' by LINX. Genes with deep deletions were marked as likelihoodMethod=='DEL'. Genes with biallelic loss were marked as biallelic==TRUE. Genes with a monoallelic pathogenic mutation were marked as biallelic==FALSE and driver=="MUTATION", and we selected only those with driverLikelihood>=0.9 (referring to the likelihood of the mutation being impactful as determined by the dndscv R package [49]). LINX determines the presence of driver events for 462 genes. Thus, there are 462 genes x 4 driver types = 1848 gene driver features in total. We then performed preliminary feature selection to reduce the computational resources required for training CUPLR. Here, one-sided Fisher's exact tests were performed and Cramer's V values were calculated. Only genes where at least one driver type had a p-value <0.01 and Cramer's V ≥0.1 were kept (203 genes x 4 driver type = 812 gene driver features).

Gene fusions belonged to 3 categories: (i) well known fusion pairs (e.g. *TMPRSS2-ERG*), (ii) immunoglobulin heavy chain (IGH) locus fusions, and (iii) fusions with promiscuous gene partners (e.g.

BCR). *IGH* fusions were grouped into a single feature as these are characteristic events in lymphoid cancers [182]. Fusions with one promiscuous gene partner were grouped by gene (e.g. *RUNX1_ETS2* and *RUNX1_RCAN1* would both fall under the *RUNX1_** feature). Fusions with two promiscuous gene partners were split into two separate features (e.g. *SLC45A3_MYC* would become the features *SLC45A3_** and **_MYC*). Only fusions that were marked as reported==TRUE by LINX (i.e. reported in literature) were selected. We then performed preliminary feature selection due to the large number of possible fusions present in our dataset (n=512). Here, one-sided Fisher's exact tests were performed and Cramer's V values were calculated. Only fusions with p-value <0.01 and Cramer's V >=0.1 were kept (46 fusions).

For the viral DNA integrations present in our dataset, we merged virus strains into 9 virus categories: adeno-associated virus (AAV), Epstein-Barr virus (EBV), hepatitis B virus (HBV), hepatitis C virus (HCV), human immunodeficiency virus (HIV), human papillomavirus (HPV), herpes simplex virus (HSV), human T-lymphotropic virus (HTLV), and Merkel cell polyomavirus (MCPyV). For example, *human papillomavirus type 16* and *human papillomavirus type 18* would be both grouped as *human papillomavirus*.

LINX chains individual SVs into SV clusters and classifies these clusters into various event types. Clusters can have one SV (for simple events such as deletions and duplications), or multiple SVs. We defined SV load as the total number of SV clusters. We quantified the presence of several SV types including: i) deletions and duplications (ResolvedType is 'DEL' or 'DUP') stratified by length (1–10kb, 10–100kb, 100kb–1Mb, 1–10Mb, >10Mb), ii) complex SV clusters (ResolvedType=='COMPLEX') stratified by the number of clusters (0-25, 25-50, 50-100, 100-200, 200-400, 400-800, >800), iii) long interspersed nuclear element (LINE) insertions (ResolvedType=='LINE'), and double minutes (ResolvedType=='DOUBLE_MINUTE'). Lastly, we also determined the number of breakends for the complex cluster with the most breakends.

CUPLR training procedure

Extraction of regional mutational density profiles

To extract the cancer specific RMD profiles from the 3071 RMD bins, a multistep procedure involving non-negative matrix factorization (NMF) (**Supplementary figure 2**) was performed prior to classifier training (**Supplementary figure 1ai**).

All NMF runs described below are performed with the nnmf() function from the NNLM R package (v0.4.4) with the parameters loss='mkl' and max.iter=2000.

For each cancer type cohort, an NMF rank search was done to determine the optimum rank (i.e. number of RMD profiles) (**Supplementary figure 2a**). For ranks 1 to 10, NMF was performed 50 times on a random subset of 100 samples from the cohort (or if the cohort contained less than 100 samples, all samples from that cohort were used) with 10% of the values randomly removed. The missing values were then imputed and the mean squared error (MSE) of these imputed values was calculated. This method of calculating MSE is described by the authors of the NNLM R package [183]. The median of the MSE across the 50 NMF iterations was then calculated. The rank search thus results in 10 MSE values across the 10 ranks searched. The optimum rank was the rank before the increase in log10(MSE) was >0.2%, and NMF was then performed using the optimum rank and without removing random values to produce the RMD profiles for the cancer type cohort (**Supplementary figure 2b**).

The above procedure thus yields a different set of RMD profiles for each cancer type cohort. However, some RMD profiles across related cancer types (e.g. pancreas and biliary cancer) may actually be equivalent RMD profiles. Hierarchical clustering (using Pearson correlation as a distance measure) was

thus performed to group similar RMD profiles across all cancer type cohorts. The resulting dendrogram was cut at a height of 0.1 (using the R function cutree()), whereby RMD profiles under this height were grouped and considered the same profile. From each of the groups, one profile was greedily selected to yield the final set of RMD profiles (**Supplementary figure 2c**).

To obtain the RMD profile contributions for each sample, the RMD bins were fitted to the RMD profiles using the nnlm() function from the NNLM R package.

Random forest ensemble training

The main component of CUPLR comprises an ensemble of binary random forests that each discriminates one cancer type (**Supplementary figure 1aii**). The below text describes the training procedure for each cancer type random forest.

First, univariate feature selection was performed to remove irrelevant features (**Supplementary figure 1aiii**). Pairwise testing was done to compare feature values from samples of the target cancer type (case group) versus the remaining samples (control group). For numeric features, p-values were determined using Wilcoxon rank sum tests, and effect sizes were calculated using Cliff's delta. For boolean features, p-values were determined using Fisher's exact tests, and effect sizes were calculated using Cramer's V. Depending on the feature, alternative hypotheses for the Wilcoxon rank sum tests and Fisher's exact tests were one or two sided. See **Supplementary data 3** for details on which features are numeric or boolean, as well as which alternative hypothesis was used. Features were kept which had $p<0.01$ and effect size ≥ 0.1. The number of features kept was capped to 100 features.

Second, random oversampling was performed for the case group which always contained fewer samples than the control group, which was randomly undersampled (**Supplementary figure 1aiv**). A grid search was performed to determine the optimal pair of 5 oversampling and 5 undersampling ratios. These ratios were automatically determined as follows: i) calculate the geometric mean between the case and control group sample sizes; ii) the resampling ratios are logarithmically spaced between the geometric mean and the case group sample size or the control group sample size. For each over-/undersampling ratio pair, stratified 10-fold cross-validation (CV) was performed, after which the area under the precision recall curve (AUPRC) was calculated. The pair with the highest AUPRC was chosen and the resampling was applied. CV and AUPRC calculations were performed using the mltoolkit R package [https://github.com/UMCUGenetics/mltoolkit].

Lastly, a random forest was trained (**Supplementary figure 1av**) using the randomForest R package (v4.6-14) with default settings. A filter is applied to the probabilities produced by the random forest based on sample gender, where breast, ovary and cervix probabilities are set to 0 for male samples, and prostate probabilities are set to 0 for female samples. Local increments were calculated for the random forest using the rfFC R package (v1.0) to enable downstream calculation of feature contributions [163].

Isotonic regression training

The entire random forest ensemble training procedure was then subjected to stratified 15-fold cross-validation which allows every sample to be excluded from the training set in order to obtain cancer type probabilities for the training samples (**Supplementary figure 1b**). These cross-validation probabilities were then used to train an ensemble of isotonic regressions using the isoreg() R function (one per cancer type random forest) to calibrate the probabilities produced by the random forest ensemble (**Supplementary figure 1c, Supplementary figure 3**).

Random forests tend to be overconfident at probabilities towards 0 and underconfident at probabilities towards 1 [148], and this bias varies between random forests (**Supplementary figure 4**). In other words, a probability of e.g. 0.8 from one random forest does not correspond to a probability of 0.8 from another random forest. Probability calibration greatly reduced this bias ensuring that predictions across the random forests are comparable (**Supplementary figure 4**).

Performance evaluation

To assess the performance of CUPLR, we used the cancer type predictions based on the isotonic regression calibrated cross-validation probabilities, as well as by applying the final model to a validation set whereby 10% of samples were held out from the full training set (**Supplementary figure 1**). Performance metrics per cancer type were defined as follows (using 'Breast' as an example):

- Recall = Fraction of Breast samples predicted as Breast
- Top-2 recall = Fraction of Breast samples where the 1st or 2nd top prediction was Breast
- Precision = Amongst samples predicted as Breast, the fraction of samples labeled as Breast

Overall performance metrics were micro-averages of the per-cancer-type metrics and defined as follows:

- Micro-averaged recall = Accuracy = Fraction of all samples correctly predicted
- Micro-averaged top-2 recall = Top-2 accuracy = Fraction of all samples where the 1st or 2nd highest probability prediction was correct
- Micro-averaged precision = Micro-average of per-cancer-type precision

Precision and recall curves for each binary random forest classifier within CUPLR are shown in **Supplementary figure 12**.

We used predictions based on calibrated cross-validation probabilities to assess the effect of excluding feature types on recall (**Supplementary figure 8**); compare the recall of CUPLR to other classifiers (**Supplementary figure 9**); assess the effect of confounding factors on recall (**Supplementary table 1**, **Supplementary figure 13**, **Supplementary Notes**); as well as to show that 30x coverage is sufficient for reliable CUPLR predictions for most cancer types, with >=60x coverage giving the most reliable predictions (**Supplementary figure 14**, **Supplementary Note 2**).

Lastly, we showed that there was likely minimal overfitting on the Hartwig and/or PCAWG datasets (i.e. batch effects) by training a CUPLR-like a model solely on Hartwig samples and another model solely on PCAWG samples, and thereafter determining performance by testing on the opposite dataset (**Supplementary figure 15**, **Supplementary Note 3**).

Data availability

For the Hartwig cohort, WGS data and corresponding metadata have been obtained from the Hartwig Medical Foundation and provided under data request number DR-104. Both WGS data and metadata is freely available for academic use from the Hartwig Medical Foundation through standardized procedures and request forms which can be found at [https://www.hartwigmedicalfoundation.nl].

Somatic variant calls, gene driver lists, copy number profiles and other core data of the PCAWG cohort generated by the Hartwig analytical pipeline are available for download at [https://dcc.icgc.org/releases/PCAWG/Hartwig]. Researchers will need to apply to the ICGC data access compliance office (https://daco.icgc-argo.org) for the ICGC portion of the dataset. Authentication of NIH eRA commons is required to access the TCGA portion of the dataset via [https://icgc.bionimbus.org]. Additional information on accessing the data, including raw read files,

can be found at [https://docs.icgc.org/pcawg/data/]. Metadata for PCAWG samples (e.g. sample whitelisting) can be found at https://dcc.icgc.org/releases/PCAWG. The extracted features for each sample and used to develop CUPLR is available at [https://doi.org/10.5281/zenodo.5939805] [184]. All other processed and raw data can be found in the **Supplementary Data** files.

Code availability

The Hartwig Medical Foundation pipeline [https://github.com/hartwigmedical/pipeline5], hosted on the Google Cloud Platform using Platinum [https://github.com/hartwigmedical/platinum], was used for germline and somatic variant calling, as well as post-processing procedures such as identification of simple and complex structural rearrangements, annotation of driver gene mutation events, and detection of gene fusions and viral DNA integrations. CUPLR can be run from the output of this pipeline, and is available as an R package on GitHub ([https://github.com/UMCUGenetics/CUPLR]; [https://doi.org/10.5281/zenodo.6637693] [185]). This repository also contains the code for data processing and generating the figures in this paper. CUPLR depends on some custom code, including mutSigExtractor for extraction of mutational signatures [https://github.com/UMCUGenetics/mutSigExtractor] and mltoolkit (only required for classifier training and not for running CUPLR; [https://github.com/UMCUGenetics/mltoolkit]).

Acknowledgements

This research was supported by an unrestricted grant (2020-Cuppen) of Stichting Hanarth Fonds, The Netherlands (received by E.C.). This publication and the underlying study have been made possible partly on the basis of the data that Hartwig Medical Foundation and the Center of Personalized Cancer Treatment (CPCT) have made available to the study.

Author contributions

L.N. performed analyses, wrote/edited the paper. A.V.H. conceived the study, performed analyses, wrote/edited the paper. E.C. edited the paper and provided discussion. E.C. and A.V.H. supervised the study. All authors proofread, made comments, and approved the paper.

Competing interests

The authors declare no competing interests.

Supplementary data

Supplementary data 1: Metadata for tumors in the Hartwig and PCAWG datasets, including cancer type label, HRD and MSI status, and inclusion/exclusion in the training and holdout sets.

Supplementary data 2: The number of training and held out test samples per cancer type per cohort.

Supplementary data 3: Descriptions and importance of each feature used by CUPLR.

Supplementary data 4: Confusion matrices and performance metrics.

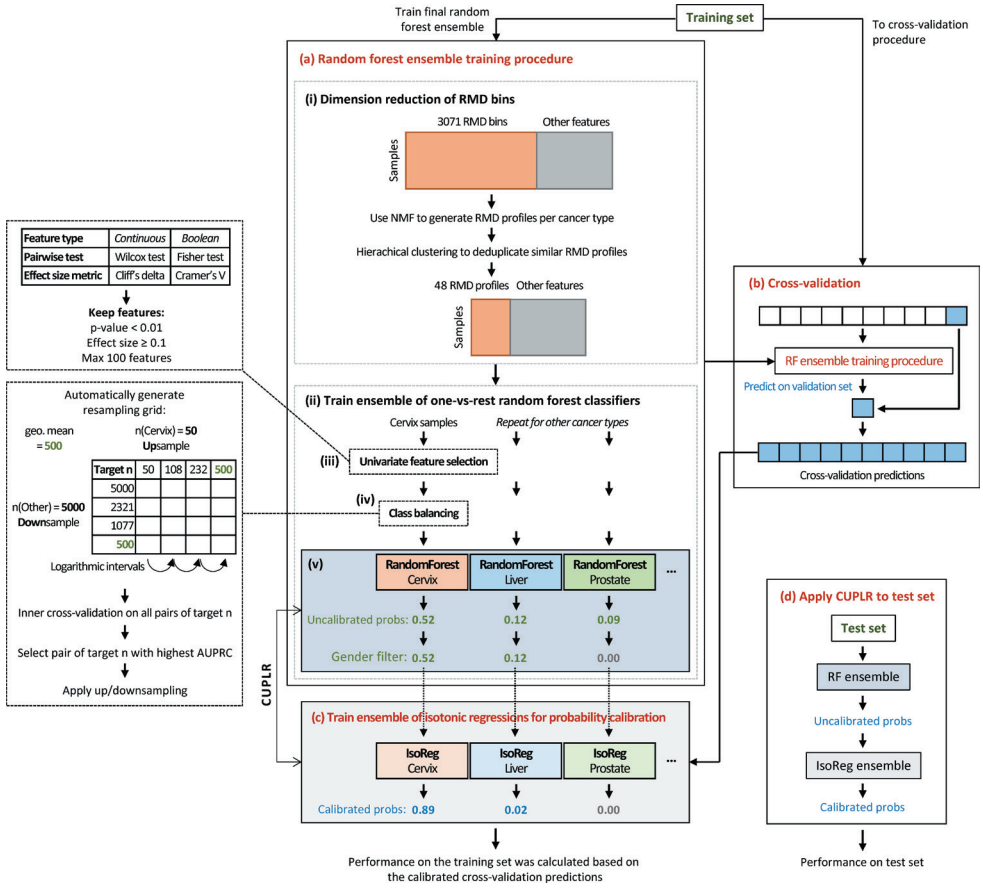
Supplementary data 5: CUPLR predictions on the training set (based on cross-validation), held out test set, and on cancer of unknown primary (CUP) samples.

Supplementary data 6: Visualization of the regional mutational density (RMD) profiles.

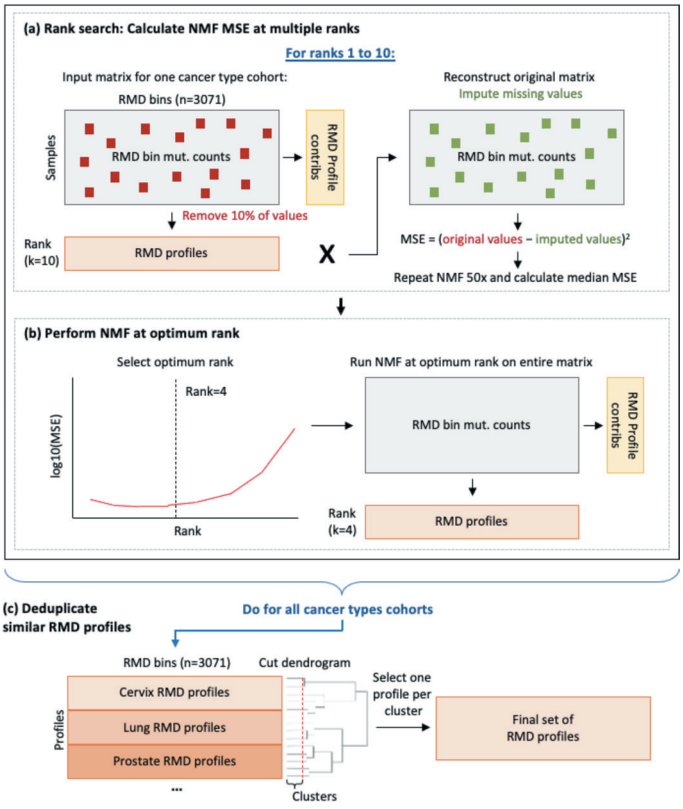
Supplementary data are available online at https://www.nature.com/articles/s41467-022-31666-w#Sec24, or by scanning the QR code below:

Supplementary information

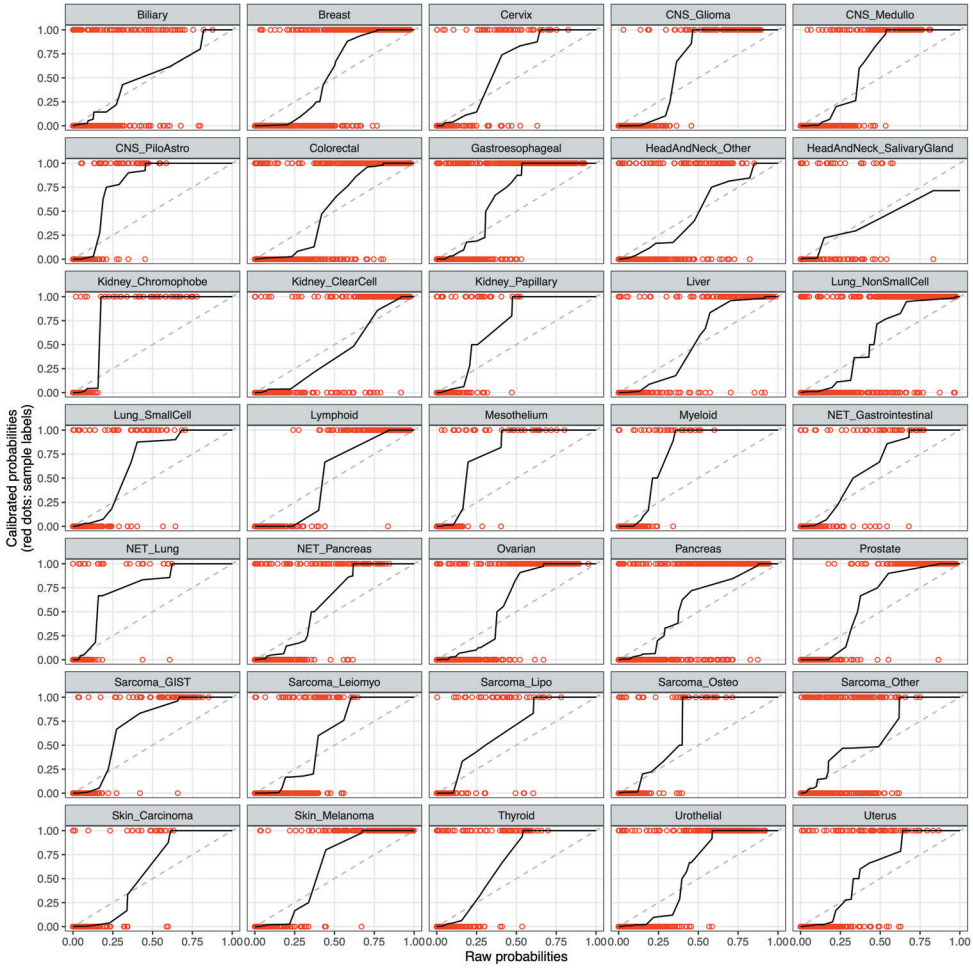
Supplementary figures



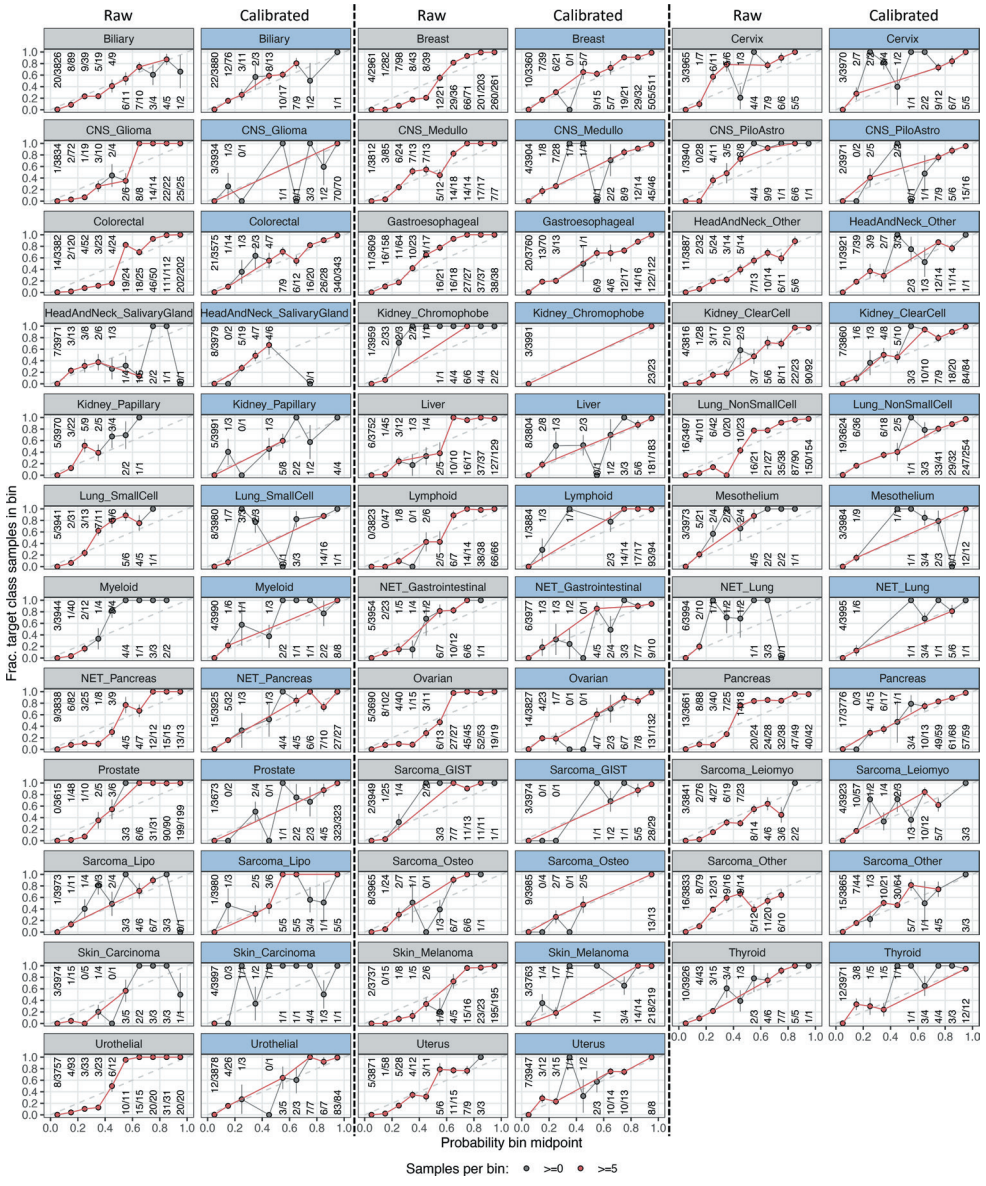
Supplementary figure 1: CUPLR training procedure. (a) An ensemble of binary random forest classifiers was trained each to discriminate one cancer type versus other cancer types. **(i)** Dimension reduction via non-negative matrix factorization (NMF) was performed on the 3071 RMD bins independently for each cancer type to ultimately produce 48 cancer type specific RMD profiles (see **Supplementary figure 2** for a detailed schematic; see **Supplementary data 6** for a visualization of each RMD profile), prior to random forest training as shown in **(ii)**. **(iii)** Univariate feature selection was performed to remove irrelevant features. (**iv**) Class resampling was performed to alleviate imbalances in the number of samples for each cancer type. (**v**) The random forests are trained. Breast, ovary and cervix probabilities from the random forest are set to 0 for male samples, and prostate probabilities are set to 0 for female samples. **(b)** The whole training procedure in **(a)** was subjected to 15-fold cross-validation to obtain cancer type probabilities for the training samples. **(c)** These probabilities were then used to train the isotonic regressions for probability calibration. The calibrated probabilities were also used to calculate cross-validation performance. **(d)** Performance was also determined by applying CUPLR to a held out test set.



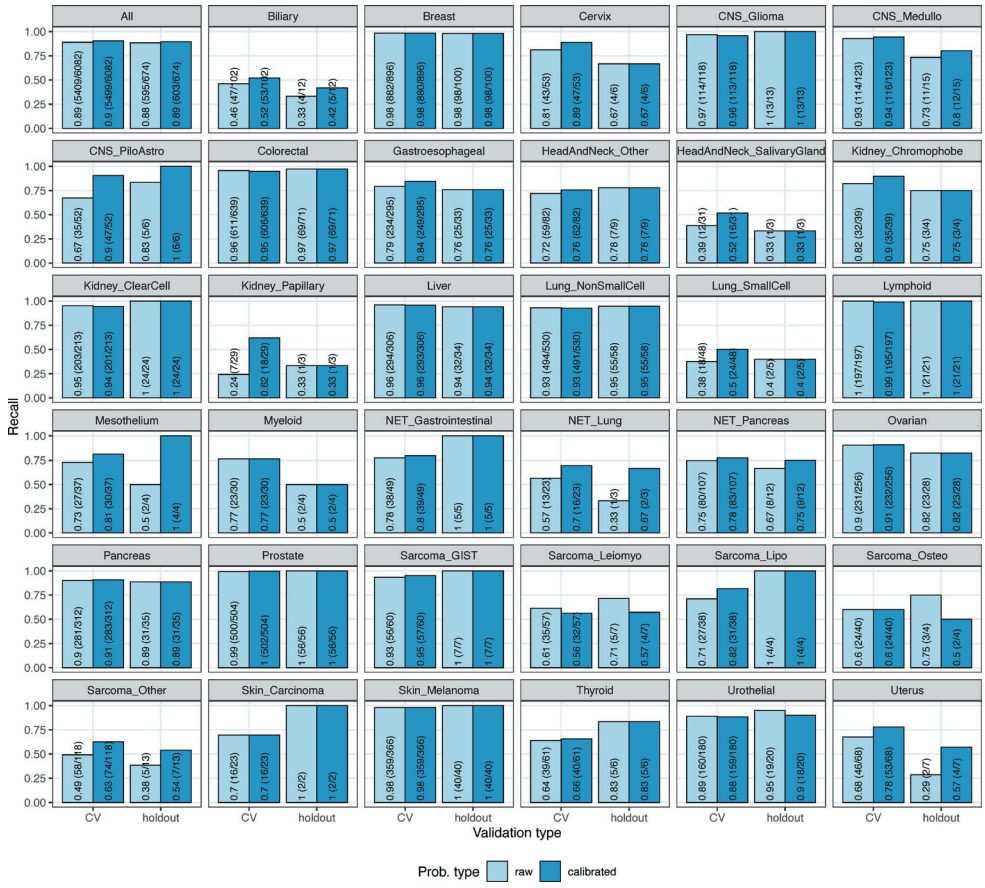
Supplementary figure 2: Extraction of regional mutational density (RMD) profiles. (a) Non-negative matrix factorization (NMF) was performed for several ranks to determine the optimum rank. For each rank, 10% of values were removed from the input matrix, NMF was performed, the original matrix was reconstructed, the missing values were imputed, and mean squared error between the original missing values and the imputed values. This was repeated 50 times and median MSE was calculated. **(b)** The optimum rank was the one at which the log10(MSE) value began to increase rapidly. NMF was performed on the entire input matrix (i.e. without removing values) to yield the RMD profiles for one cancer type cohort. **(c)** The procedures in (a) and (b) were performed for all cancer types to yield RMD profiles for all cancer types. Hierarchical clustering was performed to group profiles that were similar. One profile was selected per cluster to yield the final set of RMD profiles.



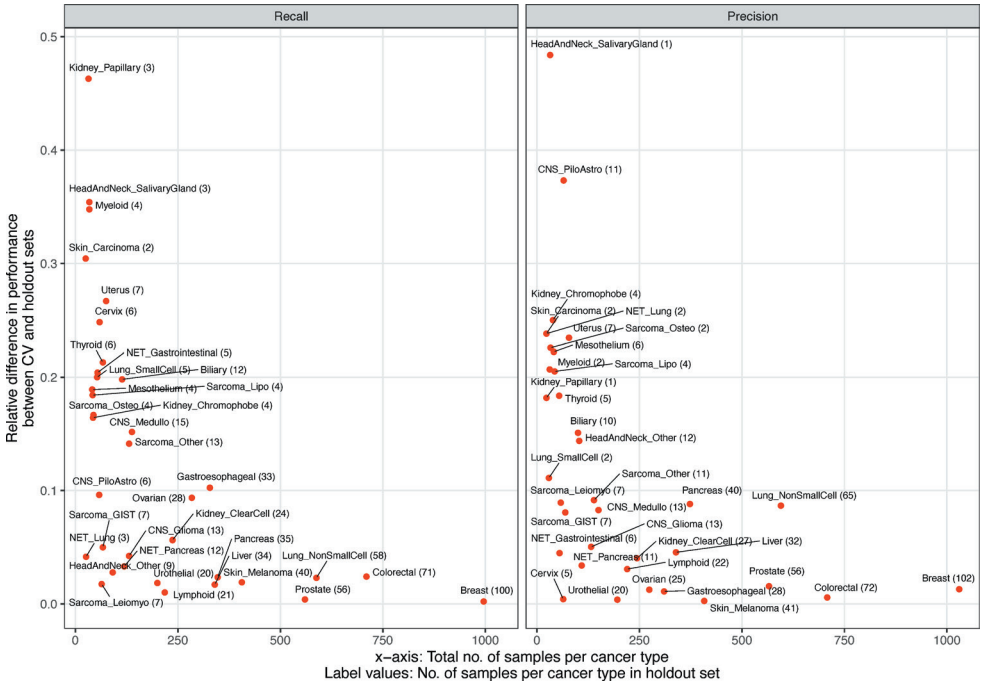
Supplementary figure 3: Isotonic regression calibration curves for each random forest in CUPLR. Red dots at y=1 are samples that were predicted by the respective cancer type random forest as that cancer type, whereas dots at y=0 are samples that were predicted as not being that cancer type. Cancer type predictions were obtained by cross-validation.



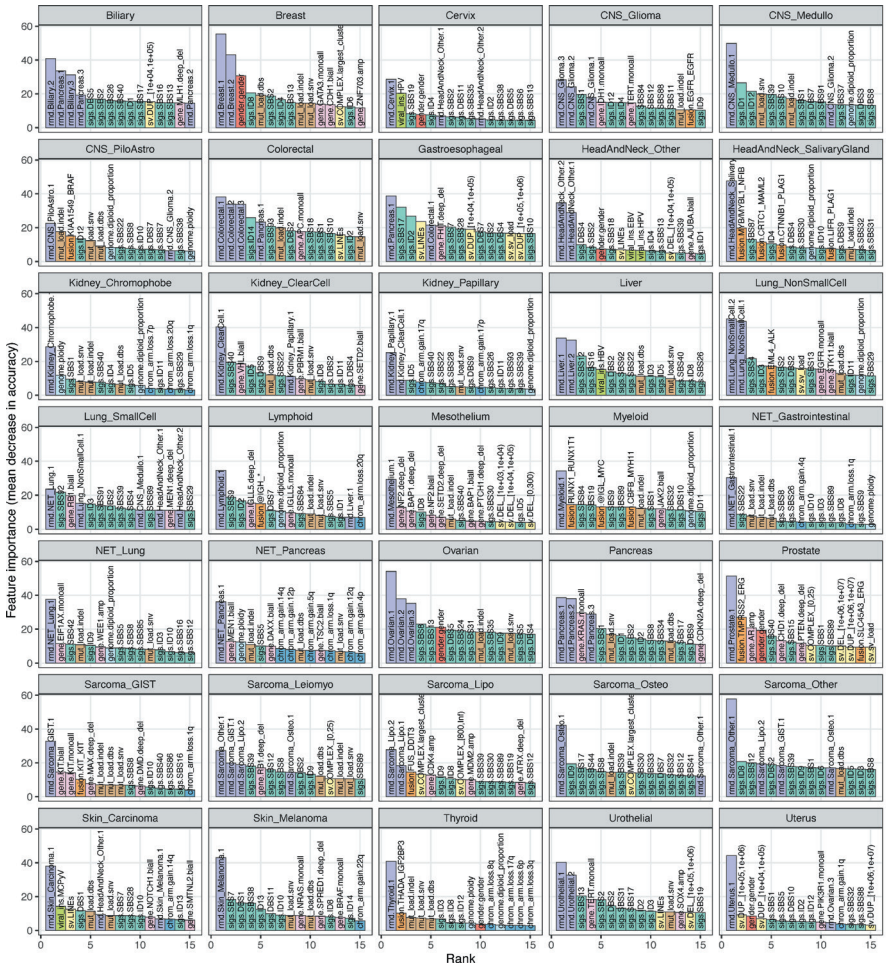
Supplementary figure 4: Reliability curves showing the probability biases before and after calibration. Grey panels show the curves before calibration and the blue panels show the curves after calibration. Each dot represents the fraction of samples of the target cancer type in a particular probability bin (a bin at e.g. 0.05 would represent probabilities between 0 and 0.1). Dots above the diagonal are probabilities where the random forest is overconfident whereas dots below the diagonal are probabilities where it is underconfident. A properly calibrated classifier has a reliability curve close to the diagonal. Two curves are shown in each panel, one filtered such that each bin has sufficient (≥5) samples for plotting a stable curve, and one where each bin has ≥0 samples representing the raw reliability curve.



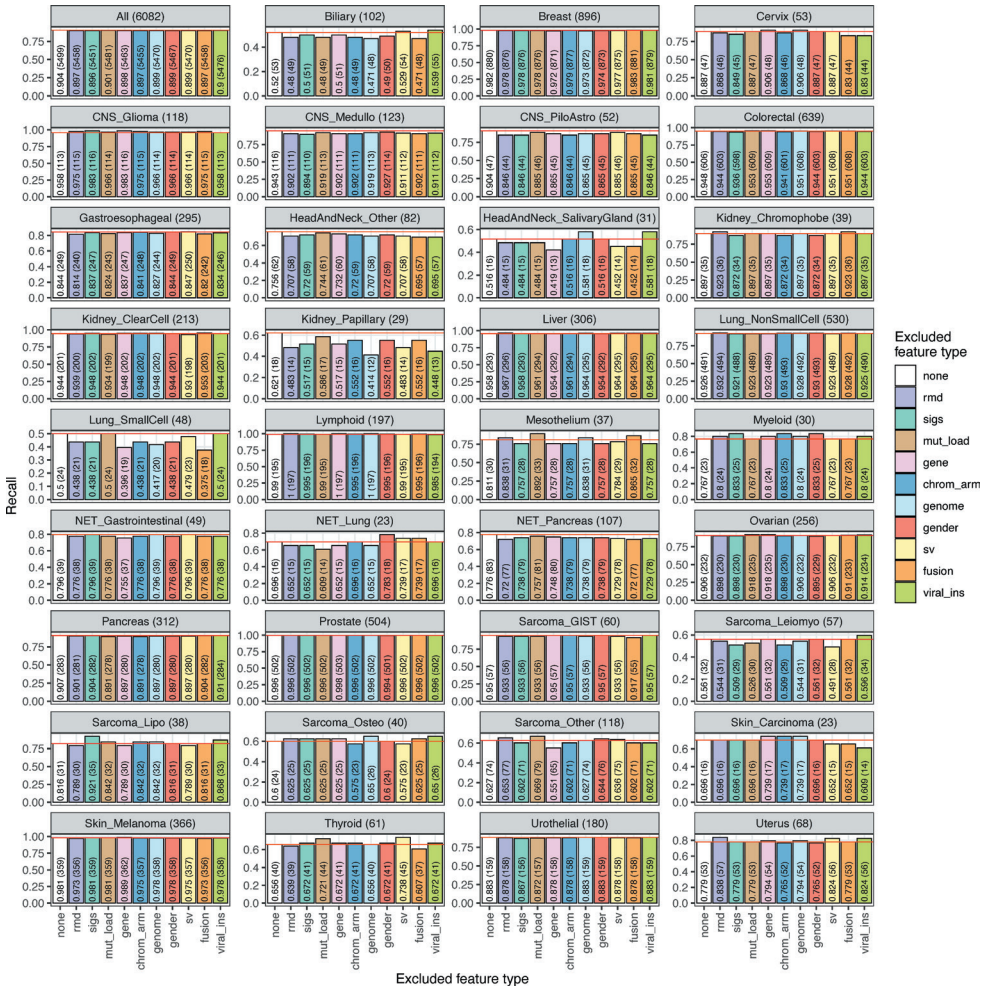
Supplementary figure 5: Performance of CUPLR before and after calibration of raw random forest probabilities. Recall (i.e. fraction of samples correctly predicted) was determined using cross validation on the training set (light blue bars) as well as by predicting on the held out test set (dark blue).



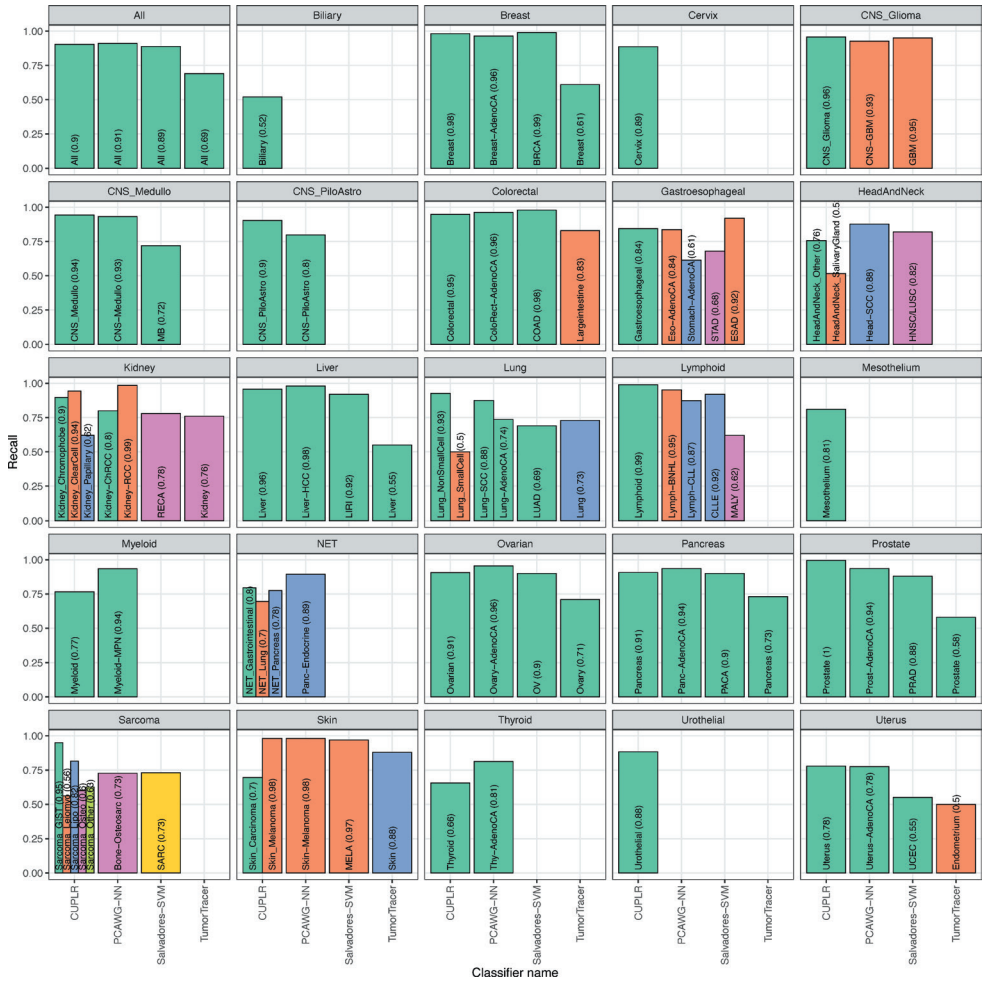
Supplementary figure 6: Relative difference in recall of CUPLR between the cross-validation (CV) predictions and predictions on the holdout test set. The relative difference ranges between 0 and 1 and was calculated using the formula: |a - b| / max(a, b); where a = CV performance and b = holdout performance. Small cancer type cohorts show higher variation between CV and holdout recall and precision.



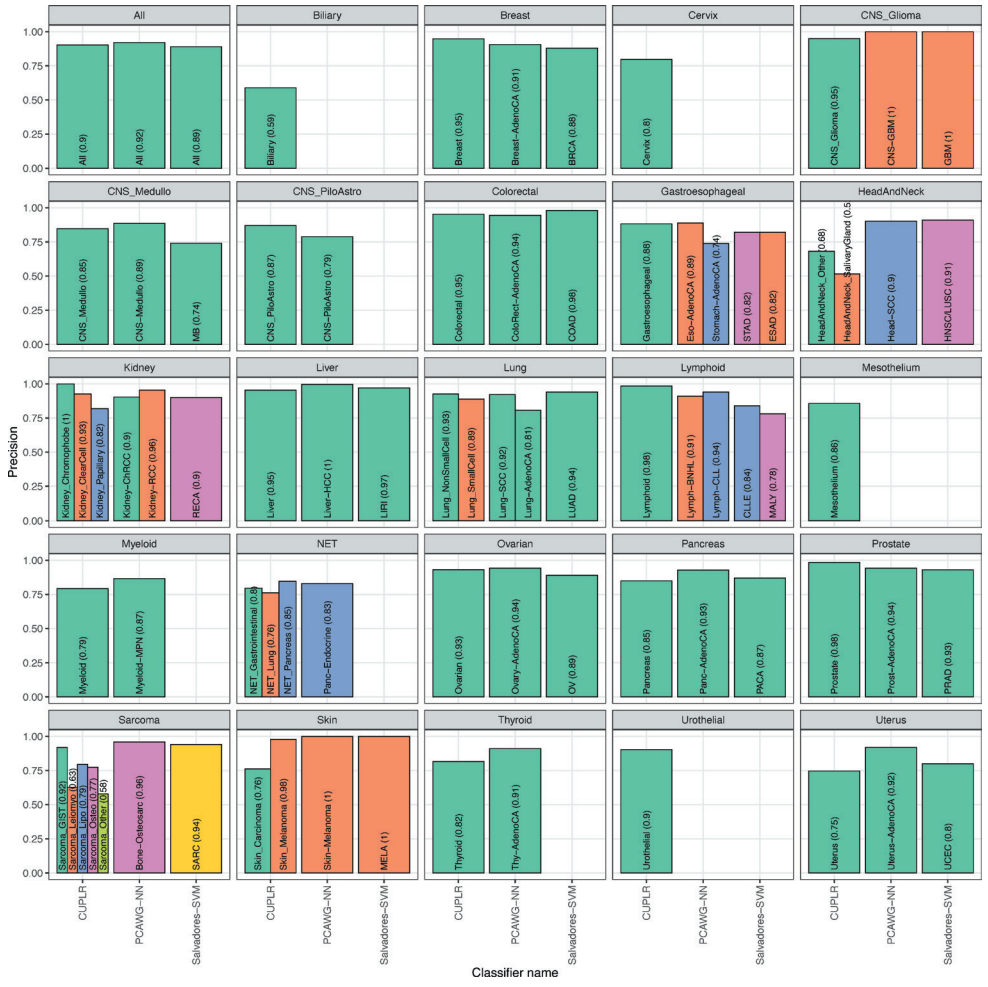
Supplementary figure 7: **Feature importances from the top 15 features for each random forest within CUPLR**. Feature importance is measured by mean decrease in accuracy across all trees in a random forest upon removing a particular feature. Feature names are in the form {feature type}.{feature name}. **Feature type definitions**; rmd: regional mutation density profiles; sigs: mutational signatures; mut_load: total number of single base substitutions, double base substitutions or indels; gene: presence of gene gain or loss of function events; chrom_arm: chromosome arm copy number fold change versus overall genome ploidy; genome: genome properties including genome ploidy, diploid proportion, whole genome duplication status; gender: sample gender as determined by copy number data; sv: structural variants; fusion: presence of gene fusions; viral_ins: presence of viral sequence insertions. See **Supplementary data 3** for the descriptions as well as feature importance values for each feature.



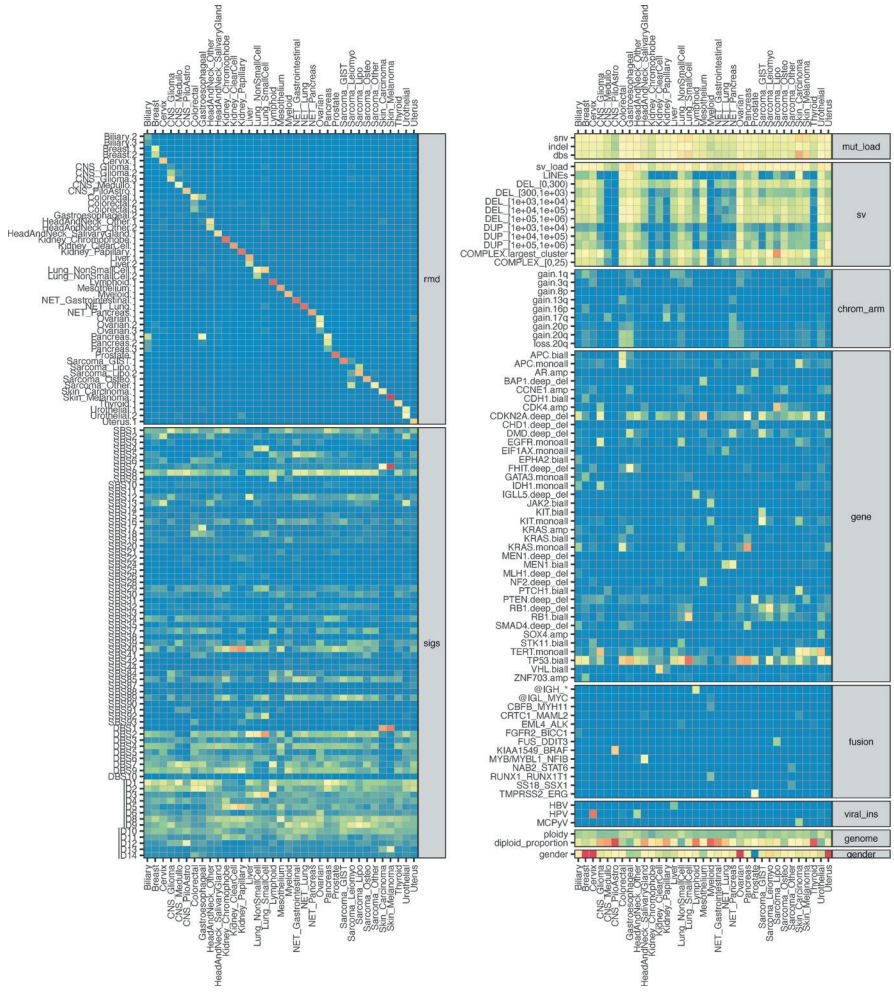
Supplementary figure 8: Cancer type prediction recall when excluding feature types from the training data. Recall (i.e. fraction of samples correctly predicted) was determined using cross validation. Panel titles indicate the cancer type and the total number of samples for that cancer type. Bars are labelled with the recall as well as the number of correctly predicted samples. The red line shows the recall when no feature types are excluded. **Feature type definitions**; rmd: regional mutation density profiles; sigs: mutational signatures; mut_load: total number of single base substitutions, double base substitutions or indels; gene: presence of gene gain or loss of function events; chrom_arm: chromosome arm copy number fold change versus overall genome ploidy; genome: genome properties including genome ploidy, diploid proportion, whole genome duplication status; gender: sample gender as determined by copy number data; sv: structural variants; fusion: presence of gene fusions; viral_ins: presence of viral sequence insertions. For a full description of each feature, see **Supplementary data 3**.



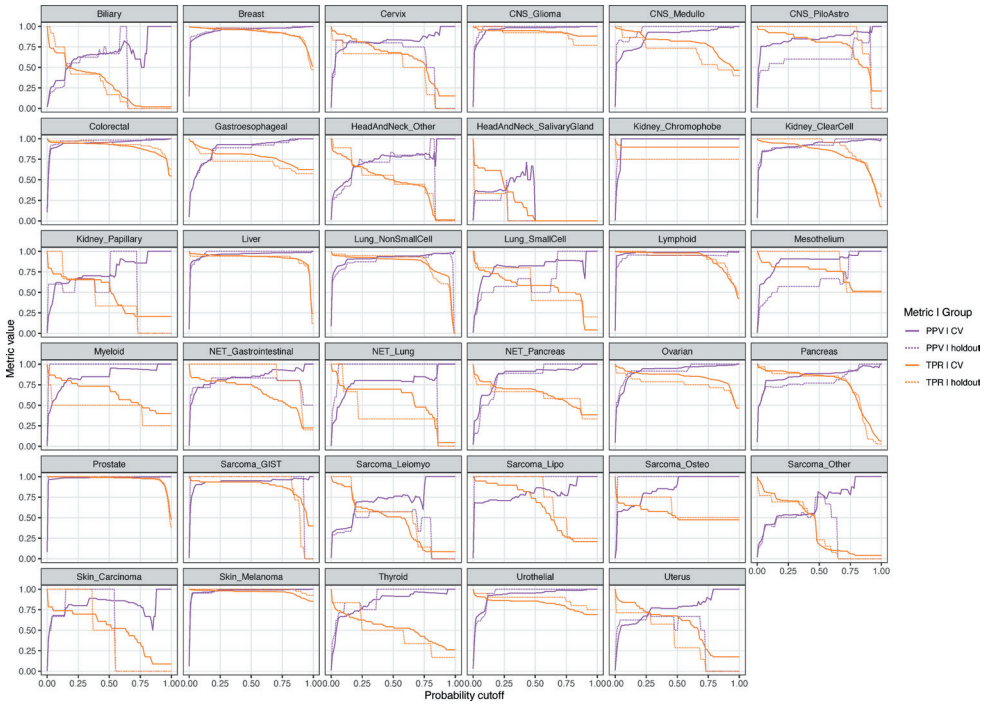
Supplementary figure 9: **Recall of CUPLR compared to other published classifiers.** Recall (i.e. fraction of samples correctly predicted) for CUPLR was determined using cross validation. Bars are labeled with the cancer type class names from the respective classifiers as well as the recall value. Cancer type class names representing the same cancer type across different classifiers are represented by the same bar colors. The names of the published classifiers refer to the following studies; PCAWG-NN: PCAWG neural network by Jiao *et al.* 2020, Salvadores-SVM: support vector machine by Salvadores *et al.* 2019, TumorTracer: Marquard *et al.* 2015.



Supplementary figure 10: **Precision of CUPLR compared to other published classifiers.** Precision for CUPLR was determined using cross validation. Bars are labeled with the cancer type class names from the respective classifiers as well as the precision value. Cancer type class names representing the same cancer type across different classifiers are represented by the same bar colors. The names of the published classifiers refer to the following studies; PCAWG-NN: PCAWG neural network by Jiao *et al.* 2020, Salvadores-SVM: support vector machine by Salvadores *et al.* 2019. No precision values were provided by Marquard *et al.* 2015 for TumorTracer.



Supplementary figure 11: **Average scaled feature value per cancer type for the top 200 features**. Feature values are scaled from 0 to 1 based on the minimum and maximum value for each feature across all samples. For gender, feature values towards 1 represent more female samples and values towards 0 more male samples. **Feature type definitions**; rmd: regional mutation density profiles; sigs: mutational signatures; mut_load: total number of single base substitutions, double base substitutions or indels; gene: presence of gene gain or loss of function events; chrom_arm: chromosome arm copy number fold change versus overall genome ploidy; genome: genome properties including genome ploidy, diploid proportion, whole genome duplication status; gender: sample gender as determined by copy number data; sv: structural variants; fusion: presence of gene fusions; viral_ins: presence of viral sequence insertions. For a full description of each feature, see **Supplementary data 3**.



Supplementary figure 12: Positive predictive value (PPV; also known as precision) and true positive rate (TPR; also known as sensitivity or recall) curves for each binary random forest classifier within CUPLR for the based on cross-validation (CV) and holdout set predictions.

Supplementary notes

Supplementary note 1: Impact of confounding factors on performance

Since the RMD profiles and mutational signatures were the most important feature types for CUPLR, we assessed whether the presence of certain confounding factors to these feature types would lead to more incorrect predictions. Firstly, this included DNA repair deficiencies including microsatellite instability (MSI) and homologous recombination deficiency (HRD) which lead to SBS/indel accumulation across the genome and could lead to different RMD profiles than would be expected for a particular cancer type. Secondly, the impact of common chemotherapies including platinum and 5-fluorouracil (5FU) was assessed as treatment induced mutations could also lead to altered RMD profiles. Furthermore, treatment associated mutational signatures could be erroneously predictive of certain cancer types (e.g. platinum signature SBS35 for ovarian cancer) despite them not being intrinsic properties of those cancer types. Lastly, smoking history was assessed for lung cancer patients to determine whether lack of smoking, and by extension the absence of SBS4 (smoking mutational signature), would reduce performance for lung cancer. We found that MSI in Gastroesophageal, Breast, and CNS_Glioma as well as HRD in Pancreas did lead to significantly more incorrect predictions ($p<0.01$, one-sided Fisher's exact test), though the number of incorrectly predicted samples was low (6, 3, 3, and 10 respectively). Overall, the presence of the majority of confounding factors did not lead to more incorrect predictions ($p\geq0.01$, one-sided Fisher's exact test; **Supplementary table 1**).

				Variable TRUE				Variable FALSE				
Variable	Class	Total	Total with variable data	Incorrect	Correct	Total	%Incorrect	Incorrect	Correct	Total	%Incorrect	p-value
has_msi	Gastroesophageal	328	328	6	1	7	85.7	48	273	321	15.0	0.00010
has_msi	Breast	996	996	3	5	8	37.5	15	973	988	1.5	0.00026
has_msi	CNS_Glioma	131	131	3	3	6	50.0	2	123	125	1.6	0.00053
has_msi	Lung_NonSmallCell	588	588	2	3	5	40.0	40	543	583	6.9	0.04341
has_msi	Urothelial	200	200	2	3	5	40.0	21	174	195	10.8	0.10217
has_msi	Uterus	75	75	4	8	12	33.3	14	49	63	22.2	0.31187
has_msi	Colorectal	710	710	2	40	42	4.8	33	635	668	4.9	0.62811
has_msi	Biliary	114	114	2	3	5	40.0	54	55	109	49.5	0.80694
has_msi	Prostate	560	560	0	17	17	0.0	2	541	543	0.4	1.00000
has_hrd	Pancreas	347	346	10	23	33	30.3	23	290	313	7.3	0.00032
has_hrd	Breast	996	989	6	145	151	4.0	9	829	838	1.1	0.01746
has_hrd	Gastroesophageal	328	320	2	3	5	40.0	45	270	315	14.3	0.15787
has_hrd	Colorectal	710	667	1	5	6	16.7	31	630	661	4.7	0.25631
has_hrd	NET_Pancreas	119	110	2	3	5	40.0	20	85	105	19.0	0.26075
has_hrd	Sarcoma_Leiomyo	64	64	3	2	5	60.0	25	34	59	42.4	0.38026
has_hrd	Lung_NonSmallCell	588	583	1	6	7	14.3	39	537	576	6.8	0.39360
has_hrd	Urothelial	200	195	1	12	13	7.7	20	162	182	11.0	0.78387
has_hrd	Biliary	114	109	3	5	8	37.5	51	50	101	50.5	0.85876
has_hrd	Ovarian	284	282	2	105	107	1.9	25	150	175	14.3	0.99998
has_hrd	Prostate	560	542	0	55	55	0.0	2	485	487	0.4	1.00000
has_smoked	Pancreas	347	83	3	24	27	11.1	2	54	56	3.6	0.19177
has_smoked	HeadAndNeck_Othe	91	8	1	4	5	20.0	0	3	3	0.0	0.62500
has_smoked	Lung_NonSmallCell	588	226	7	143	150	4.7	6	70	76	7.9	0.89876
has_smoked	Gastroesophageal	328	83	1	57	58	1.7	3	22	25	12.0	0.99312
has_smoked	Breast	996	31	0	16	16	0.0	0	15	15	0.0	1.00000
has_smoked	Liver	340	25	0	14	14	0.0	2	9	11	18.2	1.00000
has_smoked	Lung_SmallCell	53	16	7	9	16	43.8	0	0	0	0.0	1.00000
has_smoked	Lymphoid	218	55	0	38	38	0.0	1	16	17	5.9	1.00000
treated_with_platinum	Gastroesophageal	328	188	12	69	81	14.8	11	96	107	10.3	0.23647
treated_with_platinum	Breast	996	605	2	61	63	3.2	7	535	542	1.3	0.23896
treated_with_platinum	Lung_SmallCell	53	22	15	6	21	71.4	0	1	1	0.0	0.31818
treated_with_platinum	HeadAndNeck_Othe	91	25	6	12	18	33.3	1	6	7	14.3	0.33653
treated_with_platinum	Pancreas	347	51	3	9	12	25.0	6	33	39	15.4	0.35400
treated_with_platinum	Biliary	114	23	7	11	18	38.9	1	4	5	20.0	0.41377
treated_with_platinum	Ovarian	284	114	15	93	108	13.9	0	6	6	0.0	0.42016
treated_with_platinum	Lung_NonSmallCell	588	214	13	144	157	8.3	4	53	57	7.0	0.50910
treated_with_platinum	Uterus	75	17	6	10	16	37.5	0	1	1	0.0	0.64706
treated_with_platinum	Mesothelium	41	28	5	21	26	19.2	0	2	2	0.0	0.66931
treated_with_platinum	Urothelial	200	101	8	82	90	8.9	1	10	11	9.1	0.74650
treated_with_platinum	Colorectal	710	329	7	261	268	2.6	8	53	61	13.1	0.99970
treated_with_platinum	Sarcoma_Osteo	44	6	2	3	5	40.0	1	0	1	100.0	1.00000
treated_with_platinum	Prostate	560	478	0	9	9	0.0	2	467	469	0.4	1.00000
treated_with_platinum	Cervix	59	23	3	20	23	13.0	0	0	0	0.0	1.00000
treated_with_5FU	HeadAndNeck_Othe	91	25	3	3	6	50.0	4	15	19	21.1	0.19368
treated_with_5FU	Pancreas	347	51	3	10	13	23.1	6	32	38	15.8	0.41389
treated_with_5FU	Breast	996	605	3	165	168	1.8	6	431	437	1.4	0.47703
treated_with_5FU	Gastroesophageal	328	188	1	4	5	20.0	22	161	183	12.0	0.48316
treated_with_5FU	Biliary	114	23	1	4	5	20.0	7	11	18	38.9	0.91076
treated_with_5FU	Colorectal	710	329	2	80	82	2.4	13	234	247	5.3	0.92366

Supplementary table 1: Comparison of the number of correctly and incorrectly predicted samples between patients with and without: microsatellite instability (MSI), homologous recombination deficiency (HRD), smoking history, treatment with platinum, and treatment with 5-fluorouracil (5FU). One-sided Fisher's exact tests were performed to determine if there were more incorrectly predicted patients with versus without the respective variable. Only cancer types with at least 5 patients in total being positive for the respective variable (i.e. Variable TRUE: Total >= 5) are shown.

As subclonal variants are enriched for treatment induced mutations, we also compared the performance of CUPLR when the model was trained on RMD and mutational signatures that were extracted from all mutations versus only clonal mutations (i.e. treatment induced mutations excluded) (**Supplementary figure 13**). No differences in overall recall were found (all mutations: 90% vs. clonal mutations 89%), with the exception of decreased recall when only clonal mutations were used in 5 cancer types: CNS_PiloAstro (90% vs 77%), Kidney_Chromophobe (90% vs. 72%), Myeloid (77% vs. 63%), NET_Gastrointestinal (80% vs. 47%) and Sarcoma_GIST (95% vs 82%). Given the similar performance between training on all mutations versus only clonal mutations, treatment induced mutations likely have minimal impact on the performance of CUPLR for the majority of cancer types.

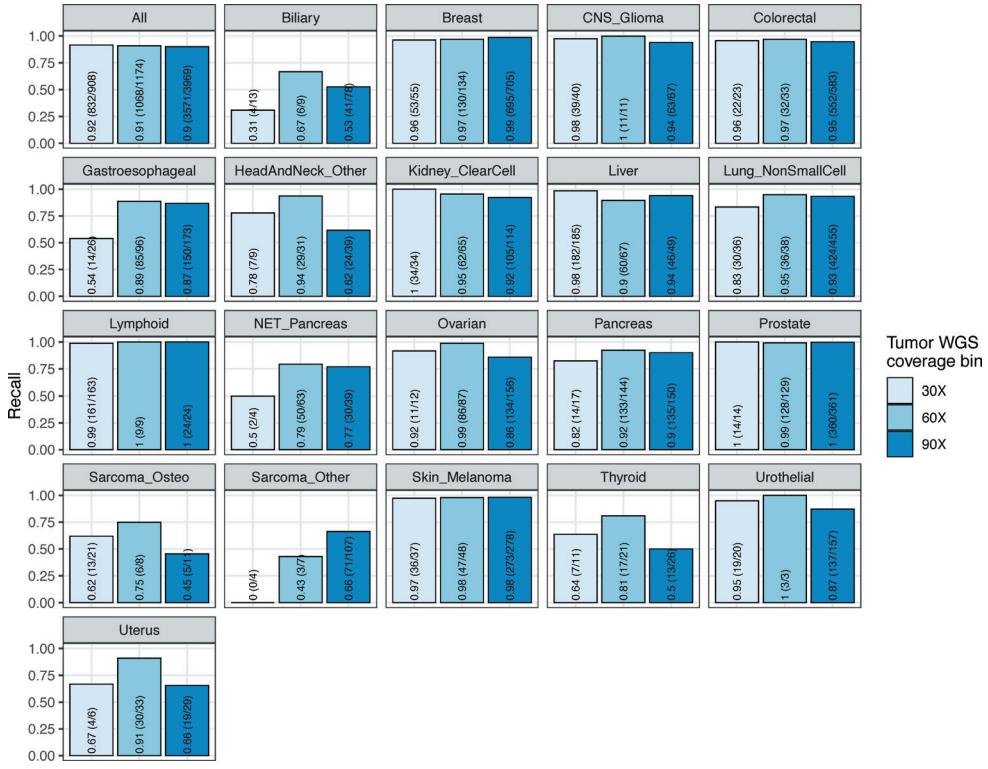
Supplementary figure 13: Comparison of cancer type prediction recall when using all mutations versus only clonal mutations to generate the regional mutational density and mutational signature features. Recall (i.e. fraction of samples correctly predicted) was determined using cross validation. Bars are labeled with the recall value, as well as in brackets the number of correctly predicted samples and the total number of samples.

Supplementary note 2: Impact of sequencing coverage on performance

Our dataset consisted of patients with tumor samples sequenced at roughly 30x, 60x and 90x coverage (with normal samples all sequenced at ~30x coverage). Thus, to assess the impact of sequencing depth, we compared the performance (based on cross-validation predictions) between these coverages for all samples, as well as for cancer types with at least 3 samples at each coverage (20 cancer types; **Supplementary figure 14**).

Overall, we found similar performance across difference sequencing depths (30x: 92%, 60x: 91%, 90x: 90%). Per cancer type, 12 cancer types had comparable recall across the 3 coverages, including Breast, CNS_Glioma, Colorectal, Kidney_ClearCell, Liver, Lung_NonSmallCell, Lymphoid, Ovarian, Pancreas, Prostate, Skin_Melanoma, Urothelial. For 4 cancer types (HeadAndNeck_Other, Sarcoma_Osteo, Thyroid and Uterus) there was no apparent correlation between coverage and recall. Lastly, recall was lower at 30x coverage for Biliary, Gastroesophageal, NET_Pancreas, with Sarcoma_Other also having lower recall at 30x and 60x.

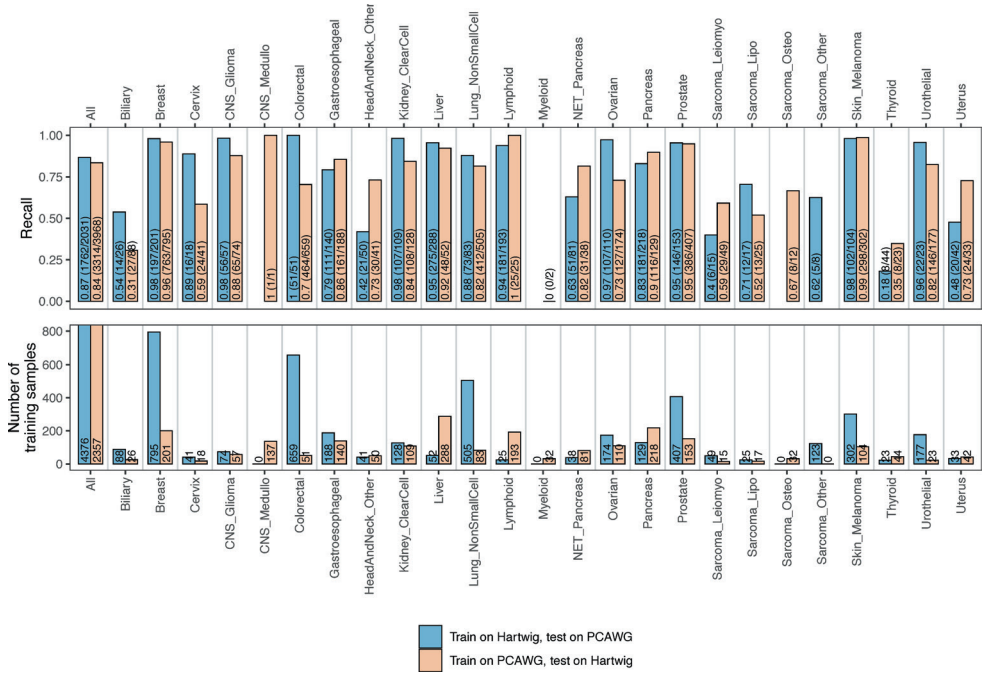
In summary, 30x coverage should be sufficient for reliable CUPLR predictions for most cancer types, but >=60x coverage would lead to the most reliable predictions.



Supplementary figure 14: Comparison of performance between samples sequenced at different coverages. Recall (i.e. fraction of samples correctly predicted) was determined by testing on the opposite dataset. Bars are labeled with the recall value, as well as in brackets the number of correctly predicted samples and the total number of samples.

Supplementary note 3: Impact of batch effects

To assess the potential for CUPLR to be overfit on the Hartwig and/or PCAWG datasets (i.e. the potential presence of batch effects), we trained a model solely on Hartwig samples (Hartwig-only) and another model solely on PCAWG samples (PCAWG-only), and determined performance by testing on the opposite dataset (**Supplementary figure 15**). Hartwig-only and PCAWG-only models were trained with equal numbers of samples to avoid better performance solely due to higher sample size.



Supplementary figure 15: Comparison of performance between a CUP classifier model trained solely on Hartwig samples (Hartwig-only) and another model solely on PCAWG samples (PCAWG-only). Only cancer types with at least 15 samples in both Hartwig and PCAWG cohorts were included for training. Recall (i.e. fraction of samples correctly predicted) was determined by testing on the opposite dataset. Bars are labeled with the recall value, as well as in brackets the number of correctly predicted samples and the total number of samples.

Overall, we found that both models achieved similar recall (Hartwig-only: 90%, PCAWG-only: 86%). For the 21 cancer types with sufficient sample sizes for training (>=15 samples in both Hartwig and PCAWG cohorts), 13 of these also had comparable recall between the two models (+/-10%). These data indicate that batch effects for most cancer types are minimal.

However, for 6 cancer types (Cervix, Colorectal, HeadAndNeck_Other, Ovarian, Sarcoma_Lipo, Thyroid), recall of the Hartwig-only model was much higher than of the PCAWG-only model. Conversely, the PCAWG-only model had higher recall than the Hartwig-only model for 2 cancer types (Biliary, Sarcoma_Leiomyo). These results indeed indicate potential technical and/or biological differences between the Hartwig and PCAWG cohorts for these cancer types. These results nevertheless highlight the importance of incorporating both Hartwig and PCAWG samples for training (e.g. to mitigate the impact from treatment associated mutations which may be more abundant in Hartwig samples).

Chapter 4

Precancerous liver diseases do not cause increased mutagenesis in liver stem cells

Luan Nguyen[1,$], Myrthe Jager[1,$], Ruby Lieshout[2], Petra E. de Ruiter[2], Mauro D. Locati[1], Nicolle Besselink[1], Bastiaan van der Roest[1], Roel Janssen[1], Sander Boymans[1], Jeroen de Jonge[2], Jan N.M. IJzermans[2], Michail Doukas[2], Monique M.A. Verstegen[2], Ruben van Boxtel[3], Luc J.W. van der Laan[2], Edwin Cuppen[1,4,*], Ewart Kuijk[1]

[1] University Medical Center Utrecht, Utrecht, The Netherlands
[2] Erasmus Medical Center, Rotterdam, The Netherlands
[3] Princess Máxima Center, Utrecht, The Netherlands
[4] Hartwig Medical Foundation, Amsterdam, The Netherlands

[$] These authors contributed equally to this work
[*] Corresponding author

Adapted from: Nature Commun Biol 4, 1301 (2021)
URL: https://doi.org/10.1038/s42003-021-02839-y
QR code to URL:

Abstract

Inflammatory liver disease increases the risk of developing primary liver cancer. The mechanism through which liver disease induces tumorigenesis remains unclear, but is thought to occur via increased mutagenesis. Here, we performed whole-genome sequencing on clonally expanded single liver stem cells cultured as intrahepatic cholangiocyte organoids (ICOs) from patients with alcoholic cirrhosis, non-alcoholic steatohepatitis (NASH), and primary sclerosing cholangitis (PSC). Surprisingly, we find that these precancerous liver disease conditions do not result in a detectable increased accumulation of mutations, nor altered mutation types in individual liver stem cells. This finding contrasts with the mutational load and typical mutational signatures reported for liver tumors, and argues against the hypothesis that liver disease drives tumorigenesis via a direct mechanism of induced mutagenesis. Disease conditions in the liver may thus act through indirect mechanisms to drive the transition from healthy to cancerous cells, such as changes to the microenvironment that favor the outgrowth of precancerous cells.

Introduction

Liver cancer is the fifth most common cancer worldwide, causing around 720,000 deaths each year [186]. Different subtypes of primary liver cancer can be recognized, of which hepatocellular carcinoma (HCC; originating from hepatocytes; originating from hepatocytes) and intrahepatic cholangiocarcinoma (CCA; originating from cholangiocytes; originating from cholangiocytes) form the largest groups, together constituting over 85% of all primary liver cancers [187]. Several factors have been linked to increased HCC risk, including chronic alcohol consumption [188], as well as metabolic associated fatty liver disease (MAFLD), and its more progressive form nonalcoholic steatohepatitis (NASH), which can be caused by obesity, diabetes, drugs/medication and metabolic conditions [189]. These factors have also been linked to risk for intrahepatic CCA [190]. Though less common, chronic inflammation and fibrosis of the biliary tracts, known as primary sclerosing cholangitis (PSC), also confers increased risk of developing both HCC and CCA [191]. Our knowledge on how these environmental conditions drive liver cancer is still incomplete [192].

Chronic alcohol consumption is thought to enhance the mutational load through the metabolite acetaldehyde, which has been reported to be directly mutagenic [193] and indirectly through the formation of reactive oxygen species [194–199]. Increased burden of somatic mutations has also been observed in non-alcoholic liver disease [200]. NASH and PSC are both characterized by chronic inflammation [201,202], which may cause the production of reactive oxygen and nitrogen species that subsequently induce DNA damage [203].

However, accurate measurements of *in vivo* induced mutations are required to confirm that accelerated mutagenesis underlies liver tumorigenesis. We previously established a sensitive method to accurately determine all somatic mutations that have been acquired throughout life in individual human adult stem cells of the liver and gastrointestinal tract [92]. We used these catalogs of somatically acquired mutations to perform mutational signature analysis, a powerful computational method for identifying mutational processes that have been active in the life history of cells [13]. Currently, 60 single base substitution (SBS) signatures, 11 double base substitution (DBS) signatures, 18 indel signatures, and 16 structural variation (SV) signatures have been identified [178,204]. These signatures are a result of endogenous mutational sources (such as apolipoprotein B mRNA editing enzyme, catalytic polypeptide-like (APOBEC) activity [205] or homologous recombination deficiency [38,103]), but also microbial impact [65], oxidative stress [206], or anti-cancer therapies [66,67]. Mutational signature analyses on healthy human stem cells revealed that mutational processes are tissue-specific and continuously active throughout life resulting in a linear accumulation of mutations with age [92], at least under 'normal' conditions. Because of the link between liver disease and liver cancer, we hypothesized that the precancerous state of liver diseases would be reflected by increased mutation rates and accumulation of mutational patterns that are characteristic to the type of DNA damage inflicted during liver tumorigenesis.

In this study, we aim to identify the mutational processes that contribute to the precancerous state in common human liver diseases. To achieve this goal, we have studied the accumulation of mutations in individual stem cells derived from livers of patients with alcoholic cirrhosis, NASH and PSC who received a liver transplantation. Surprisingly, we find that individual stem cells from liver patients did not show increased mutational burden overall, within liver cancer associated genes, nor to specific mutational signatures when compared to liver stem cells from healthy donors. Our findings suggest that environmental conditions drive liver tumorigenesis through means other than by increasing mutagenesis.

Results

Mutation rates do not increase in diseased livers

Both main liver cell types, hepatocytes and cholangiocytes, can act as liver stem cells depending on the type of tissue damage that was inflicted [207]. Cholangiocytes show a high degree of cellular plasticity during regeneration and disease and act as facultative liver stem cells during impaired hepatocyte regeneration [207,208]. Cholangiocytes can be grown as intrahepatic cholangiocyte organoids (ICO) [209] that show long-term self-renewal, differentiation and engraftment in mouse and rat models of liver failure [210,211]. In contrast, there are no suitable protocols for the long-term culture of human hepatocytes. In contrast, there are no suitable protocols for the long-term culture of human hepatocytes. We have previously exploited the proliferative capacity of individual cholangiocytes to determine mutation rates in individual liver stem cells of the healthy liver [92]. We reasoned that cholangiocytes are also suitable for the study of somatic mutation accumulation as a result of the diseased liver environment, because cholangiocytes are exposed to the same environmental conditions as the other liver cell types [212].

a

Disease status	Patients	Clones
Healthy	Healthy1	a,b
	Healthy2	a
	Healthy3	a
	Healthy4	a
	Healthy5	a,b,c
	Healthy6	a,b
	Healthy7	a
	Healthy8	a,b,c
ALC *Alcoholic cirrhosis*	ALC1	a,b,c
	ALC2	a
	ALC3	a
	ALC4	a,b,c
	ALC5	a,b,c,d,e,f
NASH *Non-alcoholic steatohepatitis*	NASH1	a,b,c,d
	NASH2	a
	NASH3	a
	NASH4	a,b,c
	NASH5	a,b,c,d
PSC *Primary sclerosing cholangitis*	PSC1	a
	PSC2	a,b,c,d,e
	PSC4	a,b
PCAWG_HCC *Hepatocellular carcinoma*	n=248	NA
PCAWG_CCA *Cholangiocarcinoma*	n=12	NA

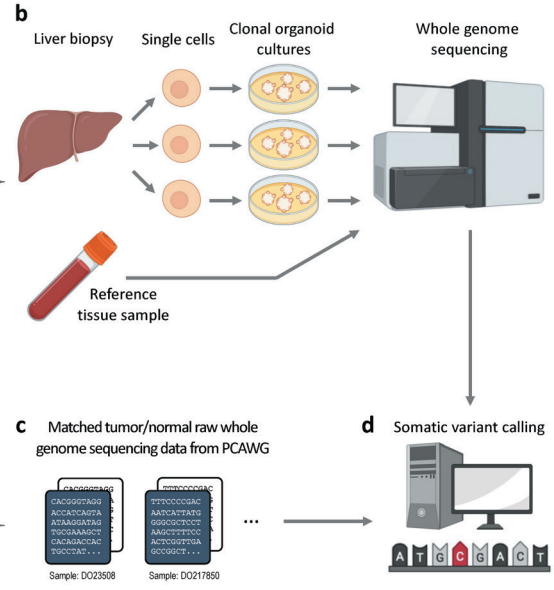


Figure 1: Samples and experimental setup. (**A**) Summary of the samples used in this study. (**B**) Liver biopsies were taken from patients with diseased livers from which clonal intrahepatic cholangiocyte stem cell-based organoid (ICO) cultures were generated. Organoid clones were subjected to whole-genome sequencing (WGS) together with a matched tissue reference sample per patient for subtraction of germline mutations and thus detection of somatic mutations. (**C**) To compare the mutation profiles in liver disease versus HCC and CCA, WGS data from primary tumor samples from the Pan-Cancer Analysis of Whole Genomes (PCAWG) consortium were also included. (**D**) Somatic variant calling of diseased and cancerous livers was performed with the same pipeline.

To determine whether a diseased liver environment leads to increased somatic mutation accumulation, we performed whole-genome sequencing (WGS) on clonal ICOs derived from liver stem cells from patients with diseased livers (**Figure 1A**, **B**; **Supplementary data 1**). Our study included: (i) 14 clones from 5 patients with cirrhosis as a result of chronic alcohol consumption; (ii) 13 clones from 5 NASH patients; and (iii) 8 clones from 3 PSC patients. Organoid establishment success rates were lower for diseased liver as compared to healthy livers and was most challenging for material obtained from PSC

patients. For each patient, a reference blood or multilineage bulk tissue sample was also sequenced to distinguish germline variants from somatic variants. The somatic mutation catalogs in diseased livers were compared to the mutation catalogs from 14 clones from liver adult stem cells derived from 7 healthy donors.

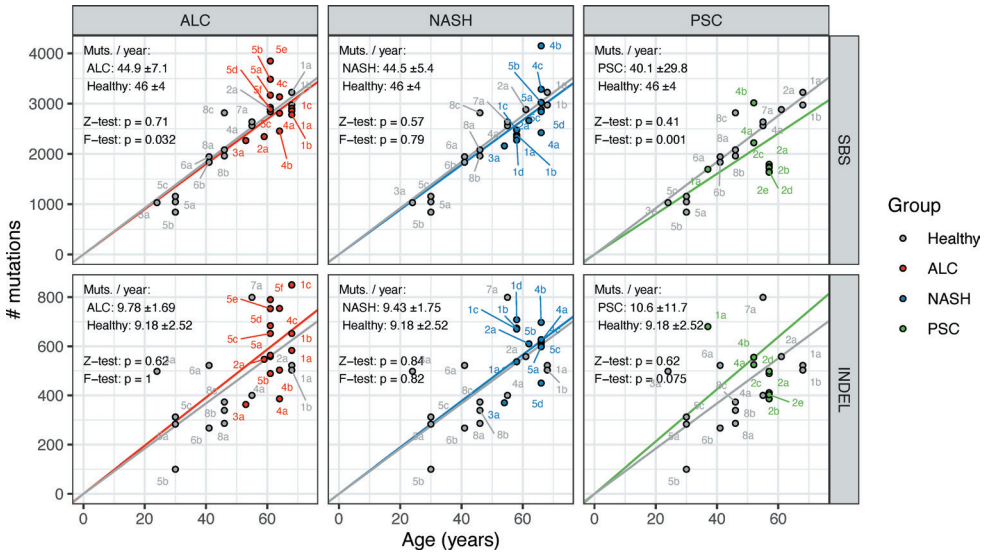


Figure 2: Accumulation of somatic single base substitutions (SBS) and small insertions/deletions (indel) in organoids derived from biopsies of healthy livers compared to those from patients with diseased livers. ALC: alcoholic cirrhosis, NASH: non-alcoholic steatohepatitis, PSC: primary sclerosing cholangitis. Each point is labelled by patient number and clone letter. Two-sided Z-tests were performed to determine whether there was a significant difference between the linear mixed effects regressions (i.e. the rate of mutation accumulation) of the disease versus healthy ICOs. One-sided F-tests were performed to determine whether there was a significant increase in variance in rate of mutation accumulation in disease samples versus healthy samples. ± values indicate the 95% confidence interval range of each regression and 'p' indicates the p-values of the Z-tests and F-tests.

In total, we identified 172,650 small mutations (single/double base substitutions or indels) in healthy (35,006), post-alcoholic (49,681), NASH (43,997), and PSC (19,521) livers (as well as 24,445 from a patient with HCC). Consistent with previous observations [92], somatic mutations accumulated linearly with age in healthy liver cells, at a rate of approximately 46 SBSs and 9 indels per year (**Figure 2**). The rate of SBS accumulation in alcoholic, NASH and PSC ICOs showed no significant differences to that of healthy ICOs (Z-test, $p \geq 0.41$), and similarly, the rate of indel accumulation in disease ICOs was also comparable to that of healthy ICOs (Z-test, $p \geq 0.62$). We observed a slight increase in variance in SBS accumulation in alcoholic cirrhosis versus healthy ICOs which may suggest that alcohol consumption leads to mutagenesis in some but not all patients, though the increase in variance was weak (F-test, $p=0.032$). Likewise, we found increased variance in SBS accumulation in PSC ICOs (F-test, $p=0.001$), though this variance is likely due to having many more ICOs originating from one patient PSC2. We also compared the accumulation of DBSs and SVs. While the number of DBSs and SVs was too low to be conclusive (≤50 DBSs and ≤25 SVs; **Supplementary figure 2**), the rate of mutation accumulation overall did not increase in diseased versus healthy ICOs. Taken together, these results suggest that chronic alcohol consumption or an inflamed liver environment does not lead to increased SBS, indel, DBS or SV accumulation in liver cells.

The mutation profile of diseased livers is similar to that of healthy livers

The presence of genome-wide patterns of mutations (also known as mutational signatures) reflects past activity of mutational processes in cells. Previously, the mutational signatures SBS12 and SBS16 have been associated with HCC [13,63], with SBS16 also being associated with alcohol consumption [213]. Additionally, SBS2 and SBS13 (APOBEC activity) were found to be active in numerous cancer types including CCA [13]. We expected that the mutational processes in diseased liver would be similar to those in cancerous liver. We thus examined whether liver disease results in increased presence of one of the above signatures or in other signatures related to liver cancer. Since the main liver cancer types are HCC and intrahepatic CCA, we selected the signatures that could be present in liver and biliary cancer based on the signature catalog from the PCAWG (Pan-Cancer Analysis of Whole Genomes) consortium [13] (see Methods for further details). We ultimately quantified the presence of 10 SBS and 7 indel signatures in our ICOs as well as in the PCAWG HCC and CCA samples (**Figure 3**). Too few DBSs and SVs were present in the diseased liver samples to perform signature analysis for these variant types (**Supplementary figure 2**).

We observed similar signature profiles in diseased and healthy ICOs, with the most predominant signatures being age-related (**Figure 3**; SBS1, SBS5, SBS40, ID1, ID5, ID8), which are present in normal cells [13]. We also found minor contributions of signatures SBS4 and ID3. These signatures are also present at a baseline level in HCC and CCA [13]. In the HCC and CCA samples from PCAWG, age-related signatures were predominant, with HCC samples also showing increased contribution of the known HCC associated signatures SBS12 and SBS16 compared to the healthy and diseased ICOs, while CCA samples showed increased contribution of APOBEC activity (SBS2, SBS13). Because these signatures were not observed in the disease ICOs, these findings argue against environmentally induced mutational processes as a force driving the transition of healthy to precancerous liver cells. It is likely that other events are required to initiate the HCC/CCA related mutational processes.

Given that certain recurrent chromosome arm gains and losses have been reported in HCC [214,215], we also examined whether liver disease leads to similar copy number alterations (CNA). We find that aside from 2 samples with polyploidization (a known phenomenon in normal liver cells [216]), the genomes of the disease ICOs were relatively stable, while the genomes of HCC and CCA were clearly more unstable with known recurrent arm gains (e.g. 1q, 8q, 17q) and losses (e.g. 8p and 17p) being observed [214,215] (**Supplementary figure 3**). These data could suggest that CNA accumulation due to liver disease does not contribute to the healthy to precancerous liver transition. However, as CNAs are rare in non-cancerous cells [58], more data would be required to validate this hypothesis.

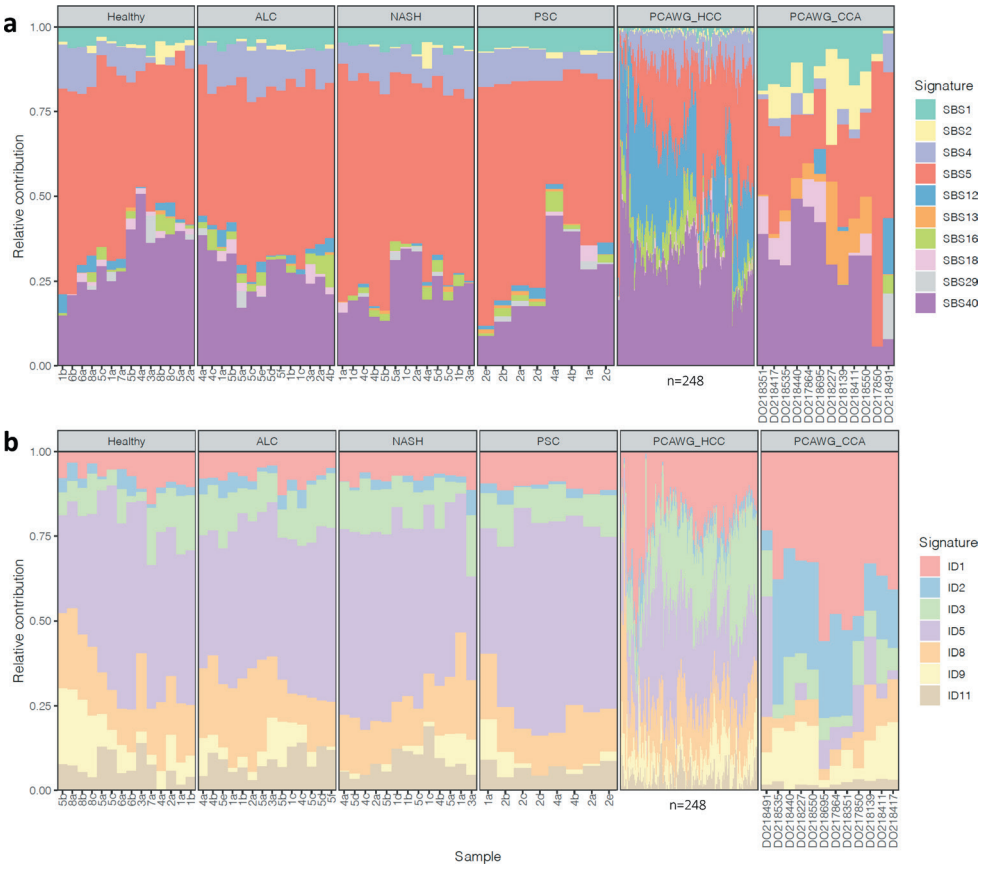


Figure 3: Relative contributions of mutational signatures in organoids derived from biopsies of healthy, diseased and cancerous livers. (**A**) single base substitution (SBS) signatures. (**B**) Indel (ID) signatures. ALC: alcoholic cirrhosis, NASH: non-alcoholic steatohepatitis, PSC: primary sclerosing cholangitis, HCC: hepatocellular carcinoma, CCA: cholangiocarcinoma. Profiles for HCC (PCAWG_HCC; n=248) and CCA (PCAWG_CCA; n=12) samples from the Pan-Cancer Analysis of Whole Genomes (PCAWG) consortium are also shown. Hierarchical clustering of samples was performed separately for SBS and ID signatures. Sample names for PCAWG_HCC samples are hidden due to the large number of samples.

Absence of driver gene mutations in ICOs

Certain mutations may confer liver stem cells a growth advantage, and in diseased livers, these cells may be able to proliferate more to regenerate lost tissue. We thus examined whether the liver disease conditions resulted in positive selection of cells with non-synonymous mutations in specific genes using the *dndscv* algorithm (details in Methods). However, amongst all of the liver disease groups, no genes were found to be enriched in non-synonymous mutations (q<0.01, **Supplementary data 2**). In line with this result, we did not observe any coding, promoter or 5'/3' untranslated region (UTR) mutations in driver genes of HCC and CCA (obtained from Intogen; see Methods), except for one missense variant in an alcoholic cirrhosis ICO sample (*TERT* c.1588C>G in sample ALC3_CLONE32) (**Figure 4**). This could potentially be explained by the cells from which our disease ICOs were derived not being actual cancer precursors but only harboring passenger mutations. We acknowledge that mutations could occur in other non-coding elements but have not examined these as their impact is currently difficult to assess [217,218]. Furthermore, we also acknowledge that our small sample sizes

limit our ability to find enriched driver gene mutations in the diseased liver ICOs. In contrast, we found enrichment of non-synonymous mutations in *TP53* in the PCAWG CCA samples, and in *TP53* and *CTNNB1* as together with 13 other genes in the PCAWG HCC samples (q<0.01, **Supplementary data 2**). While *dndscv* does not consider non-coding variants, we also observed *TERT* promoter mutations in the PCAWG HCC samples ('upstream_gene_variant', **Figure 4**). These genes have been reported as known cancer driver genes in the respective cancer types [5]. No mutations in these driver genes were found in 34% (84/284) of PCAWG HCC samples indicating that mutations in these genes are not necessarily a requirement for HCC development.

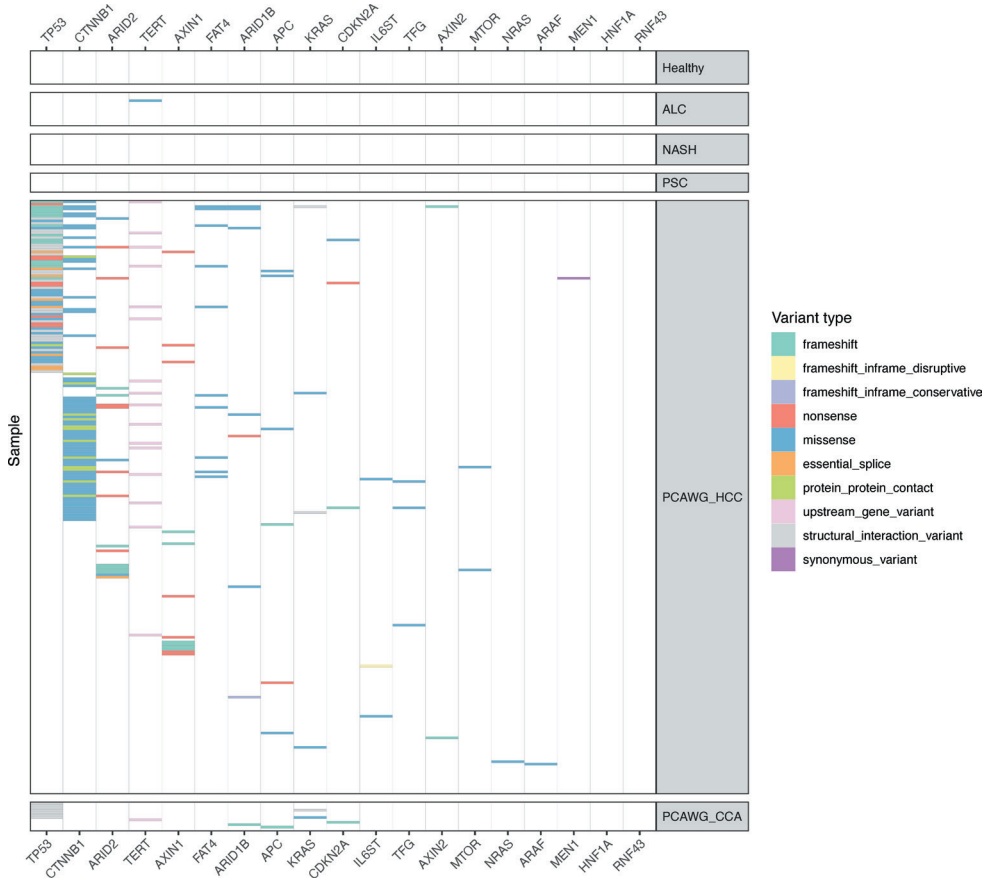


Figure 4: Non-synonymous mutations in organoids derived from biopsies of healthy, diseased and cancerous livers. ALC: alcoholic cirrhosis, NASH: non-alcoholic steatohepatitis, PSC: primary sclerosing cholangitis, HCC: hepatocellular carcinoma, CCA: cholangiocarcinoma. Profiles for HCC (PCAWG_HCC; n=248) and CCA (PCAWG_CCA; n=12) samples from the Pan-Cancer Analysis of Whole Genomes (PCAWG) consortium are also shown.

Discussion

Despite the association between liver disease and primary liver cancer (which includes HCC and CCA), the underlying mechanisms of tumorigenesis remain debated. The prevalent view is that tumorigenesis results from an increased mutational burden [193–203]. In line with this view, Brunner *et al*. [200] showed that cirrhotic liver cells from NASH patients exhibited an increase in mutational load, even though this increase was small and variance between patients was high. In contrast with these observations, we did not observe an increase in mutation rate in individual stem cells of precancerous livers. Additionally, while previous studies have associated specific mutational signatures and gene mutations in HCC to alcohol consumption [219,220], we did not observe an altered mutational landscape in our alcoholic cirrhosis ICOs.

There are several possible explanations for the unchanged mutational landscape in our liver disease stem cells. Firstly, it could be that the cholangiocytes that give rise to the ICO cultures are not the precursors to (pre-)cancerous liver cells. Hepatocellular carcinoma is derived from hepatocytes and not from cholangiocytes and the cholangiocytes that give rise to ICOs may not be representative for the cholangiocytes that have the potential to develop into cholangiocarcinoma. Secondly, there could be selection for the most stable ASCs in culture, which may not necessarily be the precancerous cells. Thirdly, it may be possible that increased mutagenesis affected only a small proportion of the liver cells and that we have sequenced too few cells to identify the hypermutated cells. Lastly, it may be that our patient cohorts were too small to detect changes in the mutational landscape, which may indeed be the case for detecting enrichment of driver gene mutations. However, we could determine via a power analysis (**Supplementary note 1**) that the sample sizes in our study were sufficient to detect changes in mutational load and mutational signature contribution similar to other studies which also used tissue derived organoids to investigate mutation accumulation [206,221,222]. It may be however possible to pick up more subtle mutational impacts by increasing sample sizes.

On the other hand, the stem cells that we studied have been exposed to the same environmental stressors as the cells that grew out as tumors (since both have been in contact with the blood stream), so when there would have been a direct mutational impact, we would expect this to be detectable in the cells that we studied. In line with our findings, Brunner *et al*. [200] also found that alcoholic cirrhosis and NASH livers did not exhibit different mutational signatures in comparison to healthy livers. While the mutational impact of NASH and PSC have not been investigated besides in the Brunner *et al*. [200] study, alcohol exposure has been shown to lead to DNA damage *in vitro* [195,198,223] and in blood cells in mice [224], but these studies do not accurately reflect the *in vivo* liver environmental conditions. It is possible that the rate of cellular endocytosis and/or diffusion is slower in the liver, resulting in less exposure to alcohol than would occur *in vitro*. Additionally, the aforementioned studies were performed on non-quiescent cells which may rely more on replicative repair, whereas liver cells are generally quiescent and likely rely on non-replicative repair which is faster at repairing alcohol induced DNA damage than replicative repair [193]. Alternatively, liver cells that acquire alcohol induced DNA damage may undergo apoptosis and be replaced by new cells as a result of liver regeneration [225], and these cells in turn lack the mutation footprint caused by the alcohol. Nevertheless, the absence of increased mutational burden in our disease ICOs may suggest that increased mutagenesis is not the primary contributing factor towards tumorigenesis.

Opposed to the view that tumorigenesis arises from mutagenesis, an alternative hypothesis proposes that chronic liver inflammation and cirrhosis (which commonly precedes primary liver cancer [226]) leads to cell death in the liver, requiring normally quiescent liver adult stem cells to proliferate at a much higher rate to regenerate the damaged liver. As a consequence, cells would accumulate more mutations, especially those caused by background mutational mechanisms related to cell proliferation (e.g. aging-associated mutational signatures). Inflammatory disease conditions would thus provide a

'fertile ground' for cells with random and potentially pre-existing (oncogenic) mutations that confer a selective growth advantage to clonally expand [49,227,228]. Such a phenomenon has been described in mouse models, whereby pancreatic cells within mice with both a pathogenic *Kras* mutation and pancreatitis transitioned into an epigenetic state similar to pancreatic ductal adenocarcinoma, while pancreatic cells in mice with only one or the other retained their original epigenetic state [229]. Additionally, Hepatitis C Virus (HCV)-induced cirrhotic livers showed an increase in the number and size of clonal patches with mutations in genes that are frequently mutated in HCC [225].

Taken together, our findings suggest that mechanisms other than direct mutagenesis drive the transition from healthy to precancerous liver, and highlights the need to explore other potential hypotheses of liver tumorigenesis, including but not limited to the 'fertile ground' hypothesis.

Methods

Human tissue material

All human tissue biopsies were obtained in the Erasmus MC - University Medical Center Rotterdam. Liver biopsies from healthy liver donors and patients with alcoholic cirrhosis, nonalcoholic steatohepatitis (NASH) or primary sclerosing cholangitis (PSC) were obtained during liver transplantation procedures. All patients were negative for viral infection and metabolic diseases. The biopsies were collected in cold organ preservation fluid (Belzer UW Cold Storage Solution, Bridge to Life, London, UK) and transported and stored at 4°C until use. The liver and tumor biopsies from the hepatocellular carcinoma patient were collected from a resected specimen and stored at -80°C until use. The acquisition of these liver and tumor biopsies for research purposes was approved by the Medical Ethical Committee of the Erasmus Medical Center (MEC-2014-060 and MEC-2013-143). Informed consent was provided by all patients involved.

Generating clonal intrahepatic cholangiocyte organoid cultures from human liver biopsies

Healthy and diseased liver tissue biopsies were washed in cold DMEM (ThermoFisher) supplemented with 1% fetal calf serum (FCS) and 1% penicillin-streptomycin (wash solution). Subsequently, the tissue was transferred to a petri dish and thoroughly minced with scalpel blades. The minced tissue was transferred to 4 ml digestion solution consisting of EBSS with Ca^{2+} / Mg^{2+} (ThermoFisher) with 1 mg/ml Collagenase type IA (Sigma, C9891) and 0.1 mg/ml DNAse I (Sigma DN25). The tissue was incubated for 30 minutes at 37°C with regular shaking. Next, the suspension was passed through a pipet to further break up the tissue and passed through a 70 µm Nylon cell strainer. The cells were washed once with wash solution, followed by two washes in Advanced DMEM F12 supplemented with 1% penicillin-streptomycin, 10mM HEPES, and 1X Glutamax (all from ThermoFisher). After the final wash, the cell pellet was resuspended in Matrigel (Corning) and plated in 40µl droplets per well in prewarmed non-adhesive 24-well plates. The plates were placed at 37°C in a humidified atmosphere and 5% CO_2. After Matrigel had solidified, 500µl liver organoid establishment medium was added to the wells. Establishment medium consisted of Advanced DMEM F12 supplemented with 1% penicillin-streptomycin, 10mM HEPES, 1X Glutamax, 10% R-Spondin conditioned medium (produced in house), B27 supplement without Vitamin A (ThermoFisher), N2 supplement (ThermoFisher), 10mM Nicotinamide (Sigma Aldrich), 1.25mM N-acetylcysteine (Sigma Aldrich), Primocin, 5µM A83-01 (Tocris Bioscience), 10µM Forskolin (Tocris Bioscience), 100ng/ml FGF-10 (Peprotech), recombinant human Noggin (Peprotech), 10 µM Rho kinase inhibitor (Abmole), hES cell cloning & recovery supplement (Stemgent), 25ng/ml HGF (Peprotech), 10nM Gastrin (Tocris), 50ng/ml human EGF (Peprotech), and 0.3nM Wnt-surrogate Fc protein (U-protein Express BV). After 2-3 days after isolation, the first

intrahepatic cholangiocyte organoid started to appear and establishment medium was switched to maintenance medium consisting of Advanced DMEM F12 supplemented with 1% penicillin-streptomycin, 10mM HEPES, 1X Glutamax, 10% R-Spondin conditioned medium, B27 supplement without Vitamin A, N2 supplement 10mM Nicotinamide, 1.25mM N-acetylcysteine, Primocin, 5µM A83-01, 10µM Forskolin, 100ng/ml FGF-10, 25ng/ml HGF (Peprotech), 10nM Gastrin (Tocris), and 50ng/ml human EGF (Peprotech). The cultures were maintained for 10-14 days after isolation, to enrich for adult stem cells. Subsequently, clonal organoid cultures were generated from these organoid cultures by FACS or by manual selection and expansion of individual organoids [230]. The organoid cultures were further expanded until there was enough material for DNA isolation. DNA was isolated from all organoid cultures, blood samples, and tissue biopsies using the Qiasymphony (Qiagen). Whole-genome sequencing libraries were generated from 200 ng of genomic DNA according to standard protocols (Illumina). The organoid cultures and control samples were sequenced paired-end (2 x 100 bp) to a depth of at least 30X coverage on the Illumina HiSeq Xten. The hepatocellular carcinoma biopsies were sequenced paired-end (2 x 100 bp) to a depth of at least 60X coverage on the Illumina HiSeq Xten. Whole-genome sequencing was performed at the Hartwig Medical Foundation in Amsterdam, the Netherlands.

Variant calling

Germline and somatic variant calling for all samples was performed using the HMF pipeline (https://github.com/hartwigmedical/pipeline; v4.8) [6]. Briefly, reads were mapped to GRCh37 using BWA-MEM v0.7.5a with duplicates being marked for filtering. Indels were realigned using GATK v3.4.46 IndelRealigner. GATK Haplotype Caller v3.4.46 was used for calling germline variants in the reference sample. For somatic SNV and indel variant calling, GATK BQSR3 was first used to recalibrate base qualities, followed by Strelka v1.0.14 for the variant calling itself. Somatic structural variant calling was performed using GRIDSS v1.8.0. Copy-number calling was performed using PURity & PLoidy Estimator (PURPLE), that combines B-allele frequency (BAF), read depth, and structural variants to estimate the purity and copy number profile of a tumor sample. For SNVs and indels, downstream analyses were performed only on variants marked as 'PASS'.

Mutation context analysis

The counts of single base substitution (SBS), double base substitution (DBS), indel and structural (SV) variant contexts were determined from somatic VCF files using the R package *mutSigExtractor* (https://github.com/UMCUGenetics/mutSigExtractor; v1.23). The mutation contexts of all mutation types are described in COSMIC (https://cancer.sanger.ac.uk/cosmic/signatures), except for the SV contexts. The 16 SV contexts were composed of the SV type (deletion, duplication, inversion, translocation) and the SV length (1–10kb, 10–100kb, 100kb–1Mb, 1–10Mb, >10Mb). Note that SV length is not applicable for translocations. The mutation context spectra for each sample group are shown in **Supplementary figure 4**, **Supplementary figure 5**, **Supplementary figure 6**, and **Supplementary figure 7**.

To perform mutational signature analysis, we selected the SBS and DBS signatures that were present in at least 10% of liver cancer (Liver-HCC) or biliary cancer (Biliary-AdenoCA) PCAWG samples (https://dcc.icgc.org/releases/PCAWG/mutational_signatures/Signatures_in_Samples/SP_Signatures_in_Samples) [13]. We then fitted the SBS and DBS mutation contexts to these selected signatures using the fitToSignatures() function from mutSigExtractor (which employs the non-negative least-squares method) to obtain absolute signature contributions. Relative signature contribution per sample was calculated by dividing the absolute contributions by the total signature contribution.

Mutation context and signature absolute contributions per sample can be found in **Supplementary data 3**.

Selection of liver and biliary cancer driver genes

A catalog of driver genes by cancer type was downloaded from Intogen (https://www.intogen.org/download; release 2020.02.01). From the Compendium_Cancer_Genes.tsv file, we selected genes where CANCER_TYPE was 'HC' or 'CH' (hepatocellular carcinoma and cholangiocarcinoma, respectively), and CGC_CANCER_GENE was 'TRUE'. Additionally, *TERT* has been reported by the PCAWG consortium [5] as a known HCC driver gene and was thus also included.

Identifying non-synonymous mutations

The mutation type of each somatic SNV/indel was determined by SnpEff (http://snpeff.sourceforge.net/; v4.3t). The following variant types were considered non-synonymous mutations: out-of-frame frameshifts, disruptive inframe frameshifts, nonsense, missense, splice variants. We also considered a mutation as non-synonymous if it was annotated as VUS, likely pathogenic, or pathogenic by ClinVar (https://www.ncbi.nlm.nih.gov/clinvar/; GRCh37, database date 2020-02-24), or if it was a hotspot mutation. The underlying code for annotating non-synonymous mutations can be found at https://github.com/UMCUGenetics/geneDriverAnnotator (v1.0).

The dndscv R package [49] was used to identify genes that were enriched for non-synonymous mutations. Briefly, this package computes the (local) background mutation rates and sequence composition of genes to calculate the background mutation rate for each gene. A likelihood ratio test is subsequently performed to identify genes that are significantly hit by nonsynonymous mutations. dndscv was run separately for each disease status group (i.e. separately for healthy ICOs, separately for NASH ICOs, etc) using all the somatic mutations from the respective group.

Statistics and reproducibility

All statistical analyses were performed in R (v4.0.3). To correlate the number of mutations with the age of each patient from which each biopsy was derived (as shown in **Figure 2** and **Supplementary figure 1**), we first assessed normality of the mutational load per mutation type per disease status group was using the Shapiro test (shapiro.test() function) (**Supplementary table 1**). This confirmed that the mutational load was normally distributed ($p>0.05$), with near normality for SBS load in PSC samples ($p=0.03$) and indel load in NASH samples ($p=0.05$). Then, the lme() function from the nlme (v3.1) package was used to fit a linear mixed effects regression, with 95% confidence intervals being calculated using the intervals() function from the nlme package. Here, 'patient' was modelled as a random effect to account for having different numbers of organoids per patient. Additionally, the intercept was fixed to zero as it was assumed that a patient has no somatic mutations at birth. A two-sided Z-test was used to calculate the difference between two regressions. The Z-statistic was first calculated using the slope (m) and standard errors (SE) of the two regressions (equation 1), which was then used to calculate a p-value using the pnorm() function (equation 2). A one-sided F-test was performed to calculate whether the variance of the regression of diseased ICOs was greater than that of the healthy ICOs. The F-statistic was first calculated by dividing the variance of the two regressions (extracted from the output of the lme() function) (equation 3), which was then used to calculate a p-value using the pf() function (equation 4).

$$Z = \frac{m_1 - m_2}{\sqrt{SE_1^2 - SE_2^2}} \quad (1)$$

$$p = 2 \times pnorm(-|Z|) \quad (2)$$

$$F = var_{disease} / var_{healthy} \quad (3)$$

$$p = 1 - pf(F) \quad (4)$$

To determine whether there was a significant increase in mutation context load in disease versus healthy ICOs (as shown in **Supplementary figure 4**, **Supplementary figure 5**, **Supplementary figure 6**, and **Supplementary figure 7**), Wilcoxon rank sum tests (using the wilcox.test() function) were performed per mutation context. Bonferroni multiple testing correction was then applied to the resulting p-values (using the p.adjust() function).

Data availability

The BAM files from the whole-genome sequencing data generated in the current study are available at EGA (https://www.ebi.ac.uk/ega/home) under accession numbers EGAS00001002983 and EGAS00001005384. BAM files from hepatocellular carcinoma and cholangiocarcinoma patients from the Pan-Cancer Analysis Whole Genomes (PCAWG) consortium were obtained under request number DACO-5333. For access to the PCAWG BAM files, researchers will need to request access via the ICGC Data Access Compliance Office (DACO; https://daco.icgc.org/). The VCF and tabular files produced from somatic variant calling are available at https://zenodo.org/record/5562381 [231].

Code availability

The code for the Hartwig Medical Foundation (HMF) germline and somatic variant calling pipeline is available at https://github.com/hartwigmedical/pipeline. The code used for data processing and generating the figures is available at https://github.com/UMCUGenetics/Diseased_livers [232].

Acknowledgments

This study was financially supported by the research program InnoSysTox (project number 114027003), by the Netherlands Organisation for Health Research and Development (ZonMw), by the Dutch Cancer Society (project number 10496) and is part of the Oncode Institute, which is partly financed by the Dutch Cancer Society and was funded by the gravitation program CancerGenomiCs.nl from the Netherlands Organisation for Scientific Research (NWO). The authors would like to thank Hartwig Medical Foundation and the Utrecht Sequencing Facility for performing the whole genome sequencing. The Utrecht Sequencing Facility is subsidized by the University Medical Center Utrecht, Hubrecht Institute, and Utrecht University.

Author contributions

R.L., J.J., J.I., M.D. and M.V. collected liver biopsies. M.J., E.K., and N.B. performed organoid culturing. L.N., M.J., M.L., B.R., R.J., and S.B. performed bioinformatic analyses. M.J., E.K., M.V., R.B., L.L., and E.C. were involved in the conceptual design of this study. L.N., E.K., M.J. and E.C. wrote the manuscript. All authors provided textual comments and have approved the manuscript. E.K., R.B., L.L., and E.C. supervised this study.

Competing interests

The authors declare no competing interests.

Supplementary data

Supplementary data 1: Sample metadata

Supplementary data 2: Genes enriched in non-synonymous mutations as determined by dndscv

Supplementary data 3: Mutation context and absolute mutational signature contributions per sample

Supplementary data are available online at: https://www.nature.com/articles/s42003-021-02839-y#Sec16, or by scanning the QR code below:

Supplementary information

Supplementary note 1: Power analysis

We performed a power analysis to assess the minimum detectable changes to mutation accumulation and mutational signature contributions given our number of samples. We used the *pwr* R package to calculate hypothetical detectable effect sizes using a significance level of 0.05 and a power of 0.8. We used the *pwr.t2n.test()* and *pwr.chisq.test()* functions to calculate Cohen's D and Cohen's W respectively.

Given that we have 8 healthy patients, 5 with alcoholic cirrhosis, 5 with NASH and 3 with PSC, we expect to be able to detect an effect size (Cohen's D) from 1.51 to 1.83 for mutation accumulation (i.e. as shown in **Figure 2**). For changes in SBS contributions (i.e. as shown in **Figure 3**), the degrees of freedom (DF) is 9 (10−1 signatures x 2−1 conditions), and thus we expect to detect an effect size (Cohen's W) of 1.10 to 1.19 when the total number of samples is 13 or 11 respectively. For changes in indel contributions, the degrees of freedom (DF) is 6 (7−1 signatures x 2−1 conditions), and thus we expect to detect an effect size (Cohen's W) of 1.02 to 1.11 when the total number of samples is 13 or 11 respectively.

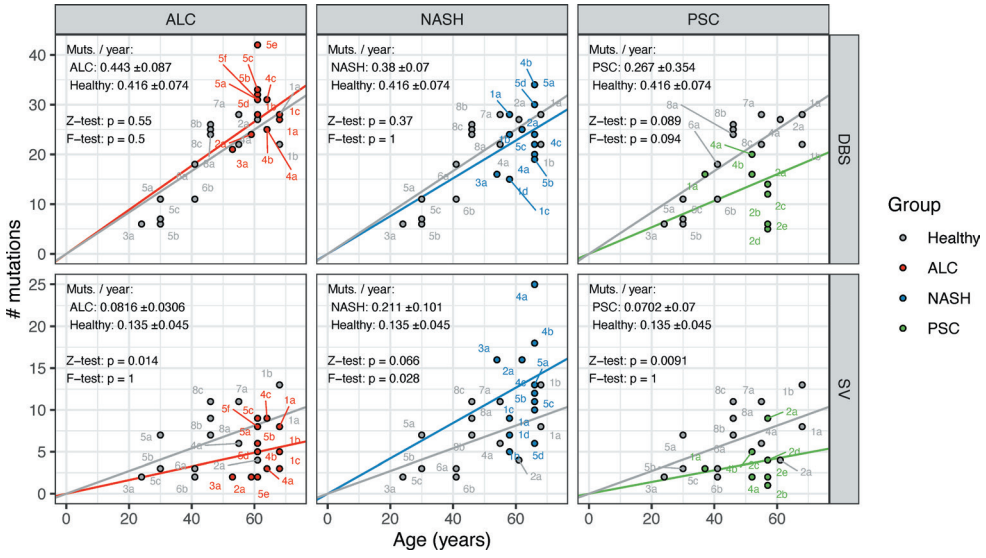
When then compared this with similar experimental setups in published studies to determine what these hypothetical effect sizes signify. Kuijk *et al. [206]* found that organoids cultured in 20% oxygen (n=5) accumulated ~1.5 times more SBSs than those cultured in 3% oxygen (n=3) (p=0.003, two-sided t-test). Here, the detectable effect size (Cohen's D) would be 2.06. Organoids cultured in 20% oxygen also showed higher relative contribution of SBS5 (~30%) than those cultured in 3% oxygen (~10%). In this comparison, the DF) was 6 = 7−1 signatures x 2−1 conditions) and the total number of samples was 8, thus yielding a detectable effect size (Cohen's W) of 1.31. Drost *et al. [221]* found that MLH1 knockout organoids (n=3) accumulated more SBSs (~25) and indels (~100) compared to wild type organoids (n=4) (~5 for both SBSs and indels). Here, the detectable effect size (Cohen's D) would be 2.22. Jager *et al.* [222] found that organoids derived from ERCC1 knockout mouse livers (n=3) showed increased relative contribution of SBS8 (~60%) compared to those derived from wild type mice (n=3) (~25%). In this comparison, the DF was 9 (= 10−1 signatures x 2−1 conditions) and the total number of samples was 3, thus yielding a detectable effect size (Cohen's W) of 1.91.

Thus, the sample sizes in our study would allow us to detect effect sizes similar to those found in the aforementioned published studies.

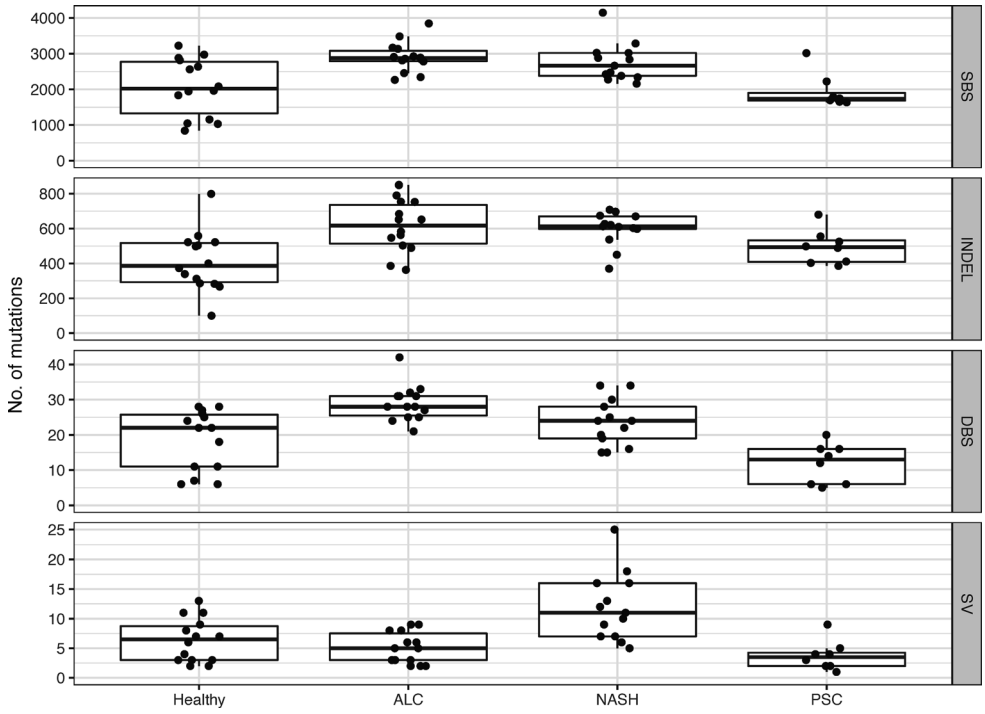
Group	Mut. type	p-value
Healthy	SBS	0.23213544
Alcohol	SBS	0.54995109
NASH	SBS	0.92665512
PSC	SBS	0.02814479
Healthy	DBS	0.09704532
Alcohol	DBS	0.92785727
NASH	DBS	0.17651379
PSC	DBS	0.38836339
Healthy	indel	0.70099886
Alcohol	indel	0.50867339
NASH	indel	0.04655043
PSC	indel	0.83045748
Healthy	SV	0.28064156
Alcohol	SV	0.10257364
NASH	SV	0.37047706
PSC	SV	1

Supplementary table 1: Shapiro test on the mutational load per mutation type per disease status group. Since for some patients multiple intrahepatic cholangiocyte organoid (ICO) clones were derived, the mean mutational load over the ICO clones for each patient was first calculated before performing the Shapiro test.

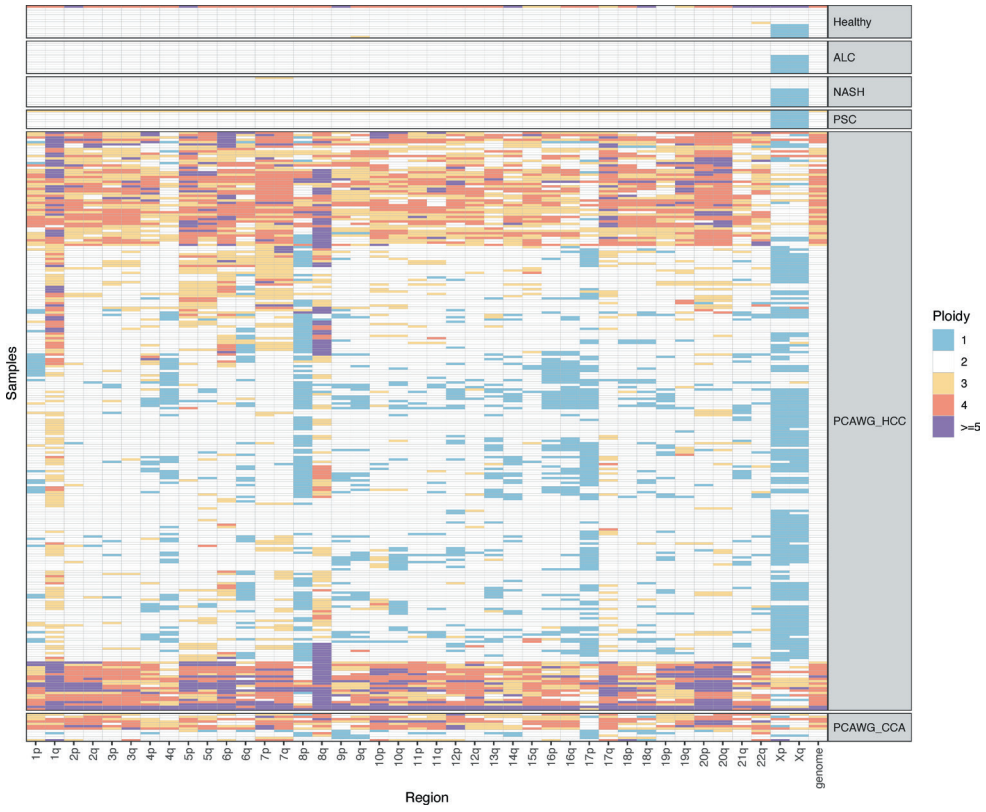
Supplementary figures



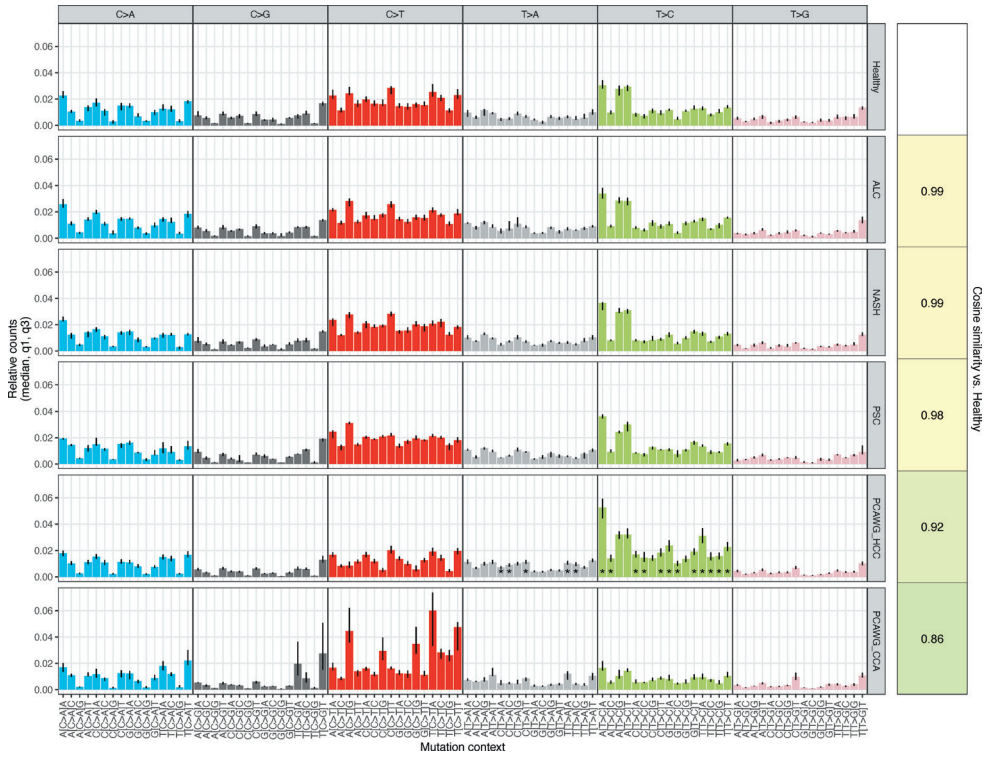
Supplementary figure 1: Accumulation of double base substitutions (DBS) and structural variants (SV) in organoids derived from biopsies of healthy livers compared to those from patients with diseased livers. ALC: alcoholic cirrhosis, NASH: non-alcoholic steatohepatitis, PSC: primary sclerosing cholangitis. Each point is labelled by patient number and clone letter. Two-sided Z-tests were performed to determine whether there was a significant difference between the linear mixed effects regressions (i.e. the rate of mutation accumulation) of the disease versus healthy ICOs. One-sided F-tests were performed to determine whether there was a significant increase in variance in rate of mutation accumulation in disease samples versus healthy samples. ± values indicate the 95% confidence interval range of each regression and 'p' indicates the p-values of the Z-tests and F-tests.



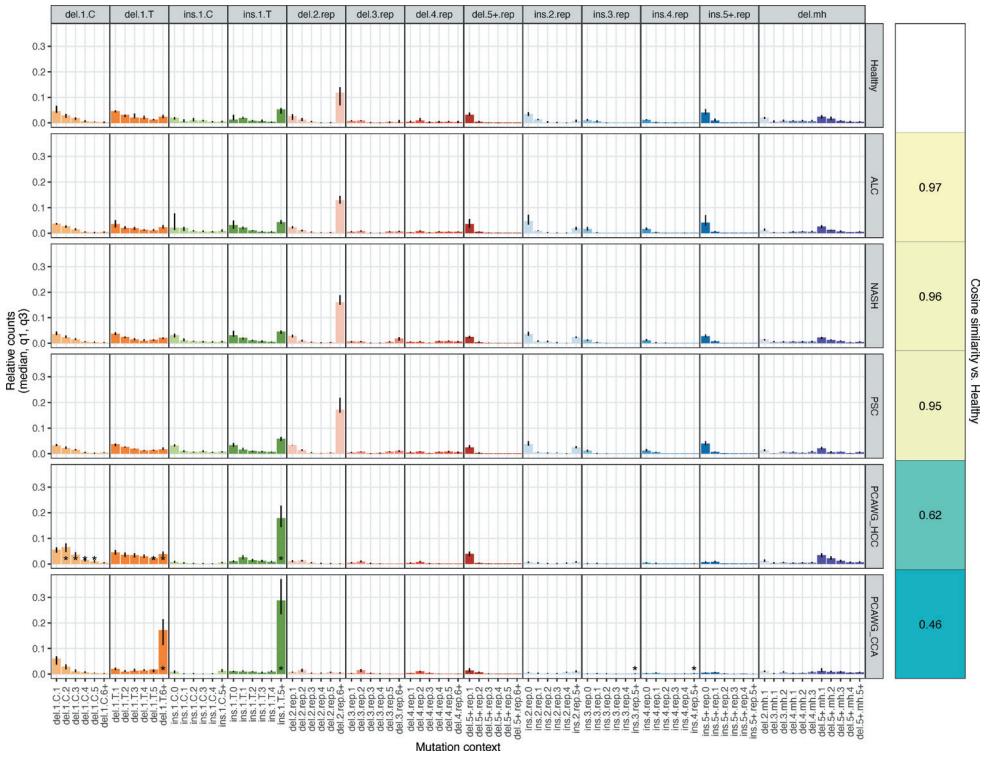
Supplementary figure 2: Mutational load per mutation type in organoids derived from biopsies of healthy livers, and livers from patients with alcoholic cirrhosis (ALC), non-alcoholic steatohepatitis (NASH), or primary sclerosing cholangitis (PSC). Each dot represents one organoid line. Boxplot boxes show the interquartile range (IQR) and whiskers show the largest/smallest values within 1.5 times the IQR. Each dot is one organoid clone.



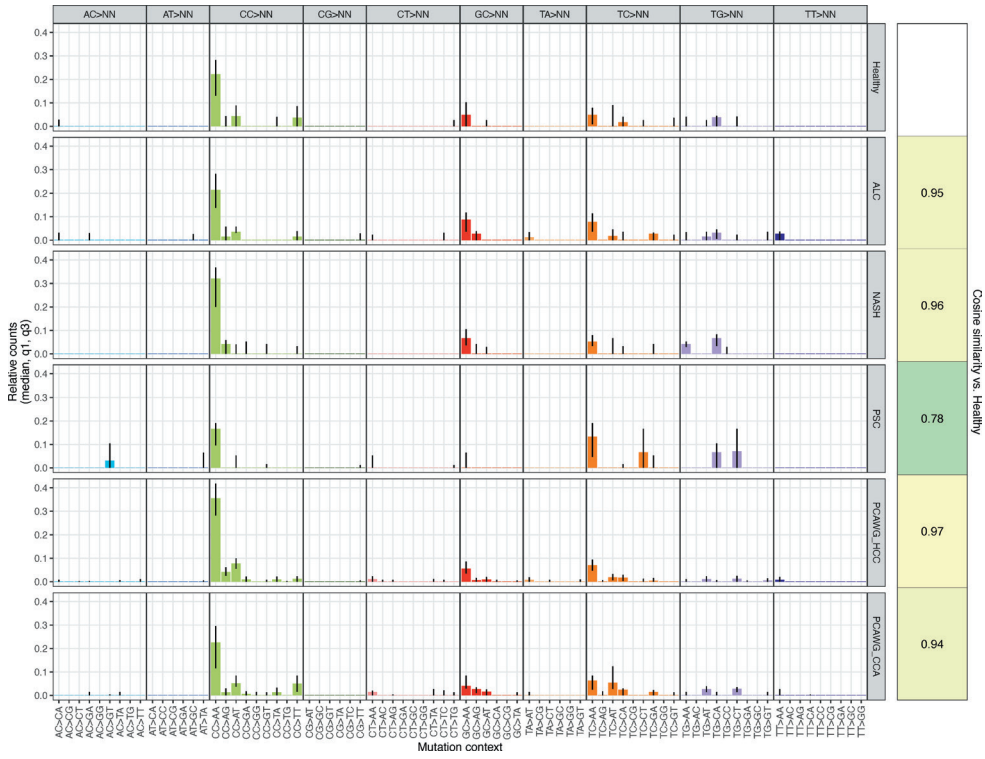
Supplementary figure 3: Chromosome arm copy number profiles in organoids derived from biopsies of healthy, diseased and cancerous livers. ALC: alcoholic cirrhosis, NASH: non-alcoholic steatohepatitis, PSC: primary sclerosing cholangitis, HCC: hepatocellular carcinoma, CCA: cholangiocarcinoma. For one 60-year-old patient with hepatocellular carcinoma biopsies from 5 locations were taken from the liver (HCC_multibiopsy). Profiles for HCC (PCAWG_HCC; n=248) and CCA (PCAWG_CCA; n=12) samples from the Pan-Cancer Analysis of Whole Genomes (PCAWG) consortium are also shown.



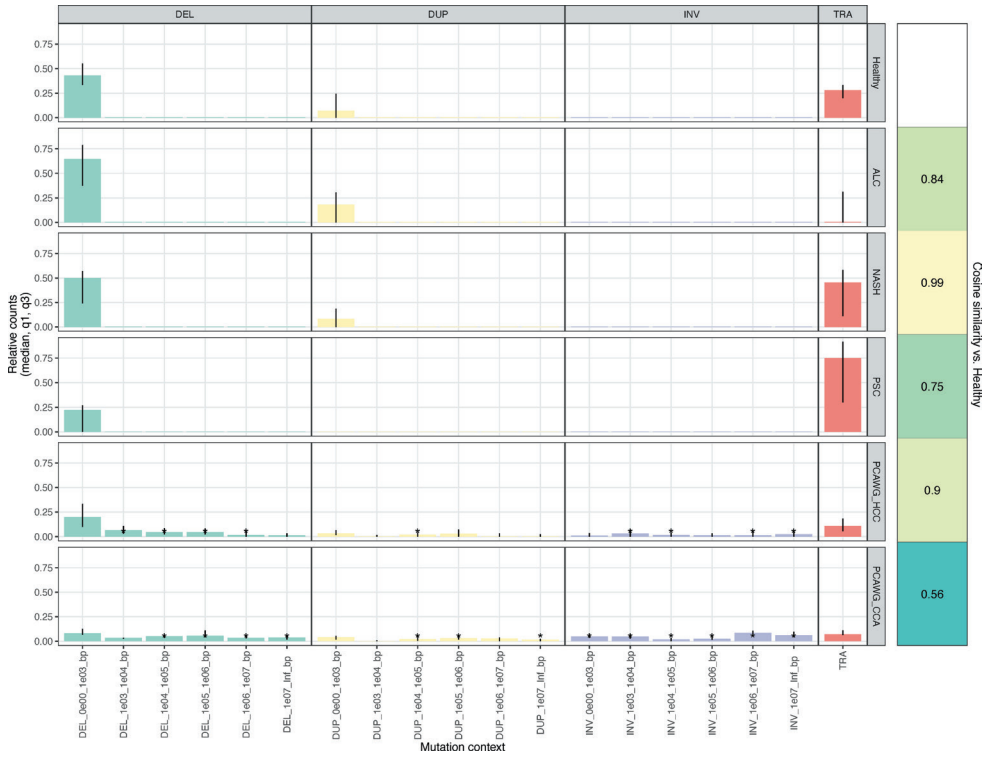
Supplementary figure 4: Trinucleotide substitution profiles in organoids derived from biopsies of healthy, diseased and cancerous livers. ALC: alcoholic cirrhosis, NASH: non-alcoholic steatohepatitis, PSC: primary sclerosing cholangitis, HCC: hepatocellular carcinoma, CCA: cholangiocarcinoma. Profiles for HCC (PCAWG_HCC; n=248) and CCA (PCAWG_CCA; n=12) samples from the Pan-Cancer Analysis of Whole Genomes (PCAWG) consortium are also shown. x-axis labels show the base substitution within the square brackets and the 5' and 3' flanking bases. Bars show the median relative mutation counts (i.e. normalized by total no. of mutations per sample), with error bars showing the 1st and 3rd quartiles. Asterisks indicate a significant increase in mutation context load in a disease/cancer sample versus healthy liver organoids (Wilcoxon rank sum test, Bonferroni adjusted p-value<0.01). Right of the bar plots, cosine similarities of the median profiles of the disease/cancer samples compared to that of the healthy liver organoids.



Supplementary figure 5: Indel profiles in organoids derived from biopsies of healthy, diseased and cancerous livers. ALC: alcoholic cirrhosis, NASH: non-alcoholic steatohepatitis, PSC: primary sclerosing cholangitis, HCC: hepatocellular carcinoma, CCA: cholangiocarcinoma. Profiles for HCC (PCAWG_HCC; n=248) and CCA (PCAWG_CCA; n=12) samples from the Pan-Cancer Analysis of Whole Genomes (PCAWG) consortium are also shown. x-axis labels are in the format (each separated by a dot): (i) deletion or insertion; (ii) no. of deleted/inserted bases; (iii) indel belongs in homopolymer repeats of C/T (C/T), multibase repeat region (rep), or has flanking microhomology (mh); (iv) no. of repeats of C/T, no. of repeat units, or number of homologous bases. Bars show the median relative mutation counts (i.e. normalized by total no. of mutations per sample), with error bars showing the 1st and 3rd quartiles. Asterisks indicate a significant increase in mutation context load in a disease/cancer sample versus healthy liver organoids (Wilcoxon rank sum test, Bonferroni adjusted p-value<0.01). Right of the bar plots, cosine similarities of the median profiles of the disease/cancer samples compared to that of the healthy liver organoids.



Supplementary figure 6: Double base substitution (DBS) profiles in organoids derived from biopsies of healthy, diseased and cancerous livers. ALC: alcoholic cirrhosis, NASH: non-alcoholic steatohepatitis, PSC: primary sclerosing cholangitis, HCC: hepatocellular carcinoma, CCA: cholangiocarcinoma. Profiles for HCC (PCAWG_HCC; n=248) and CCA (PCAWG_CCA; n=12) samples from the Pan-Cancer Analysis of Whole Genomes (PCAWG) consortium are also shown. x-axis labels indicate the dinucleotide change. Bars show the median relative mutation counts (i.e. normalized by total no. of mutations per sample), with error bars showing the 1st and 3rd quartiles. Asterisks indicate a significant increase in mutation context load in a disease/cancer sample versus healthy liver organoids (Wilcoxon rank sum test, Bonferroni adjusted p-value<0.01). Right of the bar plots, cosine similarities of the median profiles of the disease/cancer samples compared to that of the healthy liver organoids.



Supplementary figure 7: Structural variant (SV) profiles in organoids derived from biopsies of healthy, diseased and cancerous livers. ALC: alcoholic cirrhosis, NASH: non-alcoholic steatohepatitis, PSC: primary sclerosing cholangitis, HCC: hepatocellular carcinoma, CCA: cholangiocarcinoma. Profiles for HCC (PCAWG_HCC; n=248) and CCA (PCAWG_CCA; n=12) samples from the Pan-Cancer Analysis of Whole Genomes (PCAWG) consortium are also shown. x-axis labels indicate the SV stratified by mutation type and length interval. Bars show the median relative mutation counts (i.e. normalized by total no. of mutations per sample), with error bars showing the 1st and 3rd quartiles. Asterisks indicate a significant increase in mutation context load in a disease/cancer sample versus healthy liver organoids (Wilcoxon rank sum test, Bonferroni adjusted p-value<0.01). Right of the bar plots, cosine similarities of the median profiles of the disease/cancer samples compared to that of the healthy liver organoids.

Chapter 5

Pan-cancer whole genome comparison of primary and metastatic solid tumors

Francisco Martínez-Jiménez[1], Ali Movasati[1,$], Sascha Brunner[1,$], Luan Nguyen[1,$], Peter Priestley[2], Edwin Cuppen[1,3,*], Arne Van Hoeck[1]

[1] University Medical Center Utrecht, Utrecht, The Netherlands
[2] Hartwig Medical Foundation Australia, Sydney, New South Wales, Australia
[3] Hartwig Medical Foundation, Amsterdam, The Netherlands

[$] These authors contributed equally to this work
[*] Corresponding author

Submitted to Nature
Preprint URL: https://www.biorxiv.org/content/10.1101/2022.06.17.496528v1
QR code to URL:

Abstract

Metastatic cancer remains almost inevitably a lethal disease. A better understanding of disease progression and response to therapies therefore remains of utmost importance. Here, we characterize the genomic differences between early-stage untreated primary tumors and late-stage treated metastatic tumors using a harmonized pan-cancer (re-)analysis of 7,152 whole-genome-sequenced tumors. In general, our analysis shows that metastatic tumors have a low intra-tumor heterogeneity, high genomic instability and increased frequency of structural variants with comparatively a modest increase in the number of small genetic variants. However, these differences are cancer type specific and are heavily impacted by the exposure to cancer therapies. Five cancer types, namely breast, prostate, thyroid, kidney clear carcinoma and pancreatic neuroendocrine, are a clear exception to the rule, displaying an extensive transformation of their genomic landscape in advanced stages. These changes were supported by increased genomic instability and involved substantial differences in tumor mutation burden, clock-based molecular signatures and the landscape of driver alterations as well as a pervasive increase in structural variant burden. The majority of cancer types had either moderate genomic differences (e.g., cervical and colorectal cancers) or highly consistent genomic portraits (e.g., ovarian cancer and skin melanoma) when comparing early- and late-stage disease. Exposure to treatment further scars the tumor genome and introduces an evolutionary bottleneck that selects for known therapy-resistant drivers in approximately half of treated patients. Our data showcases the potential of whole-genome analysis to understand tumor evolution and provides a valuable resource to further investigate the biological basis of cancer and resistance to cancer therapies.

Introduction

Metastatic spread involves tumor cells detachment from a primary tumor, colonization of a secondary tissue, and growth in a hostile environment [233,234]. Advanced metastatic tumors are frequently able to resist aggressive treatment regimes [235]. Despite the many efforts to understand these phenomena [236–240], we still have limited knowledge of the contribution of genomic changes that equip tumors with these extraordinary capacities. Thus, it is essential to characterize genomic differences between primary and metastatic cancers and quantify their impact on therapy resistance to be able to understand and harness therapeutic interventions that establish more effective and more personalized therapies [241].

Multiple large-scale genome-sequencing efforts have been devoted to profiling the genomic landscape of primary tumorigenesis by relying on clinical panels [242] or on whole-exome sequencing [243] or whole-genome sequencing [5,103] strategies. Similarly, recent efforts have characterized large cohorts of patients with metastatic tumors [6,244]. Several cancer type specific projects have also reported analysis of primary and metastatic matched biopsies in urothelial carcinomas [245], breast cancer [246], kidney clear cell carcinoma [247] and prostate carcinoma [248,249], among others [250]. However, the ethical and logistical challenges associated with the collection of paired biopsy data hampers the extrapolation to thousands of patients with multiple cancer types.

To circumvent this issue, most large-scale comparisons between primary and metastatic tumors have relied on unmatched whole-exome data or have adopted more targeted approaches with a specific focus on driver gene landscapes [251–253]. These efforts have frequently involved separated processing pipelines for primary and metastatic cohorts, complicating the analysis of genomic features that are highly sensitive to the selected data-processing strategy [28,254]. An impressive recent study that uniformly analyzed more than 25,000 tumors [255] has provided a comprehensive overview of the genomic differences, driver alteration patterns and organotropism using clinical gene-panel sequencing as a base. However, this genomic analysis approach, which is limited to several hundreds of genes, prevented the exploration of the full spectrum of genomic alterations that play a role in tumorigenesis, such as structural variation and mutational scarring caused by intrinsic and extrinsic forces.

Here, we present a large-scale unified analysis of more than 7,000 whole-genome sequenced (WGS) paired tumor-normal samples (re-)analyzed by the same data-processing pipeline. This dataset enabled cancer type specific comparisons of whole-genome features in 22 cancer types with high representation in both primary and metastatic patients. We investigated differences in tumor clonality, genomic instability markers, whole-genome-duplication (WGD) rates, tumor mutation and structural variant (SV) burden and clock-based molecular signatures and assessed the contribution of cancer treatments to the observed differences. We also explored disparities in the driver gene landscape and their implications for therapeutic actionability. Finally, we identified known and new associations between certain driver alterations and exposure to various treatments. The harmonized genomic dataset used in our study constitutes the largest repository of uniformly analyzed cancer WGS data and is publicly available to the scientific community to better understand the full spectrum of DNA alterations as well as to study their role in tumor evolution and response to therapy.

Results

7,152 uniformly processed whole-genome-sequenced paired tumor-normal samples from patients with primary and metastatic tumors

To characterize the genomic differences between primary and metastatic tumors, we created a uniformly analyzed dataset of matched tumor and normal genomes from patients with primary and metastatic cancer. We first collated the Hartwig Medical Foundation (Hartwig) dataset, which includes 4,784 samples from 4,375 metastatic patients (from which 2,520 samples were previously described [6]). Then, we re-processed 2,835 primary tumor samples from the Pan-Cancer Analysis of Whole Genomes (PCAWG) consortium using the open-source Hartwig analytical pipeline [6,140] to harmonize somatic calling, standardize broad functional annotations of events and eliminate biases caused by processing pipelines and filter conditions (Supp. Fig. 1, Supp. Table 1). Reassuringly, per-sample comparison of the number of single base substitutions (SBSs), double base substitutions (DBSs), indels (IDs), and SVs revealed a strong agreement between our results and the consensus calls originally generated by the PCAWG consortium (see Supp. Note 1). Additionally, our processing pipeline strategy was minimally affected by differences in sequencing coverage, enabling a reasonable comparison of WGS samples from heterogeneous sources (decreased sensitivity at 60× and 38× coverage, typical for PCAWG samples, compared with 109× coverage, typical for Hartwig samples between 0% and 5% for simple mutations and between 8% and 14% for SVs respectively) (Supp. Note 1). A total of 7,152 tumor samples from 58 cancer types met the processing pipeline quality standards (see methods and Supp. Fig. 1a) and constitutes one of the largest publicly available datasets of WGS primary and metastatic tumors. The flexible nature of the processing pipeline, which does not require extensive parameter fine-tuning, enables its application to other whole-genome-sequencing projects, thereby providing an excellent opportunity for future integrative analyses that incorporate this dataset.

To explore genomic differences between primary and metastatic tumors, we focused on 22 cancer types from 14 tissues with sufficient sample representation (i.e., at least 15 unique patients in both the primary and metastatic cohorts), which totaled 5,751 tumor samples (1,916 primary and 3,835 metastatic) (Fig. 1a, Supp Fig 1a). Within this dataset, patients with metastatic tumors were slightly older at biopsy than patients with primary tumors (mean of 2.09 years older across all cancer types). In particular, the mean age at biopsy in primary and metastatic cohorts, respectively, was 59.9 and 67.9 years in patients with prostate carcinoma, 52.8 and 63.6 years in patients with thyroid carcinoma, and 49.6 and 70.3 years in patients with diffuse B cell lymphoma. Consistent gender proportions were observed across all cancer types except for thyroid adenocarcinomas, which had higher male representation in the metastatic cohort (metastatic: 72% male, 28% female; primary: 25% male, 75% female). Treatment information was available for 53% metastatic patients. This information is essential to gauge the contribution of this evolutionary bottleneck to the genomic differences between primary and metastatic tumors (Fig. 1a). Finally, biopsy location was explicitly annotated in 82.4% of metastatic patients, including 12.6% biopsies from metastatic lesions in the primary tissue (local), 16% in lymph nodes and 53.8% in distant locations. Biopsy locations were highly tumor type specific and likely reflected both the dissemination patterns of the tumors and the challenges associated with collection of clinical samples. In summary, we generated a harmonized dataset of 7,152 primary and metastatic WGS tumor samples from 58 cancer types including 22 cancer types with sufficient representation in both early and late clinical stages to allow for a systematic comparison (Supp. Table 1).

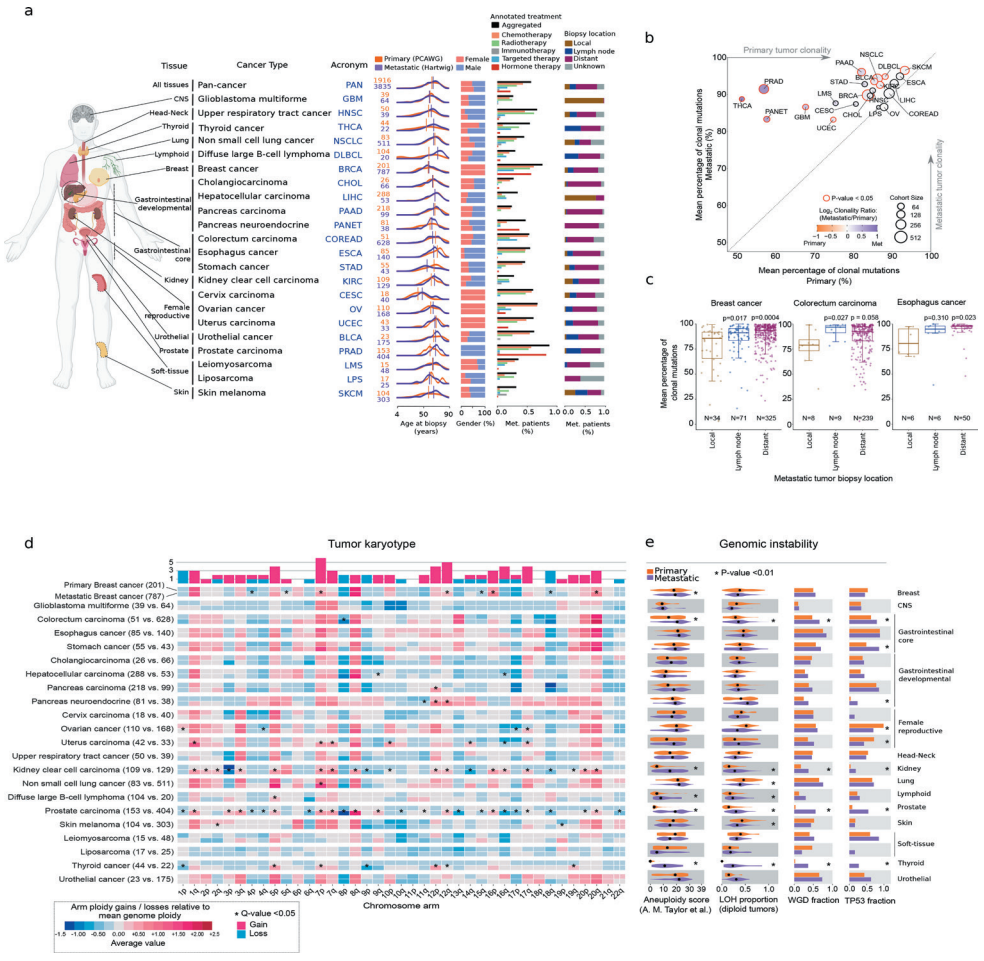


Figure 1. Overview and global genomic features of primary and metastatic tumors. **a)** Anatomic location of the 22 cancer types included in this study. Cancer types are ordered according to their tissue of origin. For each cancer type the following information is represented, from left to right: cancer type acronym, number of samples in primary and metastatic dataset (top and bottom, respectively), age at biopsy (in years), gender, type of treatment of metastatic patients and biopsy site from metastatic tumors. Partially created with BioRender.com. **b)** Mean percentage of clonal mutation in primary (x-axis) and metastatic (y-axis) tumors. Dots are coloured according to Log_2 of the clonality ratio (metastatic divided by primary). Size of dots are proportional to the total number of samples (primary and metastatic). Red edge lines represent a Mann-Whitney p-value < 0.05. **c)** Tumor clonality according to the metastatic biopsy location in breast, colorectal and esophagus cancer (from left to right). N, number of samples in the group. p, Mann-Whitney p-value if p-value < 0.05. ns, not significant. Box-plots: center line, median; box limits, first and third quartiles; whiskers, lowest/highest data points at first quartile minus/plus 1.5× IQR. **d)** Tumor karyotype. Heatmap representing the mean chromosome arm ploidy gain and losses relative to the mean genome ploidy in primary (top) and metastatic (bottom) tumors. Asterisks represent significantly different mean distributions between primary and metastatic tumors (Mann-Whitney adjusted p-value <0.05). Top bars represent the cumulative number of significant gain/losses in the metastatic cohort compared to the primary. **e)** Comparison of four genomic features between primary (top) and metastatic tumors (bottom). From left to right, aneuploidy score from ref [35], proportion of genome undergoing LOH in diploid samples, fraction of samples bearing whole genome duplication (WGD) and TP53 alterations. Black dots represent the median values. Asterisks represent Fisher's exact test p-value < 0.01 for discrete features (WGD and TP53) and Mann-Whitney p-value < 0.01 for the continuous features.

Global genomic characteristics of primary and metastatic tumors

We first explored global genomic differences between primary and metastatic tumors across the aforementioned 22 cancer types. Metastatic tumors showed an overall increase in clonality compared with their primary tumor counterparts (Fig. 1b). Particularly, 12 cancer types had a significantly higher metastatic average clonality ratio, ranging from 3.2% increased mean clonality in skin melanoma to 37% increased mean clonality in thyroid carcinoma. Interestingly, within the group of patients with metastatic breast cancer, distant and lymph node tumor biopsies showed significantly higher clonality ratios compared with local metastatic lesions (Fig. 1c). This increase in clonality was also observed in distant tumor biopsies of esophagus cancer and colorectal carcinomas (Fig. 1c). Nevertheless, the biopsy location did not influence tumor clonality in other cancer types such as non-small cell cancer and skin melanoma (Supp. Fig 1b), suggesting that patterns of tumor dissemination are highly tumor type specific[240]. Thus, our results thus support the notion that metastatic lesions generally have lower intra-tumor heterogeneity[255], which may be explained by a single major subclone seeding event from the primary cancer and/or by severe evolutionary constraints imposed by anti-cancer therapies.

Karyotype comparison revealed a generally conserved portrait, which was strongly shaped by the tissue of origin [27], and where only two cancer types showed substantial karyotypic changes (kidney clear cell and prostate carcinomas, >10 chromosome arm gains/losses) and three additional cancer types displayed moderate changes (breast, uterus, and thyroid carcinomas, >5 chromosome arm gain/losses) in the metastatic cohort compared with the primary cohort (Fig. 1d, Supp. Table 2). The majority (63 of 84) of significant discrepancies were associated with an increased frequency of chromosomal arm gains in the metastatic dataset, while only 21 discrepancies were caused by an increased frequency of chromosome arm losses. Chromosome arms 7p, 12p, 5p and 17q, which are relatively enriched in oncogenes [256], were the most recurrent chromosome arm gains. Conversely, chromosome arms 1p and 18q, highly enriched in tumor suppressor genes[256], were the most recurrent losses in the metastatic cohort (Fig. 1d).

We next investigated differences in four well-studied genomic instability markers: chromosomal aneuploidy score [35], loss of heterozygosity (LOH) genome fraction in diploid tumors, WGD [257], and *TP53* alterations [257,258] (Fig. 1e, Supp. Table 2). Four cancer types (i.e., colorectal, kidney clear cell, prostate, and thyroid) showed persistent increases in the four genomic instability markers in the metastatic cohort, whereas four additional cancer types (i.e., breast, stomach, pancreas neuroendocrine and diffuse B cell lymphoma) had some form of increased genomic instability in the metastatic cohort. Our results thus confirmed that genomic instability is a hallmark of advanced tumors [27,236,241,255,259,260] and revealed that the majority of cancer types have already acquired variable degrees of this genomic feature early in tumor evolution. However, certain cancer types, such as prostate, kidney, and thyroid, significantly increased the level of genomic instability in later evolutionary stages, which were, in turn, associated with substantial karyotypic changes.

Tumor mutation burden

The unified processing of both primary and metastatic tumor samples enables quantitative and qualitative comparison of SBSs, DBSs and ID burden, which we refer to collectively as tumor mutation burden (TMB). We found that the TMB in metastatic tumors was only moderately increased compared with primary tumors across the 22 cancer types tested (fold-change increases of 1.22 ± 0.48 for SBS, 1.52 ± 0.85 for DBS and 1.43 ± 0.55 for IDs; mean ± standard deviation [SD]). In fact, more than 60% of the cancer types (13 of 22) had no significant increase in mutation burden for any mutation type. Only five cancer types (breast, cervical, thyroid, prostate carcinoma, and pancreas neuroendocrine) had a consistent increase for the three mutation types at the metastatic stage, although the mutation profiles lacked systematic differences between primary and metastatic tumors (Fig. 2b, Supp Fig. 2a).

These results show that TMB is not necessarily indicative of tumor progression status and that the mutational spectra are tightly shaped by the mutational processes that were already active before and during primary tumor development.

Mutational processes in primary and metastatic tumors

To assess whether the TMB differences may be attributed to differential activity of environmental or endogenous mutational processes, we conducted a tissue type specific mutational signature *de novo* extraction that resulted in representative mutational signatures of 69 SBSs, 12 DBSs, and 19 IDs (Supp. Table 3). Most of these (50 of 69 SBSs, 8 of 12 DBSs, and 12 of 19 IDs) mapped onto the well-described mutational signatures in human cancer [261]. Moreover, we inferred the suspected etiology for two novel DBS mutational signatures associated with mismatch repair deficiency (MMRd) (DBS_denovo_2) and MMRd/POLE hypermutation (DBS_denovo_1) (Supp. Table 3).

By comparing the activities of all mutational processes, we found that mutations caused by cytotoxic treatments were enriched in 11 cancer types (Fig. 2c-e). Platinum-based chemotherapies (SBS31/SBS35 and DBS5) showed the strongest mutagenic effect with 429 ± 266 (mean ± SD) SBS mutations and 23 ± 15 (mean ±SD) DBS mutations on average per sample. In fact, the excess in DBS mutation burden observed in six cancer types (stomach, esophagus, cervix, upper respiratory tract, non-small cell lung, and urothelial cancer) could be fully attributed to platinum treatment (Fig. 2d top bars). Likewise, the radiotherapy ID signature [172] (ID8) was systematically enriched in multiple cancer types as a response to widespread exposure to radiation-based treatment, whereas the 5-fluorouracil [66,108] (SBS17a/b), duocarmycin [262] (SBS90) and polycyclic aromatic hydrocarbon (PAH) metabolites from chemotherapy [67] (DBS2) occasionally led to a greater mutation burden in metastatic tumors in a tumor type specific manner (Fig. 2c-e).

The systematic enrichment of SBS2/13 mutations in metastatic cancers suggests enhanced activity of APOBEC mutagenesis during the progression of advanced tumors. Specifically, our results revealed an increase in APOBEC mutation burden of 316 ± 173 (mean ± SD) mutations per sample in six metastatic tumors (breast, colorectal, stomach, kidney, prostate, and pancreas neuroendocrine) that reached statistical significance, being breast and stomach cancers the ones with the strongest increase (>500 APOBEC mutations per sample in both cancer types). Other cancer types, such as cervical and urothelial cancers, also showed enhanced APOBEC activity (>1800 mutations per sample), but they did not reach significance owing to high intrinsic APOBEC activity in the primary tumors. The metastatic breast cancer samples also had a higher percentage of APOBEC hypermutation than primary tumors (14% vs 5%, Supp. Fig. 2c, Supp. Table 3).

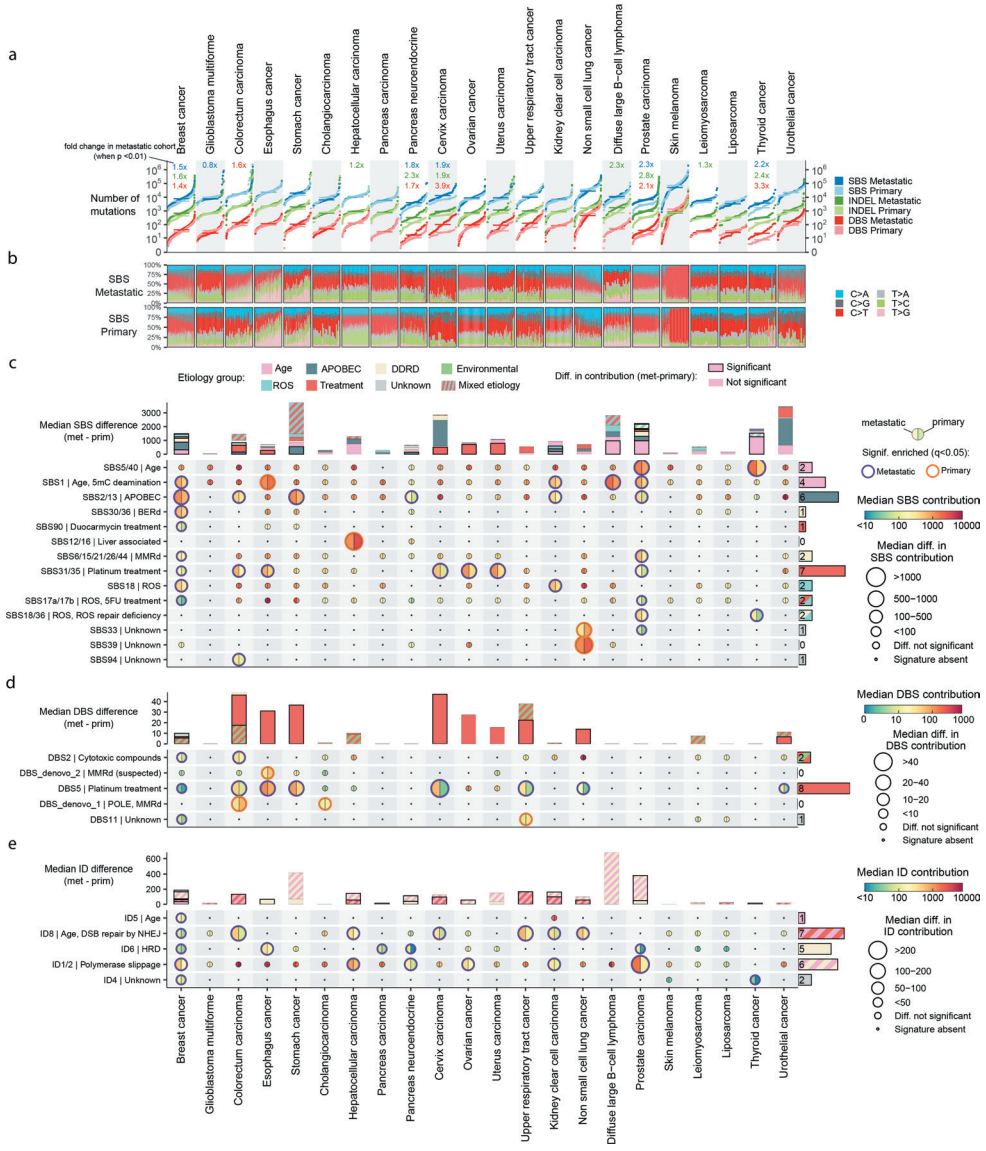


Figure 2. Tumor mutation burden and mutational processes. a) Cumulative distribution function plot (samples were ranked independently for each variant type) of tumor mutation burden for each cancer type for SBS (blue), IDs (green) and DBS (red). Horizontal lines represent median values. Fold change labels are included only when Mann-Whitney comparison rendered a significant p-value < 0.01. **b)** SBS Mutational spectra of metastatic (top) and primary (bottom) patients. Patients are ordered according to their TMB burden. **c)** Main panel, moon plot representing the mutational burden differences attributed to each mutational signature in metastatic (left) and primary (right). Edge thickness and colors represent significant differences (q<0.05, ±1.4x fold change) and the direction of the enrichment, respectively. The size of circles are proportionate to the mutation burden difference. Right bars, number of metastatic cancer types with a mutational signature significant enrichment. Top stacked bars represent the cumulative signature exposure difference. Thicker bar edge lines represent significance. Bars are coloured according to the annotated etiology. Only mutational signatures with known etiology or with at least

one cancer type with significant metastatic enrichment are included. **d)** and **e)** equivalent for DBS and IDs, respectively. Diff., difference. Muts. mutations. Sig., mutational signature. Mut. mutational. Susp., suspected.

Five metastatic cancer types also displayed more mutations from the clock-like mutational processes, including four cancer types (breast, prostate, diffuse B-cell lymphoma and kidney clear cell carcinoma) that exhibited an increased SBS1 contribution and two cancer types (prostate and thyroid) that had an increased SBS5/SBS40 mutation burden. The increase in clock-like mutations in thyroid and prostate cancers, as well as diffuse B-cell lymphomas to a lesser extent, may be explained by the greater proportion of older patients with metastatic disease compared with primary disease. However, the SBS1 enrichment was also present in cancer types in the primary and metastatic cohorts with highly similar age population distributions (see Fig. 1a). Thus, our results revealed an increase in SBS1 mutations in advanced tumor stages that cannot be explained by older patients' ages at the time of biopsy.

Additional more-focused analyses may achieve a better understanding of the observed mutational signatures with differing activity in a small subset of cancer types (such as oxidative stress-induced and DNA repair related mutations) and those lacking mechanistic etiology warrants additional more focused analyses. All contributions can be found in Supp. Table 3.

Differential SBS1 mutation rates in primary and metastatic cancers

To investigate the apparent contradiction of increased SBS1 mutation burden in the metastatic cohort, despite primary and metastatic cohorts having a similar age at biopsy, we assessed their SBS1 mutation burden by the age of biopsy separately for both cohorts (see methods). As expected, rates of SBS1 mutation acquisition were highly tissue specific[92,263], and SBS1 mutation burden increased linearly with age in the majority of cancer types in both primary and metastatic cohorts (Pearson's $R > 0.1$, 17 of 22 tumor types, Supp. Fig. 3a, Supp. Table 4). However, four cancer types (i.e., breast, prostate, kidney clear cell, and thyroid) showed an age-independent and significant enrichment of SBS1 mutations at the metastatic stage (Fig. 3a, Supp. Fig. 3a). For instance, metastatic breast cancer had a nearly uniform fold increase of 1.46 (189 ± 17 SBS1 mutations, mean ± SD) across the ages of biopsies that was independent of tumor genome ploidy and *HER2* status (Supp. Fig. 3b-c). Importantly, this pattern was highly cancer type specific and was not observed for most cancer types, including those with similar intra-tumor heterogeneity in the primary cohort (e.g., colorectal, ovarian cancer, pancreas, and stomach cancers) (Fig. 3b, Supp. Fig. 3a). Moreover, other mutational processes that operate over the evolution of the somatic tissues (e.g., clock-like mutations attributed to SBS5/SBS40 that accumulate with age in a cell-cycle independent manner [92,264]) were not enriched (Supp. Fig. 3d).

SBS1 mutation burden has been extensively correlated with estimated stem cell division rates [265]. Therefore, an increase in age and tumor type specific SBS1 mutation burden in treated metastatic tumors may indicate that these tumors have undergone a higher number of cell divisions. However, the estimated number of years to explain the SBS1 mutation burden shift (23 and 67 years for breast and prostate cancer respectively, see Supp. Table 4) shows that this cannot be the main cause. Hence, a more plausible explanation, which also supports previous observations [266–268], is that these metastatic tumors display accelerated cell division rates compared with their primary tumor counterparts (Fig. 3c). Supporting this hypothesis, metastatic tumors also had a lower normalized fraction of clonal SBS1 mutations as well as a greater fraction of SBS1 clonal late mutations (Fig. 3d, Supp. Fig. 3e). Of note, this pattern was not observed in cancer types with consistently high SBS1 mutagenic dynamics (Fig. 3e) and was indistinguishable for SBS5/SBS40 mutations (Supp. Fig. 3f).

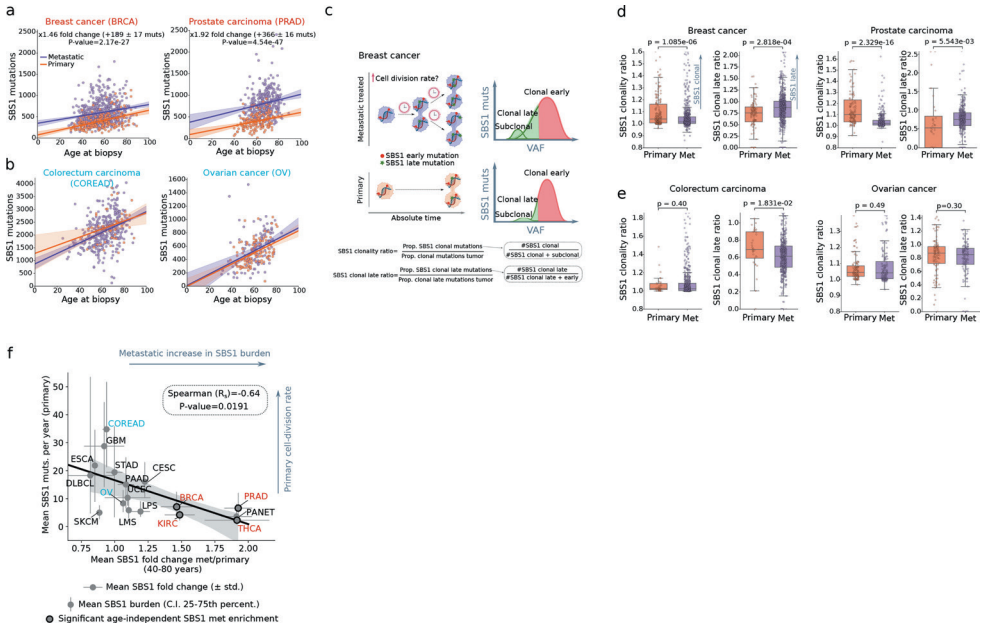


Figure 3. Cell cycle division rates in primary and metastatic tumors. a) Linear regression of the SBS1 mutation burden (y-axis) and patient's age at biopsy (x-axis) in primary and metastatic breast (left) and prostate cancers (right) and **b)** colorectal and ovarian cancer. The mean fold change, mean SBS1 increase per year and p-value are included in cancer types with an age-independent significantly different primary and metastatic distribution. The median trendline and 99% confidence intervals of the linear regression are represented as a solid line and the adjacent shaded area, respectively. **c)** Depiction illustrating increased cell division rate in metastatic breast cancer compared to primary and its expected impact on the SBS1 variant allele frequency (VAF) distribution. Partially created BioRender.com. **d)** Comparison of global SBS1 clonality ratio and clonal late ratio between primary and metastatic in breast (left) and prostate cancers (right)**. e)** Similar for colorectal and ovarian cancers. Boxplots are defined as in Fig. 1. **f)** Spearman correlation analysis of the mean SBS1 year burden of primary tumors (y-axis) and the mean metastatic SBS1 fold change (x-axis) across the 17 cancer types with linear association between age and SBS1 accumulation. Vertical error bars represent the 25th and 75th percentile, respectively. Horizontal error bars the fold change standard deviation. Cancer types with a significantly different SBS1 mutation rate are marked by thicker marker borders. Muts, mutations.

To understand the underlying factors that lead to the increased proliferation rates in some cancer types and not in others, we explored the relationship between the yearly rate of SBS1 mutation accumulation in primary tumors (a proxy of stem cell division rates [265]) and the estimated fold change of the SBS1 mutation rate in the metastatic cohort. Remarkably, we observed a strong negative association between these two factors (Spearman R_s= -0.64, p-value = 0.01, Fig. 3f), which was consistent when relying on independent measurements of primary tumor turnover rates (Supp. Fig. 3h, Supp. Table 4). This indicates that tumors with an intrinsically active turnover rate (e.g., colorectal and ovarian cancers) preserve their high proliferation rates, while others with low cell division rates (e.g., breast, prostate, kidney, and thyroid cancers), may acquire higher proliferation rates during the course of cancer progression and treatment exposure.

Structural variant burden

The total number of SVs per tumor revealed an extensive increase in SV burden in the metastatic tumors (fold change 1.9 ± 1.2, mean ± SD). In fact, 14 of 22 (64%) cancer types showed a significant median increase in total SV burden in the metastatic tumors (Fig. 4a, Supp. Table 5), which cannot be explained by differences in sequencing coverage or tumor clonality (Supp. Note 1). Importantly, these cancer types not only included the eight cancer types with increased genomic instability markers, but also six additional cancer types (i.e., esophagus, hepatocellular, pancreas carcinoma, cervix, non-small cell lung cancer, and urothelial cancer) (Fig. 4a).

Breaking down the comparison into specific SV types revealed that small (<10kb) deletions, which also contributed most to the global SV burden, showed a strong metastatic enrichment (2.15 ± 1.3 fold change in 16 of 22 cancer types with significant enrichment, mean ± SD, Fig. 4a, Supp. Fig. 4a). Moreover, larger (≥10kb) deletions and duplications had a similar pan-cancer enrichment, although with slightly lower fold changes that varied from 1.5 up to 1.9. Moreover, complex SVs with ≥20 breakpoints, encompassing events such as chromothripsis and chromoplexy, were particularly enriched in esophagus cancer (1.5-fold) and prostate cancer (3-fold). Finally, a strong cancer type specific metastatic enrichment was also noted for the long interspersed nuclear element (LINE) insertions, with an increased fold change of 12.2 and 8.5 in stomach and urothelial cancer, respectively.

Compared with TMB, the SV analyses revealed a much more widespread effect, with larger increases per metastatic cancer type that affected almost every cancer type studied, indicating that metastatic tumors appear to evolve primarily by genomic changes at the structural level.

Genomic and clinical features associated with structural variant burden

We next sought to unravel the underlying features associated with the observed increase in SV burden in metastatic tumors using linear regression models (Fig. 4b-f, Supp. Fig. 4b-f, Supp. Table 5). Our approach confirmed previously described cancer type specific driver-induced SV phenotypes, including HRd (*BRCA1/2*) [38,74], *CDK12* [269], and *MDM2* [104], which are more frequently mutated in metastatic pancreatic, prostate, and breast carcinomas, respectively. Reciprocal duplications induced by *CDK12* and *CCNE1* alterations[270] likely explain the enrichment of complex rearrangements in prostate and esophagus cancer, respectively. We also found potential novel associations, such as large deletions linked to the chromatin regulators *SETD2* and *CHD1*. Finally, the cell cycle checkpoint *CDKN2A* showed a pervasive association with deletions in breast and kidney clear cell carcinomas (Fig. 4b-d, Supp. Fig. 4b). However, this may not necessarily imply causation, because the tumor suppressor *CDKN2A* is frequently inactivated in tumors in which deletion signatures are common [178].

We also found that genomic instability features (i.e., genome ploidy and *TP53* alterations) showed a strong pan-cancer association with deletions and, to a lesser extent, with duplications (Fig. 4b-f, Supp. Fig. 4b-f), supporting the established role of WGD [257] and *TP53* [35] loss in the ubiquitous generation of large genomic alterations. As mentioned earlier (see Fig. 1e), these features were generally more prevalent in patients with metastatic tumors and thus very likely contributed to the observed SV increase in metastatic tumors.

Finally, prior exposure to radiotherapy treatment was strongly associated with small deletions in breast and prostate cancers (in agreement with radiotherapy-treated secondary malignancies [109], gliomas [172], and healthy tissue [271]). This suggests that, among common anti-cancer therapies, radiotherapy in particular significantly contributes to the SV landscape in treatment-surviving cancer cells.

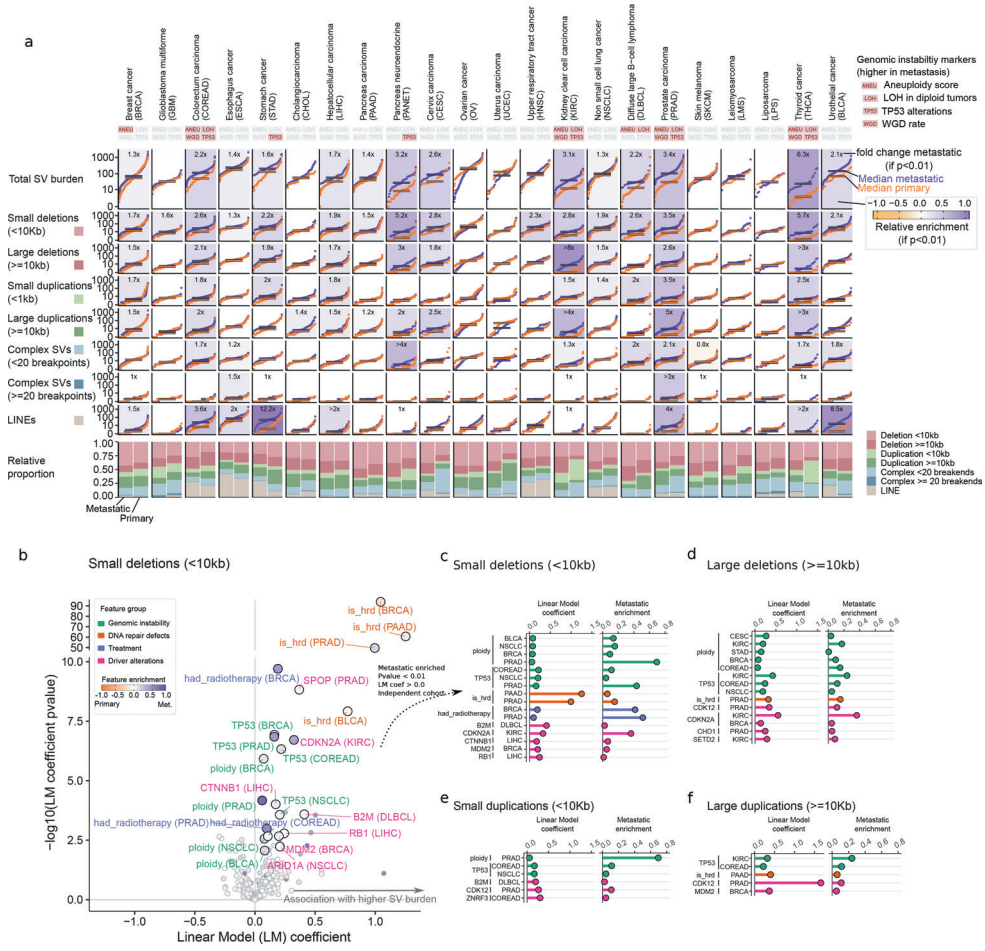


Figure 4. Structural variant burden and associated genomic features. a) Top rectangles represent the four genomic instability features defined in Fig. 1e. A red background represents significant enrichment in the metastatic cohort. S-plots, cumulative distribution function plot (samples ranked independently for each SV type) of tumor mutation burden for each cancer type for (from top to the bottom) the aggregated structural variant (SV) burden, small deletions (<10kb), large deletions (>=10kb), small duplications (<10kb), large duplications (>=10kb), complex events (<20 breakpoints), complex events (>=20 breakpoints) and LINEs insertions. Horizontal lines represent median values. Backgrounds are coloured according to the relative enrichment, defined as: $\log_{10}$(median SV type burden in metastatic tumors + 1) – $\log_{10}$(median SV type burden in primary tumors + 1). Fold change labels and coloured backgrounds are displayed when Mann-Whitney comparison renders a significant p-value < 0.01. Fold change labels are displayed with '>' when the SV burden for primary tumors is 0 (see methods for more details). For each cancer type, bottom bar plots represent the relative fraction of each SV type in the metastatic (left) and primary (right) datasets. **b)** Volcano plot representing the cancer type specific regression coefficients (x-axis) and significance (y-axis) of clinical and genomic features against the number of small deletions. Each dot represents one feature in one cancer type. Labels are coloured according to the feature category. Dots are coloured by the frequency enrichment in metastatic (purple) or primary (orange) patients. LM, linear model. Coef, coefficient. **c)** Lollipop plots representing the regression coefficients (left, relative to panel b. x-axis) and metastatic enrichment (right, relative to dots color from panel b.) of features associated with small deletions. Only significant features (LM>0.0, p-value < 0.01 and with independent significance in primary or metastatic tumors) enriched in metastatic patients (enrichment > 0.0) are displayed. **d), e)** and **f)** are identical but referring to large deletions, small duplications and large duplications, respectively.

Cancer driver gene alterations in primary and metastatic tumors

Comparison of the total number of driver gene alterations per tumor sample revealed a moderate increase in the metastatic cohort (a mean of 4.5 and 5.2 driver alterations per sample in primary and metastatic tumors, respectively), including 11 (50%) tumor types with a significant increase (Fig. 5a). Prostate adenocarcinoma (3.02 driver alterations per sample), pancreas neuroendocrine (2.16), thyroid carcinoma (1.7), and kidney clear cell carcinoma (1.63) showed the strongest increases (>1.5 driver alterations per patient), whereas the majority of cancer types showed a mean increase below 1.5 driver alterations per sample. All mutation types (amplifications, deletions, and mutations) tended to show consistent biases toward increases in metastatic tumors (Fig. 5a), with small variant mutations contributing the most (2.3 and 2.6 driver mutations per sample in primary and metastatic tumors, respectively).

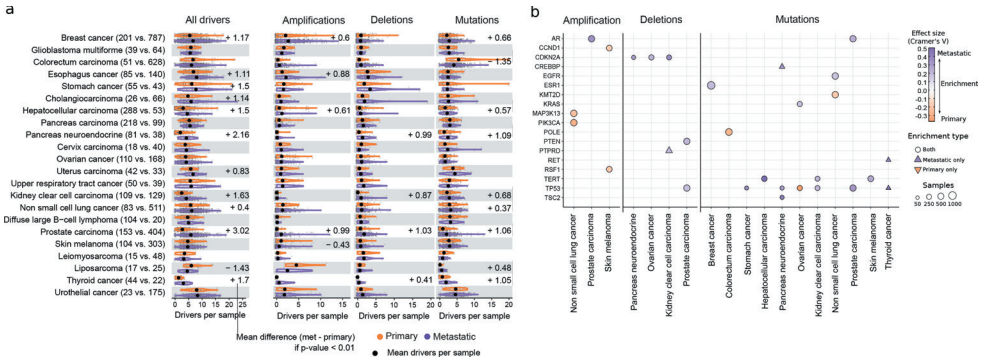


Figure 5. Driver alterations in primary and metastatic tumors. a) Cancer type specific distribution of number of driver alterations, amplifications, deletions and mutations per patient in primary (top) and metastatic (bottom). Black dots represent the mean values. Labels display mean differences (metastatic - primary) in cancer types with a significant difference. **b)** Heatmap representing the cancer genes displaying significant mutation frequency differences between primary and metastatic tumors. Circles denote mutation frequency enrichment in both cohorts while triangles facing upwards and downwards represent drivers that are exclusively enriched in metastatic and primary cohorts, respectively. Marker size is relative to the total number of mutated samples. Colors represent the direction of the enrichment.

Comparison of gene- and cancer-type frequencies revealed that only 18 genes had a significant frequency bias in at least one cancer type (29 gene and cancer-type pairs in total, Fig. 5b, Supp. Fig. 5a-b, Supp. Table 6). The majority (21 of 29, 72%) of the significant pairs had enrichment toward higher metastatic frequency, including four driver genes that were exclusively mutated in metastatic tumors (*PTPRD* in kidney clear cell carcinoma, *CREBBP* in pancreas neuroendocrine, and *RET* and *TP53* alterations in thyroid carcinoma). Moreover, most metastatic-enriched cancer drivers had a cancer type specific enrichment that included well-established resistance gene drivers associated with anti-cancer therapies, such as *AR* and *ESR1* alterations in patients with prostate and breast cancer treated with hormone deprivation therapies [170,272], and *EGFR* mutations in patients with metastatic non-small cell lung cancer often treated with anti-EGFR inhibitors. Nevertheless, three driver genes (i.e., *TP53*, *CDKN2A,* and *TERT*) showed a metastatic enrichment across multiple cancer types (Fig. 5b), indicating that alterations of these genes may enhance aggressiveness by disturbing pan-cancer hallmarks of tumorigenesis. In fact, TP53 alterations have been extensively linked to genomic instability[257,258], while *CDK2NA* and *TERT* are key regulators of cell proliferation, two pathways that are often perturbed in metastatic tumors [273] (see earlier). Finally, some driver genes were strongly enriched in primary tumors, such as *KMT2D* mutations in non-small cell lung cancer (primary 13.7%, and metastatic 2.1%) and *POLE* mutations in primary colorectal carcinomas (primary 13%, metastatic

0.5%). Of note, the higher prevalence of *POLE* mutations in primary tumors supports the increased SBS10a/b exposure (i.e., *POLE* hypermutation) in primary colorectal cancer (see Fig. 2d). Whether these alterations are indicators of better prognosis or drivers of subclonal expansion warrants further investigation.

We next investigated whether the reported driver differences may impact on potential clinical actionability. Cancer type specific comparison of therapeutically actionable variants revealed an overall greater fraction of patients with therapeutically actionable variants in the metastatic cohort, with high variability across cancer types (Supp. Fig. 6a, Supp. Table 7). Subsetting by A-on label variants (i.e., approved biomarkers in the specific cancer type) revealed a consistent pattern in which only breast cancer (driven by higher *PIK3CA* mutation frequency) and non-small cell lung cancer (*EGFR* and $KRAS^{G12C}$ alterations) showed a significant proportional increase in the metastatic cohort (Supp. Fig. 6a-b). Evidence levels representing biomarkers in experimental clinical stages (A-off label, B-on label, and B-off label) showed a modest and tumor type dependent metastatic increase, which was mainly linked to the increased alteration frequency of $KRAS^{G12X}$, *EGFR* secondary mutations, and *CDKN2A* loss in advanced tumor stages (Supp. Fig. 6b).

In conclusion, the cancer driver gene landscape is generally conserved and the observed differences are associated with hallmarks of tumor aggressiveness and resistance to anti-cancer therapies. Consequently, therapeutic options are mainly dictated by the primary tumor [274], although advanced experimental drugs may provide relevant therapeutic opportunities for metastatic tumors in the near future.

Treatment associated driver gene alterations

The high prevalence of resistance driver genes in late-stage tumors prompted us to devise a test that aimed to identify treatment enriched drivers (TEDs) that were either significantly enriched (i.e., treatment enriched) or exclusively found (i.e., treatment exclusive) in a cancer type and treatment specific manner (Fig. 6a and methods). Our analytical framework provided 56 TEDs associated with 24 treatment groups from eight cancer types (Fig. 6b bottom pie chart, Fig. 6c, Supp. Table 8). Of the identified TEDs, 28 of 56 (50%) were coding mutation drivers, 18 (30%) copy number amplifications, 8 (15%) non-coding drivers and 3 (5%) recurrent homozygous deletions (Fig. 6c, Supp. Fig. 7a-b). Reassuringly, the majority of the top hits were widely known treatment-resistance drivers, including $EGFR^{T790M}$ mutations (Fig. 6d) and *EGFR* copy number gains in patients with non-small cell lung cancer treated with *EGFR* inhibitors[275,276] (Supp. Fig. 7c), *AR*-activating mutations and gene amplifications in prostate cancer patients treated with androgen-deprivation therapy [170] (Supp. Fig. 7d-e), and *ESR1* mutations in breast cancer patients treated with aromatase inhibitors [272] (Supp. Fig. 7f), among others. Moreover, we also found that *TP53*, *KRAS* and *PIK3CA* alterations were recurrently associated with resistance to multiple treatments, which may indicate that these alterations are prognostic markers for enhanced tumor aggressiveness and plasticity rather than being a cancer type specific mechanism of drug resistance.

Our results also provided a long tail of candidate drivers of resistance, some of them with orthogonal evidence by independent reports (Fig. 6c, Supp. Fig. 7a-b). Examples of the latter group include *TYMS* amplification in breast cancer patients treated with pyrimidine antagonists [277] (Fig. 6e), *PRNC1* and *MYC* co-amplifications in prostate cancer patients treated with androgen-deprivation [278] (Supp. Fig. 7g), *SMAD4* mutations in non-small cell lung cancer treated with immunotherapy [279], and *FGFR2* promoter mutations in breast cancer patients treated with CDK4/6 inhibitors [280]. The full TEDs catalog is provided in Supp. Table 8 and constitutes a valuable resource for investigating resistance mechanisms to common cancer therapies.

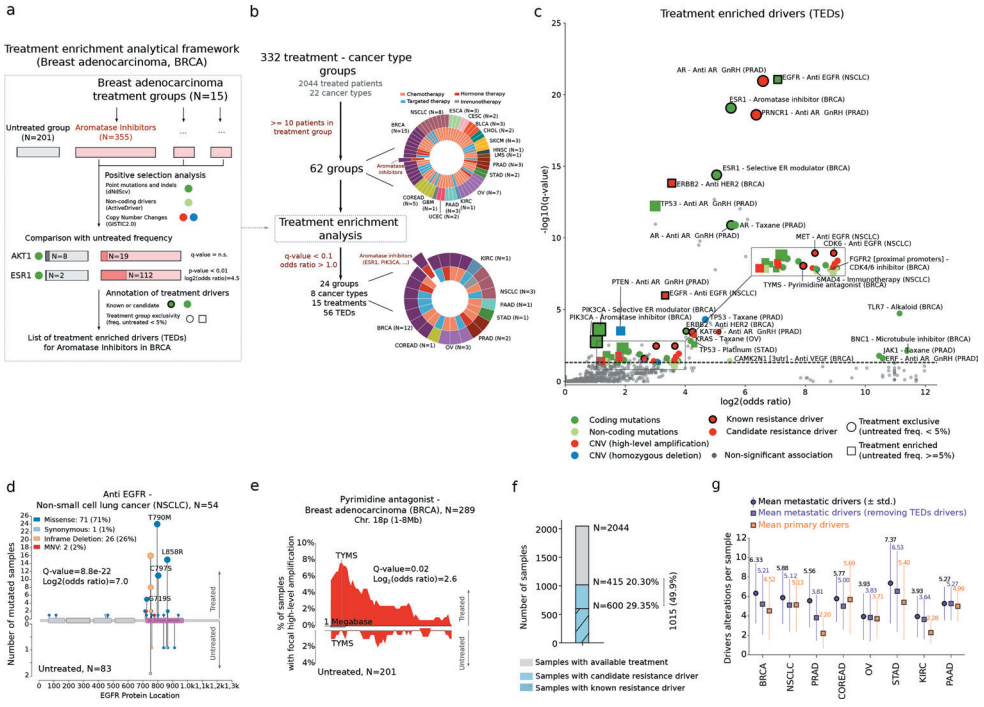


Figure 6. Treatment enriched drivers. a) Visual depiction of the analytical framework to identify treatment enriched drivers (TEDs). The example illustrates the identification of TEDs in the 355 breast cancer patients treated with aromatase inhibitors. First step, identification of cancer driver genes from coding mutations (green), non-coding mutations (soft green), copy number amplifications (red) and biallelic deletions (blue). Second, for each identified cancer driver, comparison of the mutation/copy number alteration frequency in treated and untreated patients. Third, annotation of TEDs with orthogonal evidence and type of enrichment. **b)** Left, workflow representing the number of treatment groups in each step of the analysis. Pie charts, the external layers represent the number of treatment groups analyzed (top) or with identified TEDs (bottom) coloured by cancer type. The internal layers represent the category of the corresponding treatment. The group of breast cancer patients treated with aromatase inhibitors highlighted in both charts. N, number of patients in each treatment group. **c)** Volcano plots displaying the identified TEDs. Each dot represents one cancer gene alteration type in one treatment group. X-axis displays the effect size (as $\log_2$[odds ratio]) and the y-axis the significance (-$\log_{10}$[q-value])). Circle markers represent TEDs exclusively mutated in the treatment group (squared makers otherwise). Markers are coloured according to the type of alteration. Known resistance drivers are denoted by thicker edgelines. **d)** Distribution of mutations along the *EGFR* protein sequence in non-small cell lung cancer patients treated with EGFR inhibitors (top) and untreated (bottom). Pfam domains are represented as rectangles. Mutations are coloured according to the consequence type. **e)** Distribution of highly-focal copy number gains in chromosome chr18p:1Mb-8Mb in breast cancer untreated patients (bottom) and treated with pyrimidine antagonists (top). TYMS genomic location is highlighted. **f)** Global proportion of metastatic treated patients with known and candidate TEDs. **g)** Mean number of driver alterations per metastatic patient before (purple circle) and after excluding TEDs (purple square) compared to primary patients (orange square). Vertical lines, standard deviation range and labels the mean number of driver alterations. BRCA, Breast cancer. CESC, Cervix carcinoma. CHOL, Cholangiocarcinoma. COREAD, Colorectal carcinoma. ESCA, Esophagus cancer. GBM, Glioblastoma multiforme. KIRC, Kidney clear cell carcinoma. LMS, Leiomyosarcoma. NSCLC, Non small cell lung cancer. OV, Ovarian cancer. PAAD, Pancreas carcinoma. PRAD, Prostate carcinoma. SKCM, Skin melanoma. STAD, Stomach cancer. THCA, Thyroid cancer. HNSC, Upper respiratory tract cancer. BLCA, Urothelial cancer. UCEC, Uterus carcinoma. MB, Megabase.

Overall, almost 50% of patients with metastatic disease with annotated treatment information harbored TEDs, including 30% with annotations of known resistance drivers and an additional 20% of patients with candidate resistance drivers derived from our analysis (Fig. 6f). We identified 0.70 ± 0.53 (mean ± SD) TEDs per metastatic sample across the eight cancer types that had reported TEDs (Fig. 6g), with prostate and breast cancers displaying the greatest prevalence of TEDs (i.e., 1.74 and 1.12 drivers per patient with prostate and breast cancer, respectively). Therefore, after excluding TEDs, primary and metastatic tumors had an approximately 40% reduction of their original differences in number of drivers per sample (from 5.2 to 4.9 mean drivers per sample in the metastatic cohort after excluding TEDs, compared with 4.5 mean drivers per sample in the primary cohort) (Fig. 6g, Supp. Table 8), indicating that an important proportion of the metastatic-enriched drivers are likely associated with resistance to anti-cancer therapies.

Discussion

In this study we describe a cohort of >7,000 uniformly (re-)processed WGS samples from patients with primary untreated and metastatic treated tumors. Our robust analytical pipeline enabled the processing of large sets of paired tumor-normal WGS samples from diverse sequencing platforms with high efficiency and minimal human intervention. We leveraged this dataset to perform an in-depth comparison of genomic features across 22 cancer types with high representation from patients with primary and metastatic tumors.

Our analyses revealed that metastatic tumors share common genomic traits, such as high genomic instability, low intra-tumor heterogeneity, and stronger enrichment of SVs, but fewer short mutations than primary tumors. However, the magnitude of genomic differences between primary and metastatic tumors is highly cancer type specific and is strongly influenced by exposure to cancer treatments. Overall, five cancer types (prostate, thyroid, kidney clear cell, breast, and pancreas neuroendocrine cancers) showed an intense transformation of the genomic landscape in advanced tumorigenic stages. Fueled by increased genomic instability, these cancer types displayed substantial differences in TMB, clock-based molecular signatures and driver gene landscape as well as the pervasive increase in SV burden (Fig. 7, labeled as Strong). Importantly, metastatic prostate and breast cancers regularly harbored metastatic-exclusive therapy-resistant driver gene mutations, suggesting that an important proportion of the genomic differences in metastatic tumors compared with primary tumors may be associated with clonal (re-)expansion following therapy exposure (see ref[250] supporting this notion). However, the genomic differences in kidney, thyroid, and pancreas neuroendocrine carcinomas could not be linked to genomic markers of therapy resistance, which may indicate alternative evolutionary dynamics independent of cancer treatment. Another nine cancer types (e.g., cervix and colorectal, among others) displayed moderate genomic differences, although the global genomic portrait was conserved. Finally, the genomic landscape of the eight remaining cancer types (e.g., glioblastoma, sarcoma, and ovarian cancer, among others) was highly consistent between primary and metastatic stages, and the minimal differences were mainly attributed to exposure to cancer therapies.

This study has several limitations. First, it was performed using unpaired primary and metastatic tumor biopsies. Ideally, matched biopsies from the same patient, as already implemented in cancer type specific studies [247,281,282], would be needed to more specifically address the evolutionary dynamics of treated metastatic tumors. Moreover, the sequencing depths and tumor purity ranges used in this study are suboptimal to comprehensively profile the heterogeneous landscape of subclonal alterations, which may lead to underestimation of the extent of late-active mutational processes (e.g., treatment-induced processes). Single cell-based sequencing approaches will be instrumental to further dissect mutational landscapes independent of their clonality. In addition, the sequencing depth of the primary tumor cohort was lower and more variable than that of the metastatic tumor cohort. Although we demonstrated that this does not severely impact on the overall detectability

of clonal somatic variants, individual drivers may have been occasionally missed, which may negatively impact statistical accuracy. Finally, genomic changes cannot entirely explain how tumor cells are able to colonize other organs while avoiding the strong bottlenecks imposed by the immune system [283] or by aggressive treatment regimes. Therefore, additional information from complementary tumor omics studies [284,285] and from the tumor microenvironment [286,287] will be needed to further dissect and better understand tumor evolution and resistance to cancer therapy and eventually contribute to improved management of this deadly disease.

To conclude, our dataset constitutes a valuable resource that can be leveraged to further study other aspects of tumor evolution (as illustrated by the accompanying publication [283] focusing on genetic immune escape alterations) as well for the development of machine learning tools to foster cancer diagnostics [288].

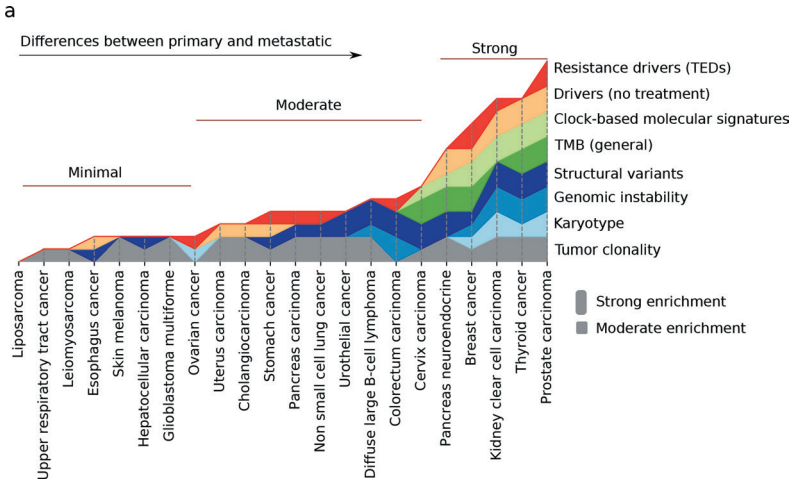


Figure 7. Pan-cancer differences between primary and metastatic tumors. Stacked plot representing the qualitative differences of the eight studied genomic features across the 22 cancer types included in this study. Cancer types are sorted in ascending order according to the cumulative number of diverging genomic features between primary and metastatic tumors. Each horizontal track represents a genomic feature. The presence (and height) of each feature for a specific cancer type correlates with the magnitude of the observed differences.

Methods

Cohort gathering and processing

We have matched tumor-normal whole genome sequencing data from cancer patients from two cohorts: the Hartwig Medical Foundation (Hartwig) and the Pan-Cancer Analysis of Whole Genomes (PCAWG) cohort. A detailed description of the Hartwig and PCAWG cohort gathering and processing as well as comprehensive documentation of the PCAWG sample reanalysis with the Hartwig somatic pipeline is described in the Supp. Note 1.

Tumor clonality analysis

Each mutation in the .vcf files is given a subclonal likelihood by PURPLE. Following PURPLE guidelines, we considered mutations with subclonal scores of 0.8 or higher to be subclonal and mutations below the 0.8 threshold to be clonal. For each sample we then computed the average

proportion of clonal mutations by dividing the number of clonal mutations by the total mutation burden (including SBS, MNVs and indels). Finally, for each cancer type we used Mann-Whitney test to assess the significance of the clonality difference between the primary and metastatic tumors. A p-value lower than 0.05 was deemed as significant.

In addition, we leveraged biopsy site data in patient reports to further investigate differences in metastatic tumor clonality according to the metastatic biopsy site. If the metastatic biopsy site was in the same organ/tissue as the primary tumor, we considered them as "Local", while if the metastatic biopsy site was reported in the lymphoid system or other organs/tissues they were dubbed as "Lymph" and "Distant", respectively. Cancer types for which there was a minimum of 5 samples available for each of the biopsy groups were selected and Mann-Whitney test was used to compare the clonality between the biopsy groups.

Karyotype

Arm-level and genome copy number (CN) was estimated as described in Taylor et al. 2018 [35]. Briefly, CN of segments determined by PURPLE were rounded to the nearest integer, then the arm coverage per CN was calculated. Arm-level CN was determined to be either the CN with coverage >50%, the most common arm CN if coverage <50%, or the genome CN if coverage <50% and the most common arm CN does match the genome CN. The most common arm CN across all arms was deemed to be the genome ploidy.

Estimated genome ploidy per sample was subtracted from the estimated chromosome arm ploidy to create a matrix of chromosome arm gains and losses relative to the estimated genome ploidy. This matrix was stratified by cohort (metastatic/primary) and a Mann-Whitney test was performed to assess the difference of the distributions for each arm in each cancer type. The resulting p-value was FDR adjusted across all arms per cancer type. A q-value < 0.05 was deemed to be significant. Mean arm gains and losses relative to genome ploidy were calculated and represented.

Genomic instability markers

To compare the differences in aneuploidy scores and the LOH proportions in each group, a Mann-Whitney test was performed per cancer type. The aneuploidy score represents the number of arms per tumor sample that deviate from the estimated genome ploidy as described in Taylor et al. 2018[35]. To compare the LOH proportions, only diploid samples were included when calculating the p-value. The LOH proportion was defined as 1 - (diploid proportion estimated by PURPLE) of the genome.

To compare the fraction of samples with a driver mutation in TP53 as well as the fraction of whole genome duplicated samples per cohort, a Fisher's exact test was performed per cancer type. Any TP53 driver alteration (non-synonymous mutation, biallelic deletion and homozygous disruption) was considered in the analysis. Whole genome duplication was defined as present if the sample had more than 10 autosomes with an estimated chromosome copy number >1.5. A p-value < 0.01 was deemed to be significant for all statistical tests.

Mutational signature analysis

Signature extraction

The number of somatic mutations falling into the 96 single nucleotide substitution (SBS), 78 double base substitutions (DBS) and 83 indel (ID) contexts (as described in the COSMIC catalog[13] https://cancer.sanger.ac.uk/signatures/) was determined using the R package mutSigExtractor (https://github.com/UMCUGenetics/mutSigExtractor, v1.23).

SigProfilerExtractor (v1.1.1) was then used (with default settings) to extract a maximum of 21 SBS, 8 DBS and 10 ID *de-novo* mutational signatures. This was performed separately for each of the 22 tissue types which had at least 30 patients in the entire dataset (aggregating primary and metastatic samples, see Supp. Table 3). Tissue types with less than 30 patients as well as metastatic patients with unknown primary location type were combined into an additional 'Other' group, resulting in a total of 23 tissue type groups for signature extraction. In order to select the optimum rank (i.e. the eventual number of signatures) for each tissue type and mutation type, we manually inspected the average stability and mean sample cosine similarity plots output by SigProfilerExtractor. This resulted in 440 *de-novo* signature profiles extracted across the 23 tissue type groups (Supp. Table 3). Least squares fitting was then performed (using the fitToSignatures() function from mutSigExtractor) to determine the per-sample contributions to each tissue type specific *de-novo* signature.

Etiology assignment

The extracted *de-novo* mutational signatures with high cosine similarity (≥0.85) to any reference COSMIC mutational signatures with known cancer type associations [13] were labeled accordingly (256 *de-novo* signatures matched to 56 COSMIC reference signatures).

For the remaining 184 unlabeled *de novo* signatures, we reasoned that there could be one or more signatures from one cancer type that are highly similar to those found in other tissue types, and that these likely represent the same underlying mutational process. We therefore performed clustering to group likely equivalent signatures. Specifically, the following steps were performed:

1. We calculated the pairwise cosine distance between each of the *de-novo* signature profiles.
2. We performed hierarchical clustering and used the base R function cutree() to group signature profiles over the range of all possible cluster sizes (min no. clusters = 2; max no. of clusters = number of signature profiles for the respective mutation type).
3. We calculated the silhouette score at each cluster size to determine the optimum number of clusters.
4. Finally, we grouped the signature profiles according to the optimum number of clusters. This yielded 27 SBS, 7 DBS, and 8 ID *de-novo* signature clusters (see Supp. Table 3).

For certain *de novo* signature clusters, we could manually assign the potential etiology based on their resemblance to signatures with known etiology described in COSMIC [13], Kucab *et al.* 2019 [67] and Signal [289]. Some clusters were an aggregate of 2 known signatures, such as SBS_denovo_clust_2 which was a combination of SBS2 and SBS13, both linked to APOBEC mutagenesis. Other clusters had characteristic peaks of known signatures, such as DBS_denovo_clust_4 which resembled DBS5 based having distinct CT>AA and CT>AC peaks. Lastly, DBS_denovo_clust_1 was annotated as POLE mutation and MMR deficiency as samples with high contribution (>150 mutations) of this cluster are frequently MSI or have POLE mutations. Likewise, DBS_denovo_clust_2 was annotated with MMR deficiency as the etiology as samples with high contribution (>250 mutations) of this cluster were all MSI. See Supp. Table 3 for a list of all the manually assigned etiologies.

Comparing the prevalence of mutational processes between primary and metastatic cancer

We then compared the activity (i.e. number of mutations contributing to) of each mutational process between primary and metastatic tumors. For each sample, we first summed the contributions of signatures of the same mutation type (i.e. SBS, DBS or ID) with the same etiology, henceforth referred to as 'etiology contribution'. Per cancer type and per etiology, we performed two-sided Mann-Whitney tests to determine whether there was a significant difference in etiology contribution of primary and metastatic tumors. Per cancer type and per mutation type, we used the p.adjust() base R function to

perform multiple testing correction using Holm's method. Next, we added a pseudocount of 1 to the contributions (to avoid dividing by zero) and calculated the median contribution log2 fold change, i.e. log2[(median contribution in metastatic tumors + 1) / (median contribution in primary tumors + 1)]. We considered the etiology contribution between primary and metastatic tumors to be significantly different when the q-value was <0.05, and $\log_2$ fold change was ≥0.4 or ≤ -0.4 (= ±x1.4).

We also determined whether there was an increase in the number of samples with high etiology contribution (i.e., hypermutators) in metastatic versus primary cohorts. For each signature, a sample was considered a hypermutator if the etiology contribution was ≥10,000 for SBS signatures, ≥500 for DBS signatures, or ≥1,000 for ID signatures. For each cancer type, for each etiology, we performed pairwise testing only for cases where there were ≥5 hypermutator samples for either metastatic or primary tumors. Each pairwise test involved calculating p-values using two-sided Fisher's exact tests, and effect sizes by multiplying Cramer's V by the sign of the log2(odds ratio) to calculate a signed Cramer's V value that ranges from -1 to +1 (indicating enrichment in primary or metastatic respectively). We then used the p.adjust() base R function to perform multiple testing correction using Bonferroni's method.

SBS1-age correlations in primary and metastatic tumors

To count the SBS1 mutations we relied on the definition from ref[180] that is based on the characteristic peaks of COSMIC SBS1 signature profile: single base CpG > TpG mutations in NpCpG context. To ensure that these counts and the downstream analyses are not affected by differential APOBEC exposure in primary and metastatic cohorts, we excluded CpG > TpG in TpCpG which is also a characteristic peak in COSMIC SBS2 signature profile. Also, for skin melanoma CpG > TpG in [C/T]pCpG which overlaps with SBS7a was excluded. To obtain the SBS5 and SBS40 counts we relied on their exposures derived from the mutational signature analyses performed in this study (explained above).

To assess the correlation between SBS1 burden and patient's age at biopsy we performed a cancer type and cohort specific linear regression (i.e., separate regression for primary and metastatic tumor samples). To avoid spurious effects caused by hypermutated tumors, samples with TMB greater than 30,000 as well as those with SBS1 burden greater than 5,000 were excluded.

For each cancer type and cohort we then computed 100 independent linear regressions by randomly selecting 75% of the available samples. We selected the median linear regression (based on the regression slope) as representative regression for further analyses. Similarly, confidence intervals were derived from the 1st and 99th percentile of the computed regressions.

To evaluate the significance of the differences between primary and metastatic representative linear regressions (hereafter referred to as linear regression for simplicity) we first filtered out cancer types that failed to show a positive correlation trend between SBS1 burden and age at biopsy in both primary and metastatic tumors (i.e., Pearson's correlation coefficient of primary and metastatic regression >0.1). Next, for each selected cancer type, we computed the regression residuals of primary and metastatic SBS1 mutation counts using, in both cases, the primary linear regression as baseline. The primary and metastatic residual distributions were then compared using a Mann-Whitney test to evaluate significance. Cancer types with a Mann-Whitney p-value < 0.01 were deemed as significant. Finally, to ensure that the differences were uniform across different age ranges (i.e., not driven by a small subset of patients) we only considered significant cancer types where the metastatic linear regression intercept is higher than the primary intercept.

SBS5/SBS40 correlations were computed following the same procedure and using the sum of SBS5 and SBS40 exposures for each tumor sample. If none of the mutations were attributed to SBS5/SBS40

mutational signatures, the aggregated value was set to zero. In the ploidy corrected analyses we divided the SBS1 mutation counts (and SBS5/SBS40 mutation counts for the SBS5/40 ploidy corrected regression, respectively) by the PURPLE estimated tumor genome ploidy.

For each cancer type the mean fold-change (fc) was defined as $\underline{fc} = \frac{1}{40}\sum_{i=40}^{80} \frac{MPredi}{PPredi}$ where $MPred_i$ and $PPred_i$ are the estimated number of SBS1 mutations for a given age i-th according to the metastatic and primary linear regressions, respectively. Similarly, the mean estimated SBS1 burden difference ($SBS1_{diff}$) was defined as: $\underline{SBS1diff} = \frac{1}{40}\sum_{i=40}^{80} MPredi - PPredi$.

Clonality and timing of clock-like mutations

SBS1 individual mutations were identified as described in the previous section. For SBS5 and SBS40 mutations, we used a maximum likelihood approach to assign individual mutations to the SBS5 and SBS40 mutational signatures in a cancer type specific manner.

For every SBS1 (and SBS5/SBS40 mutation) we then assign the clonality according to the PURPLE subclonal likelihood estimation, where only mutations with SUBCL likelihood >= 0.8 were considered as such (see above). Likewise, the molecular timing of individual mutations (i.e., clonal early and clonal late) was computed using the MutationTimeR [180] package.

For each tumor sample the SBS1 clonality ratio (or respectively SBS5/SBS40 clonality ratio) was defined as the ratio between the proportion of clonal SBS1 mutations ($\frac{SBS1\ clonal\ \ mutations}{SBS1\ mutations}$) divided by the total proportion of clonal alterations in the sample ($\frac{Total\ clonal\ mutations}{Total\ mutations}$). Similarly, the SBS1 clonal late ratio was defined as the ratio between the proportion of clonal late SBS1 mutations ($\frac{SBS1\ clonal\ \ late\ mutations}{SBS1\ clonal\ mutations}$) divided by the total proportion of clonal late alterations in the sample ($\frac{Total\ clonal\ late\ mutations}{Total\ clonal\ mutations}$), where the total of clonal SBS1/total mutations was computed as the sum of clonal late, clonal early and unassigned clonal from MutationTimerR.

Primary cell division rate and accelerated SBS1 mutagenesis in metastasis

To assess the relationship of cell division rate of primary tumors with the accelerated SBS1 mutagenesis in the metastatic setting we relied on the SBS1 burden per year as a proxy of stem-cell division rates as was previously described in ref[265]. We thus computed for each primary cancer type the average number of SBS1 per year as the number of SBS1 mutations divided by the patient's age at biopsy (only considering primary samples and excluding hypermutated samples as described above). We then used a Spearman's correlation to assess its association with the estimated mean SBS1 mutation rate fold change in metastatic tumors (see above). Additionally, to exclude potential biases in our primary cohort, we repeated the same analysis relying on an independent measurement of primary cancer SBS1 yearly accumulation. Specifically, we used the best estimated accumulation of SBS1 per year from ref[265] Supplementary Data Set 6 and regressed it to the fold change estimates for the matching cancer types present in both datasets.

Structural variant (SV) analysis

SV type definitions

LINX [270] chains one or more SVs and classifies these SV clusters into various event types ('ResolvedType'). We defined deletions and duplications as clusters with a ResolvedType of 'DEL' or 'DUP' whose start and end breakpoints are on the same chromosome (i.e. intrachromosomal). Deletions and duplications were split into those <10kb and ≥10kb in length (small and large, respectively), based on observing bimodal distributions in these lengths across cancer types (Supp.

Fig. 4a). We defined complex SVs as clusters with a 'COMPLEX' ResolvedType, an inversion ResolvedType (including: INV, FB_INV_PAIR, RECIP_INV, RECIP_INV_DEL_DUP, RECIP_INV_DUPS), or a translocation ResolvedType (including: RECIP_TRANS, RECIP_TRANS_DEL_DUP, RECIP_TRANS_DUPS, UNBAL_TRANS, UNBAL_TRANS_TI). Complex SVs were split into those with <20 and ≥20 SVs (small and large, respectively), based on observing similar unimodal distributions in the number SVs across cancer types whose tail begins at ~20 breakpoints (Supp. Fig. 4a). Lastly, we defined LINEs (long interspersed nuclear element insertions) as clusters with a ResolvedType of 'LINE'. For each sample, we counted the occurrence (i.e. SV burden) of each of the 7 SV types described above. Additionally, we determined the total SV burden by summing counts of the SV types.

Comparing SV burden between primary vs. metastatic cancer

We then compared the SV type burden between primary versus metastatic tumors as shown in Fig. 4a. Firstly, we performed Mann-Whitney tests per SV type and per cancer type to determine whether there was a significant difference in SV type burden between primary versus metastatic, and used $p<0.01$ as the significance threshold. Next, we calculated relative enrichment as follows: $\log_{10}$(median SV type burden in metastatic tumors + 1) – $\log_{10}$(median SV type burden in primary tumors + 1); and calculated fold change as follows: (median SV type burden in metastatic tumors + 1) / (median SV type burden in primary tumors + 1). When calculating relative enrichment and fold change, the pseudocount of 1 was added to avoid the log(0) and divide by zero errors, respectively. Fold changes are displayed with a '>' in Fig. 4a when the SV burden for primary tumors is 0 (i.e. when a divide by zero would occur without the pseudocount).

Identifying features associated with SV burden increase in metastatic cancer

To identify the features that could explain increased SV burden, we correlated SV burden with various tumor genomic features. This included: i) genome ploidy (determined by PURPLE); ii) HRd status (determined by CHORD [38]) and MSI status (determined by PURPLE); iii) the presence of mutations in 345 cancer associated genes (excluding fragile site genes that are often affected by CNVs [6]), henceforth referred to as 'gene status'; and iv) treatment history, including the presence of radiotherapy, the presence of one of the 79 different cancer therapies as well as the total number of treatments received. All primary samples as well as all metastatic samples without treatment information were considered to have no treatment. Genome ploidy and total number of treatments received were numeric features, whereas all of the remaining were boolean (i.e. true/false) features. In total there were 429 features.

SV type burden was transformed to $\log_{10}$(SV type burden + 1) and was correlated with the 429 features using multivariate linear regression models (LM). This was performed separately for each of the 22 cancer types, and for each of the 7 SV types, resulting in a total of 154 (=22 cancer types x 7 SV types) LM models.

Each LM model (i.e. per SV type and cancer type) involved training of three independent LMs with i) both metastatic and primary samples (primary+metastatic), ii) only Hartwig samples (metastatic-only), and iii) only primary samples (primary-only). This was done to filter out correlations between features and increased SV type burden solely due to differences in feature values between primary and metastatic tumors. We then required features that positively correlated with SV type burden in the primary+metastatic LM to independently show the same association in the metastatic-only or primary-only LMs. Only genomic features that independently showed positive correlation with the SV burden were further considered as significant (i.e., represented in the lollipots).

Each of the 3 LMs was trained as follows:

1. Remove boolean features with too few 'true' samples:

a. For the primary+metastatic LM, remove gene status features with <15 'true' samples.
b. For the metastatic-only and primary-only LMs, remove gene status features with <10 'true' samples.
c. For the remaining boolean features, remove features with <5% 'true' samples.

2. Fit a LM using the lm() base R function to correlate $\log_{10}$(SV type burden + 1) versus all features.

For each LM analysis, we used the following filtering criteria to identify the features that were correlated with increased SV type burden:

1. Only keep LM analyses where there was significant increase in SV type burden for the respective cancer type (p<0.01 as described in the previous section "Comparing SV burden between primary vs. metastatic cancer")
2. For primary+metastatic LM
   a. Regression p-value <0.01
   b. Coefficient p-value <0.01
   c. Coefficient >0
3. For metastatic-only LM or primary-only LM:
   a. Coefficient p-value <0.01
   b. Coefficient >0

Finally, to determine which features (of those correlated with increased SV type burden) were enriched in metastatic tumors compared to primary tumors (and vice versa), we calculated Cliff's delta for numeric features and Cramer's V for boolean features. Cliff's delta ranges from -1 to +1, with -1 representing complete enrichment in primary tumors, whereas +1 represents complete enrichment in metastatic tumors. Cramer's V only ranges from 0 to 1 (with 1 representing enrichment in either primary or metastatic tumors), the sign of the log(odds ratio) was assigned as the sign of the Cramer's V value so that it ranges from -1 to +1. Features with an effect size >0 were considered as those that could explain the SV burden increase in metastatic cancer when compared to primary cancer.

Driver alterations

We relied on patient specific cancer driver catalogs constructed by PURPLE [6] and LINX [270]. Only drivers with a driver likelihood > 0.5 were retained. Fusion drivers were filtered for those that were previously reported in the literature. Similarly, we manually curated the list of drivers and removed SMAD3 HOTSPOT mutations because of the high burden mutations in low mappability regions. The final driver catalog contained a total of 443 driver genes.

We then combined fusions with the LINX driver variants to calculate a patient specific number of driver events. Drivers that concern the same driver gene but a different driver type were deemed to be individual drivers (e.g. TP53 mutation and TP53 deletion in the same sample were considered as one driver event). Cancer type specific Mann-Whitney test was performed to assess differences between primary and metastatic tumors. A p-value < 0.01 was deemed to be significant.

A contingency matrix was constructed from the driver catalog, containing the frequency of driver mutations per driver type (i.e., deletion, amplification or mutations) and cancer type in each cohort (metastatic and primary). A second contingency matrix was constructed for the fusions. Partial amplifications were considered as amplifications while homologous disruptions were considered as deletions. These contingency matrices were filtered for genes which show a minimum frequency of 4 mutated samples in either the primary or the metastatic cohorts. Then a Fisher's exact test for each gene, cancer type and mutation type was performed and the p-value was adjusted for FDR per cancer type. Cramer's V and the odds ratio were used as effect size measures. An adjusted p-value < 0.01 was deemed to be significant.

Therapeutic actionability of variants

To determine the amount of actionable variants observed in each sample, we compared our variants annotated by SnpEff [290] to those derived from three different databases (OncoKB [291], CIViC [292] and CGI [293]) that were classified based on a common clinical evidence level (https://civic.readthedocs.io/en/latest/model/evidence/level.html) as described in Priestley et al. 2019 [6]. In our study we only considered A and B levels of evidence which represent variants that have been FDA approved for treatment and are currently being evaluated in a late stage clinical trial, respectively. A variant was determined to be "On-label" when the cancer type it was found in matches the cancer type the treatment was approved for or is being investigated for, and "Off-label" otherwise. Only actionable variants of the sensitive category were considered (i.e. tumors containing the variant are sensitive to a certain treatment). Sample-level actionable variants such as TMB high/low or MSI status were not evaluated, because of their tendency to overshadow the other variants, especially in the Off-label category. Further, wild-type actionable variants were not considered in this analysis for the same reason. Variants related to gene expression or methylation were not considered due to lack of available data. Additionally, we found actionable variants derived from leukemias to be very different from the solid tumors in our data set which is why we excluded them for this analysis. For the analysis of proportion of samples bearing therapeutically actionable variants we considered the highest evidence level was retained for each sample following the order A On/Off-label to B On/Off-label. To assess enrichment of actionable variants globally and on A On-label level in metastatic tumors, a Fisher's exact test was performed pancancer-wide and per cancer type. A p-value < 0.01 was deemed to be significant. Percentage changes in frequency are only shown for significant cases.

To determine which variants contribute the most to the observed significant frequency differences, individual actionable variants were tested for enrichment in metastatic tumors using a Fisher's exact test per cancer type and tier level. P-values were FDR adjusted per cancer type and a q-value < 0.05 was deemed to be significant. In Supp. Fig. 6 only actionable variants that were found at a minimum frequency of 5% in either primary or metastatic cohort and a minimum frequency difference of 5% between them were shown. However, the differences across all screened variants is available as part of Supp. Table 7.

Treatment enriched drivers

We aimed to pinpoint drivers that are potentially responsible for lack of response to certain cancer treatments in the metastatic cohort. Hence, we devised a test that identifies driver alterations that are enriched in groups of patients treated with a particular treatment type compared to the untreated group of patients from the same cancer type (see Fig. 6a for illustration of the workflow).

Treatments were grouped according to their mechanism of action so that multiple drugs with a shared mechanism of action were grouped into the mechanistic treatment category (e.g. Cisplatin, Oxaliplatin, Carboplatin as Platinum). We created 332 treatment and cancer type groups by grouping patients with treatment annotation according to their treatment record before the biopsy. One patient might be involved in multiple groups if they have received multiple lines of therapy or a simultaneous combination of multiple drugs. Only 62 treatment and cancer type groups with at least 10 patients were further considered in the analysis.

Hence, for each cancer and treatment group we performed the following steps:

1. We first performed a driver discovery analysis in treatment and cancer type specific manner. We explored three types of somatic alterations: coding mutations, non-coding mutations and copy number variants (see below for detailed description of each driver category). Driver elements from each alteration category were selected for further analysis.

2. For each driver alteration from 1) we compared the alteration frequency in the treated group to the untreated group of the same cancer type. Each driver category (coding and non coding mutations and copy number variants) were evaluated independently. We performed a Fisher's exact test to assess the significance of the frequency differences. Similarly, we computed the odds ratio of the mutation frequencies for each driver alteration. The p-values were adjusted with a multiple-testing correction using the Benjamini–Hochberg procedure (alpha=0.05). An adjusted p-value of 0.05 was used for coding mutations and copy number variants. Adjusted p-value of 0.1 was used for non-coding variants due to the overall low mutation frequency of the elements included in this category, which hampered the identification of significant differences.
3. We then annotated each driver element with information about the exclusivity in the treatment group. We labeled drivers as treatment exclusive if the mutation frequency in the untreated group was lower than 5%, we annotated as treatment enriched otherwise. Additionally, we manually curated the identified drivers with literature references of their association with each treatment category.
4. Finally, the overlap of patients in multiple treatment groups (see above) in the same cancer type prompted us to prioritize the most significant treatment association for each driver gene in a particular cancer type. In other words, for each driver gene that was deemed as significantly associated with multiple treatment groups in the same cancer type, we selected the most significant treatment association, unless a driver-treatment annotation was clearly reported in the literature.

The full catalog of TEDs and their mutation frequencies can be found in Supp. Table 8.

Coding mutations drivers

We used dNdScv[49] with default parameters to identify cancer driver genes from coding mutations. A global q-value <0.1 was used as a threshold for significance. Mutation frequencies for each driver gene were extracted from the dndscv output. We defined the mutation frequency as the number of samples bearing non-synonymous mutations.

Non-coding mutations drivers

We used ActiveDriverWGS[294] (v1.1.2, default parameters) to identify non-coding driver elements in five regulatory regions of the genome including 3'UTRs, 5'UTRs, lncRNAs, proximal promoters and splice sites. For each element category we extracted the genomic coordinates from Ensembl v101. Each regulatory region was independently tested. To select for significant hits, we filtered on adjusted p.values (FDR<0.1) and minimum of three mutated samples. We defined the mutation frequency as the number of mutated samples for each significantly mutated element in the treatment group.

Copy number variant drivers

We ran GISTIC2 [60] (v2.0.23) on each of the 62 treatment and cancer type groups using the following settings:
gistic2 -b <inputPath>-seg <inputSegmentation>-refgene hg19.UCSC.add_miR.140312.refgene.mat -genegistic 1 -gcm extreme -maxseg 4000 -broad 1 -brlen 0.98 -conf 0.95 -rx 0 -cap 3 -saveseg 0 -armpeel 1 -smallmem 0 -res 0.01 -ta 0.1 -td 0.1 -savedata 0 -savegene 1 -qvt 0.1.

The focal GISTIC peaks ($q \leq 0.1$ and < 1Mbp) were then annotated with functional elements using the coordinates from Ensembl v101. The frequency differences between treated and untreated cohorts on every gene was assessed with Fisher exact test as described above. For this, we first calculated the focal amplification and deep depletion status of every gene within each sample. A gene was amplified when the ploidy level of the gene was 2.5 ploidy levels higher than its genome-wide mean ploidy level

(as measured by PURPLE), and deleted when the gene ploidy level was lower than 0.3 (i.e. deep deletion). We observed that the majority of the peaks contained multiple significant gene candidates (after multiple correction $q < 0.05$) and therefore we retained the gene most closely positioned to the peak summit, which is the most significantly enriched region across the treated samples. Next, we also found recurrent peaks across multiple treatment groups per cancer type that are not, or less, present in the untreated control group because most of the Hartwig samples have received multiple treatment types. We therefore merged peaks with overlapping ranges to produce a single peak per genomic region per cancer type. For each collapsed peak we selected the treatment type showing the lowest q-value for the gene near the peak summit. Deletion and amplification peaks were processed separately.

Group level aggregation of treatment resistance associated variants

To estimate the contribution of TEDs to the total number of drivers per sample in the metastatic cohort, we excluded any TED from the catalog of driver mutations (see above Driver alterations section) in a cancer type, gene and driver type specific manner.

Data availability

Metastatic WGS data and metadata from the Hartwig Medical Foundation are freely available for academic use through standardized procedures. Request forms can be found at https://www.hartwigmedicalfoundation.nl.

Somatic variant calls, gene driver lists, copy number profiles and other core data of the PCAWG cohort generated by the Hartwig analytical pipeline are available for download at https://dcc.icgc.org/releases/PCAWG/Hartwig. Researchers will need to apply to the ICGC data access compliance office (https://daco.icgc-argo.org) for the ICGC portion of the dataset. Authentication of NIH eRA commons is required to access the TCGA portion of the dataset via https://icgc.bionimbus.org. Additional information on accessing the data, including raw read files, can be found at https://docs.icgc.org/pcawg/data/.

Code availability

The Hartwig analytical processing pipeline is available at (https://github.com/hartwigmedical/pipeline5) and implemented in Platinum (https://github.com/hartwigmedical/platinum).

The source code to reproduce the analysis of the manuscript will be made public in this repository https://github.com/UMCUGenetics/PCAWG_Hartwig_comparison upon manuscript acceptance in a peer-review journal.

Acknowledgements

This publication and the underlying study have been made possible partly on the basis of the data that Hartwig Medical Foundation and the Center of Personalised Cancer Treatment (CPCT) have made available to the study.

We thank Roel Janssen for the technical assistance in the collection and processing of the PCAWG raw sequencing data using the Hartwig tumor analytical pipeline. We thank Joaquin Mateo for his valuable scientific input. We also thank Lincoln Stein and Linda Xiang for their assistance in the publication of the re-processed ICGC part of the PCAWG dataset. Similarly, we thank Robert Grossman, Christopher Meyer and Trevar Simmons for their assistance in the publication of the re-

processed TCGA part of the PCAWG. We also like to thank Paul Wolfe and other staff of the Hartwig Medical Foundation team for aligning the processing of PCAWG and Hartwig dataset.

Author contributions

Conceptualization, FMJ, AvH and EC. Methodology, FMJ, AM, SB, LN, PP, AvH. Software, FMJ, AM, SB, LN, PP, AvH. Validation, FMJ, AM, SB, LN, PP, AvH. Formal Analysis, FMJ, AM, SB, LN, PP, AvH and EC. Investigation Resources, FMJ, AM, SB, LN, PP, AvH . Data Curation, FMJ, AM, SB, LN, PP, AvH. Writing Original Draft, FMJ and AvH. Writing - Review & Editing, FMJ, AM, SB, LN, PP, AvH and EC. Visualization, FMJ, AM, SB, LN, AvH and EC. Supervision, FMJ, AvH and EC. Project Administration, AvH and EC. Funding Acquisition EC.

Conflicts of interest

The authors do not declare any conflicts of interest.

Supplementary files

Supplementary Table 1. Primary and metastatic cohorts metadata.

Supplementary Table 2. Karyotype and genomic instability measurements.

Supplementary Table 3. Mutational signatures.

Supplementary Table 4. SBS1 mutation rate.

Supplementary Table 5. Structural variants.

Supplementary Table 6. Driver alterations.

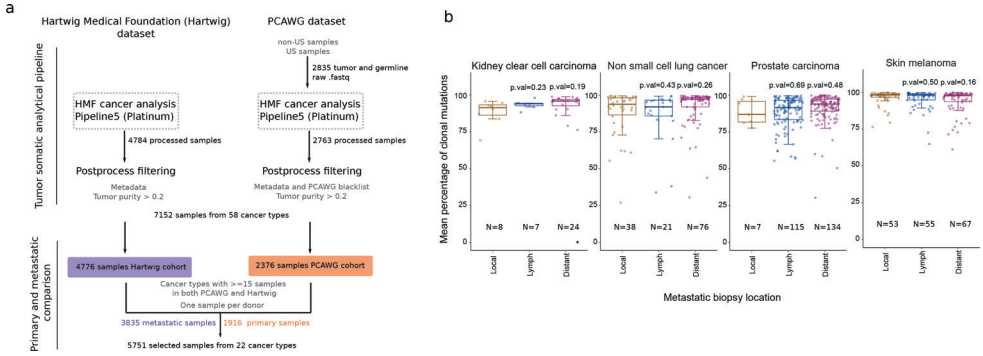
Supplementary Table 7. Therapeutic actionability of variants.

Supplementary Table 8. Treatment associated drivers (TEDs).

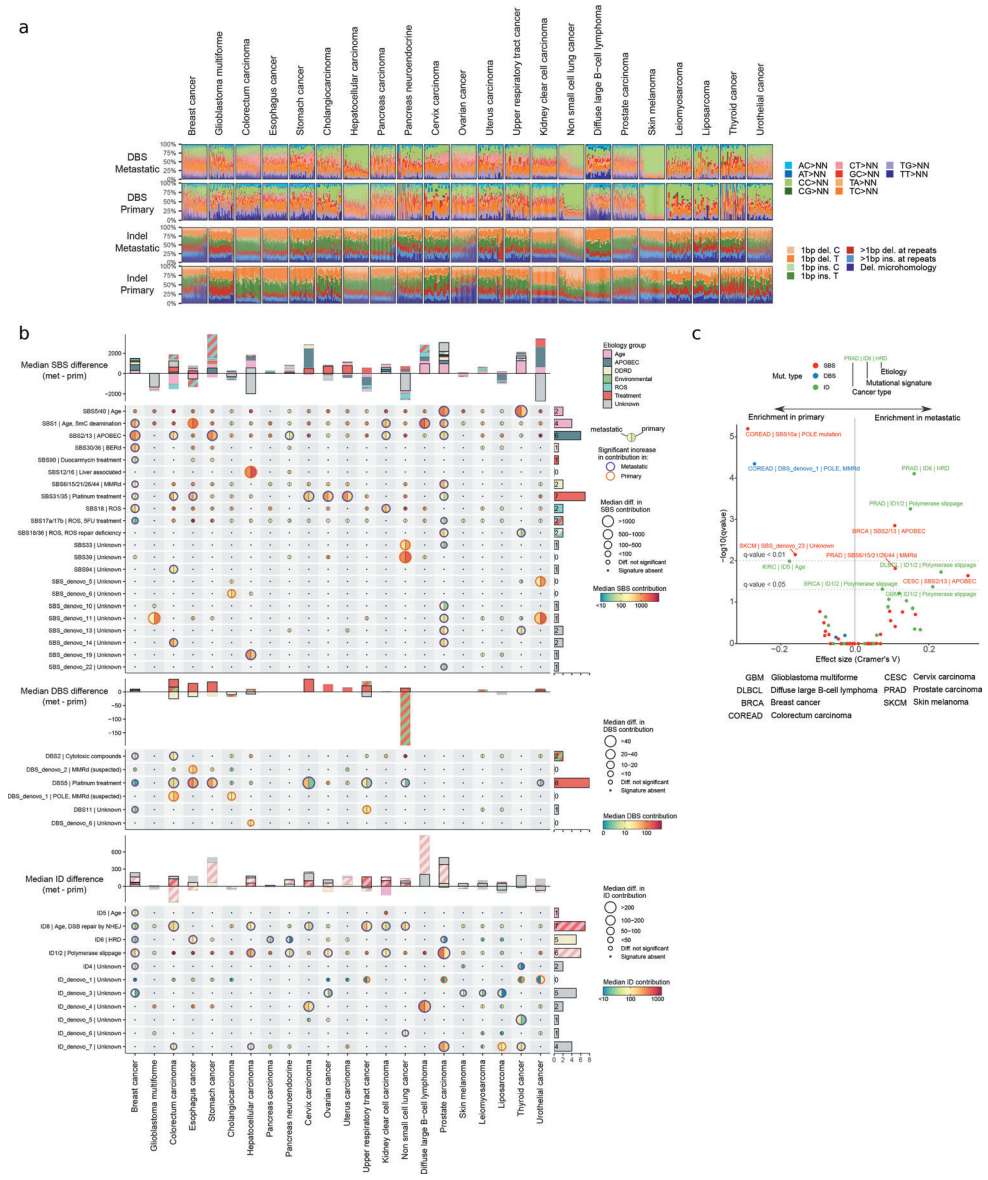
Supplementary Note 1. Application of Hartwig analytical pipeline to the PCAWG dataset.

Supplementary files are available online at:
https://www.biorxiv.org/content/10.1101/2022.06.17.496528v1.supplementary-material,
or by scanning the QR code below:

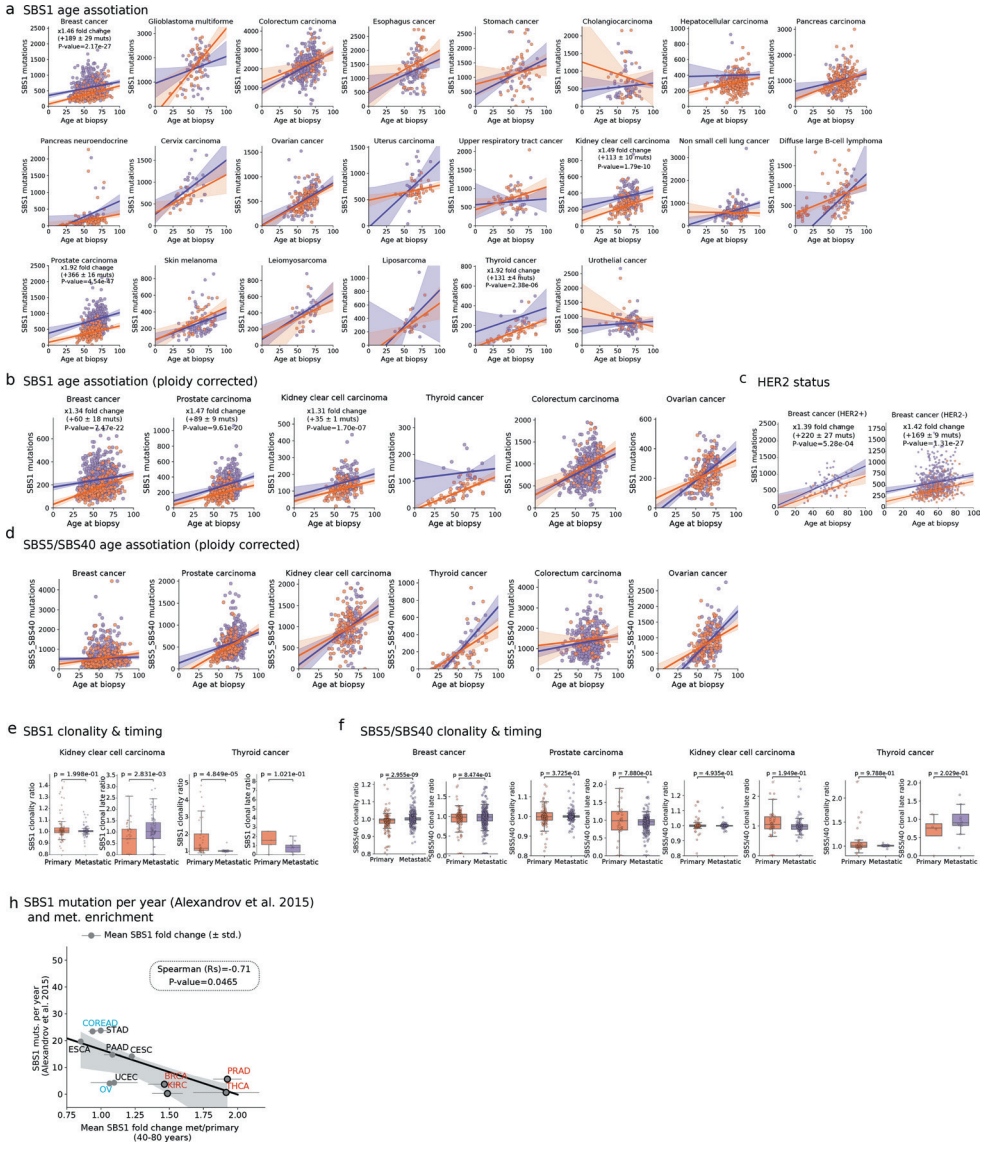
Supplementary figures



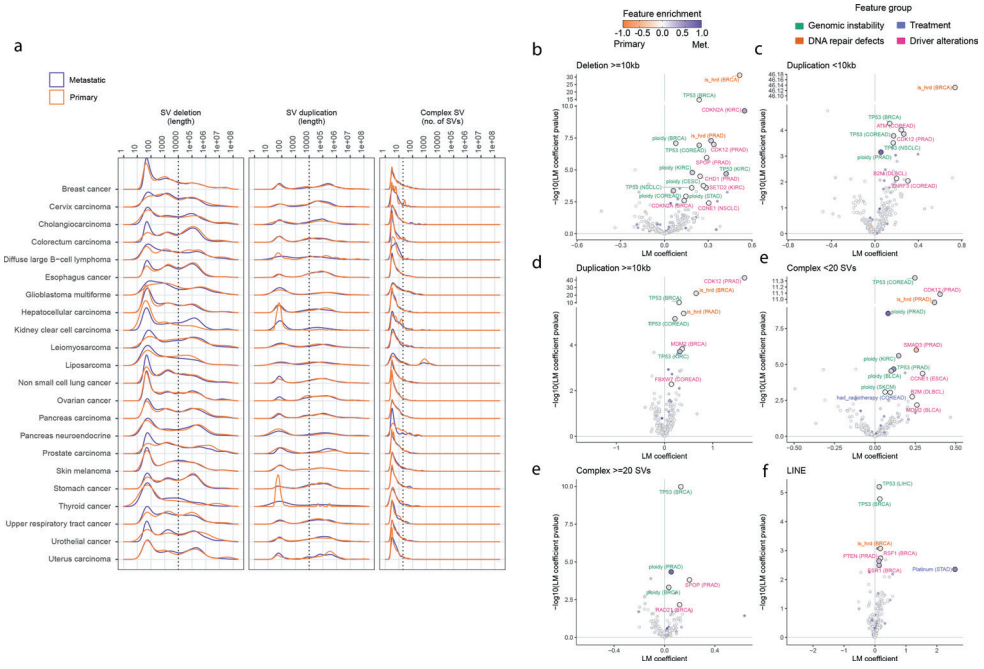
Supplementary Figure 1. Cohort overview and global genomic features. a) Workflow of the unified processing pipeline used in this study for Hartwig (left) and PCAWG (right) WGS samples. First, PCAWG tumor and matched normal raw sequencing files were gathered and re-processed using the Hartwig tumor analytical pipeline. Next, the output of tumor samples that were correctly processed by the pipeline were further subjected to a strict quality control filtering. As a result, a total of 7,152 samples from 58 cancer types compose the harmonized dataset. 5,751 patient tumor samples from 22 cancer types with sufficient representation in both primary and metastatic datasets were selected for this study. **b)** Tumor clonality according to the metastatic biopsy location in kidney clear cell carcinoma, non-small cell lung cancer, prostate carcinomas and skin melanoma. N, number of samples in the group. p, Mann-Whitney p-value if p-value < 0.05. ns, not significant. Box-plots: center line, median; box limits, first and third quartiles; whiskers, lowest/highest data points at first quartile minus/plus 1.5× IQR.



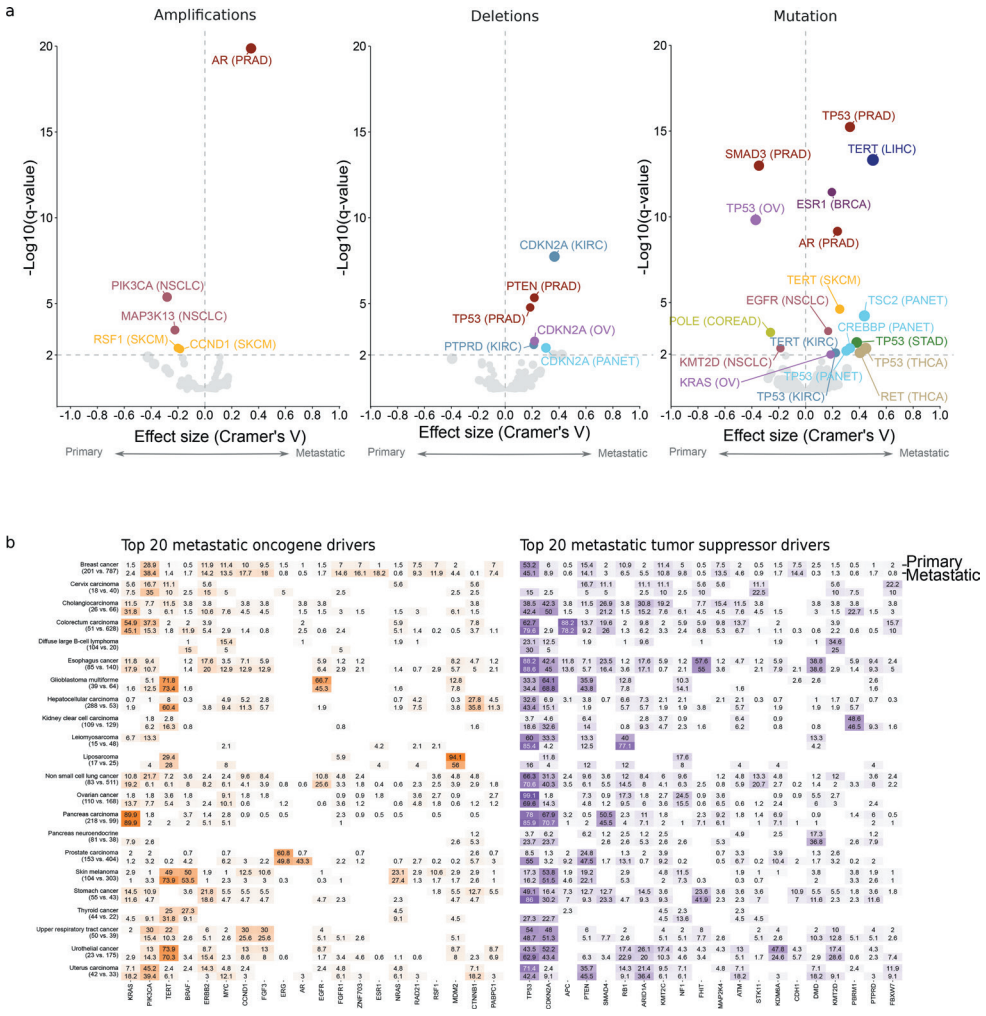
Supplementary Figure 2. Mutation burden and mutational signatures. a) DBS (top) and ID (bottom) mutational spectra of metastatic and primary tumors. Patients are ordered according to their TMB burden. **b)** From top to bottom, moon plot representing the SBS, DBS and ID burden differences attributed to each mutational signature. Top stacked bars represent the cumulative signature exposure difference, including mutational signatures enriched in primary tumors (negative values). Thicker bar edge lines represent significance. Bars are coloured according to the annotated etiology. All mutational signatures, independent of their annotated etiology, are included. Diff., difference. Muts. mutations. Sig., mutational signature. Mut. mutational. Susp., suspected. **c)** Volcano plot representing the mutational signature hypermutation (>10,000 mutations for SBS, >500 for DBS, and >1000 for ID) prevalence comparison between primary and metastatic patients. Y-axis, $\log_{10}$(adjusted p-value). X-axis, effect size as Cramer's V. Each dot represents a mutational signature in a cancer type. Dots are coloured according to the mutation type.



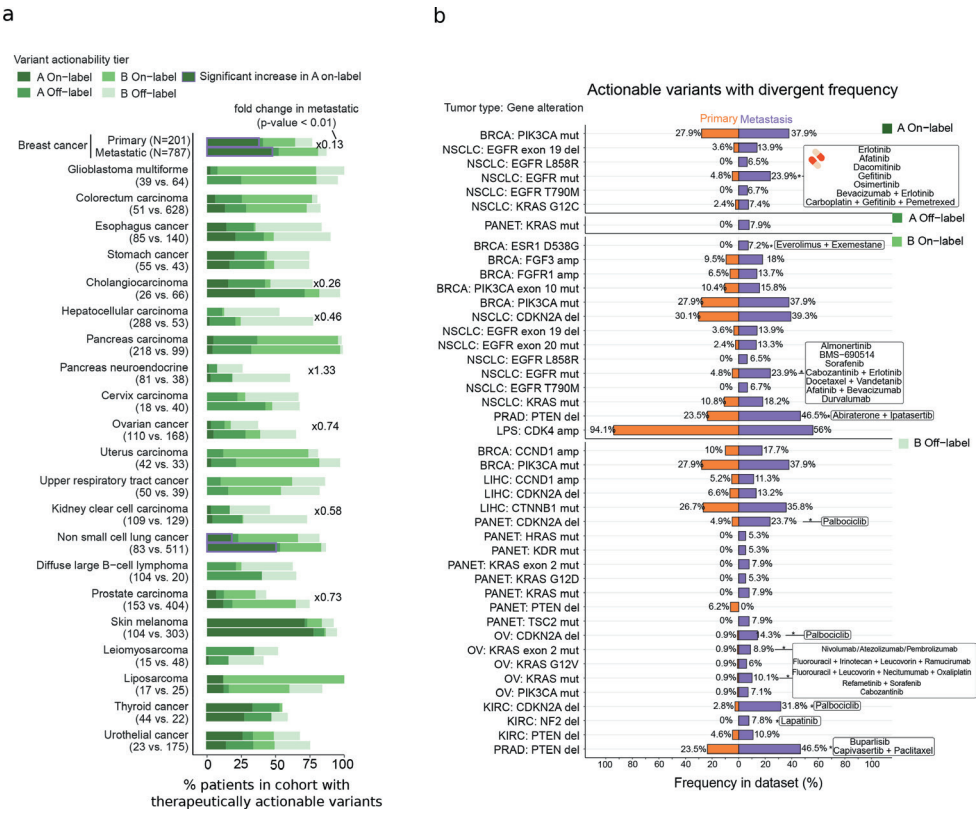
Supplementary Figure 3. SBS1 mutation rates in primary and metastatic tumors. a) Linear regression of the SBS1 mutation burden (y-axis) and patient's age at biopsy (x-axis) in primary and metastatic cancer across the 22 cancer types. The mean fold change, mean SBS1 increase per year and p-value are only represented in cancer types with an age-independent significantly different primary and metastatic distribution. **b)** Relative to a) for ploidy corrected SBS1 in the tumor types of interest. **c)** Independent linear regressions for HER2+ (left) and HER2 (right) breast cancer samples. ERBB2 amplification status was used to annotate cancer subtypes. **d)** Relative to a) for ploidy corrected SBS5/40 counts in the tumor types of interest. **e)** Equivalent to Fig. 3d for primary and metastatic in kidney clear carcinoma (left) and thyroid cancers (right). **f)** Relative to Fig. 3d, but using ploidy corrected SBS5/40 clonality ratio and clonal late ratio in the cancer types of interest. Boxplots are defined as in Fig. 1. **h)** Relative to Fig. 3f but using SBS1 year mutation rate from ref [265]. Muts, mutations.



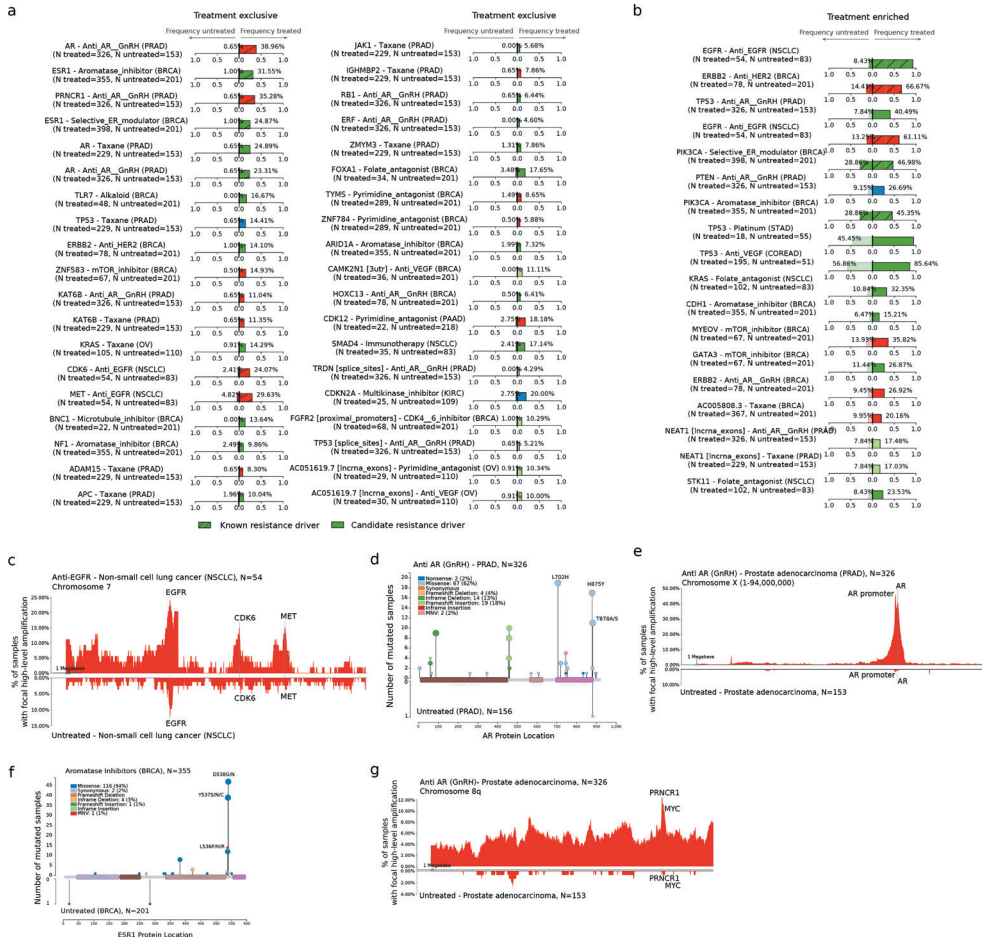
Supplementary Figure 4. Structural variant burden. a) SV length frequency distribution of deletions (left panel) and duplications (middle panel). Right panel shows the frequency distribution of the number of linked breakpoints for complex SVs. Dashed vertical lines represent the chosen threshold to separate between short and large deletions, duplications and complex SVs, respectively. **b)** Volcano plot representing the cancer type specific regression coefficients (x-axis) and significance (y-axis) of clinical and genomic features against the number of large deletions. Each dot represents one feature in one cancer type. Labels are coloured according to the feature category. Dots are coloured by the frequency enrichment in metastatic (purple) or primary (orange) patients. LM, linear model. Coef, coefficient. Similar panels are displayed for **c)** short duplications, **d)** large duplications, **e)** short complex SVs, large complex SVs and **f)** LINEs.



Supplementary Figure 5. Driver landscape and drivers per patient. a) Volcano plots representing the cancer type specific enrichment (x-axis) and significance (y-axis) of driver genes between primary and metastatic cohorts. From left to right, amplification drivers, biallelically deleted drivers and mutated driver genes. BRCA, Breast cancer. CESC, Cervix carcinoma. CHOL, Cholangiocarcinoma. COREAD, Colorectal carcinoma. ESCA, Esophagus cancer. GBM, Glioblastoma multiforme. KIRC, Kidney clear cell carcinoma. LMS, Leiomyosarcoma. NSCLC, Non small cell lung cancer. OV, Ovarian cancer. PAAD, Pancreas carcinoma. PRAD, Prostate carcinoma. SKCM, Skin melanoma. STAD, Stomach cancer. THCA, Thyroid cancer. HNSC, Upper respiratory tract cancer. BLCA, Urothelial cancer. UCEC, Uterus carcinoma. **b)** Comparison fraction of mutated patients for the top 20 most frequently mutated (including all types of alterations) oncogenes (left) and tumor suppressor genes (right) in the metastatic cohort. Top numbers represent primary frequency, bottom metastatic.



Supplementary Figure 6. Therapeutic actionability of variants. a) Cancer type specific fraction of primary (top) and metastatic (bottom) patients with reported therapeutically actionable variants. For each patient the variant with the greatest level of evidence was considered. Bars are coloured according to the variant actionability tiers. Fold change labels are displayed in cancer types with a significant proportional increase (Fisher's exact test p-value <0.01). Purple edgelines highlight significant increase in metastatic A-on label fraction patients. **b)** Primary (left) and metastatic (right) alteration frequency of actionable variants with a high discrepancy (>5% frequency difference) from cancer types with a global significant increase of actionable variants in metastatic patients from panel a). Asterisk, Fisher's exact test p-value <0.01. Text boxes include the associated treatments for alterations with a significant mutation frequency increase in metastatic patients.



Supplementary Figure 7. Treatment enriched drivers. a) Side by side alteration frequency comparison between treated (right bar) and untreated (left bar) patients for all treatment-exclusive and **b)** treatment enriched TEDs. **c)** Distribution of highly-focal copy number gains in chromosome 7 in non-small cell lung cancer untreated patients (bottom) and treated with anti-EGFR (top). *EGFR*, *MET* and *CDK6* genomic locations are highlighted. Each bin represents 100Kbs. **d)** Distribution of mutations along the *AR* protein sequence in prostate cancer patients treated with androgen deprivation (top) and untreated (bottom). Pfam domains are represented as rectangles. Mutations are coloured according to the consequence type. **e)** Distribution of highly-focal copy number gains in chromosome X in prostate untreated patients (bottom) and treated with androgen deprivation(top). *AR* coding region and the promoter region are highlighted. Each bin represents 100Kbs. **f)** Distribution of mutations along the *ESR1* protein sequence in breast cancer patients treated with aromatase inhibitors (top) and untreated (bottom). **g)** Similar to e) but representing MYC and PRNCR1 co-amplifications in chromosome 8q.

Chapter 6

General discussion

Introduction

Cancer is a result of the sequential accumulation of mutations. Some of these are driver mutations which provide the tumor cell a growth advantage over cells that surround it, while the remainder are passenger mutations and have no impact on tumor cell fitness. The pool of mutations that these cells carry is highly dependent on the cancer type and subtype. In this thesis, we have leveraged large whole-genome sequencing (WGS) datasets to characterize these mutational differences, allowing us to build classifiers for patient stratification and guide treatment decisions (**Chapter 2** and **Chapter 3**), and study how different cancers develop (**Chapter 4** and **Chapter 5**). Here, we will discuss the work in this thesis in a broader context, address its potential limitations, and provide future directions.

Gaps and biases in pan-cancer WGS datasets

In this thesis we have made extensive use of the Hartwig Medical Foundation (Hartwig) [6] and Pan-cancer Analysis of Whole Genomes (PCAWG) [5] pan-cancer datasets. It is important to be aware however that these datasets originate from developed countries where patients are typically of European descent, with the Hartwig cohort containing patient data from the Netherlands [6], and the PCAWG cohort containing patient data from the Global North countries such as the US, UK, Australia, Japan, Germany and France [5]. This is a prevailing bias with sequencing data in general and not unique to the Hartwig/PCAWG datasets [295,296]. Findings based on these two datasets thus may not necessarily apply to patients from underrepresented ethnicities. However, this is an issue that will only be resolved with time as WGS becomes cheaper and more common in less developed countries.

While the Hartwig/PCAWG datasets are among the largest publicly available pan-cancer WGS datasets (with >4700 metastatic and >2800 primary tumor samples respectively), not all cancer types were represented equally. Cancer types highly common in the general population such as breast, prostate, and skin cancer naturally had hundreds of samples in both Hartwig/PCAWG cohorts, whereas rare cancer types such as gallbladder, thymic, and testicular cancer had <10 samples [5,6,297,298]. The low sample sizes of rare cancer types prevented tumor tissue of origin classification by Cancer of Unknown Primary Location Resolver (CUPLR) in **Chapter 3**. Insufficient sample sizes meant that population statistics, such as the frequency homologous recombination deficiency (HRD) in **Chapter 2**, were less reliable or unavailable for certain cancer types. Many cancer types were also present in one dataset but not the other, preventing comparisons in HRD frequency between primary versus metastatic cancer. This issue also prevented certain cancer types from being included in our broader comparison of primary versus metastatic cancer in **Chapter 5**. This was especially unfortunate for subtypes of more common cancer types that had sufficient sample size (≥15 samples) in one dataset but not the other, such as small cell lung cancer, kidney chromophobe and papillary carcinomas, and acute myeloid and chronic lymphocytic leukemias [5,6,297,298].

A solution to increasing sample sizes for rarer cancer types would be to include existing tumor WGS cohorts, such as the Genomics England (GEL) cohort consisting of >12000 tumors from 19 cancer types [299,300]. Notably, the GEL cohort contains >1400 bone/soft-tissue tumors which are in general underrepresented in the Hartwig/PCAWG cohorts (~350 tumors combined). The GEL cohort also contains >1000 lung and >1300 kidney tumors, which could greatly increase the sample size for lung and kidney cancer subtypes underrepresented in the Hartwig/PCAWG cohorts, including small cell lung cancer, and kidney papillary and chromophobe carcinomas (each subtype with around 50 tumors or less in Hartwig/PCAWG). Other datasets focused on specific cancer types could also be used to increase sample sizes for the rarer cancer types in Hartwig/PCAWG cohorts, such as the Children's Oncology Group (COG) cohort with 1699 pediatric leukaemias and solid tumors [301], and the CoMMpass cohort with 765 multiple myeloma samples [302]. Ideally, the WGS data from these cohorts will be processed with the same pipeline for variant calling and functional annotation of events (i.e. the

Hartwig pipeline) to eliminate pipeline biases, as was done to harmonize the PCAWG dataset with the Hartwig dataset in **Chapter 5**. This however does come with bureaucratic and technical hurdles. Firstly, access to the (privacy sensitive) raw sequencing data will be required, which usually involves a lengthy data request (as with the COG [303] and CoMMpass [304] cohorts), though access to the GEL cohort is even more complex, requiring membership to the Genomics England Clinical Interpretation Partnership (GECIP) and data analysis within a restricted environment [305]. Then, the pipeline will need to be optimized and validated to account for differences in sequence approaches (e.g. read depths or platform) before being applied to all samples in each cohort. Despite the time and money investment required, the inclusion of other tumor WGS cohorts is a feasible short term solution to generate a more comprehensive pan-cancer dataset.

Challenges in structural variant detection

Paired end short read sequencing (typically Illumina sequencing) is currently the standard technique for WGS of tumors, and were used for generating the Hartwig and PCAWG dataset. With short read sequencing, reads are eventually mapped to a reference genome and pieced together to form the continuous genomic sequence of the sample [4]. However, the <300bp long reads typically produced by Illumina sequencing are too short to span long repetitive regions of the genome (e.g. telomeres, centromeres, and stretches of segmental duplications), resulting in ambiguities during mapping, and complicating the detection of structural variants (SV) (but also small variants) in these regions [306]. Short reads also cannot span the majority of SVs [307], meaning the presence of SVs must be solely inferred using for example read depth, paired end read discordance, and split read information [308]. While this information can be used to detect simple SVs (i.e. deletions, duplications, inversions, insertions and translocations), the presence of complex SVs must be further inferred by clustering simple SV breakpoints which are statistically unlikely to have occurred independently as a single event [127,178].

Further complications arise when trying to identify large complex SV events such as chromothripsis, as definitions for such events require assumptions and arbitrary thresholds to be made. PCAWG defined chromothripsis as an interleaved SV cluster with ≥7 segments oscillating between 2 copy number (CN) states [145]. Other studies only required 10 segmental CN oscillations within 50 Mb [309,310] or within a single chromosome [311,312]. However, since short reads cannot span the derivative chromosome resulting from chromothripsis, we cannot be certain whether segmental CN oscillations originate from: i) a single event (i.e. truly chromothripsis), ii) template-switching DNA-replication errors which can generate a similar pattern [313], iii) multiple distinct events, or iv) a combination of the above. For similar reasons, detection of breakage-fusion-bridge (BFB) is problematic, which is characterized by multiple inverted duplications (also known as foldback inversions) and a 'staircase' amplification CN profile [30]. In the primary versus metastatic cancer comparison in **Chapter 5**, we therefore opted to quantify complex SVs stratified by number of segments rather than directly quantify chromothripsis or BFB given the difficulty in accurately identifying these events.

Long read sequencing technologies, such as those from Oxford Nanopore Technologies and Pacific Biosciences (PacBio), could potentially allow for better detection of SVs [306]. However, because long read sequencing has still lower nucleotide level accuracy compared to short read sequencing [314], the combination of short and long read sequencing would currently be required to properly identify large SVs while retaining the ability to call small mutations and SV breakpoints at nucleotide resolution. In general, Nanopore sequencing offers longer reads compared to PacBio sequencing [306], with mean read lengths of ~10kb using a MinION sequencer [315–317]. While Nanopore read lengths still cannot span all SVs nor all repetitive regions of the genome, the longer read length reduces ambiguity during SV calling, especially for smaller complex SVs, such as extrachromosomal circular DNA

(ecDNA) which are usually <1Mb in size [318], or inverted duplications where the duplication and inversion segments are usually <1Mb and <10Mb respectively [178]. However, to be able to reap the benefits of improved SV detection from long reads, tools for calling SVs from long reads will still need to mature [319]. Nevertheless, the identification of large complex SV events will remain a challenge unless it is eventually possible to sequence a chromosome using only a few ultra-long reads (or even a single read).

Limitations of mutational signature analysis

Mutational signatures describe the combinations of mutation types that result from specific mutational processes. They have been widely used in recent years for studying cancer [6,13,67,74,75,244,320], and have also been crucial to our analyses in **Chapter 3**, **Chapter 4** and **Chapter 5**. *De novo* signature extraction and signature fitting are the basis of mutational signature analysis. *De novo* signature extraction aims to identify the underlying patterns of somatic mutations in a cohort of samples, and was used to build the COSMIC catalog of reference signatures [64]. It is typically based around non-negative matrix factorization (NMF), using frameworks such as SigProfilerExtractor [321]. Signature fitting (typically using the least-squares method) can be used to quantify the presence of signatures from an existing set of signatures (e.g. from the COSMIC catalog), and can be used when a sample cohort has too few samples for a robust *de novo* signature extraction [107,321].

De novo signature extraction for a pan-cancer cohort is best performed separately per tissue type. This is because performing NMF on a full pan-cancer cohort (with tissue type sample size imbalances) would predominantly yield signatures present in tissue types with the most samples. Nevertheless, performing NMF on tissue types with more samples still results in more sensitive signature extraction, and enables the detection of less prevalent signatures. For these reasons, mutational signature landscapes described by the PCAWG consortium [13] as well as in **Chapter 5** are more accurate and complete for larger tissue type cohorts, and less so for smaller ones. By extension, the COSMIC catalog is thus also biased towards signatures from larger tissue type cohorts that are more easily extracted. This bias should thus be taken into consideration for analyses based on signature fitting on the COSMIC catalog (or other existing catalogs).

It is also important to be aware that NMF is ultimately a dimension reduction algorithm [322]. For instance, NMF could be used to reduce the 96 trinucleotide context mutation counts from a sample cohort into 5 underlying patterns of these contexts (i.e. mutational signatures) that are present in these samples. A major challenge with NMF is to choose the number of signatures to extract (also known as NMF rank), as this parameter is not known upfront. Typically, NMF is performed on a range of ranks, and a rank is chosen based on the optimum of one or more performance metrics (e.g. signature stability) [321]. In practice manual assessment is also required for rank selection. In **Chapter 5**, we used SigProfilerExtractor which automatically suggests an optimum rank, but we often had to choose either, i) a higher rank so that we would extract expected *de novo* signatures matching to known COSMIC signatures (e.g. SBS1, an age associated signature expected in all samples), or ii) lower ranks to prevent extraction of multiple similar signatures that should be in fact one signature.

The issue of extracting multiple similar signatures due to choosing too high of a rank ('over-extraction') may have led to the split of several of the original 30 COSMIC signatures into 'new' signatures despite these not being confirmed to represent distinct mutational processes [13], such as with SBS5 into SBS5 and SBS40 (associated with age), and SBS7 into SBS7a and SBS7b (UV light exposure). The ongoing search for more novel SBS signatures also increases the risk of over-extraction. This becomes more problematic as the number of SBS signatures approaches 96, the point at which these signatures are no longer a dimension reduction of the 96 trinucleotide contexts. Recently, Degasperi *et al.* [299] used >12000 tumors from the GEL pan-cancer cohort to extract 82 SBS signatures, which have been

reported in the SIGNAL catalog [323]. However, when examining the 40 signatures that were novel, some were identical to existing COSMIC signatures, such as SBS95/SBS96 to SBS1 (linked to 5-methylcytosine deamination), and SBS98 to SBS15 (mismatch repair deficiency). Furthermore, 9 novel signatures were only present in 1 sample in the entire cohort, and had no associated etiology. This suggests that some of these signatures may have no biological significance and are a technical artifact due to over-extraction.

We may thus be reaching the limit of new SBS signatures that can be extracted from the 96 trinucleotide contexts. Future signature discovery efforts should therefore focus on double base substitution (DBS) and indel (ID) signatures, as these have been less extensively researched. The Degasperi *et al.* study [299] has already reported a set of novel DBS signatures but these still require further investigation to determine their etiologies. Nevertheless, the 1536 possible pentanucleotide contexts (base substitution with 2 flanking 5' and 3' bases) could be used instead of the 96 trinucleotides as a basis for extraction of extended SBS signatures [324,325]. Aside from the flanking bases, somatic point mutations could be stratified by at the mutation locus, such as replication timing [179] or position on the DNA groove [326]. However, having too many mutation contexts leads to too few mutations being attributed to each context (which is especially problematic for the pentanucleotide contexts), potentially preventing robust extraction of *de novo* signatures. It is therefore important to use features that can meaningfully stratify mutations, such as replication timing, chromatin accessibility, or degree of mutation clustering [179].

Lastly, regional mutational density (RMD; the mutation frequency per 1Mb bin across the genome) has only thus far been directly used for tumor tissue of origin classification [90,91], but the potential for RMD signatures remains largely unexplored. In **Chapter 3**, we have used NMF reduce >3000 RMD bins to <50 rudimentary RMD signatures. A more robust extraction and analysis of RMD signatures could yield novel biological insights. RMD could be for example used to study the activity of different DNA repair pathways across the genome, as increased mutation frequency in certain parts of the genome is (partially) thought to be due to reduced accessibility to DNA repair complexes [144]. RMD may also have the potential for machine learning based classification of DNA repair deficiencies other than HRD and MMRD (for these, models already exist [38,74,75]), such as for base excision repair (BER) or nucleotide excision repair (NER) deficiency.

Contributing factors to cancer development beyond driver mutations

The relationship between increased driver mutation burden and increased cancer risk is well established [327–329]. However, the presence of driver mutations is likely not enough to initiate cancer, as these mutations are also found in normal cells [330]. A study by Martincorena *et al.* [331] showed that cells in normal skin carried more mutations in *NOTCH1* than in skin cancers, and these normal cells also carried mutations in other driver genes such as *TP53* and *FAT1*. The presence of driver mutations in normal cells has also been shown in numerous other tissues [332–335]. As with cancer initiation, the presence of driver mutations alone may also not be sufficient for the progression from primary to metastatic cancer. In **Chapter 5**, we find that metastatic tumors showed increased burden of passenger small mutations and SVs, mainly as a result of treatment exposure and cancer type specific endogenous mutational processes. However the driver mutation landscape of metastatic tumors remained relatively unchanged when compared to primary tumors, suggesting that factors other than driver mutations contribute to metastatic progression. An explanation is that the distinction between driver and passenger mutations is merely a matter of definition. Whether a mutation is a driver can depend on the computational method used for driver detection. Methods based on identifying mutations occurring more significantly than expected by chance (e.g. dNdScv [49], used in **Chapter 5**) may yield different driver mutations than a method based on detecting linear clustering of mutations

(e.g. OncodriveCLUSTL [52]). Additionally, a computational study [336] showed that passenger mutations can have weak driver potential, with another study [337] showing that a mutation can be a driver or passenger depending on the context, such as initial cell properties or current selection pressures. Mutations thus likely exist on a spectrum of deleteriousness rather than the dichotomy of driver and passenger mutations.

Aside from mutations, altered epigenetic regulation of gene expression likely also drives cancer development, with several studies on breast [338], bladder [339], and gastric [340] cancer patients showing that cancer tissue and the adjacent normal tissue share similar patterns of methylation. Altered methylation can be a result of environmental pressures such as exposure to carcinogens [341]. For example, aberrant methylation of lung cancer associated genes was found in the upper respiratory tract of smokers [342,343], while high alcohol intake (together with low dietary folate intake) was linked with methylation of colorectal cancer associated genes [344]. Besides carcinogen exposure, inflammation has also been shown to lead to altered methylation. Chronic inflammation was shown in a mouse colon cancer model to lead to hypermethylation of several genes important in gastrointestinal homeostasis and repair, some of which are also hypermethylated in human colorectal cancer and mouse intestinal adenomas [345]. Likewise, Ammerpohl *et al.* showed that liver cirrhosis (which commonly precedes liver cancer [226]) shared a similar methylation pattern with liver cancers but not with normal liver tissue [346].

We hypothesized in **Chapter 4** that chronic liver inflammation triggers quiescent cells to proliferate to repair the damaged liver, providing the opportunity for cells with new or pre-existing mutations that confer a selective growth advantage to clonally expand [49,227,228]. By extension, clonal expansion of cells with advantageous epigenetic alterations may also occur as a result of damaging environmental pressures that trigger cell proliferation for tissue repair. Besides inflammation, this could also include carcinogen exposure, or immune destruction of cancer cells (that eventually results in metastasis capable cancer cells). Thus, despite not finding increased mutation accumulation in cirrhotic livers in **Chapter 4**, additional transcriptomic and methylation analyses could show an altered epigenetic landscape tending towards liver cancer, as was reported by Ammerpohl *et al.* [346]. However, such analyses would require many more samples (ideally from unique patients) than we have included, due to the technical variability of quantifying RNA expression and methylation (but also due to biological variation) [347,348]. Additionally, while we identify driver mutations linked to treatment resistance but not those directly involved in metastasis in **Chapter 5**, a pan-cancer comparison of the transcriptomes and methylomes of primary versus metastatic tumors may provide more insight into the mechanisms of metastasis (as well as treatment resistance), and in particular, the possible of driver impact of the additional SVs we find in metastatic cancer. Ideally, longitudinal samples would be compared to minimize inter-sample and inter-cohort variability, though this would be logistically challenging especially because decades may pass between appearance of the primary tumor and eventual metastasis. All procedures, such as storage of (preferably frozen) patient material, data generation (e.g. DNA sequencing), and data analyses, would need to be scaled up as well as standardized to avoid technical variation. Such a long term, large scale study would likely need to happen in the context of routine clinical practice, with the standardization of procedures not only streamlining diagnostics, but also future research by eliminating the need for dataset harmonization efforts (as was done by the PCAWG consortium themselves, but also by us in **Chapter 5**).

Interpretable machine learning models are required for cancer diagnostics

Despite the widespread application of machine learning (ML) in cancer genomics [349], it has yet to be widely used in clinical practice [97]. A potential reason is that most ML models are built with performance in mind, but not interpretability [54,74,75,84–88,90,91,93–96]. Interpretable models aim

to justify how they come to a particular prediction, which is essential if ML is to be adopted into clinical practice. For some models [74,90], feature importances are reported which do somewhat reveal the inner workings of the models, but not for individual predictions. Truly interpretable models provide an explanation as to which features contributed to each individual prediction. BoostDM (a driver mutation predictor) achieves this using shapley additive explanations (SHAP values) [56], whereas CUPLR uses random forest feature contributions (**Chapter 3**) [350]. In the context of clinical diagnostics, feature explanations can be used to reject a bad ML prediction (based on existing knowledge) that otherwise would have led to a patient being given the wrong treatment. For this reason, future clinically oriented ML models should be built with methods that provide interpretable predictions.

Simpler classification algorithms also tend to be inherently interpretable compared to more complex ones [77]. For example, logistic regression coefficients directly indicate how much the odds of the outcome (e.g. HRD) change for a unit of change in the feature (e.g. number of microhomology deletions). Complex algorithms, such as neural networks (NNs) and by extension deep learning, sacrifice interpretability but can model more complex relationships between labels and features. NNs particularly excel when specific architectures are used for specialized tasks. For example, convolutional neural networks (CNN) identify and combine simple patterns for classification (think of the individual strokes used to write the letter 'A'), and have been used to predict the effects of noncoding variants [86]. On the other hand, recurrent neural networks (RNN) have 'memory' and are useful for sequential data, such as for base calling from Nanopore sequencing data [351]. However, NNs are sometimes seemingly used only to ride on the popularity of deep learning [90,93,352]. These studies only used basic feed forward NNs, and did not require nor take advantage of specialized NN architectures. When possible, the use of simpler interpretable ML algorithms is preferable over deep learning. Depending on the application, simple algorithms such as logistic regression or support vector machines may not be sufficiently performant as they can only capture linear (or pseudolinear) feature-label relationships. Tree based algorithms (e.g. random forest or gradient boosting machine) provide a good middle ground in terms of interpretability and capturing complex feature-label relationships in genomic data. Besides choice of algorithm, careful feature engineering can also facilitate model interpretability. With CUPLR (**Chapter 3**) for example, dimension reduction of the RMD features produced RMD meta-features that could be annotated by their direct cancer type associations.

Lastly, ML model performance evaluation using cross-validation (CV) and/or using independent test sets are confounded by the biases of the datasets used, such as sequencing procedure or cohort inclusion criteria. As such, CV and independent testing likely provides an optimistic estimate of real world performance (i.e. on external data). Ultimately, for ML models to be adopted in clinical practice, their performance (ideally based on patient treatment response) must also be validated in practice within prospective clinical trials or programs.

Concluding remarks

WGS has enabled the detection of the full spectrum of mutations found in tumor genomes, from point mutations to SVs. We have exploited these mutations to develop two diagnostic classifiers CHORD and CUPLR which have the potential to improve future treatment decisions. PARP inhibitor therapy for HRD patients is currently limited to specific cancer types (e.g. breast, ovarian, and pancreatic cancer [102,119]). With CHORD, we found that HRD the occurrence and mutational footprint of HRD is not tissue type specific, meaning that many additional patients could still benefit from PARP inhibitor therapy. Likewise, with CUPLR we could identify the tumor tissue of origin for over half of patients diagnosed as having a cancer of unknown primary, indicating that many of such patients could then be eligible for standard of care treatments for specific cancer types. In studying the genomes of numerous tumors and tissues, we however still do not fully understand how cancer develops from the perspective of mutations, possibly because we cannot yet fully assess the impact of passenger

mutations or SVs (particularly large SVs), or because we need to look beyond mutations towards epigenetic or environmental mechanisms of tumorigenesis. We do however find that advanced metastatic cancers more frequently harbor treatment resistance driver mutations compared to primary cancers. To treat metastatic cancer, biomarkers for predicting treatment resistance will need to be identified, which would inform when a change of course of therapy is required. Nevertheless, treatment resistance may be a consequence of broadly applied standard of care treatments. A more effective strategy may be to apply personalized treatment regimens early on (upon diagnosis), with the routine use of WGS enabling early detection of actionable biomarkers (e.g. HRD).

Addendum

References

1. ENCODE Consortium. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012;489: 57–74.
2. Hanahan D, Weinberg RA. Hallmarks of cancer: the next generation. Cell. 2011;144: 646–674.
3. Martincorena I, Campbell PJ. Somatic mutation in cancer and normal cells. Science. 2015;349: 1483–1489.
4. Shendure J, Balasubramanian S, Church GM, Gilbert W, Rogers J, Schloss JA, et al. DNA sequencing at 40: past, present and future. Nature. 2017;550: 345–353.
5. ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium. Pan-cancer analysis of whole genomes. Nature. 2020;578: 82–93.
6. Priestley P, Baber J, Lolkema MP, Steeghs N, de Bruijn E, Shale C, et al. Pan-cancer whole-genome analyses of metastatic solid tumours. Nature. 2019;575: 210–216.
7. Ballouz S, Dobin A, Gillis JA. Is it time to change the reference genome? Genome Biol. 2019;20: 159.
8. Raphael BJ, Dobson JR, Oesper L, Vandin F. Identifying driver mutations in sequenced cancer genomes: computational approaches to enable precision medicine. Genome Med. 2014;6: 5.
9. Van Hoeck A, Tjoonk NH, van Boxtel R, Cuppen E. Portrait of a cancer: mutational signature analyses for cancer diagnostics. BMC Cancer. 2019;19: 457.
10. Vogelstein B, Kinzler KW. Cancer genes and the pathways they control. Nat Med. 2004;10: 789–799.
11. Soussi T, Wiman KG. TP53: an oncogene in disguise. Cell Death Differ. 2015;22: 1239–1249.
12. Wang Q, Pierce-Hoffman E, Cummings BB, Alföldi J, Francioli LC, Gauthier LD, et al. Landscape of multi-nucleotide variants in 125,748 human exomes and 15,708 genomes. Nat Commun. 2020;11: 2539.
13. Alexandrov LB, Kim J, Haradhvala NJ, Huang MN, Tian Ng AW, Wu Y, et al. The repertoire of mutational signatures in human cancer. Nature. 2020;578: 94–101.
14. Preston BD, Albertson TM, Herr AJ. DNA replication fidelity and cancer. Semin Cancer Biol. 2010;20: 281–293.
15. Sfeir A, Symington LS. Microhomology-Mediated End Joining: A Back-up Survival Mechanism or Dedicated Pathway? Trends Biochem Sci. 2015;40: 701–714.
16. Chu D, Wei L. Nonsynonymous, synonymous and nonsense mutations in human cancer-related genes undergo stronger purifying selections than expectation. BMC Cancer. 2019;19: 359.
17. Davies H, Bignell GR, Cox C, Stephens P, Edkins S, Clegg S, et al. Mutations of the BRAF gene in human cancer. Nature. 2002;417: 949–954.
18. Liu J, Kang R, Tang D. The KRAS-G12C inhibitor: activity and resistance. Cancer Gene Ther. 2021. doi:10.1038/s41417-021-00383-9
19. Baeissa HM, Pearl FMG. Identifying the Impact of Inframe Insertions and Deletions on Protein Function in Cancer. J Comput Biol. 2020;27: 786–795.
20. Diederichs S, Bartsch L, Berkmann JC, Fröse K, Heitmann J, Hoppe C, et al. The dark matter of the cancer genome: aberrations in regulatory elements, untranslated regions, splice sites, non-coding RNA and synonymous mutations. EMBO Mol Med. 2016;8: 442–457.
21. Bell RJA, Rube HT, Xavier-Magalhães A, Costa BM, Mancini A, Song JS, et al. Understanding TERT Promoter Mutations: A Common Path to Immortality. Mol Cancer Res. 2016;14: 315–323.
22. Kong-Beltran M, Seshagiri S, Zha J, Zhu W, Bhawe K, Mendoza N, et al. Somatic mutations lead to an oncogenic deletion of met in lung cancer. Cancer Res. 2006;66: 283–289.
23. Kutchko KM, Sanders W, Ziehr B, Phillips G, Solem A, Halvorsen M, et al. Multiple conformations are a conserved and regulatory feature of the RB1 5' UTR. RNA. 2015;21: 1274–1285.
24. Carvalho CMB, Lupski JR. Mechanisms underlying structural variant formation in genomic disorders. Nat Rev Genet. 2016;17: 224–238.

25. Yi K, Ju YS. Patterns and mechanisms of structural variations in human cancer. Exp Mol Med. 2018;50: 1–11.

26. Ho SS, Urban AE, Mills RE. Structural variation in the sequencing era. Nat Rev Genet. 2020;21: 171–189.

27. Ben-David U, Amon A. Context is everything: aneuploidy in cancer. Nat Rev Genet. 2020;21: 44–62.

28. Cameron DL, Baber J, Shale C, Valle-Inclan JE, Besselink N, van Hoeck A, et al. GRIDSS2: comprehensive characterisation of somatic structural variation using single breakend variants and structural variant phasing. Genome Biol. 2021;22: 202.

29. Tanaka H, Watanabe T. Mechanisms Underlying Recurrent Genomic Amplification in Human Cancers. Trends Cancer Res. 2020;6: 462–477.

30. Greenman CD, Cooke SL, Marshall J, Stratton MR, Campbell PJ. Modeling the evolution space of breakage fusion bridge cycles with a stochastic folding process. J Math Biol. 2016;72: 47–86.

31. Korbel JO, Campbell PJ. Criteria for inference of chromothripsis in cancer genomes. Cell. 2013;152: 1226–1236.

32. Baca SC, Prandi D, Lawrence MS, Mosquera JM, Romanel A, Drier Y, et al. Punctuated evolution of prostate cancer genomes. Cell. 2013;153: 666–677.

33. Kim H, Nguyen N-P, Turner K, Wu S, Gujar AD, Luebeck J, et al. Extrachromosomal DNA is associated with oncogene amplification and poor outcome across multiple cancers. Nat Genet. 2020;52: 891–897.

34. Tanaka H, Yao M-C. Palindromic gene amplification--an evolutionarily conserved role for DNA inverted repeats in the genome. Nat Rev Cancer. 2009;9: 216–224.

35. Taylor AM, Shih J, Ha G, Gao GF, Zhang X, Berger AC, et al. Genomic and Functional Approaches to Understanding Cancer Aneuploidy. Cancer Cell. 2018;33: 676–689.e3.

36. Turner KM, Deshpande V, Beyter D, Koga T, Rusert J, Lee C, et al. Extrachromosomal oncogene amplification drives tumour evolution and genetic heterogeneity. Nature. 2017;543: 122–125.

37. Hoebeeck J, van der Luijt R, Poppe B, De Smet E, Yigit N, Claes K, et al. Rapid detection of VHL exon deletions using real-time quantitative PCR. Lab Invest. 2005;85: 24–33.

38. Nguyen L, W M Martens J, Van Hoeck A, Cuppen E. Pan-cancer landscape of homologous recombination deficiency. Nat Commun. 2020;11: 5584.

39. Rodriguez-Martin B, Alvarez EG, Baez-Ortega A, Zamora J, Supek F, Demeulemeester J, et al. Pan-cancer analysis of whole genomes identifies driver rearrangements promoted by LINE-1 retrotransposition. Nat Genet. 2020;52: 306–319.

40. Zapatka M, Borozan I, Brewer DS, Iskar M, Grundhoff A, Alawi M, et al. The landscape of viral associations in human cancers. Nat Genet. 2020;52: 320–330.

41. Mertens F, Johansson B, Fioretos T, Mitelman F. The emerging complexity of gene fusions in cancer. Nat Rev Cancer. 2015;15: 371–381.

42. Tomlins SA, Laxman B, Varambally S, Cao X, Yu J, Helgeson BE, et al. Role of the TMPRSS2-ERG gene fusion in prostate cancer. Neoplasia. 2008;10: 177–188.

43. Soda M, Choi YL, Enomoto M, Takada S, Yamashita Y, Ishikawa S, et al. Identification of the transforming EML4-ALK fusion gene in non-small-cell lung cancer. Nature. 2007;448: 561–566.

44. Haferlach C, Dicker F, Kohlmann A, Schindela S, Weiss T, Kern W, et al. AML with CBFB-MYH11 rearrangement demonstrate RAS pathway alterations in 92% of all cases including a high frequency of NF1 deletions. Leukemia. 2010;24: 1065–1069.

45. Lorenz S, Barøy T, Sun J, Nome T, Vodák D, Bryne J-C, et al. Unscrambling the genomic chaos of osteosarcoma reveals extensive transcript fusion, recurrent rearrangements and frequent novel TP53 aberrations. Oncotarget. 2016;7: 5273–5288.

46. International Cancer Genome Consortium, Hudson TJ, Anderson W, Artez A, Barker AD, Bell C, et al. International network of cancer genome projects. Nature. 2010;464: 993–998.

47. Tomczak K, Czerwińska P, Wiznerowicz M. The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. Contemp Oncol. 2015;19: A68–77.

48. Martínez-Jiménez F, Muiños F, Sentís I, Deu-Pons J, Reyes-Salazar I, Arnedo-Pac C, et al. A compendium

of mutational cancer driver genes. Nat Rev Cancer. 2020;20: 555–572.

49. Martincorena I, Raine KM, Gerstung M, Dawson KJ, Haase K, Van Loo P, et al. Universal Patterns of Selection in Cancer and Somatic Tissues. Cell. 2017;171: 1029–1041.e21.

50. Weghorn D, Sunyaev S. Bayesian inference of negative and positive selection in human cancers. Nat Genet. 2017;49: 1785–1788.

51. Lawrence MS, Stojanov P, Polak P, Kryukov GV, Cibulskis K, Sivachenko A, et al. Mutational heterogeneity in cancer and the search for new cancer-associated genes. Nature. 2013;499: 214–218.

52. Arnedo-Pac C, Mularoni L, Muiños F, Gonzalez-Perez A, Lopez-Bigas N. OncodriveCLUSTL: a sequence-based clustering method to identify cancer drivers. Bioinformatics. 2019;35: 4788–4790.

53. Tokheim C, Bhattacharya R, Niknafs N, Gygax DM, Kim R, Ryan M, et al. Exome-Scale Discovery of Hotspot Mutation Regions in Human Cancer Using 3D Protein Structure. Cancer Res. 2016;76: 3719–3731.

54. Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, Shendure J. A general framework for estimating the relative pathogenicity of human genetic variants. Nat Genet. 2014;46: 310–315.

55. Mularoni L, Sabarinathan R, Deu-Pons J, Gonzalez-Perez A, López-Bigas N. OncodriveFML: a general framework to identify coding and non-coding regions with cancer driver mutations. Genome Biol. 2016;17: 128.

56. Muiños F, Martínez-Jiménez F, Pich O, Gonzalez-Perez A, Lopez-Bigas N. In silico saturation mutagenesis of cancer genes. Nature. 2021;596: 428–432.

57. Zack TI, Schumacher SE, Carter SL, Cherniack AD, Saksena G, Tabak B, et al. Pan-cancer patterns of somatic copy number alteration. Nat Genet. 2013;45: 1134–1140.

58. Beroukhim R, Mermel CH, Porter D, Wei G, Raychaudhuri S, Donovan J, et al. The landscape of somatic copy-number alteration across human cancers. Nature. 2010;463: 899–905.

59. Beroukhim R, Getz G, Nghiemphu L, Barretina J, Hsueh T, Linhart D, et al. Assessing the significance of chromosomal aberrations in cancer: methodology and application to glioma. Proc Natl Acad Sci U S A. 2007;104: 20007–20012.

60. Mermel CH, Schumacher SE, Hill B, Meyerson ML, Beroukhim R, Getz G. GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. Genome Biol. 2011;12: R41.

61. Rheinbay E, Nielsen MM, Abascal F, Wala JA, Shapira O, Tiao G, et al. Analyses of non-coding somatic drivers in 2,658 cancer whole genomes. Nature. 2020;578: 102–111.

62. Nik-Zainal S, Alexandrov LB, Wedge DC, Van Loo P, Greenman CD, Raine K, et al. Mutational processes molding the genomes of 21 breast cancers. Cell. 2012;149: 979–993.

63. Alexandrov LB, Nik-Zainal S, Wedge DC, Aparicio SAJR, Behjati S, Biankin AV, et al. Signatures of mutational processes in human cancer. Nature. 2013;500: 415–421.

64. COSMIC. Catalogue Of Somatic Mutations In Cancer. 2020. Available: https://cancer.sanger.ac.uk/signatures/

65. Pleguezuelos-Manzano C, Puschhof J, Rosendahl Huber A, van Hoeck A, Wood HM, Nomburg J, et al. Mutational signature in colorectal cancer caused by genotoxic pks+ E. coli. Nature. 2020;580: 269–273.

66. Christensen S, Van der Roest B, Besselink N, Janssen R, Boymans S, Martens JWM, et al. 5-Fluorouracil treatment induces characteristic T>G mutations in human cancer. Nat Commun. 2019;10: 4571.

67. Kucab JE, Zou X, Morganella S, Joel M, Nanda AS, Nagy E, et al. A Compendium of Mutational Signatures of Environmental Agents. Cell. 2019;177: 821–836.e16.

68. Tomkova M, Tomek J, Kriaucionis S, Schuster-Böckler B. Mutational signature distribution varies with DNA replication timing and strand asymmetry. Genome Biol. 2018;19: 129.

69. Volkova NV, Meier B, González-Huici V, Bertolini S, Gonzalez S, Vöhringer H, et al. Mutational signatures are jointly shaped by DNA damage and repair. Nat Commun. 2020;11: 2169.

70. Secrier M, Li X, de Silva N, Eldridge MD, Contino G, Bornschein J, et al. Mutational signatures in esophageal adenocarcinoma define etiologically distinct subgroups with therapeutic relevance. Nat Genet. 2016;48: 1131–1141.

71. Li X, Wu WKK, Xing R, Wong SH, Liu Y, Fang X, et al. Distinct Subtypes of Gastric Cancer Defined by Molecular Characterization Include Novel Mutational Signatures with Prognostic Capability. Cancer Res. 2016;76: 1724–1732.

72. Schulze K, Imbeaud S, Letouzé E, Alexandrov LB, Calderaro J, Rebouissou S, et al. Exome sequencing of hepatocellular carcinomas identifies new mutational signatures and potential therapeutic targets. Nat Genet. 2015;47: 505–511.

73. Connor AA, Denroche RE, Jang GH, Timms L, Kalimuthu SN, Selander I, et al. Association of Distinct Mutational Signatures With Correlates of Increased Immune Activity in Pancreatic Ductal Adenocarcinoma. JAMA Oncol. 2017;3: 774–783.

74. Davies H, Glodzik D, Morganella S, Yates LR, Staaf J, Zou X, et al. HRDetect is a predictor of BRCA1 and BRCA2 deficiency based on mutational signatures. Nat Med. 2017;23: 517–525.

75. Zou X, Koh GCC, Nanda AS, Degasperi A, Urgo K, Roumeliotis TI, et al. A systematic CRISPR screen defines mutational mechanisms underpinning signatures caused by replication errors and endogenous DNA damage. Nat Cancer. 2021;2: 643–657.

76. Heath AP, Ferretti V, Agrawal S, An M, Angelakos JC, Arya R, et al. The NCI Genomic Data Commons. Nat Genet. 2021;53: 257–262.

77. Sidey-Gibbons JAM, Sidey-Gibbons CJ. Machine learning in medicine: a practical introduction. BMC Med Res Methodol. 2019;19: 64.

78. Bareche Y, Venet D, Ignatiadis M, Aftimos P, Piccart M, Rothe F, et al. Unravelling triple-negative breast cancer molecular heterogeneity using an integrative multiomic analysis. Ann Oncol. 2018;29: 895–902.

79. van Dessel LF, van Riet J, Smits M, Zhu Y, Hamberg P, van der Heijden MS, et al. The genomic landscape of metastatic castration-resistant prostate cancers reveals multiple distinct genotypes with potential clinical impact. Nat Commun. 2019;10: 5251.

80. Cancer Genome Atlas Research Network. Comprehensive molecular characterization of urothelial bladder carcinoma. Nature. 2014;507: 315–322.

81. Lever J, Krzywinski M, Altman N. Principal component analysis. Nat Methods. 2017;14: 641–642.

82. Libbrecht MW, Noble WS. Machine learning applications in genetics and genomics. Nat Rev Genet. 2015;16: 321–332.

83. Eraslan G, Avsec Ž, Gagneur J, Theis FJ. Deep learning: new computational modelling techniques for genomics. Nat Rev Genet. 2019;20: 389–403.

84. Rentzsch P, Schubach M, Shendure J, Kircher M. CADD-Splice-improving genome-wide variant effect prediction using deep learning-derived splice scores. Genome Med. 2021;13: 31.

85. Kumar S, Harmanci A, Vytheeswaran J, Gerstein MB. SVFX: a machine learning framework to quantify the pathogenicity of structural variants. Genome Biol. 2020;21: 274.

86. Zhou J, Troyanskaya OG. Predicting effects of noncoding variants with deep learning-based sequence model. Nat Methods. 2015;12: 931–934.

87. Han Y, Yang J, Qian X, Cheng W-C, Liu S-H, Hua X, et al. DriverML: a machine learning algorithm for identifying driver genes in cancer sequencing studies. Nucleic Acids Res. 2019;47: e45.

88. Luo P, Ding Y, Lei X, Wu F-X. deepDriver: Predicting Cancer Driver Genes Based on Somatic Mutations Using Deep Convolutional Neural Networks. Front Genet. 2019;10: 13.

89. Way GP, Sanchez-Vega F, La K, Armenia J, Chatila WK, Luna A, et al. Machine Learning Detects Pan-cancer Ras Pathway Activation in The Cancer Genome Atlas. Cell Rep. 2018;23: 172–180.e3.

90. Jiao W, Atwal G, Polak P, Karlic R, Cuppen E, PCAWG Tumor Subtypes and Clinical Translation Working Group, et al. A deep learning system accurately classifies primary and metastatic cancers using passenger mutation patterns. Nat Commun. 2020;11: 728.

91. Salvadores M, Mas-Ponte D, Supek F. Passenger mutations accurately classify human tumors. PLoS Comput Biol. 2019;15: e1006953.

92. Blokzijl F, de Ligt J, Jager M, Sasselli V, Roerink S, Sasaki N, et al. Tissue-specific mutation accumulation in human adult stem cells during life. Nature. 2016;538: 260–264.

93. Menden MP, Iorio F, Garnett M, McDermott U, Benes CH, Ballester PJ, et al. Machine learning prediction

of cancer cell sensitivity to drugs based on genomic and chemical properties. PLoS One. 2013;8: e61318.

94. Mucaki EJ, Zhao JZL, Lizotte DJ, Rogan PK. Predicting responses to platin chemotherapy agents with biochemically-inspired machine learning. Signal Transduct Target Ther. 2019;4: 1.

95. Chang Y, Park H, Yang H-J, Lee S, Lee K-Y, Kim TS, et al. Cancer Drug Response Profile scan (CDRscan): A Deep Learning Model That Predicts Drug Effectiveness from Cancer Genomic Signature. Sci Rep. 2018;8: 8857.

96. Dorman SN, Baranova K, Knoll JHM, Urquhart BL, Mariani G, Carcangiu ML, et al. Genomic signatures for paclitaxel and gemcitabine resistance in breast cancer derived by machine learning. Mol Oncol. 2016;10: 85–100.

97. Kourou K, Exarchos TP, Exarchos KP, Karamouzis MV, Fotiadis DI. Machine learning applications in cancer prognosis and prediction. Comput Struct Biotechnol J. 2015;13: 8–17.

98. Lord CJ, Ashworth A. BRCAness revisited. Nat Rev Cancer. 2016;16: 110–120.

99. Heeke AL, Pishvaian MJ, Lynce F, Xiu J, Brody JR, Chen W-J, et al. Prevalence of Homologous Recombination-Related Gene Mutations Across Multiple Cancer Types. JCO Precis Oncol. 2018;2018. doi:10.1200/PO.17.00286

100. Audeh MW, Carmichael J, Penson RT, Friedlander M, Powell B, Bell-McGuinn KM, et al. Oral poly(ADP-ribose) polymerase inhibitor olaparib in patients with BRCA1 or BRCA2 mutations and recurrent ovarian cancer: a proof-of-concept trial. Lancet. 2010;376: 245–251.

101. Mateo J, Carreira S, Sandhu S, Miranda S, Mossop H, Perez-Lopez R, et al. DNA-Repair Defects and Olaparib in Metastatic Prostate Cancer. N Engl J Med. 2015;373: 1697–1708.

102. Hoppe MM, Sundar R, Tan DSP, Jeyasekharan AD. Biomarkers for Homologous Recombination Deficiency in Cancer. J Natl Cancer Inst. 2018;110: 704–713.

103. Nik-Zainal S, Davies H, Staaf J, Ramakrishna M, Glodzik D, Zou X, et al. Landscape of somatic mutations in 560 breast cancer whole-genome sequences. Nature. 2016;534: 47–54.

104. Degasperi A, Amarante TD, Czarnecki J, Shooter S, Zou X, Glodzik D, et al. A practical framework and online tool for mutational signature analyses show inter-tissue variation and driver dependencies. Nat Cancer. 2020;1: 249–263.

105. Nones K, Johnson J, Newell F, Patch AM, Thorne H, Kazakoff SH, et al. Whole-genome sequencing reveals clinically relevant insights into the aetiology of familial breast cancers. Ann Oncol. 2019;30: 1071–1079.

106. Staaf J, Glodzik D, Bosch A, Vallon-Christersson J, Reuterswärd C, Häkkinen J, et al. Whole-genome sequencing of triple-negative breast cancers in a population-based clinical study. Nat Med. 2019;25: 1526–1533.

107. Maura F, Degasperi A, Nadeu F, Leongamornlert D, Davies H, Moore L, et al. A practical guide for mutational signature analysis in hematological malignancies. Nat Commun. 2019;10: 2969.

108. Pich O, Muiños F, Lolkema MP, Steeghs N, Gonzalez-Perez A, Lopez-Bigas N. The mutational footprints of cancer therapies. Nat Genet. 2019;51: 1732–1740.

109. Behjati S, Gundem G, Wedge DC, Roberts ND, Tarpey PS, Cooke SL, et al. Mutational signatures of ionizing radiation in second malignancies. Nat Commun. 2016;7: 12605.

110. Polak P, Kim J, Braunstein LZ, Karlic R, Haradhavala NJ, Tiao G, et al. A mutational signature reveals alterations underlying deficient homologous recombination repair in breast cancer. Nat Genet. 2017;49: 1476–1486.

111. Jonsson P, Bandlamudi C, Cheng ML, Srinivasan P, Chavan SS, Friedman ND, et al. Tumour lineage shapes BRCA-mediated phenotypes. Nature. 2019;571: 576–579.

112. Póti Á, Gyergyák H, Németh E, Rusz O, Tóth S, Kovácsházi C, et al. Correlation of homologous recombination deficiency induced mutational signatures with sensitivity to PARP inhibitors and cytotoxic agents. Genome Biol. 2019;20: 240.

113. Chun J, Buechelmaier ES, Powell SN. Rad51 paralog complexes BCDX2 and CX3 act at different stages in the BRCA1-BRCA2-dependent homologous recombination pathway. Mol Cell Biol. 2013;33: 387–395.

114. Zhao W, Steinfeld JB, Liang F, Chen X, Maranon DG, Jian Ma C, et al. BRCA1-BARD1 promotes RAD51-

mediated homologous DNA pairing. Nature. 2017;550: 360–365.

115. Cantor SB, Guillemette S. Hereditary breast cancer and the BRCA1-associated FANCJ/BACH1/BRIP1. Future Oncol. 2011;7: 253–261.

116. Castillo A, Paul A, Sun B, Huang TH, Wang Y, Yazinski SA, et al. The BRCA1-interacting protein Abraxas is required for genomic stability and tumor suppression. Cell Rep. 2014;8: 807–817.

117. Folias A, Matkovic M, Bruun D, Reid S, Hejna J, Grompe M, et al. BRCA1 interacts directly with the Fanconi anemia protein FANCA. Hum Mol Genet. 2002;11: 2591–2597.

118. Quigley DA, Dang HX, Zhao SG, Lloyd P, Aggarwal R, Alumkal JJ, et al. Genomic Hallmarks and Structural Variation in Metastatic Prostate Cancer. Cell. 2018;174: 758–769.e9.

119. Golan T, Hammel P, Reni M, Van Cutsem E, Macarulla T, Hall MJ, et al. Maintenance Olaparib for Germline BRCA-Mutated Metastatic Pancreatic Cancer. N Engl J Med. 2019;381: 317–327.

120. Pilarski R. The Role of BRCA Testing in Hereditary Pancreatic and Prostate Cancer Families. Am Soc Clin Oncol Educ Book. 2019;39: 79–86.

121. Chopra N, Tovey H, Pearson A, Cutts R, Toms C, Proszek P, et al. Homologous recombination DNA repair deficiency and PARP inhibition activity in primary triple negative breast cancer. Nat Commun. 2020;11: 2662.

122. Arora K, Barbieri CE. Molecular Subtypes of Prostate Cancer. Curr Oncol Rep. 2018;20: 58.

123. Tariq N-U-A, McNamara MG, Valle JW. Biliary tract cancers: current knowledge, clinical candidates and future challenges. Cancer Manag Res. 2019;11: 2623–2642.

124. Sakai W, Swisher EM, Karlan BY, Agarwal MK, Higgins J, Friedman C, et al. Secondary mutations as a mechanism of cisplatin resistance in BRCA2-mutated cancers. Nature. 2008;451: 1116–1120.

125. Edwards SL, Brough R, Lord CJ, Natrajan R, Vatcheva R, Levine DA, et al. Resistance to therapy caused by intragenic deletion in BRCA2. Nature. 2008;451: 1111–1115.

126. Nangalia J, Campbell PJ. Genome Sequencing during a Patient's Journey through Cancer. N Engl J Med. 2019;381: 2145–2156.

127. Cameron DL, Baber J, Shale C, Papenfuss AT, Valle-Inclan JE, Besselink N, et al. GRIDSS, PURPLE, LINX: Unscrambling the tumor genome via integrated analysis of structural variation and copy number. bioRxiv. 2019. p. 781013. doi:10.1101/781013

128. Anderson GG, Weiss LM. Determining tissue of origin for metastatic cancers: meta-analysis and literature review of immunohistochemistry performance. Appl Immunohistochem Mol Morphol. 2010;18: 3–8.

129. Pavlidis N, Pentheroudakis G. Cancer of unknown primary site. Lancet. 2012;379: 1428–1435.

130. Greco FA. Molecular diagnosis of the tissue of origin in cancer of unknown primary site: useful in patient management. Curr Treat Options Oncol. 2013;14: 634–642.

131. Dietlein F, Eschner W. Inferring primary tumor sites from mutation spectra: a meta-analysis of histology-specific aberrations in cancer-derived cell lines. Hum Mol Genet. 2014;23: 1527–1537.

132. Marquard AM, Birkbak NJ, Thomas CE, Favero F, Krzystanek M, Lefebvre C, et al. TumorTracer: a method to identify the tissue of origin from the somatic mutations of a tumor specimen. BMC Med Genomics. 2015;8: 58.

133. Ramaswamy S, Tamayo P, Rifkin R, Mukherjee S, Yeang CH, Angelo M, et al. Multiclass cancer diagnosis using tumor gene expression signatures. Proc Natl Acad Sci U S A. 2001;98: 15149–15154.

134. Rosenfeld N, Aharonov R, Meiri E, Rosenwald S, Spector Y, Zepeniuk M, et al. MicroRNAs accurately identify cancer tissue origin. Nat Biotechnol. 2008;26: 462–469.

135. Meiri E, Mueller WC, Rosenwald S, Zepeniuk M, Klinke E, Edmonston TB, et al. A second-generation microRNA-based assay for diagnosing tumor tissue origin. Oncologist. 2012;17: 801–812.

136. Laprovitera N, Riefolo M, Porcellini E, Durante G, Garajova I, Vasuri F, et al. MicroRNA expression profiling with a droplet digital PCR assay enables molecular diagnosis and prognosis of cancers of unknown primary. Mol Oncol. 2021;15: 2732–2751.

137. Zhao Y, Pan Z, Namburi S, Pattison A, Posner A, Balachander S, et al. CUP-AI-Dx: A tool for inferring cancer tissue of origin and molecular subtype using RNA gene-expression data and artificial intelligence.

EBioMedicine. 2020;61: 103030.

138. Grewal JK, Tessier-Cloutier B, Jones M, Gakkhar S, Ma Y, Moore R, et al. Application of a Neural Network Whole Transcriptome-Based Pan-Cancer Method for Diagnosis of Primary and Metastatic Cancers. JAMA Netw Open. 2019;2: e192597.

139. Moran S, Martínez-Cardús A, Sayols S, Musulén E, Balañá C, Estival-Gonzalez A, et al. Epigenetic profiling to classify cancer of unknown primary: a multicentre, retrospective analysis. Lancet Oncol. 2016;17: 1386–1395.

140. Roepman P, de Bruijn E, van Lieshout S, Schoenmaker L, Boelens MC, Dubbink HJ, et al. Clinical Validation of Whole Genome Sequencing for Cancer Diagnostics. J Mol Diagn. 2021;23: 816–833.

141. Trans-Omics for Precision Medicine (TOPMed) Program. [cited 31 Jan 2022]. Available: https://www.nhlbi.nih.gov/science/trans-omics-precision-medicine-topmed-program

142. 100,000 Genomes Project Pilot Investigators, Smedley D, Smith KR, Martin A, Thomas EA, McDonagh EM, et al. 100,000 Genomes Pilot on Rare-Disease Diagnosis in Health Care - Preliminary Report. N Engl J Med. 2021;385: 1868–1880.

143. Gaziano JM, Concato J, Brophy M, Fiore L, Pyarajan S, Breeling J, et al. Million Veteran Program: A mega-biobank to study genetic influences on health and disease. J Clin Epidemiol. 2016;70: 214–223.

144. Schuster-Böckler B, Lehner B. Chromatin organization is a major influence on regional mutation rates in human cancer cells. Nature. 2012;488: 504–507.

145. Cortés-Ciriano I, Lee JJ-K, Xi R, Jain D, Jung YL, Yang L, et al. Comprehensive analysis of chromothripsis in 2,658 human cancers using whole-genome sequencing. Nat Genet. 2020;52: 331–341.

146. Martínez-Jiménez F, Movasati A, Brunner S, Nguyen L, Priestley P, Cuppen E, et al. Pan-cancer whole genome comparison of primary and metastatic solid tumors. bioRxiv. 2022. p. 2022.06.17.496528. doi:10.1101/2022.06.17.496528

147. Polak P, Karlić R, Koren A, Thurman R, Sandstrom R, Lawrence M, et al. Cell-of-origin chromatin organization shapes the mutational landscape of cancer. Nature. 2015;518: 360–364.

148. Niculescu-Mizil A, Caruana R. Predicting good probabilities with supervised learning. Proceedings of the 22nd international conference on Machine learning. New York, NY, USA: Association for Computing Machinery; 2005. pp. 625–632.

149. Lim YK, Padma R, Foo L, Chia YN, Yam P, Chia J, et al. Survival outcome of women with synchronous cancers of endometrium and ovary: a 10 year retrospective cohort study. J Gynecol Oncol. 2011;22: 239–243.

150. Henson DE, Schwartz AM, Nsouli H, Albores-Saavedra J. Carcinomas of the pancreas, gallbladder, extrahepatic bile ducts, and ampulla of vater share a field for carcinogenesis: a population-based study. Arch Pathol Lab Med. 2009;133: 67–71.

151. Sell S, Dunsford HA. Evidence for the stem cell origin of hepatocellular carcinoma and cholangiocarcinoma. Am J Pathol. 1989;134: 1347–1363.

152. Oronsky B, Ma PC, Morgensztern D, Carter CA. Nothing But NET: A Review of Neuroendocrine Tumors and Carcinomas. Neoplasia. 2017;19: 991–1002.

153. Miyai K, Schwartz MR, Divatia MK, Anton RC, Park YW, Ayala AG, et al. Adenoid cystic carcinoma of breast: Recent advances. World J Clin Cases. 2014;2: 732–741.

154. Cortes-Ciriano I, Lee S, Park W-Y, Kim T-M, Park PJ. A molecular portrait of microsatellite instability across multiple cancers. Nat Commun. 2017;8: 15180.

155. Alexandrov LB, Ju YS, Haase K, Van Loo P, Martincorena I, Nik-Zainal S, et al. Mutational signatures associated with tobacco smoking in human cancer. Science. 2016;354: 618–622.

156. Sasaki T, Rodig SJ, Chirieac LR, Jänne PA. The biology and treatment of EML4-ALK non-small cell lung cancer. Eur J Cancer. 2010;46: 1773–1780.

157. Faulkner C, Ellis HP, Shaw A, Penman C, Palmer A, Wragg C, et al. BRAF Fusion Analysis in Pilocytic Astrocytomas: KIAA1549-BRAF 15-9 Fusions Are More Frequent in the Midline Than Within the Cerebellum. J Neuropathol Exp Neurol. 2015;74: 867–872.

158. Göransson M, Andersson MK, Forni C, Ståhlberg A, Andersson C, Olofsson A, et al. The myxoid

liposarcoma FUS-DDIT3 fusion oncoprotein deregulates NF-kappaB target genes by interaction with NFKBIZ. Oncogene. 2009;28: 270–278.

159. Psyrri A, DiMaio D. Human papillomavirus in cervical and head-and-neck cancer. Nat Clin Pract Oncol. 2008;5: 24–31.

160. Broccolo F, Ciccarese G, Rossi A, Anselmi L, Drago F, Toniolo A. Human papillomavirus (HPV) and Epstein-Barr virus (EBV) in keratinizing versus non- keratinizing squamous cell carcinoma of the oropharynx. Infect Agent Cancer. 2018;13: 32.

161. Tu T, Budzinska MA, Shackel NA, Urban S. HBV DNA Integration: Molecular Mechanisms and Clinical Implications. Viruses. 2017;9. doi:10.3390/v9040075

162. Dworkin AM, Tseng SY, Allain DC, Iwenofu OH, Peters SB, Toland AE. Merkel cell polyomavirus in cutaneous squamous cell carcinoma of immunocompetent individuals. J Invest Dermatol. 2009;129: 2868–2874.

163. Palczewska A, Palczewski J, Robinson RM, Neagu D. Interpreting random forest classification models using a feature contribution method. arXiv [cs.LG]. 2013. Available: http://arxiv.org/abs/1312.1121

164. Kasar S, Kim J, Improgo R, Tiao G, Polak P, Haradhvala N, et al. Whole-genome sequencing reveals activation-induced cytidine deaminase signatures during indolent chronic lymphocytic leukaemia evolution. Nat Commun. 2015;6: 8866.

165. Stephens PJ, Tarpey PS, Davies H, Van Loo P, Greenman C, Wedge DC, et al. The landscape of cancer genes and mutational processes in breast cancer. Nature. 2012;486: 400–404.

166. Hosen MI, Sheikh M, Zvereva M, Scelo G, Forey N, Durand G, et al. Urinary TERT promoter mutations are detectable up to 10 years prior to clinical diagnosis of bladder cancer: Evidence from the Golestan Cohort Study. EBioMedicine. 2020;53: 102643.

167. Cheung LWT, Hennessy BT, Li J, Yu S, Myers AP, Djordjevic B, et al. High frequency of PIK3R1 and PIK3R2 mutations in endometrial cancer elucidates a novel mechanism for regulation of PTEN protein stability. Cancer Discov. 2011;1: 170–185.

168. Waks AG, Winer EP. Breast Cancer Treatment: A Review. JAMA. 2019;321: 288–300.

169. Brooks RA, Fleming GF, Lastra RR, Lee NK, Moroney JW, Son CH, et al. Current recommendations and recent progress in endometrial cancer. CA Cancer J Clin. 2019;69: 258–279.

170. Watson PA, Arora VK, Sawyers CL. Emerging mechanisms of resistance to androgen receptor inhibitors in prostate cancer. Nat Rev Cancer. 2015;15: 701–711.

171. Zundelevich A, Dadiani M, Kahana-Edwin S, Itay A, Sella T, Gadot M, et al. ESR1 mutations are frequent in newly diagnosed metastatic and loco-regional recurrence of endocrine-treated breast cancer and carry worse prognosis. Breast Cancer Res. 2020;22: 16.

172. Kocakavuk E, Anderson KJ, Varn FS, Johnson KC, Amin SB, Sulman EP, et al. Radiotherapy is associated with a deletion signature that contributes to poor outcomes in patients with cancer. Nat Genet. 2021;53: 1088–1096.

173. Lu MY, Chen TY, Williamson DFK, Zhao M, Shady M, Lipkova J, et al. AI-based pathology predicts origins for cancers of unknown primary. Nature. 2021;594: 106–110.

174. Ghobrial IM, Detappe A, Anderson KC, Steensma DP. The bone-marrow niche in MDS and MGUS: implications for AML and MM. Nat Rev Clin Oncol. 2018;15: 219–233.

175. Katz D, Palmerini E, Pollack SM. More Than 50 Subtypes of Soft Tissue Sarcoma: Paving the Path for Histology-Driven Treatments. Am Soc Clin Oncol Educ Book. 2018;38: 925–938.

176. Brierley J, O'Sullivan B, Asamura H, Byrd D, Huang SH, Lee A, et al. Global Consultation on Cancer Staging: promoting consistent understanding and use. Nat Rev Clin Oncol. 2019;16: 763–771.

177. Koh G, Degasperi A, Zou X, Momen S, Nik-Zainal S. Mutational signatures: emerging concepts, caveats and clinical applications. Nat Rev Cancer. 2021;21: 619–637.

178. Li Y, Roberts ND, Wala JA, Shapira O, Schumacher SE, Kumar K, et al. Patterns of somatic structural variation in human cancer genomes. Nature. 2020;578: 112–121.

179. Vöhringer H, Van Hoeck A, Cuppen E, Gerstung M. Learning mutational signatures and their multidimensional genomic properties with TensorSignatures. Nat Commun. 2021;12: 3628.

180. Gerstung M, Jolly C, Leshchiner I, Dentro SC, Gonzalez S, Rosebrock D, et al. The evolutionary history of 2,658 cancers. Nature. 2020;578: 122–128.

181. Lee-Six H, Olafsson S, Ellis P, Osborne RJ, Sanders MA, Moore L, et al. The landscape of somatic mutation in normal colorectal epithelial cells. Nature. 2019;574: 532–537.

182. Schmitz R, Renné C, Rosenquist R, Tinguely M, Distler V, Menestrina F, et al. Insights into the multistep transformation process of lymphomas: IgH-associated translocations and tumor suppressor gene mutations in clonally related composite Hodgkin's and non-Hodgkin's lymphomas. Leukemia. 2005;19: 1452–1458.

183. Lin X, Boutros PC. Optimization and expansion of non-negative matrix factorization. BMC Bioinformatics. 2020;21: 7.

184. Nguyen L. CUPLR features, HMF and PCAWG samples. 2022. doi:10.5281/zenodo.5939805

185. Nguyen L. UMCUGenetics/cuplr: 2022. doi:10.5281/zenodo.6637693

186. Jemal A, Bray F, Center MM, Ferlay J, Ward E, Forman D. Global cancer statistics. CA Cancer J Clin. 2011;61: 69–90.

187. Altekruse SF, Devesa SS, Dickie LA, McGlynn KA, Kleiner DE. Histological classification of liver and intrahepatic bile duct cancers in SEER registries. J Registry Manag. 2011;38: 201–205.

188. Testino G, Leone S, Borro P. Alcohol and hepatocellular carcinoma: a review and a point of view. World J Gastroenterol. 2014;20: 15943–15954.

189. Loomba R, Friedman SL, Shulman GI. Mechanisms and disease consequences of nonalcoholic fatty liver disease. Cell. 2021;184: 2537–2564.

190. Wongjarupong N, Assavapongpaiboon B, Susantitaphong P, Cheungpasitporn W, Treeprasertsuk S, Rerknimitr R, et al. Non-alcoholic fatty liver disease as a risk factor for cholangiocarcinoma: a systematic review and meta-analysis. BMC Gastroenterol. 2017;17: 149.

191. Razumilava N, Gores GJ, Lindor KD. Cancer surveillance in patients with primary sclerosing cholangitis. Hepatology. 2011;54: 1842–1852.

192. Duan X-Y, Zhang L, Fan J-G, Qiao L. NAFLD leads to liver cancer: do we have sufficient evidence? Cancer Lett. 2014;345: 230–234.

193. Hodskinson MR, Bolner A, Sato K, Kamimae-Lanning AN, Rooijers K, Witte M, et al. Alcohol-derived DNA crosslinks are repaired by two distinct mechanisms. Nature. 2020;579: 603–608.

194. Tamura M, Ito H, Matsui H, Hyodo I. Acetaldehyde is an oxidative stressor for gastric epithelial cells. J Clin Biochem Nutr. 2014;55: 26–31.

195. Novitskiy G, Traore K, Wang L, Trush MA, Mezey E. Effects of ethanol and acetaldehyde on reactive oxygen species production in rat hepatic stellate cells. Alcohol Clin Exp Res. 2006;30: 1429–1435.

196. van Loon B, Markkanen E, Hübscher U. Oxygen as a friend and enemy: How to combat the mutational potential of 8-oxo-guanine. DNA Repair . 2010;9: 604–616.

197. Obe G, Ristow H. Mutagenic, cancerogenic and teratogenic effects of alcohol. Mutat Res. 1979;65: 229–259.

198. Helander A, Lindahl-Kiessling K. Increased frequency of acetaldehyde-induced sister-chromatic exchanges in human lymphocytes treated with an aldehyde dehydrogenase inhibitor. Mutat Res Lett. 1991;264: 103–107.

199. Matsuda T, Kawanishi M, Matsui S, Yagi T, Takebe H. Specific tandem GG to TT base substitutions induced by acetaldehyde are due to intra-strand crosslinks between adjacent guanine bases. Nucleic Acids Res. 1998;26: 1769–1774.

200. Brunner SF, Roberts ND, Wylie LA, Moore L, Aitken SJ, Davies SE, et al. Somatic mutations and clonal dynamics in healthy and cirrhotic human liver. Nature. 2019;574: 538–542.

201. Michelotti GA, Machado MV, Diehl AM. NAFLD, NASH and liver cancer. Nature Reviews Gastroenterology & Hepatology. 2013. pp. 656–665. doi:10.1038/nrgastro.2013.183

202. Dyson JK, Beuers U, Jones DEJ, Lohse AW, Hudson M. Primary sclerosing cholangitis. Lancet. 2018;391: 2547–2559.

203. Kawanishi S, Ohnishi S, Ma N, Hiraku Y, Murata M. Crosstalk between DNA Damage and Inflammation in the Multiple Steps of Carcinogenesis. Int J Mol Sci. 2017;18. doi:10.3390/ijms18081808

204. Cosmic. COSMIC. 2020. Available: https://cancer.sanger.ac.uk/signatures/

205. Chan K, Roberts SA, Klimczak LJ, Sterling JF, Saini N, Malc EP, et al. An APOBEC3A hypermutation signature is distinguishable from the signature of background mutagenesis by APOBEC3B in human cancers. Nat Genet. 2015;47: 1067–1072.

206. Kuijk E, Jager M, van der Roest B, Locati MD, Van Hoeck A, Korzelius J, et al. The mutational impact of culturing human pluripotent and adult stem cells. Nat Commun. 2020;11: 2493.

207. Post Y, Clevers H. Defining Adult Stem Cell Function at Its Simplest: The Ability to Replace Lost Cells through Mitosis. Cell Stem Cell. 2019;25: 174–183.

208. Raven A, Lu W-Y, Man TY, Ferreira-Gonzalez S, O'Duibhir E, Dwyer BJ, et al. Cholangiocytes act as facultative liver stem cells during impaired hepatocyte regeneration. Nature. 2017;547: 350–354.

209. Marsee A, Roos FJM, Verstegen MMA, HPB Organoid Consortium, Gehart H, de Koning E, et al. Building consensus on definition and nomenclature of hepatic, pancreatic, and biliary organoids. Cell Stem Cell. 2021;28: 816–832.

210. Huch M, Gehart H, van Boxtel R, Hamer K, Blokzijl F, Verstegen MMA, et al. Long-term culture of genome-stable bipotent stem cells from adult human liver. Cell. 2015;160: 299–312.

211. Kuijk EW, Rasmussen S, Blokzijl F, Huch M, Gehart H, Toonen P, et al. Generation and characterization of rat liver stem cell lines and their engraftment in a rat model of liver failure. Sci Rep. 2016;6: 22154.

212. Sia D, Villanueva A, Friedman SL, Llovet JM. Liver Cancer Cell of Origin, Molecular Class, and Effects on Patient Prognosis. Gastroenterology. 2017;152: 745–761.

213. Li XC, Wang MY, Yang M, Dai HJ, Zhang BF, Wang W, et al. A mutational signature associated with alcohol consumption and prognostically significantly mutated driver genes in esophageal squamous cell carcinoma. Ann Oncol. 2018;29: 938–944.

214. Kim T-M, Yim S-H, Shin S-H, Xu H-D, Jung Y-C, Park C-K, et al. Clinical implication of recurrent copy number alterations in hepatocellular carcinoma and putative oncogenes in recurrent gains on 1q. Int J Cancer. 2008;123: 2808–2815.

215. Court CM, Hou S, Liu L, Winograd P, DiPardo BJ, Liu SX, et al. Somatic copy number profiling from hepatocellular carcinoma circulating tumor cells. NPJ Precis Oncol. 2020;4: 16.

216. Donne R, Saroul-Aïnama M, Cordier P, Celton-Morizur S, Desdouets C. Polyploidy in liver development, homeostasis and disease. Nat Rev Gastroenterol Hepatol. 2020;17: 391–405.

217. Gasperini M, Tome JM, Shendure J. Towards a comprehensive catalogue of validated and target-linked human enhancers. Nat Rev Genet. 2020;21: 292–310.

218. Doni Jayavelu N, Jajodia A, Mishra A, Hawkins RD. Candidate silencer elements for the human and mouse genomes. Nat Commun. 2020;11: 1061.

219. Letouzé E, Shinde J, Renault V, Couchy G, Blanc J-F, Tubacher E, et al. Mutational signatures reveal the dynamic interplay of risk factors and cellular processes during liver tumorigenesis. Nat Commun. 2017;8: 1315.

220. Fujimoto A, Furuta M, Totoki Y, Tsunoda T, Kato M, Shiraishi Y, et al. Whole-genome mutational landscape and characterization of noncoding and structural mutations in liver cancer. Nat Genet. 2016;48: 500–509.

221. Drost J, van Boxtel R, Blokzijl F, Mizutani T, Sasaki N, Sasselli V, et al. Use of CRISPR-modified human stem cell organoids to study the origin of mutational signatures in cancer. Science. 2017;358: 234–238.

222. Jager M, Blokzijl F, Kuijk E, Bertl J, Vougioukalaki M, Janssen R, et al. Deficiency of nucleotide excision repair is associated with mutational signature observed in cancer. Genome Res. 2019;29: 1067–1077.

223. Voordeckers K, Colding C, Grasso L, Pardo B, Hoes L, Kominek J, et al. Ethanol exposure increases mutation rate through error-prone polymerases. Nat Commun. 2020;11: 3664.

224. Garaycoechea JI, Crossan GP, Langevin F, Mulderrig L, Louzada S, Yang F, et al. Alcohol and endogenous aldehydes damage chromosomes and mutate stem cells. Nature. 2018;553: 171–177.

225. Zhu M, Lu T, Jia Y, Luo X, Gopal P, Li L, et al. Somatic Mutations Increase Hepatic Clonal Fitness and

Regeneration in Chronic Liver Disease. Cell. 2019. doi:10.1016/j.cell.2019.03.026

226. Massarweh NN, El-Serag HB. Epidemiology of Hepatocellular Carcinoma and Intrahepatic Cholangiocarcinoma. Cancer Control. 2017;24: 1073274817729245.

227. Hernandez-Gea V, Toffanin S, Friedman SL, Llovet JM. Role of the microenvironment in the pathogenesis and treatment of hepatocellular carcinoma. Gastroenterology. 2013;144: 512–527.

228. Zhu L, Finkelstein D, Gao C, Shi L, Wang Y, López-Terrada D, et al. Multi-organ Mapping of Cancer Risk. Cell. 2016;166: 1132–1146.e7.

229. Alonso-Curbelo D, Ho Y-J, Burdziak C, Maag JLV, Morris JP 4th, Chandwani R, et al. A gene-environment-induced epigenetic program initiates tumorigenesis. Nature. 2021;590: 642–648.

230. Jager M, Blokzijl F, Sasselli V, Boymans S, Janssen R, Besselink N, et al. Measuring mutation accumulation in single human adult stem cells by whole-genome sequencing of organoid cultures. Nat Protoc. 2018;13: 59–78.

231. Nguyen L. Liver diseases that predispose patients to liver cancer do not result in increased mutagenesis in liver cholangiocyte stem cells. 2021. doi:10.5281/zenodo.5562381

232. Nguyen L. UMCUGenetics/Diseased_livers: natcomm_submission. 2021. doi:10.5281/zenodo.5564971

233. Lambert AW, Pattabiraman DR, Weinberg RA. Emerging Biological Principles of Metastasis. Cell. 2017;168: 670–691.

234. Massagué J, Obenauf AC. Metastatic colonization by circulating tumour cells. Nature. 2016;529: 298–306.

235. Alexander S, Friedl P. Cancer invasion and resistance: interconnected processes of disease progression and therapy failure. Trends Mol Med. 2012;18: 13–26.

236. Turajlic S, Swanton C. Metastasis as an evolutionary process. Science. 2016;352: 169–175.

237. Massagué J, Batlle E, Gomis RR. Understanding the molecular mechanisms driving metastasis. Mol Oncol. 2017;11: 3–4.

238. Welch DR, Hurst DR. Defining the Hallmarks of Metastasis. Cancer Res. 2019;79: 3011–3027.

239. Fares J, Fares MY, Khachfe HH, Salhab HA, Fares Y. Molecular principles of metastasis: a hallmark of cancer revisited. Signal Transduct Target Ther. 2020;5: 28.

240. Birkbak NJ, McGranahan N. Cancer Genome Evolutionary Trajectories in Metastasis. Cancer Cell. 2020;37: 8–19.

241. Weiss F, Lauffenburger D, Friedl P. Towards targeting of shared mechanisms of cancer metastasis and therapy resistance. Nat Rev Cancer. 2022;22: 157–173.

242. Cheng DT, Mitchell TN, Zehir A, Shah RH, Benayed R, Syed A, et al. Memorial Sloan Kettering-Integrated Mutation Profiling of Actionable Cancer Targets (MSK-IMPACT): A Hybridization Capture-Based Next-Generation Sequencing Clinical Assay for Solid Tumor Molecular Oncology. J Mol Diagn. 2015;17: 251–264.

243. Cancer Genome Atlas Research Network, Weinstein JN, Collisson EA, Mills GB, Shaw KRM, Ozenberger BA, et al. The Cancer Genome Atlas Pan-Cancer analysis project. Nat Genet. 2013;45: 1113–1120.

244. Pleasance E, Titmuss E, Williamson L, Kwan H, Culibrk L, Zhao EY, et al. Pan-cancer analysis of advanced patient tumors reveals interactions between therapy and genomic landscapes. Nat Cancer. 2020;1: 452–468.

245. Faltas BM, Prandi D, Tagawa ST, Molina AM, Nanus DM, Sternberg C, et al. Clonal evolution of chemotherapy-resistant urothelial carcinoma. Nat Genet. 2016;48: 1490–1499.

246. Rueda OM, Sammut S-J, Seoane JA, Chin S-F, Caswell-Jin JL, Callari M, et al. Dynamics of breast-cancer relapse reveal late-recurring ER-positive genomic subgroups. Nature. 2019;567: 399–404.

247. Turajlic S, Xu H, Litchfield K, Rowan A, Chambers T, Lopez JI, et al. Tracking Cancer Evolution Reveals Constrained Routes to Metastases: TRACERx Renal. Cell. 2018;173: 581–594.e12.

248. Abida W, Armenia J, Gopalan A, Brennan R, Walsh M, Barron D, et al. Prospective Genomic Profiling of Prostate Cancer Across Disease States Reveals Germline and Somatic Alterations That May Affect Clinical Decision Making. JCO Precis Oncol. 2017;2017. doi:10.1200/PO.17.00029

249. Mateo J, Seed G, Bertan C, Rescigno P, Dolling D, Figueiredo I, et al. Genomics of lethal prostate cancer at diagnosis and castration resistance. J Clin Invest. 2020;130: 1743–1751.

250. Hu Z, Li Z, Ma Z, Curtis C. Multi-cancer analysis of clonality and the timing of systemic spread in paired primary tumors and metastases. Nat Genet. 2020;52: 701–708.

251. Bertucci F, Ng CKY, Patsouris A, Droin N, Piscuoglio S, Carbuccia N, et al. Genomic characterization of metastatic breast cancers. Nature. 2019;569: 560–564.

252. Robinson DR, Wu Y-M, Lonigro RJ, Vats P, Cobain E, Everett J, et al. Integrative clinical genomics of metastatic cancer. Nature. 2017;548: 297–303.

253. Zehir A, Benayed R, Shah RH, Syed A, Middha S, Kim HR, et al. Mutational landscape of metastatic cancer revealed from prospective clinical sequencing of 10,000 patients. Nat Med. 2017;23: 703–713.

254. Garcia-Prieto CA, Martínez-Jiménez F, Valencia A, Porta-Pardo E. Detection of oncogenic and clinically actionable mutations in cancer genomes critically depends on variant calling tools. Bioinformatics. 2022. doi:10.1093/bioinformatics/btac306

255. Nguyen B, Fong C, Luthra A, Smith SA, DiNatale RG, Nandakumar S, et al. Genomic characterization of metastatic patterns from prospective clinical sequencing of 25,000 patients. Cell. 2022;185: 563–575.e11.

256. Watkins TBK, Lim EL, Petkovic M, Elizalde S, Birkbak NJ, Wilson GA, et al. Pervasive chromosomal instability and karyotype order in tumour evolution. Nature. 2020;587: 126–132.

257. Gemble S, Wardenaar R, Keuper K, Srivastava N, Nano M, Macé A-S, et al. Genetic instability from a single S phase after whole-genome duplication. Nature. 2022. doi:10.1038/s41586-022-04578-4

258. Donehower LA, Soussi T, Korkut A, Liu Y, Schultz A, Cardenas M, et al. Integrated Analysis of TP53 Gene and Pathway Alterations in The Cancer Genome Atlas. Cell Rep. 2019;28: 1370–1384.e5.

259. Gerlinger M, McGranahan N, Dewhurst SM, Burrell RA, Tomlinson I, Swanton C. Cancer: evolution within a lifetime. Annu Rev Genet. 2014;48: 215–236.

260. Bakhoum SF, Ngo B, Laughney AM, Cavallo J-A, Murphy CJ, Ly P, et al. Chromosomal instability drives metastasis through a cytosolic DNA response. Nature. 2018;553: 467–472.

261. Forbes SA, Beare D, Boutselakis H, Bamford S, Bindal N, Tate J, et al. COSMIC: somatic cancer genetics at high-resolution. Nucleic Acids Res. 2017;45: D777–D783.

262. Boot A, Ng AWT, Chong FT, Ho S-C, Yu W, Tan DSW, et al. Characterization of colibactin-associated mutational signature in an Asian oral squamous cell carcinoma and in other mucosal tumor types. Genome Res. 2020;30: 803–813.

263. Moore L, Cagan A, Coorens THH, Neville MDC, Sanghvi R, Sanders MA, et al. The mutational landscape of human somatic and germline cells. Nature. 2021;597: 381–386.

264. Abascal F, Harvey LMR, Mitchell E, Lawson ARJ, Lensing SV, Ellis P, et al. Somatic mutation landscapes at single-molecule resolution. Nature. 2021;593: 405–410.

265. Alexandrov LB, Jones PH, Wedge DC, Sale JE, Campbell PJ, Nik-Zainal S, et al. Clock-like mutational processes in human somatic cells. Nat Genet. 2015;47: 1402–1407.

266. Klein CA. Parallel progression of primary tumours and metastases. Nat Rev Cancer. 2009;9: 302–312.

267. Berges RR, Vukanovic J, Epstein JI, CarMichel M, Cisek L, Johnson DE, et al. Implication of cell kinetic changes during the progression of human prostatic cancer. Clin Cancer Res. 1995;1: 473–480.

268. Tubiana M. Tumor cell proliferation kinetics and tumor growth rate. Acta Oncol. 1989;28: 113–121.

269. Menghi F, Barthel FP, Yadav V, Tang M, Ji B, Tang Z, et al. The Tandem Duplicator Phenotype Is a Prevalent Genome-Wide Cancer Configuration Driven by Distinct Gene Mutations. Cancer Cell. 2018;34: 197–210.e5.

270. Shale C, Cameron DL, Baber J, Wong M, Cowley MJ, Papenfuss AT, et al. Unscrambling cancer genomes via integrated analysis of structural variation and copy number. Cell Genomics. 2022; 100112.

271. Kuijk E, Kranenburg O, Cuppen E, van Hoeck A. Common anti-cancer therapies induce somatic mutations in stem cells of healthy tissue. 2022 [cited 29 Apr 2022]. doi:10.21203/rs.3.rs-1435993/v1

272. Jeselsohn R, Buchwalter G, De Angelis C, Brown M, Schiff R. ESR1 mutations—a mechanism for acquired endocrine resistance in breast cancer. Nat Rev Clin Oncol. 2015;12: 573–583.

273. Varn FS, Johnson KC, Martinek J, Huse JT, Nasrallah MP, Wesseling P, et al. Glioma progression is shaped by genetic evolution and microenvironment interactions. Cell. 2022;185: 2184–2199.e16.

274. van de Haar J, Hoes LR, Roepman P, Lolkema MP, Verheul HMW, Gelderblom H, et al. Limited evolution of the actionable metastatic cancer genome under therapeutic pressure. Nat Med. 2021;27: 1553–1563.

275. Yun C-H, Mengwasser KE, Toms AV, Woo MS, Greulich H, Wong K-K, et al. The T790M mutation in EGFR kinase causes drug resistance by increasing the affinity for ATP. Proc Natl Acad Sci U S A. 2008;105: 2070–2075.

276. Nathanson DA, Gini B, Mottahedeh J, Visnyei K, Koga T, Gomez G, et al. Targeted therapy resistance mediated by dynamic regulation of extrachromosomal mutant EGFR DNA. Science. 2014;343: 72–76.

277. Watson RG, Muhale F, Thorne LB, Yu J, O'Neil BH, Hoskins JM, et al. Amplification of thymidylate synthetase in metastatic colorectal cancer patients pretreated with 5-fluorouracil-based chemotherapy. Eur J Cancer. 2010;46: 3358–3364.

278. Yang L, Lin C, Jin C, Yang JC, Tanasa B, Li W, et al. lncRNA-dependent mechanisms of androgen-receptor-regulated gene activation programs. Nature. 2013;500: 598–602.

279. Wasserman I, Lee LH, Ogino S, Marco MR, Wu C, Chen X, et al. SMAD4 Loss in Colorectal Cancer Patients Correlates with Recurrence, Loss of Immune Infiltrate, and Chemoresistance. Clin Cancer Res. 2019;25: 1948–1956.

280. Formisano L, Lu Y, Servetto A, Hanker AB, Jansen VM, Bauer JA, et al. Aberrant FGFR signaling mediates resistance to CDK4/6 inhibitors in ER+ breast cancer. Nat Commun. 2019;10: 1373.

281. Jamal-Hanjani M, Wilson GA, McGranahan N, Birkbak NJ, Watkins TBK, Veeriah S, et al. Tracking the Evolution of Non-Small-Cell Lung Cancer. N Engl J Med. 2017;376: 2109–2121.

282. GLASS Consortium. Glioma through the looking GLASS: molecular evolution of diffuse gliomas and the Glioma Longitudinal Analysis Consortium. Neuro Oncol. 2018;20: 873–884.

283. Martínez-Jiménez F, Priestley P, Shale C, Baber J, Rozemuller E, Cuppen E. Genetic immune escape landscape in primary and metastatic cancer. bioRxiv. 2022. p. 2022.02.23.481444. doi:10.1101/2022.02.23.481444

284. Au L, Hatipoglu E, Robert de Massy M, Litchfield K, Beattie G, Rowan A, et al. Determinants of anti-PD-1 response and resistance in clear cell renal cell carcinoma. Cancer Cell. 2021;39: 1497–1518.e11.

285. Kim N, Kim HK, Lee K, Hong Y, Cho JH, Choi JW, et al. Single-cell RNA sequencing demonstrates the molecular and cellular reprogramming of metastatic lung adenocarcinoma. Nat Commun. 2020;11: 2285.

286. Longo SK, Guo MG, Ji AL, Khavari PA. Integrating single-cell and spatial transcriptomics to elucidate intercellular tissue dynamics. Nat Rev Genet. 2021;22: 627–644.

287. Moldoveanu D, Ramsay L, Lajoie M, Anderson-Trocme L, Lingrand M, Berry D, et al. Spatially mapping the immune landscape of melanoma using imaging mass cytometry. Sci Immunol. 2022;7: eabi5072.

288. Nguyen L, van Hoeck A, Cuppen E. Machine learning-based tissue of origin classification for cancer of unknown primary diagnostics using genome-wide mutation features. bioRxiv. 2021. p. 2021.10.05.463244. doi:10.1101/2021.10.05.463244

289. Signal. [cited 18 May 2022]. Available: https://signal.mutationalsignatures.com/explore/study/6?mutationType=1

290. Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly . 2012;6: 80–92.

291. Chakravarty D, Gao J, Phillips S, Kundra R, Zhang H, Wang J, et al. OncoKB: A Precision Oncology Knowledge Base. JCO Precision Oncology. 2017. pp. 1–16. doi:10.1200/po.17.00011

292. Griffith M, Spies NC, Krysiak K, McMichael JF, Coffman AC, Danos AM, et al. CIViC is a community knowledgebase for expert crowdsourcing the clinical interpretation of variants in cancer. Nat Genet. 2017;49: 170–174.

293. Tamborero D, Rubio-Perez C, Deu-Pons J, Schroeder MP, Vivancos A, Rovira A, et al. Cancer Genome Interpreter annotates the biological and clinical relevance of tumor alterations. Genome Med. 2018;10: 25.

294. Zhu H, Uusküla-Reimand L, Isaev K, Wadi L, Alizada A, Shuai S, et al. Candidate Cancer Driver Mutations

in Distal Regulatory Elements and Long-Range Chromatin Interaction Networks. Mol Cell. 2020;77: 1307–1321.e10.

295. Nugent A, Conatser KR, Turner LL, Nugent JT, Sarino EMB, Ricks-Santi LJ. Reporting of race in genome and exome sequencing studies of cancer: a scoping review of the literature. Genet Med. 2019;21: 2676–2680.

296. Pramesh CS, Badwe RA, Bhoo-Pathy N, Booth CM, Chinnaswamy G, Dare AJ, et al. Priorities for cancer research in low- and middle-income countries: a global perspective. Nat Med. 2022;28: 649–657.

297. Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, et al. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. CA Cancer J Clin. 2021;71: 209–249.

298. Siegel RL, Miller KD, Fuchs HE, Jemal A. Cancer statistics, 2022. CA Cancer J Clin. 2022;72: 7–33.

299. Degasperi A, Zou X, Amarante TD, Martinez-Martinez A, Koh GCC, Dias JML, et al. Substitution mutational signatures in whole-genome–sequenced cancers in the UK population. Science. 2022;376: abl9283.

300. Turnbull C. Introducing whole-genome sequencing into routine cancer care: the Genomics England 100 000 Genomes Project. Ann Oncol. 2018;29: 784–787.

301. Ma X, Liu Y, Liu Y, Alexandrov LB, Edmonson MN, Gawad C, et al. Pan-cancer genome and transcriptome analyses of 1,699 paediatric leukaemias and solid tumours. Nature. 2018;555: 371–376.

302. Hoang PH, Dobbins SE, Cornish AJ, Chubb D, Law PJ, Kaiser M, et al. Whole-genome sequencing of multiple myeloma reveals oncogenic pathways are targeted somatically through multiple mechanisms. Leukemia. 2018;32: 2459–2470.

303. COG data sharing. In: Children's Oncology Group [Internet]. 21 Jan 2016 [cited 16 May 2022]. Available: https://childrensoncologygroup.org/data-sharing

304. Obtaining access to controlled data. In: Genomic Data Commons [Internet]. [cited 16 May 2022]. Available: https://gdc.cancer.gov/access-data/obtaining-access-controlled-data

305. Genomics England Clinical Interpretation Partnership. In: Genomics England [Internet]. 9 Sep 2021 [cited 5 May 2022]. Available: https://www.genomicsengland.co.uk/research/academic

306. Logsdon GA, Vollger MR, Eichler EE. Long-read human genome sequencing and its applications. Nat Rev Genet. 2020;21: 597–614.

307. Chaisson MJP, Sanders AD, Zhao X, Malhotra A, Porubsky D, Rausch T, et al. Multi-platform discovery of haplotype-resolved structural variation in human genomes. Nat Commun. 2019;10: 1784.

308. Mahmoud M, Gobet N, Cruz-Dávalos DI, Mounier N, Dessimoz C, Sedlazeck FJ. Structural variant calling: the long and the short of it. Genome Biol. 2019;20: 246.

309. Voronina N, Wong JKL, Hübschmann D, Hlevnjak M, Uhrig S, Heilig CE, et al. The landscape of chromothripsis across adult cancer types. Nat Commun. 2020;11: 2320.

310. Chudasama P, Mughal SS, Sanders MA, Hübschmann D, Chung I, Deeg KI, et al. Integrative genomic and transcriptomic analysis of leiomyosarcoma. Nat Commun. 2018;9: 144.

311. Gröbner SN, Worst BC, Weischenfeldt J, Buchhalter I, Kleinheinz K, Rudneva VA, et al. The landscape of genomic alterations across childhood cancers. Nature. 2018;555: 321–327.

312. Rausch T, Jones DTW, Zapatka M, Stütz AM, Zichner T, Weischenfeldt J, et al. Genome sequencing of pediatric medulloblastoma links catastrophic DNA rearrangements with TP53 mutations. Cell. 2012;148: 59–71.

313. Liu P, Erez A, Nagamani SCS, Dhar SU, Kołodziejska KE, Dharmadhikari AV, et al. Chromosome catastrophes involve replication mechanisms generating complex genomic rearrangements. Cell. 2011;146: 889–903.

314. Rang FJ, Kloosterman WP, de Ridder J. From squiggle to basepair: computational approaches for improving nanopore sequencing read accuracy. Genome Biol. 2018;19: 90.

315. Bowden R, Davies RW, Heger A, Pagnamenta AT, de Cesare M, Oikkonen LE, et al. Sequencing of human genomes with nanopore technology. Nat Commun. 2019;10: 1869.

316. Jain M, Koren S, Miga KH, Quick J, Rand AC, Sasani TA, et al. Nanopore sequencing and assembly of a human genome with ultra-long reads. Nat Biotechnol. 2018;36: 338–345.

317. Carter J-M, Hussain S. Robust long-read native DNA sequencing using the ONT CsgG Nanopore system. Wellcome Open Res. 2017;2: 23.

318. Koche RP, Rodriguez-Fos E, Helmsauer K, Burkert M, MacArthur IC, Maag J, et al. Extrachromosomal circular DNA drives oncogenic genome remodeling in neuroblastoma. Nat Genet. 2020;52: 29–34.

319. Valle-Inclan JE, Besselink NJM, de Bruijn E, Cameron DL, Ebler J, Kutzera J, et al. A multi-platform reference for somatic structural variation detection. bioRxiv. 2020. p. 2020.10.15.340497. doi:10.1101/2020.10.15.340497

320. Lefebvre C, Bachelot T, Filleron T, Pedrero M, Campone M, Soria J-C, et al. Mutational Profile of Metastatic Breast Cancers: A Retrospective Analysis. PLoS Med. 2016;13: e1002201.

321. Ashiqul Islam SM, Díaz-Gay M, Wu Y, Barnes M, Vangara R, Bergstrom EN, et al. Uncovering novel mutational signatures by de novo extraction with SigProfilerExtractor. bioRxiv. 2022. p. 2020.12.13.422570. doi:10.1101/2020.12.13.422570

322. Wang Y-X, Zhang Y-J. Nonnegative Matrix Factorization: A Comprehensive Review. IEEE Trans Knowl Data Eng. 2013;25: 1336–1353.

323. SIGNAL, Degasperi et al 2022 study. [cited 5 May 2022]. Available: https://signal.mutationalsignatures.com/explore/study/6.

324. Poon SL, McPherson JR, Tan P, Teh BT, Rozen SG. Mutation signatures of carcinogen exposure: genome-wide detection and new opportunities for cancer prevention. Genome Med. 2014;6: 24.

325. Wong JKL, Aichmüller C, Schulze M, Hlevnjak M, Elgaafary S, Lichter P, et al. Association of mutation signature effectuating processes with mutation hotspots in driver genes and non-coding regions. Nat Commun. 2022;13: 178.

326. Karolak A, Levatić J, Supek F. A framework for mutational signature analysis based on DNA shape parameters. PLoS One. 2022;17: e0262495.

327. Zhang X, Simon R. Estimating the number of rate limiting genomic changes for human breast cancer. Breast Cancer Res Treat. 2005;91: 121–124.

328. Tomasetti C, Marchionni L, Nowak MA, Parmigiani G, Vogelstein B. Only three driver gene mutations are required for the development of lung and colorectal cancers. Proc Natl Acad Sci U S A. 2015;112: 118–123.

329. Iranzo J, Martincorena I, Koonin EV. Cancer-mutation network and the number and specificity of driver mutations. Proc Natl Acad Sci U S A. 2018;115: E6010–E6019.

330. Kennedy SR, Zhang Y, Risques RA. Cancer-Associated Mutations but No Cancer: Insights into the Early Steps of Carcinogenesis and Implications for Early Cancer Detection. Trends Cancer Res. 2019;5: 531–540.

331. Martincorena I, Roshan A, Gerstung M, Ellis P, Van Loo P, McLaren S, et al. Tumor evolution. High burden and pervasive positive selection of somatic mutations in normal human skin. Science. 2015;348: 880–886.

332. Martincorena I, Fowler JC, Wabik A, Lawson ARJ, Abascal F, Hall MWJ, et al. Somatic mutant clones colonize the human esophagus with age. Science. 2018;362: 911–917.

333. Keogh MJ, Wei W, Aryaman J, Walker L, van den Ameele J, Coxhead J, et al. High prevalence of focal and multi-focal somatic genetic variants in the human brain. Nat Commun. 2018;9: 4257.

334. Anglesio MS, Papadopoulos N, Ayhan A, Nazeran TM, Noë M, Horlings HM, et al. Cancer-Associated Mutations in Endometriosis without Cancer. N Engl J Med. 2017;376: 1835–1848.

335. Yizhak K, Aguet F, Kim J, Hess JM, Kübler K, Grimsby J, et al. RNA sequence analysis reveals macroscopic somatic clonal expansion across normal tissues. Science. 2019;364. doi:10.1126/science.aaw0726

336. Kumar S, Warrell J, Li S, McGillivray PD, Meyerson W, Salichos L, et al. Passenger Mutations in More Than 2,500 Cancer Genomes: Overall Molecular Functional Impact and Consequences. Cell. 2020;180: 915–927.e16.

337. Gatenby RA, Cunningham JJ, Brown JS. Evolutionary triage governs fitness in driver and passenger mutations and suggests targeting never mutations. Nat Commun. 2014;5: 5499.

338. Teschendorff AE, Gao Y, Jones A, Ruebner M, Beckmann MW, Wachter DL, et al. DNA methylation

outliers in normal breast tissue identify field defects that are enriched in cancer. Nat Commun. 2016;7: 10478.

339. Wolff EM, Chihara Y, Pan F, Weisenberger DJ, Siegmund KD, Sugano K, et al. Unique DNA methylation patterns distinguish noninvasive and invasive urothelial cancers and establish an epigenetic field defect in premalignant tissue. Cancer Res. 2010;70: 8169–8178.

340. Ando T, Yoshida T, Enomoto S, Asada K, Tatematsu M, Ichinose M, et al. DNA methylation of microRNA genes in gastric mucosae of gastric cancer patients: its possible involvement in the formation of epigenetic field defect. Int J Cancer. 2009;124: 2367–2374.

341. Feinberg AP, Koldobskiy MA, Göndör A. Epigenetic modulators, modifiers and mediators in cancer aetiology and progression. Nat Rev Genet. 2016;17: 284–299.

342. Zöchbauer-Müller S, Lam S, Toyooka S, Virmani AK, Toyooka KO, Seidl S, et al. Aberrant methylation of multiple genes in the upper aerodigestive tract epithelium of heavy smokers. Int J Cancer. 2003;107: 612–616.

343. Russo AL, Thiagalingam A, Pan H, Califano J, Cheng K-H, Ponte JF, et al. Differential DNA hypermethylation of critical genes mediates the stage-specific tobacco smoke-induced neoplastic progression of lung cancer. Clin Cancer Res. 2005;11: 2466–2470.

344. van Engeland M, Weijenberg MP, Roemen GMJM, Brink M, de Bruïne AP, Goldbohm RA, et al. Effects of dietary folate and alcohol intake on promoter methylation in sporadic colorectal cancer: the Netherlands cohort study on diet and cancer. Cancer Res. 2003;63: 3133–3137.

345. Abu-Remaileh M, Bender S, Raddatz G, Ansari I, Cohen D, Gutekunst J, et al. Chronic inflammation induces a novel epigenetic program that is conserved in intestinal adenomas and in colorectal cancer. Cancer Res. 2015;75: 2120–2130.

346. Ammerpohl O, Pratschke J, Schafmayer C, Haake A, Faber W, von Kampen O, et al. Distinct DNA methylation patterns in cirrhotic liver and hepatocellular carcinoma. Int J Cancer. 2012;130: 1319–1328.

347. Maksimovic J, Gagnon-Bartsch JA, Speed TP, Oshlack A. Removing unwanted variation in a differential methylation analysis of Illumina HumanMethylation450 array data. Nucleic Acids Res. 2015;43: e106.

348. McIntyre LM, Lopiano KK, Morse AM, Amin V, Oberg AL, Young LJ, et al. RNA-seq: technical variability and sampling. BMC Genomics. 2011;12: 293.

349. Huang K, Xiao C, Glass LM, Critchlow CW, Gibson G, Sun J. Machine learning applications for therapeutic tasks with genomics data. Patterns (N Y). 2021;2: 100328.

350. Palczewska A, Palczewski J, Marchese Robinson R, Neagu D. Interpreting Random Forest Classification Models Using a Feature Contribution Method. In: Bouabana-Tebibel T, Rubin SH, editors. Integration of Reusable Systems. Cham: Springer International Publishing; 2014. pp. 193–218.

351. Boža V, Brejová B, Vinař T. DeepNano: Deep recurrent neural networks for base calling in MinION nanopore reads. PLoS One. 2017;12: e0178751.

352. Sharifi-Noghabi H, Zolotareva O, Collins CC, Ester M. MOLI: multi-omics late integration with deep neural networks for drug response prediction. Bioinformatics. 2019;35: i501–i509.

Summary

Somatic mutations are those that accumulate in the genome of an individual over their lifetime, and have been proposed to contribute to aging and cancer development. There are several types of mutations, ranging from those that result in a single or a double base substitution (SBS and DBS respectively), several bases being inserted or deleted (indel), or those that alter large segments of DNA which are referred to as structural variants (SV). Some mutations enable cells to grow and divide uncontrollably, leading to cancer. These mutations (typically 4-5 per cancer genome) are called 'driver' mutations and occur in a set of 'cancer driver genes'. The remaining mutations are called 'passenger' mutations, having no contribution to cancer development. Nevertheless, genome-wide patterns of these passenger mutations (also known as 'mutational signatures') can be used to infer the presence of mutational processes that have been active in cancer cells, including environmental processes such as ultraviolet radiation exposure, or endogenous processes such as errors during DNA repair.

Whole-genome sequencing (WGS) is a technique that can be used to determine the DNA sequence of the entire genome of a tumor sample, which can then be used to identify the driver and passenger mutations in that sample. In recent years, organizations such as the Hartwig Medical Foundation (Hartwig) and the Pan-cancer Analysis of Whole Genomes consortium (PCAWG) have performed WGS on the tumors from thousands of cancer patients across numerous cancer types to improve our insights on cancer. In this thesis, we have extensively used the Hartwig and PCAWG datasets to develop machine learning tools for cancer diagnostics (**Chapter 2** and **Chapter 3**), as well as to study metastatic cancer progression (**Chapter 5**). In **Chapter 4**, we performed WGS on samples from diseased liver patients to study their relationship with liver tumorigenesis.

One application of machine learning for WGS-based cancer diagnostics is to detect homologous recombination deficiency (HRD). HRD is common in cancer cells, and refers to an impaired homologous recombination (HR) pathway that is required for accurate repair of double stranded DNA breaks. In clinical practice, HRD is often detected by identifying mutations that inactivate HR pathway genes (typically *BRCA1* and *BRCA2*), and cancer patients in which HRD is detected can be treated with PARP inhibitors. An alternative method to detect HRD is to identify the characteristic types of passenger mutations across the whole genome that are a consequence of HRD cells resorting to other less accurate DNA double strand break repair pathways. In **Chapter 2**, we have used the Hartwig and PCAWG datasets to develop CHORD (Classifier of HOmologous Recombination Deficiency), a machine learning classifier that uses the genome-wide presence of microhomology deletions and structural duplications to detect HRD. We then used CHORD to identify and analyze the HRD patients within the Hartwig and PCAWG cohorts, and found that besides *BRCA1* and *BRCA2,* mutations in two other HR genes (*RAD51C* and *PALB2*) were also a frequent genetic cause of HRD. However, we found that gene mutation based testing for HRD (as is done in the clinic) would miss ~40% of HRD patients which show no clear genetic inactivation of the HR genes. CHORD has been integrated into the patient diagnostic report at the Hartwig Medical Foundation to guide treatment decisions. Nevertheless, follow up studies will be required to assess whether HRD patients (especially those without an HR gene mutation) benefit from PARP inhibitors (or other HRD-specific cancer treatments).

Machine learning can also aid with determining the tumor tissue of origin. A tumor that develops for example in the breast (i.e. the tissue of origin) may eventually metastasize and spread elsewhere in the body (e.g. to the lungs). The tissue of origin of a metastatic tumor is typically determined via histological staining in the clinic, and greatly determines the standard of care treatment options available to a patient. However, for ~3% of diagnosed cancer patients, the primary tumor site remains undetermined, which are considered cancers of unknown primary (CUP). To aid with resolving these cases, in **Chapter 3** we have used the Hartwig and PCAWG datasets to develop CUPLR (Cancer of Unknown Primary Location Resolver), a machine learning classifier that uses features based on both driver and

passenger mutations to be able to distinguish 35 cancer types with ~90% precision. For each tumor sample prediction, the classifier produces a graphical output that not only shows the probability of each cancer type, but also an explanation as to which mutation features most contributed to the prediction. The feature explanations can be used to confirm a prediction based on existing knowledge. For example, if CUPLR predicts a sample to be a prostate cancer, where androgen receptor (*AR*) mutation (characteristic of prostate cancers) was a highly contributing feature, we can be confident that the prediction is indeed correct. Using this logic, we could resolve the tumor tissue of origin for 82/141 (58%) of CUP patients in the Hartwig cohort, who then would benefit from therapies approved for the respective cancer types.

Besides developing tools for cancer diagnostics, we have also studied how cancer develops. Inflammatory liver disease has been thought to result in the accumulation of mutations which could in turn lead to liver cancer. To investigate this (**Chapter 4**), we isolated stem cells taken from biopsies taken from the livers of healthy patients, as well as from patients with alcoholic cirrhosis, non-alcoholic steatohepatitis (NASH), and primary sclerosing cholangitis (PSC). WGS was then performed on these stem cells which allowed us to estimate the number of mutations in these stem cells, and by proxy the number of mutations in the patients' livers. Surprisingly, we found that the diseased livers did not show more mutation accumulation than healthy livers, suggesting that mutations do not directly drive the development of liver cancer. It is likely that other mechanisms, such as the tumor microenvironment, contribute to the transition of healthy cells into cancerous cells in the context of liver disease.

We also sought to understand the mutational factors that lead to the progression of primary tumors (early stage cancer) to metastatic tumors (late stage cancer). In **Chapter 5**, we therefore compared of the mutation landscape of primary tumors (PCAWG cohort) versus metastatic tumors (Hartwig cohort) in 22 cancer types. Metastatic tumors generally showed higher genomic instability, with minor increases in the number of small mutations (SBSs, DBSs, and indels), but notable increases in the number of SVs, particularly of structural deletions and duplications. In this study we also showed that cancer therapy exposure contributed to the mutation landscape but also influenced the selection of tumor cells since cancer therapy resistance mutations were found in approximately half of the treated cancer genomes. Nevertheless, most of the observations were cancer type dependent, with 5 cancer types (breast, prostate, thyroid, kidney clear carcinoma, and pancreatic neuroendocrine) showing a substantial transformation in the metastatic cancer genome, mainly driven by higher genomic instability. These 5 cancer types also showed mutational signatures indicative of higher rates of cell division in metastatic tumors, suggesting that these tumors acquired the ability to proliferate faster during the course of cancer progression.

Lastly, **Chapter 6** reflects on the findings and advances presented in this thesis. We discuss the limitations of using WGS data, including biases in WGS datasets, the challenges in detecting SVs, and the complexity of determining the contribution of each mutation towards cancer. Nevertheless, the strength of WGS is its ability to characterize the full spectrum of mutations which we have used to develop two tools (CHORD and CUPLR) for improving cancer diagnostics, which could ultimately provide patients with additional (and importantly, personalized) treatment options.

Samenvatting

Somatische mutaties zijn mutaties die zich gedurende hun leven ophopen in het genoom van een individu en waarvan is voorgesteld dat ze bijdragen aan veroudering en de ontwikkeling van kanker. Er zijn verschillende soorten mutaties, variërend van mutaties die resulteren in een enkele of dubbele basensubstitutie (respectievelijk SBS en DBS), meerdere basen die worden ingevoegd of verwijderd (indel), of mutaties die grote DNA-segmenten veranderen die worden aangeduid als structurele varianten (SV). Sommige mutaties zorgen ervoor dat cellen ongecontroleerd kunnen groeien en delen, wat leidt tot kanker. Deze mutaties (meestal 4-5 per kankergenoom) worden 'driver'-mutaties genoemd en komen voor in een reeks 'cancer driver-genen'. De overige mutaties worden ‘passagiersmutaties’ genoemd en leveren geen bijdrage aan de ontwikkeling van kanker. Niettemin kunnen genoombrede patronen van deze passagiersmutaties (ook bekend als 'mutatiesignaturen') worden gebruikt om mutatieprocessen die actief zijn geweest in kankercellen vast te stellen, waaronder omgevingsprocessen zoals blootstelling aan ultraviolette straling, of endogene processen zoals fouten tijdens DNA-reparatie.

Whole-genome sequencing (WGS) is een techniek die kan worden gebruikt om de DNA-sequentie van het gehele genoom van een tumormonster te bepalen, die vervolgens kan worden gebruikt om de driver- en passagiersmutaties in dat monster te identificeren. In de afgelopen jaren hebben organisaties zoals de Hartwig Medical Foundation (Hartwig) en het consortium Pan-cancer Analysis of Whole Genomes (PCAWG) WGS uitgevoerd op de tumoren van duizenden kankerpatiënten van verschillende soorten kanker om onze inzichten over kanker te verbeteren. In dit proefschrift hebben we de Hartwig en PCAWG datasets uitgebreid gebruikt om machine learning tools te ontwikkelen voor kankerdiagnostiek (**Hoofdstuk 2** en **Hoofdstuk 3**), en om de progressie van uitgezaaide kanker te bestuderen (**Hoofdstuk 5**). In **Hoofdstuk 4** hebben we WGS uitgevoerd op monsters van zieke leverpatiënten om hun relatie met kankerontwikkeling te bestuderen.

Een toepassing van machine learning voor op WGS gebaseerde kankerdiagnostiek is het detecteren van homologe recombinatie deficiëntie (HRD). HRD komt veel voor in kankercellen en verwijst naar een verstoorde homologe recombinatie (HR) route die nodig is voor nauwkeurig herstel van dubbelstrengs DNA-breuken. In de klinische praktijk wordt HRD vaak gedetecteerd door mutaties te identificeren die HR-pathway-genen inactiveren (meestal BRCA1 en BRCA2), en kankerpatiënten waarbij HRD wordt gedetecteerd kunnen worden behandeld met PARP-remmers. Een alternatieve methode om HRD te detecteren, is het identificeren van de karakteristieke typen passagiersmutaties in het hele genoom die een gevolg zijn van HRD-cellen die hun toevlucht nemen tot andere, minder nauwkeurige DNA-herstelpaden voor dubbelstrengsbreuken. In **Hoofdstuk 2** hebben we de Hartwig en PCAWG datasets gebruikt om CHORD (Classifier of HOmologous Recombination Deficiency) te ontwikkelen, een machine learning classifier die gebruik maakt van de genoombrede aanwezigheid van microhomologiedeleties en structurele duplicaties om HRD te detecteren. Vervolgens hebben we CHORD gebruikt om de HRD-patiënten binnen de Hartwig- en PCAWG-cohorten te identificeren en te analyseren, en ontdekten we dat, naast BRCA1 en BRCA2, mutaties in twee andere HR-genen (RAD51C en PALB2) ook een veelvoorkomende genetische oorzaak van HRD waren. We ontdekten echter dat op genmutatie gebaseerde testen voor HRD (zoals gedaan in de kliniek) ~40% van de HRD-patiënten zouden missen die geen duidelijke genetische inactivatie van de HR-genen vertonen. CHORD is geïntegreerd in het diagnostisch rapport van de patiënt bij de Hartwig Medical Foundation om behandelbeslissingen te begeleiden. Niettemin zullen vervolgonderzoeken nodig zijn om te beoordelen of HRD-patiënten (vooral die zonder een HR-genmutatie) baat hebben bij PARP-remmers (of andere HRD-specifieke kankerbehandelingen).

Machine learning kan ook helpen bij het bepalen van het weefsel waar de tumor van oorsprong vandaan komt. Een tumor die zich bijvoorbeeld in de borst ontwikkelt (d.w.z. het weefsel van

oorsprong) kan uiteindelijk uitzaaien en zich elders in het lichaam verspreiden (bijvoorbeeld naar de longen). Het weefsel van oorsprong van een gemetastaseerde tumor wordt gewoonlijk bepaald via histologische kleuring in de kliniek en bepaalt in hoge mate de standaardbehandelingsopties die beschikbaar zijn voor een patiënt. Voor ~3% van de gediagnosticeerde kankerpatiënten blijft de primaire tumorplaats echter onbepaald, wat wordt beschouwd als primaire tumor onbekend (cancer of unknown primary in het Engels; CUP). Om te helpen bij het oplossen van deze gevallen, hebben we in **hoofdstuk 3** de Hartwig- en PCAWG-datasets gebruikt om CUPLR (Cancer of Unknown Primary Location Resolver) te ontwikkelen, een machine learning classifier die eigenschappen gebruikt die zijn gebaseerd op zowel driver- als passagiersmutaties om onderscheid te kunnen maken tussen 35 verschillende kankertypes met ~90% precisie. Voor elke voorspelling van tumormonsters produceert de classifier een grafische uitvoer die niet alleen de waarschijnlijkheid van elk kankertype laat zien, maar ook een verklaring welke mutatiekenmerken het meest hebben bijgedragen aan de voorspelling. De functieverklaringen kunnen worden gebruikt om een voorspelling op basis van bestaande kennis te bevestigen. Als CUPLR bijvoorbeeld voorspelt dat een monster een prostaatkanker is, waarbij de androgeenreceptor (AR)-mutatie (kenmerk van prostaatkanker) een sterk bijdragende eigenschap was, kunnen we erop vertrouwen dat de voorspelling inderdaad correct is. Met behulp van deze logica konden we het tumorweefsel van oorsprong bepalen voor 82/141 (58%) van de CUP-patiënten in het Hartwig-cohort, die dan baat zouden hebben bij therapieën die zijn goedgekeurd voor de respectieve kankertypes.

Naast het ontwikkelen van instrumenten voor kankerdiagnostiek, hebben we ook onderzocht hoe kanker zich ontwikkelt. Men denkt dat een inflammatoire leverziekte resulteert in de accumulatie van mutaties die op hun beurt kunnen leiden tot leverkanker. Om dit te onderzoeken (**Hoofdstuk 4**), hebben we stamcellen geïsoleerd uit biopsies van de lever van gezonde patiënten, evenals van patiënten met alcoholische cirrose, niet-alcoholische steatohepatitis (NASH) en primaire scleroserende cholangitis (PSC). WGS werd vervolgens uitgevoerd op deze stamcellen, waardoor we het aantal mutaties in deze stamcellen konden schatten, en bij benadering het aantal mutaties in de levers van de patiënten. Verrassend genoeg ontdekten we dat de zieke levers niet meer mutatie-accumulatie vertoonden dan gezonde levers, wat suggereert dat mutaties niet direct de ontwikkeling van leverkanker stimuleren. Het is waarschijnlijk dat andere mechanismen, zoals de micro-omgeving van de tumor, bijdragen aan de overgang van gezonde cellen naar kankercellen in de context van leverziekte.

We probeerden ook de mutatiefactoren te begrijpen die leiden tot de progressie van primaire tumoren (kanker in een vroeg stadium) tot uitgezaaide tumoren (kanker in een laat stadium). In **Hoofdstuk 5** hebben we daarom het mutatielandschap van primaire tumoren (PCAWG cohort) vergeleken met gemetastaseerde tumoren (Hartwig cohort) in 22 kankertypes. Gemetastaseerde tumoren vertoonden over het algemeen een hogere genomische instabiliteit, met een kleine toename van het aantal kleine mutaties (SBS's, DBS'en en indels), maar een opmerkelijke toename van het aantal SV's, met name van structurele deleties en duplicaties. In deze studie toonden we ook aan dat blootstelling aan kankertherapie het mutatielandschap beïnvloedde. Ook de selectie van tumorcellen werd beïnvloed, aangezien mutaties in resistentie tegen kankertherapie werden gevonden in ongeveer de helft van de behandelde kankergenomen. Desalniettemin waren de meeste waarnemingen afhankelijk van het kankertype, waarbij 5 kankertypes (borstkanker, prostaatkanker, schildklierkanker, niervrij carcinoom en neuro-endocriene kanker van de pancreas) een substantiële transformatie vertoonden in het gemetastaseerde kankergenoom, voornamelijk gedreven door hogere genomische instabiliteit. Deze 5 kankertypes vertoonden ook mutatiesignaturen die wijzen op hogere celdelingssnelheden bij uitgezaaide tumoren, wat suggereert dat deze tumoren het vermogen kregen om sneller te prolifereren in de loop van kankerprogressie.

Ten slotte reflecteert **Hoofdstuk 6** op de bevindingen en vorderingen die in dit proefschrift worden gepresenteerd. We bespreken de beperkingen van het gebruik van WGS-gegevens, inclusief

systematische fouten in WGS-datasets, de uitdagingen bij het detecteren van SVs en de complexiteit van het bepalen van de bijdrage van elke mutatie aan kanker. Desalniettemin is de kracht van WGS het vermogen om het volledige spectrum van mutaties te karakteriseren, een kracht die wij hebben gebruikt om twee instrumenten (CHORD en CUPLR) te ontwikkelen voor het verbeteren van de kankerdiagnostiek, die patiënten uiteindelijk aanvullende (en belangrijker nog, gepersonaliseerde) behandelingsopties zouden kunnen bieden.

Acknowledgements

There are many people who have helped me through my PhD journey, and I would like to thank them all here!

First I would like to thank my assessment committee, prof. **Gerrit Meijer,** prof. **Berend Snel**, prof. **Emile Voest**, prof. **Marcel Reinders**, and prof. **Susanne Lens**, for taking the time to critically read this thesis. **Edwin**, **Jeroen**, **Ruben**, and **Arne**, thank you for your advice during my evaluation meetings.

To the Cuppen group: **Edwin**, **Arne**, **Ewart**, **Roel**, **Judith**, **Myrthe**, **Chris**, **Ellen**, **Jose**, **Sascha**, **Ali**, **Fran**, **Ies**, **Nicolle**, **Sander**, **Lisanne**, **Robin**, **Ivo**, **Sharon**, **Sjors**, **Francis**, **Joep**, it was great to work together in such a fun and relaxed atmosphere!

Edwin, you were always approachable and provided valuable feedback (personal and academic), even when you left the UMCU to work full time at Hartwig. I liked your flexible pragmatic approach to things, especially when needing to deal with reviewer comments, or dealing with bureaucracy (which you of course very much enjoy :)). I am still impressed at how you can so quickly understand ideas (especially for topics outside your expertise like machine learning) and then give critical comments, whether it be reviewing full manuscripts in less than 2 hours, or during work discussions even when you were on your laptop! I always saw you as an entrepreneur disguised as a biologist, and I'm grateful for your work together with everyone at Hartwig, since without it much of my PhD would not have been possible. Thanks for taking me on as a master's student and eventually as your last PhD student!

Arne, thanks for your guidance over the past 5 years! There were many times that I would've gone off in the wrong direction if you hadn't put me back on the right track. It was cool to have an easy going supervisor who was also into metal. The music festivals we went to were really fun; like Graspop (I have to thank you and your parents for letting me stay over, beats camping any day), and Le Guess Who. One big thing that I've learnt from you is to get things done quickly without being too perfect, though I still find your copy pasted code, or that massive Excel sheet for detecting biallelic gene inactivation funny :P. Good luck with the renovations to your house. I hope you, Lore, and your kids Twan, Lowie, Mona (hope I spelled their names correctly!) will be able to live in it and enjoy it soon!

Ewart, you make some pretty good dad jokes, and it was fun to talk to you about music since both of us had less than mainstream tastes. I was never able to resist your persuasiveness, like when you convinced me to join the Stelvio for Life multiple times, or when you convinced me to jump into the ice cold lake that one time. Thanks for motivating me through the liver footprint project; as you know it was a bit of a struggle for me at times :). You always seemed to have endless biology knowledge and scientific ideas, so I think you will make a great assistant professor, and I wish you the best of luck there!

Nicolle, it was always nice to chat to you about Sabaton, Amon Amarth, and other metal bands; your lab troubles or achievements; video games; Vincent's new wheels (I agree with him that they're cool); or any other random topic. As we moved to the east side office in the de Ridder group, it was good to know that you were almost always in the office so that there was someone to talk to when I came to the office physically.

To both **Nicolle** and **Ewart**, thank you for being my paranymphs!

To the primary versus metastatic crew, **Arne**, **Fran**, **Sascha**, **Ali**, it was impressive that we managed to do the work of a consortium with just the 5 of us! It was a pleasure to work on such a big project with you. For Fran and Ali, though most of our interactions were online, it was nice to meet in a much

more casual setting in Sevilla. Overall it was the perfect fun “conclusion” to the project (though I hope that the reviewers will be kind to us).

Roel, I have learnt a lot of my advanced knowledge and interesting history about programming from you, and enjoyed our talks about obscure programming topics and techniques (like metaprogramming), as well as our conversations about guitar and music. Thanks for your work with processing the PCAWG dataset, as a good part of my PhD depended on a harmonized Hartwig/PCAWG dataset!

Jose, I still find it funny that although you supervised me for my master’s literature review, we eventually found out that we were basically the same age. Though finishing your PhD as an adoptee was tough for you, you definitely brought much liveliness to the Cuppen group! **Judith**, it was nice to have long chats with you now and then as I came into the office in the morning, and thanks for kickstarting the discussion in my thesis! **Sander**, you rival Ewart with how bad (good) your dad jokes and puns are :P. **Robin**, good luck with venturing into Nanostring and programming! **Ivo**, thanks for your beer and liquor wisdom, especially at our retreats!

To the de Ridder group, **Jeroen**, **Emmy**, **Lucia, Joske, Joanna Wolthuis**, **Joanna von Berg**, **Liting**, **Luca**, **Marc**, **Alessio, Myrthe**, **Roy**, **Alexandra, Brent**, **Dieter**, **Anke**, **Marleen**, thanks for adopting me and many of the other Cuppen group members into your group when the Cuppen group was disbanded.

Jeroen, you were a big help with explaining and clarifying machine learning concepts and preventing me from falling into the common traps, especially at the beginning of my master’s internship when I was still a machine learning noob.

Emmy, you were quickly upgraded to friend soon after our first bouldering session together at Sterk! We of course did a lot of climbing together after that which I very much enjoyed! Thanks for all the random conversations during the coffees, lunches, and dinners; listening to my useless (probably inaccurate) YouTube/internet knowledge; and of course your disses :), like when I was ‘giving up’ when climbing hard routes. I hope you and Axel will build a great social network in Barcelona and have a great time there. I look forward to seeing you in Barcelona, and you visiting us back in the Netherlands!

Joske, you brought a lot of energy to the de Ridder group, but also to the CMM, especially during borrels when they were still a thing before COVID. I remember when I still used to sit in the 2nd floor office, I could hear you escorting the de Ridder people to the coffee corner from a mile away. When I didn’t want to get any work done and just wanted to chat, I could always count on you :).

Adrien, though I already wanted to learn programming at the time, the first Python (or was it R?) course I had with you really cemented my interest in programming, and with your other courses, eventually bioinformatics. You are a great teacher (no bias from your cookie bribes) who is easy to talk to and who can explain concepts clearly. I wish you the best of luck with designing and coordinating courses that inspire future students!

Lucia, it was nice to have those morning chats the rare times we were both in the office. It was a pity that our office days didn’t often overlap that much, making it that we had way too much to talk about after not seeing each other for a month or something :P.

Joanna Wolthuis, it was always fun to talk to you about video games, anime, to hear about your chickens, and to see your amazing spontaneous drawings! **Carlo**, it was great to discuss trivia and hypothetical scenarios with you, as well as listen to all your cycling wisdom! **Joanna von Berg**, from the few courses we took together a long time ago until now, I always saw you as a really knowledgeable person who knew how to stand your ground during discussions. **Marleen**, it was a pleasure to work

with you together with Jesko on the SV classifier project. **Dieter**, thanks for checking the Google translated Dutch summary in my thesis, and for being genuinely interested in the research I've done!

To those who joined the Stelvio for Life events, **Edwin**, **Arne**, **Ewart**, **Myrthe**, **Ellen**, **Sascha**, **Nicolle**, **Sander**, **Sharon**, **Joep**, including non-Cuppen group people **Emmy**, **Carlo**, **Liting**, it was definitely some of the highlights over the years. It was a fun experience to make the road trip to Italy together, and eventually struggle up the mountain on our bikes or legs. Though for some of you, **Arne**, **Ewart**, **Emmy**, **Carlo**, getting up the mountain seemed like more of a walk in the park :P.

Lastly, to the **CMM** and the **Genetics** department: Thanks for all the borrels, retreats, masterclasses, and other events! **Wout**, even though we barely crossed paths at the Stratenum, it is great to have you as a friend and climbing buddy! I look forward to more climbing trainings and trips in the future, as well as dinners with board games and co-op video games! **Flip**, thank you for aiding us through all IT and IT-related administrative issues! **Laura**: I had fun at our bouldering and climbing sessions, and it was nice to chat with you at the coffee table (when you didn't have meetings in your agenda :P). **Aniek**: It was nice to have such casual talks with you now and then by the coffee corner on the 2nd floor!

List of publications

Part of this thesis

Nguyen, L., W. M. Martens, J., Van Hoeck, A. & Cuppen, E. Pan-cancer landscape of homologous recombination deficiency. Nat. Commun. 11, 5584 (2020).

Nguyen, L., Van Hoeck, A. & Cuppen, E. Machine learning-based tissue of origin classification for cancer of unknown primary diagnostics using genome-wide mutation features. Nat. Commun. 13, 4013 (2022).

Nguyen, L., Jager, M., Lieshout, R., de Ruiter, P. E., Locati, M. D., Besselink, N., van der Roest, B., Janssen, R., Boymans, S., de Jonge, J., IJzermans, J. N. M., Doukas, M., Verstegen, M. M. A., van Boxtel, R., van der Laan, L. J. W., Cuppen, E. & Kuijk, E. Precancerous liver diseases do not cause increased mutagenesis in liver stem cells. Commun Biol 4, 1301 (2021).

Martínez-Jiménez, F., Movasati, A., Brunner, S., **Nguyen, L.**, Priestley, P., Cuppen, E. & Van Hoeck, A. Pan-cancer whole genome comparison of primary and metastatic solid tumors. bioRxiv 2022.06.17.496528 (2022). doi:10.1101/2022.06.17.496528. *Submitted to Nature*.

A

Other publications

Meijer, T. G., **Nguyen, L.**, Van Hoeck, A., Sieuwerts, A. M., Verkaik, N. S., Ladan, M. M., Ruigrok-Ritstier, K., van Deurzen, C. H. M., van de Werken, H. J. G., Lips, E. H., Linn, S. C., Memari, Y., Davies, H., Nik-Zainal, S., Kanaar, R., Martens, J. W. M., Cuppen, E., Jager, A. & van Gent, D. C. Functional RECAP (REpair CAPacity) assay identifies homologous recombination deficiency undetected by DNA-based BRCAness tests. Oncogene 41, 3498–3506 (2022).

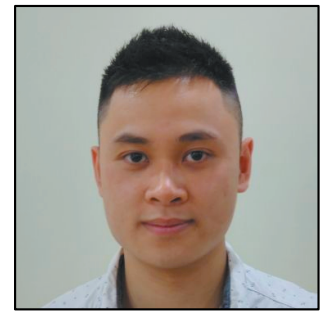
de Witte, C. J., Kutzera, J., van Hoeck, A., **Nguyen, L.**, Boere, I. A., Jalving, M., Ottevanger, P. B., van Schaik-van de Mheen, C., Stevense, M., Kloosterman, W. P., Zweemer, R. P., Cuppen, E. & Witteveen, P. O. Distinct Genomic Profiles Are Associated with Treatment Response and Survival in Ovarian Cancer. Cancers 14, (2022).

Nieboer, M. M., **Nguyen, L.** & de Ridder, J. Predicting pathogenic non-coding SVs disrupting the 3D genome in 1646 whole cancer genomes using multiple instance learning. Sci. Rep. 11, 14411 (2021).

de Witte, C. J., Espejo Valle-Inclan, J., Hami, N., Lõhmussaar, K., Kopper, O., Vreuls, C. P. H., Jonges, G. N., van Diest, P., **Nguyen, L.**, Clevers, H., Kloosterman, W. P., Cuppen, E., Snippert, H. J. G., Zweemer, R. P., Witteveen, P. O. & Stelloo, E. Patient-Derived Ovarian Cancer Organoids Mimic Clinical Response and Exhibit Heterogeneous Inter- and Intrapatient Drug Responses. Cell Rep. 31, 107762 (2020).

Angus, L., Smid, M., Wilting, S. M., van Riet, J., Van Hoeck, A., **Nguyen, L.**, Nik-Zainal, S., Steenbruggen, T. G., Tjan-Heijnen, V. C. G., Labots, M., van Riel, J. M. G. H., Bloemendal, H. J., Steeghs, N., Lolkema, M. P., Voest, E. E., van de Werken, H. J. G., Jager, A., Cuppen, E., Sleijfer, S. & Martens, J. W. M. The genomic landscape of metastatic breast cancer highlights changes in mutation and signature frequencies. Nat. Genet. 51, 1450–1458 (2019).

Curriculum Vitae



Luan Ngoc Nguyen was born on the 5th of January 1992 in Ho Chi Minh City, Vietnam. At the age of 5 he moved together with his parents to Australia and grew up there until the age of 13, after which he (together with his family) moved back to Ho Chi Minh City. There he attended secondary school at ABC international school from 2005-2006, after which he moved to New Zealand for high school at Auckland International College (AIC) from 2007-2009. In August 2009, he moved to the Netherlands to start the Life Sciences bachelor's program at the Hogeschool van Arnhem en Nijmegen (HAN). He graduated from the bachelor's program in September 2012, and thereafter worked as a proteomics and mass spectrometry technician at the Radboud Institute for Molecular Life Sciences (RIMLS) in Nijmegen until October 2015. He then started the master's program Molecular and Cellular Life Sciences (MCLS) at Utrecht University in February 2016. Though he primarily had experience with wet lab research, he eventually made the transition to computational dry lab research when he joined the group of prof. Edwin Cuppen at the University Medical Center Utrecht (UMCU) in February 2016 for his 2nd master's internship. He completed this internship as well as the MCLS master's program in August 2018, after which he continued as a PhD student at the Cuppen group. There, his research involved developing machine learning tools for cancer diagnostics, as well as understanding cancer development, the results of which are presented in this thesis.