Learning to better understand:

Novel bioinformatics algorithms for cancer research

Marleen M. Nieboer

ISBN: 978-94-6416-721-4
Design and layout by: Marleen Nieboer. Trees, DNA, laptop and cell icons designed by Freepik.
Printed by: Ridderprint | www.ridderprint.nl

Learning to better understand: Novel bioinformatics algorithms for cancer research

Leren om beter te begrijpen:
Nieuwe bioinformatica algoritmes voor onderzoek naar kanker

(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

dinsdag 31 augustus 2021 des middags te 2.15 uur

door

Marleentje Martha Nieboer

geboren op 27 november 1993
te Winschoten

Promotor:
Prof.dr. E.P.J.G. Cuppen

Copromotor:
Dr. J. de Ridder

Contents

1

Introduction

1

Cancer is one of the leading causes of death worldwide. However, despite years of ongoing effort and a myriad of developed treatments, a cure has yet to be found. Although it may seem from the millions of published studies on cancer that science has yielded enough information to put the pieces of the puzzle together, the enormous amount of data is not fully understood. Ever since the first human genome sequence was completed in 2002, a large number of projects have followed to make an effort in characterizing the genome sequences of different cancers. Through these studies, it became clear that cancers differ from our healthy cells by accumulating somatic mutations in the DNA. Mutations are found in different kinds and shapes. One highly-studied type is the *Single Nucleotide Variant* (SNV), in which a single base of DNA is switched to another. When more than 1000 base pairs of DNA are either deleted or duplicated, such mutation is called a *Copy Number Variant* (CNV). Any other change affecting more than 50 base pairs that is either inserted, deleted, duplicated, inverted, or translocated (meaning a break is introduced into the DNA sequence, and then stitched together with another piece of DNA elsewhere), is called a *Structural Variant* (SV). These changes can originate from external exposure such as UV radiation or smoking, but can also be left behind as a result of incorrect DNA repair[1].

The reason that these mutations are harmful, or *pathogenic*, to our cells is because they can have an effect on the expression of our genes. For example, a mutation inactivating a tumor suppressor gene such as TP53 can cause the cell to lose control and start dividing rapidly, with cancer being the result thereof[2]. Cancers can start out with as little as one mutation, but can accumulate thousands more mutations as development progresses[3]. However, not all mutations are actually pathogenic. Instead, per cancer only about on average 4-5 mutations have the ability to *drive* the development of the cancer, whereas the rest are passengers that do not have any functional effects[3, 4]. Although this observation may sound like good news for developing anti-cancer treatment, the big problem is that we are not yet very good at identifying these driver mutations in a patient. Therefore, the *driver genes*, which are the genes that are involved in the development of cancer when these are affected by a *driver mutation*, may remain undetected, complicating the process of selecting optimal anti-cancer therapy.

Our current knowledge of cancer-driving mutations is incomplete

One of the most challenging parts of finding suitable treatment for cancer is that each patient accumulates vastly different mutations. Although some cancer types appear to more often be characterized by specific driver mutations, these can be entirely different between even patients with the same cancer type[3]. Despite that many common drivers have been identified and collected by efforts such as the Cancer Gene Census (CGC)[5], the catalogue is far from complete, and there remains a group of patients in which no obvious drivers can be characterized with our current knowledge[6]. Aiming to solve this problem, clever computational methods have been developed to utilize our knowledge of biology to make predictions of which mutations are most likely cancer drivers.

Cancer driver mutation prediction methods

The overarching idea behind predicting which mutations drive cancer is to use existing knowledge to make predictions about the consequences of a mutation. Initially, many

driver prediction studies were focused on mutations directly affecting genes, resulting in changes to the protein product.

One strategy is to identify the genes that are mutated with protein-altering changes more often across patients than would be expected by random chance, which is applied by methods such as MutSig[7], MuSiC[8] and oncodriveCLUST[9] (Fig 1A). A big challenge for these statistics-based methods is to define a proper background rate of passenger mutations, which often varies per cancer type or even per patient, and is influenced by a lot of factors including GC content, gene density, repeat regions and copy number[6, 10, 11].

On the other hand, machine learning-based methods, such as CADD[12], FATHMM[13] or CHASM[14], use existing knowledge by training a classifier using known driver mutations and likely passenger mutations as the respective positive and negative classes (Fig 1A). While these methods further improve our ability to predict drivers, defining the class labels is often challenging. Although sufficient examples can be provided for the negative class, examples of driver mutations for the positive class are a lot sparser[15]. Furthermore, mutations in the negative class are often considered to be passengers only because they have not (yet) been identified as drivers, therefore potentially introducing noisy labels.

Although all of these approaches have been used to successfully discover drivers in patients, their main strength is the identification of common drivers[16, 17]. It is now acknowledged that cancers follow the 'long-tail phenomenon', where only a small number of 'mountains' of common mutations exist, in contrast to a large number of 'hills' of infrequent mutations that vary between patients and cancer types[18]. For example, studies on ovarian and breast cancer identified rare and infrequent mutations in ERBB2 and BRAF that were overlooked by statistics-based approaches, but had clinical relevance[19, 20]. These findings led to the advancement of prediction methods designed to handle rare drivers (Fig 1B).

Methods to predict rare driver mutations

With the growth of efforts such as The Cancer Genome Atlas (TCGA) and Pan Cancer Analysis of Whole Genomes (PCAWG) generating data from different –omics categories than just Whole Genome Sequencing (WGS), many new possibilities have been opened up for identifying (rare) driver mutations. For example, combining expression data with CNVs in glioblastoma patients enabled the discovery of an overexpression of BRAF, which itself was not mutated, through an amplification of its interacting partner EGFR[21]. Similarly, other interaction partners of EGFR, including FGF11, PIK3R1, and PRKACB, also showed increased expression. Such findings suggest that combining multi-omics data can benefit the identification of rare drivers, as more data sources should suggest a high likelihood of the mutation being a driver than would be expected for a passenger. Driver prediction methods integrating multi-omics data range from combining SNVs with SVs[22] with expression data[23, 24] to including methylation and DNAseI hypersensitivity data[25] (Fig 1B).

Another class of methods uses gene or protein interaction networks to identify drivers that are not necessarily mutated themselves, but may instead be deregulated due to mutations in upstream pathway partners[26–28]. These network-based methods were also

1

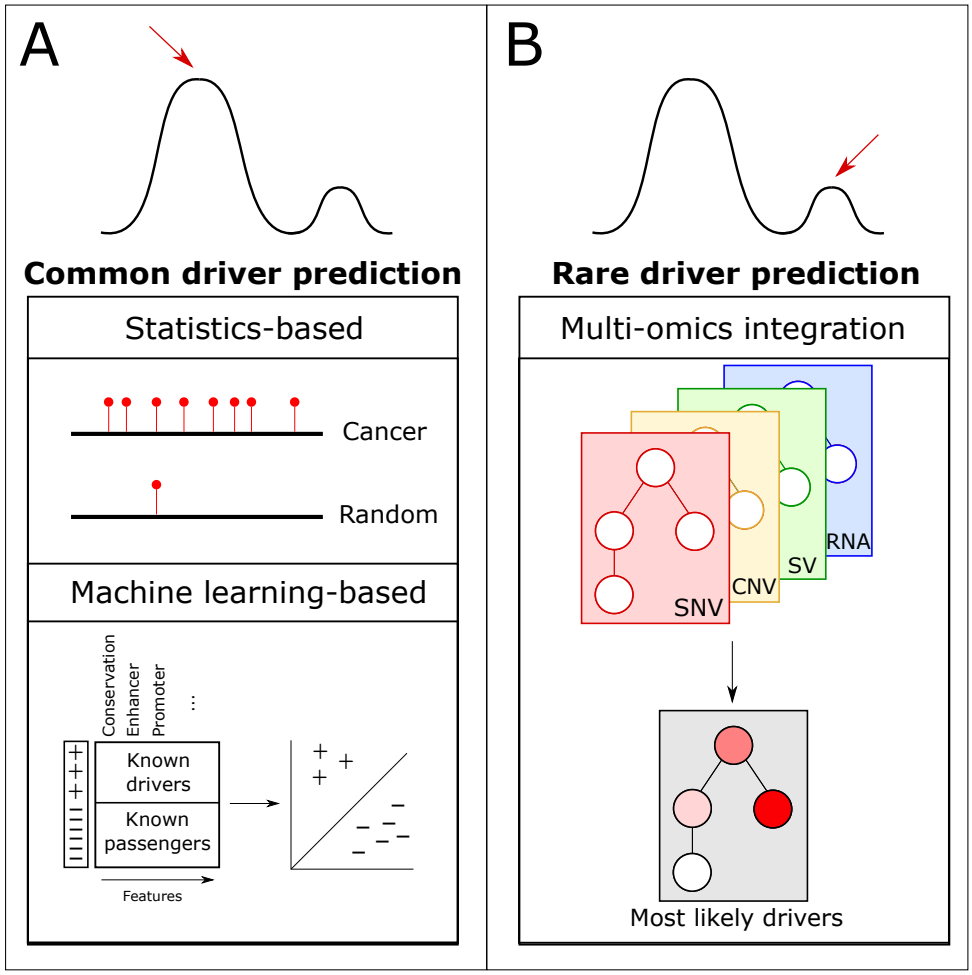


Fig. 1. overview of driver prediction methods. (A) Prediction methods for common drivers are often either based on statistics, or machine learning. (B) Prediction methods for rare drivers focus more on integrating data from multi-omics sources.

shown to achieve higher predictive performance when including more data sources such as expression data[17, 21, 29]. However, one large downside of most existing approaches is that a lot of the interaction networks are often noisy and incomplete. Deepdriver is a machine learning method based on deep convolutional neural networks that aims to overcome this problem by characterizing similarity networks between genes based on expression data[15]. Yet, expression data are often still not available for every single patient in a cohort, leaving a remaining challenge in predicting rare driver mutations in some patients.

Driver prediction in the non-coding genome

A lot of the methods discussed so far have focused on predicting driver mutations in the coding part of the genome, while the vast majority of mutations occur in the non-coding part of the genome[30]. For a long time, the non-coding genome, which makes up about 98% of all base pairs[31], was believed to be junk. Instead, with the increase in the number of WGS data, it has now been discovered that this part of the genome carries out a lot of regulatory functions, which can be disrupted by mutations to indirectly affect the expression of genes. For example, although the TERT gene is rarely mutated itself, non-coding mutations in the TERT promoter are highly recurrent in bladder, thyroid, skin and central nervous system cancers, resulting in overexpression of the TERT gene[32]. While still underrepresented compared to coding-based methods, more and more methods are being developed to predict the pathogenicity of mutations in non-coding regions[10, 13, 33–35]. Notable mentions of recent methods include DeepSEA[36] and ExPecto[37], which use machine learning to learn which genomic features, such as histone modifications, transcription factor binding profiles and enhancers, characterize the regions surrounding and overlapping pathogenic SNVs.

However, such studies are especially not yet extensively performed for non-coding SVs, which despite being seemingly more consequential due to their size, have been largely overlooked. As these SVs are rarely recurrent between patients, their role in cancer remains not fully understood[10]. But recently, more and more evidence is found that these may actually exert their function by disrupting the 3D genome[38].

Non-coding mutations may drive cancer by disrupting the 3D genome

For a long time, the genome sequence was treated as a linear structure. However, it is now known that it is not always the nearest enhancer in linear distance that regulates a gene, and instead the DNA can fold to bring enhancers close to genes that may be as much as 1 Mb apart[39–41]. Recently, the development of the so called 'C' techniques (including 3C, 4C, Hi-C) have enabled us to study such interactions between regions of DNA on a large scale[42]. From these studies, it became apparent that the way the DNA is folded, called the 3D structure, may be of enormous importance for the regulation of genes[43, 44]. On a global scale, the DNA is divided into compartments, where 'A' compartments contain open, active chromatin, whereas 'B' compartments consist of closed, inactive chromatin[45]. Within these compartments, structures exist wherein DNA interacts more frequently with each other than with DNA elsewhere in the genome. These structures are called Topologically Associated Domains (TADs), which typically range from about 200 Kb to 1 Mb in size[45, 46] (Fig 2A, TADs). What is highly interesting is that most interactions between enhancers and gene promoters are confined to take place *within* TADs, which is ensured by the presence of *boundaries* between adjacent TADs (Fig 2A, DNA interactions). These boundaries are enriched for motifs of the CCCTC-binding factor (CTCF), also called CTCF sites[47]. Therefore, it is believed that TADs are formed as the result of a process called *loop extrusion*, in which the DNA is pulled through a ring of cohesin until two convergent CTCF sites are encountered, forming *chromatin loops*[48] (Fig 2B). This idea is supported by the finding that the bound-

1

aries of TADs are often characterized by clusters of such convergent CTCF sites[49] (Fig 2A, CTCF binding sites). Chromatin loops are also formed within a TAD, regulating gene expression on an even smaller scale[50]. Due to the complexity of structures that exist to maintain proper regulation, it is not unthinkable that disrupting the 3D genome can have major consequences for the cell. In a study across 1,962 whole genomes, Liu *et al* identified 21 CTCF sites that were mutated more frequently than expected by random chance, and could thus be explained as potential sites for cancer drivers[51]. In addition, Hnisz *et al* found that T-ALL patients frequently contained microdeletions of CTCF loops nearby well-known T-ALL oncogenes[52]. Similarly, non-coding SVs may also exert their pathogenic effects by disrupting 3D genome structures.

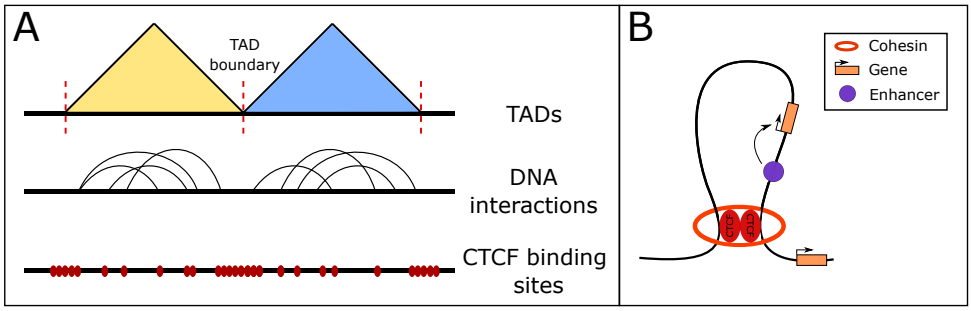


Fig. 2. (A) schematic overview of TADs. DNA interactions between genes and e.g. enhancers are restricted by TAD boundaries, which show an enrichment of CTCF binding sites. These binding sites are also found within TADs, where they regulate the formation of CTCF loops as illustrated in (B). After a loop is formed, genes can only interact with regulators inside the loop.

Non-coding SVs can disrupt TADs to drive cancer

In a study conducted in 2015 by Lupiañez et al, the authors observed that the presence of SVs overlapping with TAD boundaries in the WNT6/IHH/EPHA4/PAX3 region could lead to limb malformation in humans and mice[53]. Using Hi-C techniques to determine the interactions between genomic regions, the authors discovered that the presence of these SVs enabled contact between the genes in the TADs and enhancers in adjacent TADs, by interfering with the separating boundaries. Interestingly, different SV types disrupting other TAD boundaries in the region lead to different limb malformation phenotypes. The study made it possible to construct a model of how TAD boundary-disrupting non-coding SVs can enable spurious contacts between genes and regulatory elements, such as enhancers (Fig 3). Although this study, together with following research, proved that *germline* non-coding SVs can lead to developmental disorders and congenital phenotypes[54–57], similar evidence has been found that *somatic* non-coding SVs that interfere with TAD boundaries may also play a role in the development of cancer[52, 58–61]. Despite the clear impact that SVs can exert through disrupting TADs, it remains unclear if non-coding SVs may also drive disease by disrupting other aspects of the 3D genome, such as CTCF loops. Although Despang *et al* showed in mice that removing single CTCF sites resulted in less disruption of gene expression than when the entire TAD structures were affected[49], it is not known if these smaller effects on gene

expression would be enough to cause cancer in humans.

Furthermore, it is not yet known what the overall contribution of non-coding SVs affecting 3D structures are in the cancer genome, compared to other types of mutations. How often are cancers driven by non-coding SVs, rather than, for example, (non-coding) SNVs or CNVs? Altogether, the true impact of non-coding SVs, including those that do not affect 3D structures, has not yet been fully elucidated, and thus remain an understudied mutation type. Therefore, understanding the mechanisms by which non-coding SVs play a role in cancer remain an important research topic, thereby potentially enabling novel treatment approaches in the clinic.

Subclonal heterogeneity negatively affects our ability to treat cancer

Although driver prediction methods have enabled us to identify novel targets for anticancer therapies, another problem that is holding us back in treating the disease is heterogeneity.

Rather than being one mass of cells that all have the same mutations, many cancers typically present as a heterogeneous population of cells[62]. One striking example of such heterogeneity is type II Testicular Germ Cell Cancer, which may contain tissue from all 3 germ layers[63]. Heterogeneity is comprised by cells with different genetic makeup, also called the *subclones*. Each time a cancer cell divides and acquires a new mutation a new subclone is formed, which also inherits all mutations from its parent subclone[64] (Fig 4A). These subclones continue to divide and form cell populations that either expand or decrease over time, depending on if the acquired mutation was beneficial for the growth of the cancer[65, 66] (Fig 4B). Some mutations may, in fact, kill cancer cells. As a result, cancers often present as a mixture of these different subclones. This characteristic poses a problem for treatment, which often focuses on targeting a specific mutation[67]. While targeting mutations that occurred *early* on during cancer development may be good candidates for treatment as these are likely present in all subclones, cancer cells have developed a way to circumvent these therapies. Mutations that are harmful enough to cause cancer may be too harmful for the cancer cells to stay alive themselves. Therefore, subclones that either *remove* these mutations, or *gain* mutations that can offset the harmful effects, may conquer selective advantage over the other subclones in the tumor[68, 69]. Any treatment may then eradicate a large part of the cancer, but not the subclones that became *resistant* to the therapy. Often, these are the subclones that will lead to recurrence of the cancer, and may eventually metastasize[3] (Fig 4A, treatment). Thus, knowing which mutations are present in which subclones is essential to select a combination of treatments to ensure that all cancer cells are eradicated[67].

However, cancer biopsies are usually only taken at one point in time, sometimes when the cancer has already developed to an advanced stage[70]. Since cancer cells are typically sequenced in bulk, the resulting mutations represent an average across the entire biopsy, losing information about which individual subclones are present in that sample[71]. Such samples also complicate driver prediction, as subclonal driver mutations can be present in very low frequencies that may be missed by mutation detection algorithms. Furthermore, it is often not possible to take a biopsy of the whole cancer,

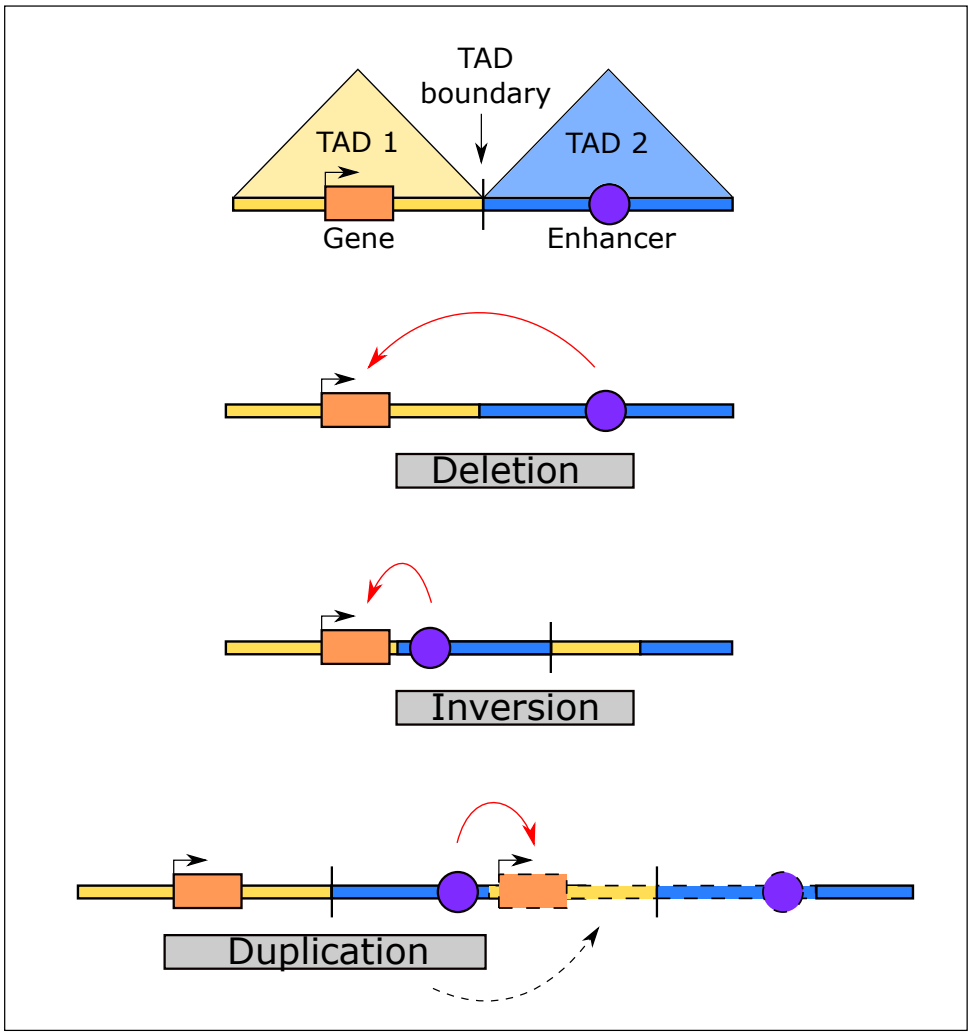


Fig. 3. schematic model of how SVs can disrupt TAD boundaries to alter interactions between genes and regulatory elements. Deletions remove boundaries, enabling previously blocked interactions. Inversions crossing boundaries can bring regulatory elements in close proximity to previously inaccessible genes. Duplications can form 'neo-TADs' in which genes and regulators from different TADs can come in close contact.

as a result of which subclonal mutations may be undetected. Due to the potential lack of mutations leading to treatment resistance being present in all tumor cells, studying cancers as a single, partial sample may result in an incorrect overview of which drivers to target in treatment. Therefore, taking the tumor subclones into account may lead to more effective treatment results[72]. Over the past decade, many computational methods have been designed that aim to infer the subclones and their evolutionary history from single cancer samples.

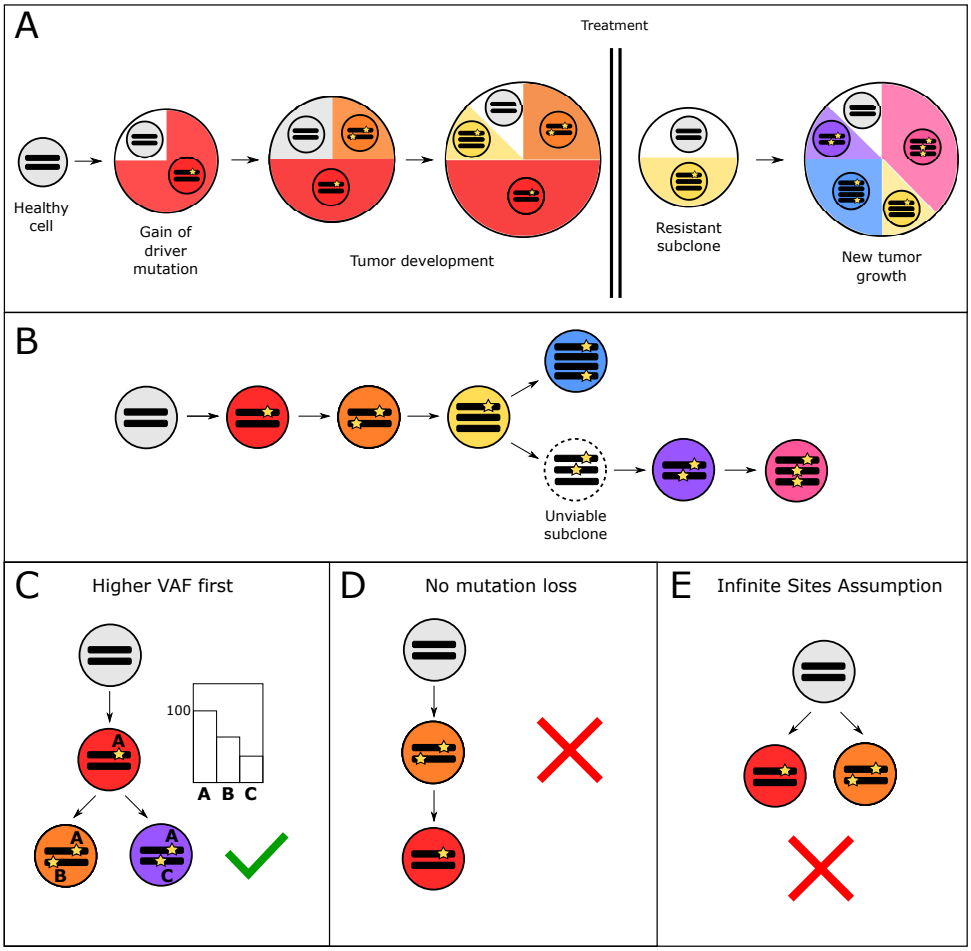


Fig. 4. (A) schematic of tumor development. Existing cells divide to form new subclones. Treatment may eradicate all but resistant subclones, which may then continue forming new subclones. (B) New subclones develop from existing subclones over time, represented as a phylogenetic tree. Due to selective pressure, subclones may disappear. (C) Mutations with higher VAF are expected to have originated early on, as these should be present in nearly all subclones. (D) If mutations are assumed to not be lost, certain phylogenies are impossible. (E) Under the ISA, mutations cannot be gained twice independently.

Methods to reconstruct the evolutionary history of cancer

Within cancer samples, a number of interesting patterns were observed that became essential in inferring subclonal evolution[73]. For example, the frequency in which mutations are present in samples, called the *Variant Allele Frequency* (VAF), highly varies between mutations. As these mutations occurred in copy number neutral regions (meaning that the copy number equals 2), it suggests that not every mutation is present in every cell in the sample. Given that mutations are inherited from the parent cell in each division, the mutations that are present with high VAF are expected to have originated early

during tumor development (Fig 4C). Additionally, it is unlikely for cancers to remove mutations once these have been introduced, unless for example the entire chromosome arm is lost, as these were at least not detrimental to tumor growth (Fig 4D). Finally, due to the sheer number of base pairs in the genome, it is not expected that any mutation would be gained twice independently in two cells, which is also known as the *Infinite Sites Assumption* (ISA) (Fig 4E). These observations proved to be helpful in determining the subclones present in a sample and inferring a phylogenetic tree representing their evolutionary history (Fig 4B), by a process known as *deconvolution*[74].

Deconvolution methods

Certain methods applying deconvolution to identify subclones focus only on SNVs[73, 75–79], while others use only CNVs[80–84]. Methods relying solely on SNVs are restricted to using mutations in copy number neutral regions, as no assumptions can be made about losses due to chromosomal events. Overall, combining different mutation types has been shown to yield the best predictive performance[85]. However, a number of remaining issues plague these existing methods. First of all, cancer genomes are not routinely sequenced at very high depth (typically < 30X). Therefore, it is difficult to distinguish true passenger mutations from noise with high certainty, whereas these are of importance especially when inferring subclones that have not necessarily each gained novel driver mutations. Finally, the lack of ground truth data makes it difficult to validate the true performance of deconvolution-based methods[86].

Sampling-based methods

To partially overcome the need for deconvolution, alternative strategies have focused on taking samples from multiple tumor sites[73, 78, 79, 87, 88]. If certain mutations are not present in every sample, these methods can conclude that the tumor must have formed different branches of subclones over time[78]. However, a limitation of these methods is that the multiple samples themselves often still represent a heterogeneous combination of subclones. Although single-cell sequencing-based methods have shown great promise to overcome a large portion of the deconvolution problem by sampling individual subclones directly, it remains a technological challenge to sequence all cells in a tumor tissue at sufficient quality[64]. Therefore, reconstructing subclonal tumor evolution remains an open problem in understanding cancer development.

Contributions of this thesis

Despite the major progress that has been made in the field of understanding the driving factors of cancer, we are still far away from being able to provide every patient with a successful treatment. In this thesis, we focused on using computational models and multi-omics data to fill in more of the gap of knowledge about how, and which, mutations can drive cancer.

Our first model, described in **chapter 2**, aims to identify better targets for cancer treatment by constructing a phylogenetic timeline of subclones using copy numbers, SNVs and allele frequencies acquired using deep targeted sequencing of physically separated tumor samples. In contrast to existing methods using multiple samples, our ap-

proach specifically utilizes microdissected samples with reduced heterogeneity to overcome the problems with noise encountered during deconvolution.

In **chapter 3**, we annotated TAD boundary-disrupting non-coding SVs with a large amount of regulatory data and showed that aberrant gene expression in these TADs can be explained by altered interactions between genes and regulatory elements. We used these annotations to build a machine learning model to predict pathogenic TAD boundary-disrupting non-coding SVs. To overcome the lack of a ground truth pathogenic non-coding SV set, we used expression data between mutated and non-mutated patients to define our labels. These models enabled us to explore the mechanisms by which non-coding SVs disrupt gene expression in cancer.

In **chapter 4**, we built further on the machine learning method described in **chapter 3** to learn more about the overall role of non-coding SVs in cancer. We applied our method to 12 cancer types, revealing that although non-coding SVs have similar mechanisms across cancer types, their contribution to the development and growth of cancer highly varies between cancer types. We furthermore stepped away from TADs to explore in more detail the role of non-coding SVs on disrupting CTCF loops as a driver mechanism.

Chapter 5 focuses on using machine learning to identify driver genes that are not directly affected by known pathogenic mutations, but rather through upstream disturbances, non-coding mutations or variants of unknown significance. As only WGS data is available in this study, the labels of the negative set are difficult to define as it can only be assumed that a gene is not affected if there is no evidence of mutations in the WGS data. However, this leads to a problem where exactly the disrupted, but non-mutated, genes that we wish to identify contaminate the negative set. As a result, it is actually a good thing for classifiers to report more false positives, making metrics such as AUCPR unreliable. We introduce a swap-one-patient-out CV approach to measure how well classifiers identify false positives as an alternative metric to AUCPR.

Lastly, in **chapter 6** we discuss our findings and suggest improvements for future studies. Although we are now a step closer to understanding the role of mutations in cancer development, a journey of many more miles towards optimal treatment still lies ahead.

References

[1] B. Meier, N. V. Volkova, Y. Hong, P. Schofield, P. J. Campbell, M. Gerstung, and A. Gartner, *Mutational signatures of DNA mismatch repair deficiency in C. elegans and human cancers,* Genome Research **28**, 666 (2018).

[2] D. Hanahan and R. A. Weinberg, *Hallmarks of Cancer: The Next Generation,* Cell **144**, 646 (2011).

[3] M. R. Stratton, P. J. Campbell, and P. A. Futreal, *The cancer genome,* Nature **458**, 719 (2009).

[4] *Pan-cancer analysis of whole genomes,* Nature **578**, 82 (2020).

[5] Z. Sondka, S. Bamford, C. G. Cole, S. A. Ward, I. Dunham, and S. A. Forbes, *The COSMIC Cancer Gene Census: describing genetic dysfunction across all human cancers,* Nature Reviews Cancer **18**, 696 (2018).

[6] D. Tamborero, A. Gonzalez-Perez, C. Perez-Llamas, J. Deu-Pons, C. Kandoth, J. Reimand, M. S. Lawrence, G. Getz, G. D. Bader, L. Ding, and N. Lopez-Bigas, *Comprehensive identification of mutational cancer driver genes across 12 tumor types,* Scientific Reports **3**, 2650 (2013).

[7] M. S. Lawrence, P. Stojanov, P. Polak, G. V. Kryukov, K. Cibulskis, A. Sivachenko, S. L. Carter, C. Stewart, C. H. Mermel, S. a. Roberts, A. Kiezun, P. S. Hammerman, A. McKenna, Y. Drier, L. Zou, A. H. Ramos, T. J. Pugh, N. Stransky, E. Helman, J. Kim, C. Sougnez, L. Ambrogio, E. Nickerson, E. Shefler, M. L. Cortés, D. Auclair, G. Saksena, D. Voet, M. Noble, D. DiCara, P. Lin, L. Lichtenstein, D. I. Heiman, T. Fennell, M. Imielinski, B. Hernandez, E. Hodis, S. Baca, A. M. Dulak, J. Lohr, D.-A. Landau, C. J. Wu, J. Melendez-Zajgla, A. Hidalgo-Miranda, A. Koren, S. a. McCarroll, J. Mora, R. S. Lee, B. Crompton, R. Onofrio, M. Parkin, W. Winckler, K. Ardlie, S. B. Gabriel, C. W. M. Roberts, J. a. Biegel, K. Stegmaier, A. J. Bass, L. a. Garraway, M. Meyerson, T. R. Golub, D. a. Gordenin, S. Sunyaev, E. S. Lander, and G. Getz, *Mutational heterogeneity in cancer and the search for new cancer-associated genes.* Nature **499**, 214 (2013).

[8] N. D. Dees, Q. Zhang, C. Kandoth, M. C. Wendl, W. Schierding, D. C. Koboldt, T. B. Mooney, M. B. Callaway, D. Dooling, E. R. Mardis, R. K. Wilson, and L. Ding, *MuSiC: Identifying mutational significance in cancer genomes,* Genome Research **22**, 1589 (2012).

[9] D. Tamborero, A. Gonzalez-Perez, and N. Lopez-Bigas, *OncodriveCLUST: exploiting the positional clustering of somatic mutations to identify cancer genes,* Bioinformatics **29**, 2238 (2013).

[10] E. Rheinbay, M. M. Nielsen, F. Abascal, J. A. Wala, O. Shapira, G. Tiao, H. Hornshøj, J. M. Hess, R. I. Juul, Z. Lin, L. Feuerbach, R. Sabarinathan, T. Madsen, J. Kim, L. Mularoni, S. Shuai, A. Lanzós, C. Herrmann, Y. E. Maruvka, C. Shen, S. B. Amin, P. Bandopadhayay, J. Bertl, K. A. Boroevich, J. Busanovich, J. Carlevaro-Fita, D. Chakravarty, C. W. Y. Chan, D. Craft, P. Dhingra, K. Diamanti, N. A. Fonseca, A. Gonzalez-Perez, Q. Guo, M. P. Hamilton, N. J. Haradhvala, C. Hong, K. Isaev, T. A. Johnson, M. Juul, A. Kahles, A. Kahraman, Y. Kim, J. Komorowski, K. Kumar, S. Kumar, D. Lee, K.-V. Lehmann, Y. Li, E. M. Liu, L. Lochovsky, K. Park, O. Pich, N. D. Roberts, G. Saksena, S. E. Schumacher, N. Sidiropoulos, L. Sieverling, N. Sinnott-Armstrong, C. Stewart, D. Tamborero, J. M. C. Tubio, H. M. Umer, L. Uusküla-Reimand, C. Wadelius, L. Wadi, X. Yao, C.-Z. Zhang, J. Zhang, J. E. Haber, A. Hobolth, M. Imielinski, M. Kellis, M. S. Lawrence, C. von Mering, H. Nakagawa, B. J. Raphael, M. A. Rubin, C. Sander, L. D. Stein, J. M. Stuart, T. Tsunoda, D. A. Wheeler, R. Johnson, J. Reimand, M. Gerstein, E. Khurana, P. J. Campbell, N. López-Bigas, J. Weischenfeldt, R. Beroukhim, I. Martincorena, J. S. Pedersen, and G. Getz, *Analyses of non-coding somatic drivers in 2,658 cancer whole genomes,* Nature **578**, 102 (2020).

[11] L. Mularoni, R. Sabarinathan, J. Deu-Pons, A. Gonzalez-Perez, and N. López-Bigas, *OncodriveFML: a general framework to identify coding and non-coding regions with cancer driver mutations,* Genome Biology **17**, 128 (2016).

[12] P. Rentzsch, D. Witten, G. M. Cooper, J. Shendure, and M. Kircher, *CADD: predicting the deleteriousness of variants throughout the human genome,* Nucleic Acids Research **47**, D886 (2019).

[13] M. F. Rogers, H. A. Shihab, M. Mort, D. N. Cooper, T. R. Gaunt, and C. Campbell, *FATHMM-XF: accurate prediction of pathogenic point mutations via extended features,* Bioinformatics **34**, 511 (2018).

[14] H. Carter, S. Chen, L. Isik, S. Tyekucheva, V. E. Velculescu, K. W. Kinzler, B. Vogelstein, and R. Karchin, *Cancer-Specific High-Throughput Annotation of Somatic Mutations: Computational Prediction of Driver Missense Mutations,* Cancer Research **69**, 6660 (2009).

[15] P. Luo, Y. Ding, X. Lei, and F.-X. Wu, *deepDriver: Predicting Cancer Driver Genes Based on Somatic Mutations Using Deep Convolutional Neural Networks,* Frontiers in Genetics **10** (2019), 10.3389/fgene.2019.00013.

[16] S. Althubaiti, A. Karwath, A. Dallol, A. Noor, S. S. Alkhayyat, R. Alwassia, K. Mineta, T. Gojobori, A. D. Beggs, P. N. Schofield, G. V. Gkoutos, and R. Hoehndorf, *Ontology-based prediction of cancer driver genes,* Scientific Reports **9**, 17405 (2019).

[17] J. P. Hou and J. Ma, *DawnRank: discovering personalized driver genes in cancer,* Genome Medicine **6**, 56 (2014).

[18] L. D. Wood, D. W. Parsons, S. Jones, J. Lin, T. Sjoblom, R. J. Leary, D. Shen, S. M. Boca, T. Barber, J. Ptak, N. Silliman, S. Szabo, Z. Dezso, V. Ustyanksky, T. Nikolskaya, Y. Nikolsky, R. Karchin, P. A. Wilson, J. S. Kaminker, Z. Zhang, R. Croshaw, J. Willis, D. Dawson, M. Shipitsin, J. K. V. Willson, S. Sukumar, K. Polyak, B. H. Park, C. L. Pethiyagoda, P. V. K. Pant, D. G. Ballinger, A. B. Sparks, J. Hartigan, D. R. Smith, E. Suh, N. Papadopoulos, P. Buckhaults, S. D. Markowitz, G. Parmigiani, K. W. Kinzler, V. E. Velculescu, and B. Vogelstein, *The Genomic Landscapes of Human Breast and Colorectal Cancers,* Science **318**, 1108 (2007).

[19] The Cancer Genome Atlas Research Network, *Integrated genomic analyses of ovarian carcinoma,* Nature **474**, 609 (2011).

[20] S. P. Shah, A. Roth, R. Goya, A. Oloumi, G. Ha, Y. Zhao, G. Turashvili, J. Ding, K. Tse, G. Haffari, A. Bashashati, L. M. Prentice, J. Khattra, A. Burleigh, D. Yap, V. Bernard, A. McPherson, K. Shumansky, A. Crisan, R. Giuliany, A. Heravi-Moussavi, J. Rosner, D. Lai, I. Birol, R. Varhol, A. Tam, N. Dhalla, T. Zeng, K. Ma, S. K. Chan, M. Griffith, A. Moradian, S.-W. G. Cheng, G. B. Morin, P. Watson, K. Gelmon, S. Chia, S.-F. Chin, C. Curtis, O. M. Rueda, P. D. Pharoah, S. Damaraju, J. Mackey, K. Hoon, T. Harkins, V. Tadigotla, M. Sigaroudinia, P. Gascard, T. Tlsty, J. F. Costello, I. M. Meyer, C. J. Eaves, W. W. Wasserman, S. Jones, D. Huntsman, M. Hirst, C. Caldas, M. A. Marra,

and S. Aparicio, *The clonal and mutational evolution spectrum of primary triple-negative breast cancers,* Nature **486**, 395 (2012).

[21] A. Bashashati, G. Haffari, J. Ding, G. Ha, K. Lui, J. Rosner, D. G. Huntsman, C. Caldas, S. A. Aparicio, and S. P. Shah, *DriverNet: uncovering the impact of somatic driver mutations on transcriptional networks in cancer,* Genome Biology **13**, R124 (2012).

[22] C. Dong, Y. Guo, H. Yang, Z. He, X. Liu, and K. Wang, *iCAGES: integrated CAncer GEnome Score for comprehensively prioritizing driver genes in personal cancer genomes,* Genome Medicine **8**, 135 (2016).

[23] D. Tamborero, N. Lopez-Bigas, and A. Gonzalez-Perez, *Oncodrive-CIS: A Method to Reveal Likely Driver Genes Based on the Impact of Their Copy Number Changes on Expression,* PLoS ONE **8**, e55489 (2013).

[24] U. D. Akavia, O. Litvin, J. Kim, F. Sanchez-Garcia, D. Kotliar, H. C. Causton, P. Pochanard, E. Mozes, L. A. Garraway, and D. Pe'er, *An Integrated Approach to Uncover Drivers of Cancer,* Cell **143**, 1005 (2010).

[25] P. Dhingra, A. Martinez-Fundichely, A. Berger, F. W. Huang, A. N. Forbes, E. M. Liu, D. Liu, A. Sboner, P. Tamayo, D. S. Rickman, M. A. Rubin, and E. Khurana, *Identification of novel prostate cancer drivers using RegNetDriver: a framework for integration of genetic and epigenetic alterations with tissue-specific regulatory network,* Genome Biology **18**, 141 (2017).

[26] E. Cerami, E. Demir, N. Schultz, B. S. Taylor, and C. Sander, *Automated Network Analysis Identifies Core Pathways in Glioblastoma,* PLoS ONE **5**, e8918 (2010).

[27] A. Cho, J. E. Shim, E. Kim, F. Supek, B. Lehner, and I. Lee, *MUFFINN: cancer gene discovery via network analysis of somatic mutation data,* Genome Biology **17**, 129 (2016).

[28] M. A. Reyna, M. D. M. Leiserson, and B. J. Raphael, *Hierarchical HotNet: identifying hierarchies of altered subnetworks,* Bioinformatics **34**, i972 (2018).

[29] D. Bertrand, K. R. Chng, F. G. Sherbaf, A. Kiesel, B. K. H. Chia, Y. Y. Sia, S. K. Huang, D. S. Hoon, E. T. Liu, A. Hillmer, and N. Nagarajan, *Patient-specific driver gene prediction and risk assessment through integrated network analysis of cancer omics profiles,* Nucleic Acids Research **43**, e44 (2015).

[30] T. N. Cuykendall, M. A. Rubin, and E. Khurana, *Non-coding genetic variation in cancer,* Current Opinion in Systems Biology **1**, 9 (2017).

[31] E. Perenthaler, S. Yousefi, E. Niggl, and T. S. Barakat, *Beyond the Exome: The Non-coding Genome and Enhancers in Neurodevelopmental Disorders and Malformations of Cortical Development,* Frontiers in Cellular Neuroscience **13** (2019), 10.3389/fncel.2019.00352.

[32] J. Vinagre, A. Almeida, H. Pópulo, R. Batista, J. Lyra, V. Pinto, R. Coelho, R. Celestino, H. Prazeres, L. Lima, M. Melo, A. G. da Rocha, A. Preto, P. Castro, L. Castro, F. Pardal, J. M. Lopes, L. L. Santos, R. M. Reis, J. Cameselle-Teijeiro, M. Sobrinho-Simões, J. Lima, V. Máximo, and P. Soares, *Frequency of TERT promoter mutations in human cancers,* Nature Communications **4**, 2185 (2013).

[33] G. R. S. Ritchie, I. Dunham, E. Zeggini, and P. Flicek, *Functional annotation of noncoding sequence variants,* Nature Methods **11**, 294 (2014).

[34] A. R. Soltis, C. L. Dalgard, H. B. Pollard, and M. D. Wilkerson, *MutEnricher: a flexible toolset for somatic mutation enrichment analysis of tumor whole genomes,* BMC Bioinformatics **21**, 338 (2020).

[35] W. McLaren, L. Gil, S. E. Hunt, H. S. Riat, G. R. S. Ritchie, A. Thormann, P. Flicek, and F. Cunningham, *The Ensembl Variant Effect Predictor,* Genome Biology **17**, 122 (2016).

[36] J. Zhou and O. G. Troyanskaya, *Predicting effects of noncoding variants with deep learning–based sequence model,* Nature Methods **12**, 931 (2015).

[37] J. Zhou, C. L. Theesfeld, K. Yao, K. M. Chen, A. K. Wong, and O. G. Troyanskaya, *Deep learning sequence-based ab initio prediction of variant effects on expression and disease risk,* Nature Genetics **50**, 1171 (2018).

[38] M. Spielmann, D. G. Lupiáñez, and S. Mundlos, *Structural variation in the 3D genome,* Nature Reviews Genetics **19**, 453 (2018).

[39] D. Shlyueva, G. Stampfel, and A. Stark, *Transcriptional enhancers: from properties to genome-wide predictions,* Nature Reviews Genetics **15**, 272 (2014).

[40] Q. Cao, C. Anyansi, X. Hu, L. Xu, L. Xiong, W. Tang, M. T. S. Mok, C. Cheng, X. Fan, M. Gerstein, A. S. L. Cheng, and K. Y. Yip, *Reconstruction of enhancer–target networks in 935 samples of human primary cells, tissues and cell lines,* Nature Genetics **49**, 1428 (2017).

[41] S. Schoenfelder, M. Furlan-Magaril, B. Mifsud, F. Tavares-Cadete, R. Sugar, B.-M. Javierre, T. Nagano, Y. Katsman, M. Sakthidevi, S. W. Wingett, E. Dimitrova, A. Dimond, L. B. Edelman, S. Elderkin, K. Tabbada, E. Darbo, S. Andrews, B. Herman, A. Higgs, E. LeProust, C. S. Osborne, J. A. Mitchell, N. M. Luscombe, and P. Fraser, *The pluripotent regulatory circuitry connecting promoters to their long-range interacting elements,* Genome Research **25**, 582 (2015).

[42] N. Naumova, E. M. Smith, Y. Zhan, and J. Dekker, *Analysis of long-range chromatin interactions using Chromosome Conformation Capture,* Methods **58**, 192 (2012).

[43] D. U. Gorkin, D. Leung, and B. Ren, *The 3D Genome in Transcriptional Regulation and Pluripotency,* Cell Stem Cell **14**, 762 (2014).

[44] P. H. L. Krijger and W. de Laat, *Regulation of disease-associated gene expression in the 3D genome,* Nature Reviews Molecular Cell Biology **17**, 771 (2016).

[45] M. J. Rowley and V. G. Corces, *Organizational principles of 3D genome architecture,* Nature Reviews Genetics **19**, 789 (2018).

[46] J. R. Dixon, S. Selvaraj, F. Yue, A. Kim, Y. Li, Y. Shen, M. Hu, J. S. Liu, and B. Ren, *Topological domains in mammalian genomes identified by analysis of chromatin interactions,* Nature **485**, 376 (2012).

[47] B. Bonev and G. Cavalli, *Organization and function of the 3D genome,* Nature Reviews Genetics **17**, 661 (2016).

[48] G. Fudenberg, M. Imakaev, C. Lu, A. Goloborodko, N. Abdennur, and L. A. Mirny, *Formation of Chromosomal Domains by Loop Extrusion,* Cell Reports **15**, 2038 (2016).

[49] A. Despang, R. Schöpflin, M. Franke, S. Ali, I. Jerković, C. Paliou, W.-L. Chan, B. Timmermann, L. Wittler, M. Vingron, S. Mundlos, and D. M. Ibrahim, *Functional dissection of the Sox9–Kcnj2 locus identifies nonessential and instructive roles of TAD architecture,* Nature Genetics **51**, 1263 (2019).

[50] J. E. Phillips-Cremins, M. E. Sauria, A. Sanyal, T. I. Gerasimova, B. R. Lajoie, J. S. Bell, C.-T. Ong, T. A. Hookway, C. Guo, Y. Sun, M. J. Bland, W. Wagstaff, S. Dalton, T. C. McDevitt, R. Sen, J. Dekker, J. Taylor, and V. G. Corces, *Architectural Protein Subclasses Shape 3D Organization of Genomes during Lineage Commitment,* Cell **153**, 1281 (2013).

[51] E. M. Liu, A. Martinez-Fundichely, B. J. Diaz, B. Aronson, T. Cuykendall, M. MacKay, P. Dhingra, E. W. Wong, P. Chi, E. Apostolou, N. E. Sanjana, and E. Khurana, *Identification of Cancer Drivers at CTCF Insulators in 1,962 Whole Genomes,* Cell Systems **8**, 446 (2019).

[52] D. Hnisz, A. S. Weintraub, D. S. Day, A.-L. Valton, R. O. Bak, C. H. Li, J. Goldmann, B. R. Lajoie, Z. P. Fan, A. A. Sigova, J. Reddy, D. Borges-Rivera, T. I. Lee, R. Jaenisch, M. H. Porteus, J. Dekker, and R. A. Young, *Activation of proto-oncogenes by disruption of chromosome neighborhoods,* Science **351**, 1454 (2016).

[53] D. G. Lupiáñez, K. Kraft, V. Heinrich, P. Krawitz, F. Brancati, E. Klopocki, D. Horn, H. Kayserili, J. M. Opitz, R. Laxova, F. Santos-Simarro, B. Gilbert-Dussardier, L. Wittler, M. Borschiwer, S. A. Haas, M. Osterwalder, M. Franke, B. Timmermann, J. Hecht, M. Spielmann, A. Visel, and S. Mundlos, *Disruptions of Topological Chromatin Domains Cause Pathogenic Rewiring of Gene-Enhancer Interactions,* Cell **161**, 1012 (2015).

[54] E. Giorgio, D. Robyr, M. Spielmann, E. Ferrero, E. Di Gregorio, D. Imperiale, G. Vaula, G. Stamoulis, F. Santoni, C. Atzori, L. Gasparini, D. Ferrera, C. Canale, M. Guipponi, L. A. Pennacchio, S. E. Antonarakis, A. Brussino, and A. Brusco, *A large genomic deletion leads to enhancer adoption by the lamin B1 gene: a second path to autosomal dominant adult-onset demyelinating leukodystrophy (ADLD),* Human Molecular Genetics **24**, 3143 (2015).

[55] C. Redin, H. Brand, R. L. Collins, T. Kammin, E. Mitchell, J. C. Hodge, C. Hanscom, V. Pillalamarri, C. M. Seabra, M.-A. Abbott, O. A. Abdul-Rahman, E. Aberg, R. Adley, S. L. Alcaraz-Estrada, F. S. Alkuraya, Y. An, M.-A. Anderson, C. Antolik, K. Anyane-Yeboa, J. F. Atkin, T. Bartell, J. A. Bernstein, E. Beyer, I. Blumenthal, E. M. H. F. Bongers, E. H. Brilstra, C. W. Brown, H. T. Brüggenwirth, B. Callewaert, C. Chiang, K. Corning, H. Cox, E. Cuppen, B. B. Currall, T. Cushing, D. David, M. A. Deardorff, A. Dheedene, M. D'Hooghe, B. B. A. de Vries, D. L. Earl, H. L. Ferguson, H. Fisher, D. R. FitzPatrick, P. Gerrol, D. Giachino, J. T. Glessner, T. Gliem, M. Grady, B. H. Graham, C. Griffis, K. W. Gripp, A. L. Gropman, A. Hanson-Kahn, D. J. Harris, M. A. Hayden, R. Hill, R. Hochstenbach, J. D. Hoffman, R. J. Hopkin, M. W. Hubshman, A. M. Innes, M. Irons, M. Irving, J. C. Jacobsen, S. Janssens, T. Jewett, J. P. Johnson, M. C. Jongmans, S. G. Kahler, D. A. Koolen, J. Korzelius, P. M. Kroisel, Y. Lacassie, W. Lawless, E. Lemyre, K. Leppig, A. V. Levin, H. Li, H. Li, E. C. Liao, C. Lim, E. J. Lose, D. Lucente, M. J. Macera, P. Manavalan, G. Mandrile, C. L. Marcelis, L. Margolin, T. Mason, D. Masser-Frye, M. W. McClellan, C. J. Z. Mendoza, B. Menten, S. Middelkamp, L. R. Mikami, E. Moe, S. Mohammed, T. Mononen, M. E. Mortenson, G. Moya, A. W. Nieuwint, Z. Ordulu, S. Parkash, S. P. Pauker, S. Pereira, D. Perrin, K. Phelan, R. E. P. Aguilar, P. J. Poddighe, G. Pregno, S. Raskin, L. Reis, W. Rhead, D. Rita, I. Renkens, F. Roelens, J. Ruliera, P. Rump, S. L. P. Schilit, R. Shaheen, R. Sparkes, E. Spiegel, B. Stevens, M. R. Stone, J. Tagoe, J. V. Thakuria, B. W. van Bon, J. van de Kamp, I. van Der Burgt, T. van Essen, C. M. van Ravenswaaij-Arts, M. J. van Roosmalen, S. Vergult, C. M. L. Volker-Touw, D. P. Warburton, M. J. Waterman, S. Wiley, A. Wilson, M. d. l. C. A. Yerena-de Vega, R. T. Zori, B. Levy, H. G. Brunner, N. de Leeuw, W. P. Kloosterman, E. C. Thorland, C. C. Morton, J. F. Gusella, and M. E. Talkowski, *The genomic landscape of balanced cytogenetic abnormalities associated with human congenital anomalies,* Nature Genetics **49**, 36 (2017).

[56] M. Franke, D. M. Ibrahim, G. Andrey, W. Schwarzer, V. Heinrich, R. Schöpflin, K. Kraft, R. Kempfer, I. Jerković, W.-L. Chan, M. Spielmann, B. Timmermann, L. Wittler, I. Kurth, P. Cambiaso, O. Zuffardi, G. Houge, L. Lambie, F. Brancati, A. Pombo, M. Vingron, F. Spitz, and S. Mundlos, *Formation of new chromatin domains determines pathogenicity of genomic duplications,* Nature **538**, 265 (2016).

[57] Y. Zhang, L. Yang, M. Kucherlapati, F. Chen, A. Hadjipanayis, A. Pantazi, C. A. Bristow, E. A. Lee, H. S. Mahadeshwar, J. Tang, J. Zhang, S. Seth, S. Lee, X. Ren, X. Song, H. Sun, J. Seidman, L. J. Luquette, R. Xi, L. Chin, A. Protopopov, W. Li, P. J. Park, R. Kucherlapati, and C. J. Creighton, *A Pan-Cancer Compendium of Genes Deregulated by Somatic Genomic Rearrangement across More Than 1,400 Cases,* Cell Reports **24**, 515 (2018).

[58] J. Weischenfeldt, T. Dubash, A. P. Drainas, B. R. Mardin, Y. Chen, A. M. Stütz, S. M. Waszak, G. Bosco, A. R. Halvorsen, B. Raeder, T. Efthymiopoulos, S. Erkek, C. Siegl, H. Brenner, O. T. Brustugun, S. M. Dieter, P. A. Northcott, I. Petersen, S. M. Pfister, M. Schneider, S. K. Solberg, E. Thunissen, W. Weichert, T. Zichner, R. Thomas, M. Peifer, A. Helland, C. R. Ball, M. Jechlinger, R. Sotillo, H. Glimm, and J. O. Korbel, *Pan-cancer analysis of somatic copy-number alterations implicates IRS4 and IGF2 in enhancer hijacking,* Nature Genetics **49**, 65 (2017).

[59] J. R. Dixon, J. Xu, V. Dileep, Y. Zhan, F. Song, V. T. Le, G. G. Yardımcı, A. Chakraborty, D. V. Bann, Y. Wang, R. Clark, L. Zhang, H. Yang, T. Liu, S. Iyyanki, L. An, C. Pool, T. Sasaki, J. C. Rivera-Mulia, H. Ozadam, B. R. Lajoie, R. Kaul, M. Buckley, K. Lee, M. Diegel, D. Pezic, C. Ernst, S. Hadjur, D. T. Odom, J. A. Stamatoyannopoulos, J. R. Broach, R. C. Hardison, F. Ay, W. S. Noble, J. Dekker, D. M. Gilbert, and F. Yue, *Integrative detection and analysis of structural variation in cancer genomes,* Nature Genetics **50**, 1388 (2018).

[60] A.-L. Valton and J. Dekker, *TAD disruption as oncogenic driver,* Current Opinion in Genetics and Development **36**, 34 (2016).

[61] K. C. Akdemir, V. T. Le, S. Chandran, Y. Li, R. G. Verhaak, R. Beroukhim, P. J. Campbell, L. Chin, J. R. Dixon, and P. A. Futreal, *Disruption of chromatin folding domains by somatic genomic rearrangements in human cancer,* Nature Genetics **52**, 294 (2020).

[62] A. Marusyk and K. Polyak, *Tumor heterogeneity: Causes and consequences,* Biochimica et Biophysica Acta (BBA) - Reviews on Cancer **1805**, 105 (2010).

[63] J. Lobo, A. Gillis, C. Jerónimo, R. Henrique, and L. Looijenga, *Human Germ Cell Tumors are Developmental Cancers: Impact of Epigenetics on Pathobiology and Clinic,* International Journal of Molecular Sciences **20**, 258 (2019).

[64] M. J. Williams, A. Sottoriva, and T. A. Graham, *Measuring Clonal Evolution in Cancer with Genomics,* Annual Review of Genomics and Human Genetics **20**, 309 (2019).

[65] P. Nowell, *The clonal evolution of tumor cell populations,* Science **194**, 23 (1976).

[66] M. Greaves and C. C. Maley, *Clonal evolution in cancer,* Nature **481**, 306 (2012).

[67] F. Janku, *Tumor heterogeneity in the clinic: is it a real problem?* Therapeutic Advances in Medical Oncology **6**, 43 (2014).

[68] X. Zhou, Q. Hao, and H. Lu, *Mutant p53 in cancer therapy—the barrier or the path,* Journal of Molecular Cell Biology **11**, 293 (2019).

[69] I. Dagogo-Jack and A. T. Shaw, *Tumour heterogeneity and resistance to cancer therapies,* Nature Reviews Clinical Oncology **15**, 81 (2018).

[70] S. Turajlic, N. McGranahan, and C. Swanton, *Inferring mutational timing and reconstructing tumour evolutionary histories,* Biochimica et Biophysica Acta (BBA) - Reviews on Cancer **1855**, 264 (2015).

[71] S. Malikic, K. Jahn, J. Kuipers, S. C. Sahinalp, and N. Beerenwinkel, *Integrative inference of subclonal tumour evolution from single-cell and bulk sequencing data,* Nature Communications **10**, 2750 (2019).

[72] K. Tomlinson and L. Oesper, *Parameter, noise, and tree topology effects in tumor phylogeny inference,* BMC Medical Genomics **12**, 184 (2019).

[73] W. Jiao, S. Vembu, A. G. Deshwar, L. Stein, and Q. Morris, *Inferring clonal evolution of tumors from single nucleotide somatic mutations.* BMC bioinformatics **15**, 35 (2014), arXiv:1210.3384 .

[74] D. Lee, Y. Park, and S. Kim, *Towards multi-omics characterization of tumor heterogeneity: a comprehensive review of statistical and machine learning approaches,* Briefings in Bioinformatics (2020), 10.1093/bib/bbaa188.

[75] F. Strino, F. Parisi, M. Micsinai, and Y. Kluger, *TrAp: a tree approach for fingerprinting subclonal tumor composition,* Nucleic Acids Research **41**, e165 (2013).

[76] A. Roth, J. Khattra, D. Yap, A. Wan, E. Laks, J. Biele, G. Ha, S. Aparicio, A. Bouchard-Côté, and S. P. Shah, *PyClone: statistical inference of clonal population structure in cancer,* Nature Methods **11**, 396 (2014).

[77] C. a. Miller, B. S. White, N. D. Dees, M. Griffith, J. S. Welch, O. L. Griffith, R. Vij, M. H. Tomasson, T. a. Graubert, M. J. Walter, M. J. Ellis, W. Schierding, J. F. DiPersio, T. J. Ley, E. R. Mardis, R. K. Wilson, and L. Ding, *SciClone: Inferring Clonal Architecture and Tracking the Spatial and Temporal Patterns of Tumor Evolution,* PLoS Computational Biology **10**, e1003665 (2014).

[78] M. El-Kebir, L. Oesper, H. Acheson-Field, and B. J. Raphael, *Reconstruction of clonal trees and tumor composition from multi-sample sequencing data,* Bioinformatics **31**, i62 (2015).

[79] V. Popic, R. Salari, I. Hajirasouliha, D. Kashef-Haghighi, R. B. West, and S. Batzoglou, *Fast and scalable inference of multi-sample cancer lineages,* Genome Biology **16**, 91 (2015).

[80] L. Oesper, A. Mahmoody, and B. J. Raphael, *Inferring intra-tumor heterogeneity from high-throughput DNA sequencing data,* Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) **7821 LNBI**, 171 (2013).

[81] G. Ha, A. Roth, J. Khattra, J. Ho, D. Yap, L. M. Prentice, N. Melnyk, A. McPherson, A. Bashashati, E. Laks, J. Biele, J. Ding, A. Le, J. Rosner, K. Shumansky, M. a. Marra, C. B. Gilks, D. G. Huntsman, J. N. McAlpine, S. Aparicio, and S. P. Shah, *TITAN: inference of copy number architectures in clonal cell populations from tumor whole-genome sequence data.* Genome research **24**, 1881 (2014).

[82] Z. Yu, A. Li, and M. Wang, *CLImAT-HET: detecting subclonal copy number alterations and loss of heterozygosity in heterogeneous tumor samples from whole-genome sequencing data,* BMC Medical Genomics **10**, 15 (2017).

[83] Y. Li and X. Xie, *Deconvolving tumor purity and ploidy by integrating copy number alterations and loss of heterozygosity,* Bioinformatics **30**, 2121 (2014).

[84] R. F. Schwarz, A. Trinh, B. Sipos, J. D. Brenton, N. Goldman, and F. Markowetz, *Phylogenetic Quantification of Intra-tumour Heterogeneity,* PLoS Computational Biology **10**, e1003535 (2014).

[85] A. G. Deshwar, S. Vembu, C. K. Yung, G. Jang, L. Stein, and Q. Morris, *PhyloWGS: Reconstructing subclonal composition and evolution from whole-genome sequencing of tumors,* Genome Biology **16**, 35 (2015).

[86] W. M. Ismail, E. Nzabarushimana, and H. Tang, *Algorithmic approaches to clonal reconstruction in heterogeneous cell populations,* Quantitative Biology **7**, 255 (2019).

[87] S. Malikic, A. W. McPherson, N. Donmez, and C. S. Sahinalp, *Clonality inference in multiple tumor samples using phylogeny,* Bioinformatics **31**, 1349 (2015).

[88] H. Zare, J. Wang, A. Hu, K. Weber, J. Smith, D. Nickerson, C. Song, D. Witten, C. A. Blau, and W. S. Noble, *Inferring Clonal Composition from Multiple Sections of a Breast Cancer,* PLoS Computational Biology **10**, e1003703 (2014).

2

TargetClone: a multi-sample approach for reconstructing subclonal evolution of tumors

Marleen M. Nieboer, Lambert C. J. Dorssers, Roy Straver, Leendert H. J. Looijenga, Jeroen de Ridder

Adapted from PloS One 13.11 (2018). https://doi.org/10.1371/journal.pone.0208002

Abstract

Most tumors are composed of a heterogeneous population of subclones. A more detailed insight into the subclonal evolution of these tumors can be helpful to study progression and treatment response. Problematically, tumor samples are typically very heterogeneous, making deconvolving individual tumor subclones a major challenge. To overcome this limitation, reducing heterogeneity, such as by means of microdissections, coupled with targeted sequencing, is a viable approach. However, computational methods that enable reconstruction of the evolutionary relationships require unbiased read depth measurements, which are commonly challenging to obtain in this setting. We introduce TargetClone, a novel method to reconstruct the subclonal evolution tree of tumors from single-nucleotide polymorphism allele frequency and somatic single-nucleotide variant measurements. Furthermore, our method infers copy numbers, alleles and the fraction of the tumor component in each sample. TargetClone was specifically designed for targeted sequencing data obtained from microdissected samples. We demonstrate that our method obtains low error rates on simulated data. Additionally, we show that our method is able to reconstruct expected trees in a testicular germ cell cancer and ovarian cancer dataset. The TargetClone package including tree visualization is written in Python and is publicly available at https://github.com/UMCUGenetics/targetclone.

Introduction

Tumors develop from the accumulation of somatic mutations over time. In a tumor, often various subclonal populations with (partially) overlapping mutation patterns coexist. These subclones are formed through an evolutionary process [1–3]. Reconstructing the subclonal evolution is important, as it can assist in characterizing the mutations driving tumor development and progression, and can be helpful to decipher the mechanisms underlying treatment response [4, 5].

A number of algorithms have been developed to reconstruct subclonal evolution trees from rapidly emerging next-generation sequencing data (Fig S1). The existing methods can coarsely be divided into two categories, those based on somatic single-nucleotide variants (SNVs) and those based on somatic copy number variations (CNVs). Somatic SNV-based methods, such as LICHeE, PhyloSub, TrAp and AncesTree, are most often based on two important assumptions; the sum-rule assumption and infinite sites assumption (ISA) [6–9]. Based on the sum rule, a branched tree, rather than a linear tree, can be ruled out if the sum of the variant allele frequency (VAF) of SNVs in the child subclones is larger than the VAF of SNVs in the parent [7]. Under the ISA, somatic SNVs are not expected to be gained twice independently. Furthermore, somatic SNVs are not expected to be lost once gained. An important limitation is that the VAF is affected by CNVs. As a result, SNV-based methods are restricted to using somatic SNVs in copy number-neutral regions. To overcome potential loss of information due to these restrictions, alternative methods, such as CNTMD, ThetA, TITAN, MEDICC, CloneCNA and CLImAT-HET, have been developed that aim to either infer the copy numbers of subclones, or reconstruct (subclonal) evolution trees from CNVs inferred from e.g. read depth information [10–15]. Additionally, the PhyloWGS algorithm combines somatic SNVs and CNVs to further increase the tree reconstruction accuracy [16]. However, us-

ing read depth to determine the copy number of individual subclones in heterogeneous tumor populations is a challenging problem, as such populations consist of several subclones and non-tumor cells mixed in different unknown fractions [3, 15, 17]. It is therefore hard to distinguish between CNVs and differences in subclonal fraction, and multiple combinations of subclonal fraction and subclonal CNVs may explain the overall read depth profile.

While single-cell sequencing approaches largely mitigate the problem of sample heterogeneity, it is currently not yet possible to sample accurate representations of the entire subclonal diversity using these techniques [18–20]. Therefore, an interesting alternative is to perform microdissections to obtain multiple samples of the same tumor (Fig S2), while at the same time reducing sample heterogeneity [21–23]. However, the typical low read depth of whole genome sequencing (WGS) data complicates the inference of somatic SNVs and CNVs in any sample, and in microdissections in particular [16, 24, 25]. Targeted sequencing-based approaches, including whole exome sequencing (WES), have resulted in a higher coverage, but lead to variable and biased read depth across the genome that may limit accurate detection of CNVs [17, 26–31]. Currently, no methods exist that can be used to unravel subclonality directly from the uncorrected read depth data measured with targeted sequencing. Here, we present TargetClone, a method to reconstruct subclonal evolution of tumors from only SNP allele frequencies and somatic SNVs, which does not rely on read depth or CNVs and thus does not require additional corrections. TargetClone is geared towards inferring trees from targeted sequencing data from microdissected samples.

TargetClone is mainly based on three assumptions. First, it assumes that the input samples contain one major tumor subclone, which have for example been acquired through microdissection as was discussed in the previous paragraph. Contamination with other subclones is allowed, as long as one subclone is dominant in the sample. Second, due to the existence of evolutionary relationships between all subclones in a tumor sample, the subclones are expected to exhibit (partial) overlap in their mutation patterns [6, 9, 32]. In combination with the assumption that somatic mutations accumulate over time and are not lost, we assume that subclones with major overlapping mutation patterns are more closely related than subclones with very distinct mutation patterns (vertical dependency) (Fig 1A) [7, 8]. Thus, we can add direction to the subclonal evolution trees, as the parent of a subclone should have a smaller set of mutations. Third, as our method aims to reconstruct evolutionary trees, we integrate the horizontal dependency assumption to more accurately estimate evolutionary distances between subclones as was previously described in MEDICC [13]. The horizontal dependency works by assuming that two adjacent measurements on the genome are likely dependent, and thus have a high probability of being affected by the same CNV event (Fig 1A).

We demonstrate the performance of our method on simulated data and two real data cases. The first real dataset consists of four type II non-seminomatous (NS) Testicular Germ Cell Cancers (TGCC) with intrinsic resistance to chemotherapy [33]. For each tumor, multiple histological elements have been macro- and microdissected. Allele frequencies (AF) and somatic SNVs were measured with targeted sequencing [23]. Second, we aimed to demonstrate that TargetClone can also be applied to another tumor type. Thus, we ran our method on a dataset consisting of multiple primary tumor

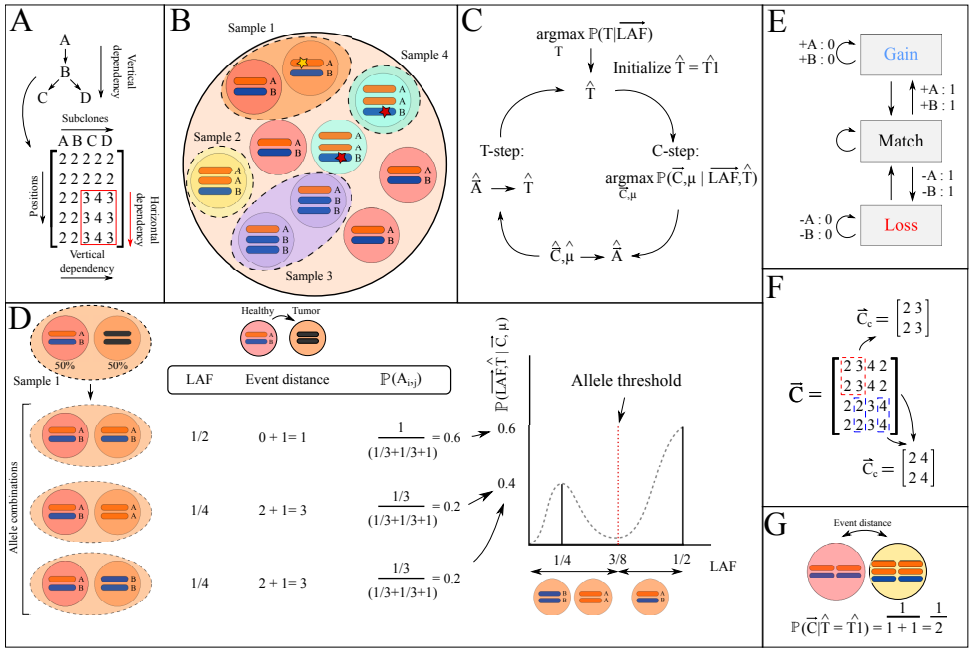


Fig. 1. Overview of the TargetClone methodology. (A) A vertical dependency exists between subclones A-D. In red a horizontally dependent region is highlighted. (B) Multiple subclones with different somatic mutation patterns are sampled from a tumor. Sample 1 contains a mixture of tumor cells and healthy cells, while sample 2, 3 and 4 only contain tumor cells. A star indicates a somatic SNV. (C) General overview of the iterative optimization used in TargetClone. (D) A sample containing healthy cells and a tumor component (each present in 50% of the sample) with a copy number of 2 can be explained by 3 possible scenarios (see left) that each result in a different LAF measurement. Each scenario is scored using the event distance to generate a probability distribution (right). The alleles of the tumor component can be derived from the probability distribution. (E) The FST used to compute the event distance between subclones. Every allele can be gained or lost, which is assigned a distance of 1. If the adjacent position is affected by the same event, the distance is not increased further, which is indicated by the loops to the same state. (F) Two ($\vec{C}_c$) with a different combination of parent and child subclone are highlighted with the blue and red dashed lines. (G) Computation of $\mathbb{P}(\vec{C}|\hat{T})$ for two adjacent alleles. Using the horizontal dependency, the event distance equals 1.

and metastasis samples with reduced heterogeneity of an ovarian cancer patient [34]. In this dataset, the AF were measured using a SNP array, and somatic SNVs were measured using targeted sequencing.

Materials and methods

Definitions

The method accepts m purified samples of the tumor bulk, which can be obtained through e.g. microdissection. As a result of the reduced heterogeneity, we make the assumption that samples consist of one major tumor subclone and are potentially mixed with healthy cells (Fig 1B), although we later show that TargetClone is robust to moderate levels of contamination from other subclones. The fraction of the major tumor subclone

in the sample is denoted as the scalar μ, and hence, the fraction of healthy cells in the sample can be computed as 1 - μ. Each sample can have a different μ.

We assume that the AF have been measured at n heterozygous Single-Nucleotide Polymorphism (SNP) positions in the matched healthy genome that are informative for detecting allelic imbalance. In this text, the term AF measurements will refer to the fraction of the non-reference allele measured at these SNP positions. Furthermore, we assume that in every sample the AF of somatic SNVs have been measured, which will be referred to as somatic SNV measurements. The AF measurements of the SNPs and the AF measurements of the somatic SNVs are used as input to TargetClone.

The AF are represented in a matrix $\vec{AF} = [AF_{i,j}]$, where $AF_{i,j}$ represents the measured AF at SNP position i in subclone j. From the AF measurements, lesser allele frequency (LAF) measurements are computed as 1 - $AF_{i,j}$ for every $AF_{i,j}$ larger than 0.5. The LAF measurements are represented in matrix $L\vec{A}F$, which is in the same format as matrix $\vec{AF}$.

The copy numbers of the subclones can be represented in matrix $\vec{C} = [C_{i,j}]$, where $C_{i,j} \in \mathbb{N}$ represents the copy number of subclone j at AF measurement position i. Consistent with the assumption that every sample may contain healthy cell admixture, the first column of $\vec{C}$ will always contain the copy numbers of healthy cells, which are assumed to be 2 (Fig 1A). Similar to the copy numbers, the alleles of the m samples can be represented in a matrix $\vec{A} = [A_{i,j}]$. $A_{i,j}$ denotes the alleles that are present at this AF measurement position, which will be referred to as allele A (reference) or B (variant). For example, $A_{i,j}$ could be AB or ABB. The total number of alleles equals the copy number at each position. The first column of $\vec{A}$ also represents the alleles of healthy cells, which are assumed to be AB. The rows in $\vec{C}$ and $\vec{A}$ are ordered by AF measurement position on the genome. The ordering of the columns (with the exception of the first column) is arbitrary.

TargetClone outputs estimates of the copy numbers ($\vec{C}$) and alleles ($\vec{A}$), the tumor fraction (μ) per sample, and an estimate of the subclonal evolution tree (T), which describes the relations between the input samples and an estimated distance between these.

Model

The objective of TargetClone is to infer the subclonal evolution tree T from the AF and somatic SNV measurements (Fig. 1C):

$$\arg\max_{T} \mathbb{P}(T|\vec{AF}, S\vec{N}V) \tag{1}$$

Eq 1 is optimized using an iterative heuristic model, consisting of a **T-** and **C-step**:

T-step: a tree $\hat{T}$ is inferred from $\hat{\vec{A}}$, the AF measurements, and somatic SNV measurements. $\hat{\vec{A}}$ can be estimated from $\hat{\vec{C}}$ and $\hat{\mu}$, which are both inferred by the model in the **C-step**.

C-step: we maximize the likelihood of observing $\vec{C}$ and μ given the LAF measurements per sample, which are derived from the AF measurements, and the current esti-

mate of the subclonal evolution tree $\hat{T}$:

$$\arg\max_{\vec{C},\mu} \mathbb{P}(\vec{C},\mu|L\vec{A}F,\hat{T}) \tag{2}$$

The model is initiated with an estimate of the subclonal evolution tree, $\hat{T}1$. By default, we assume that all subclones have a healthy cell as the last known common precursor. Thus, in our initial tree the healthy cell is set as the parent of every tumor subclone. However, starting the model from a different precursor with allele compositions other than AB is also possible.

We demonstrate that starting the model with different initial trees does not affect the results, showing that the method is robust for different starting points.

The **T** and **C steps** are repeated iteratively until $\hat{T}$ has converged. The tree is considered converged when the edges and the total distance between all subclones equals that of a tree that has been reconstructed in any previous iteration.

C-step

Eq 2 can be rewritten as the following using Bayes' rule (see Supplementary Methods for the full derivation):

$$\mathbb{P}(\vec{C},\mu|L\vec{A}F,\hat{T}) \propto \mathbb{P}(L\vec{A}F|\vec{C},\mu,\hat{T})\mathbb{P}(\vec{C}|\ \hat{T}) \tag{3}$$

The computation of $\mathbb{P}(L\vec{A}F|\vec{C},\mu,\hat{T})$ and $\mathbb{P}(\vec{C}|\ \hat{T})$ are explained below.

Computing $\mathbb{P}(L\vec{A}F|\vec{C},\mu,\hat{T})$

For a single measurement position i and some subclone j, $\mathbb{P}(LAF_{i,j}|C_{i,j},\mu,\hat{T})$ is computed by enumerating all possible alleles that can result from the copy number $C_{i,j}$, which can easily be performed for realistic $C_{i,j}$. For example, if $C_{i,j}=2$, the tumor subclone can contain the alleles AA, BB or AB (see Fig 1D), which we will denote as the set Q. Subsequently, the LAF measured at position i in subclone j is computed for every element in Q (formula in Supplementary Methods).

Under the assumption that subclone j is derived from its parent in the current estimate of the tree $\hat{T}$, not all alleles are equally likely to occur. For example, in case a subclone with 4 copies (AABB) is transformed into a subclone with 3 copies, it is more likely to result in ABB, which only requires a loss of one A allele, than BBB, which would require a loss of two A alleles and a gain of one B allele. To quantify this, we assume that the probability of observing $A_{i,j}$ depends on $A_{i,p(j)}$, where *p(j)* denotes the parent of subclone *j*, which is provided in $\hat{T}$, as follows:

$$\mathbb{P}(A_{i,j}|A_{i,p(j)}) = \frac{\frac{1}{ED(A_{i,j},A_{i,p(j)})+1}}{\sum_q^Q \frac{1}{ED(q,A_{i,p(j)})+1}} \tag{4}$$

Here, the event distance (ED) is computed as the total number of alleles that are different between the parent and the subclone at position *i*. A distance of one is counted for every loss or gain of an allele. The total event distance is computed as the sum of the event distance at every position. $\mathbb{P}(A_{i,j}|A_{i,p(j)})$ is normalized based on the event distance to all other alleles in the set Q. A pseudocount of one is added to avoid divisions by zero. In

conclusion, the event distance allows us to distinguish between for example AB or AABB, which both result in the same LAF measurement.

Following a previously published model, we assume that sequencing noise follows a Gaussian distribution [35]. This assumption requires that the sequencing depth is larger than 1000x. We model the overall probability distribution $\mathbb{P}(LAF_{i,j}|C_{i,j},\mu,\hat{T})$ as a Gaussian mixture model (detailed in Supplementary Methods, see Fig 1D), where the means are equal to the LAFs resulting from each allele combination in Q, and the noise component is estimated from the LAF measurements in the normal samples of our real TGCC dataset. The interval of the distribution is limited between 0 and 0.5 to adequately model LAF measurements.

So far, we have only considered a single position i and ignored the fact that a horizontal dependency exists between adjacent measurement positions. To incorporate this dependency, we calculate $\mathbb{P}(\vec{LAF_c}|\vec{C_c},\mu,\hat{T})$. $\vec{C_c}$ is a submatrix of $\vec{C}$, containing $\vec{C_{i,j}}$, $\vec{C_{i+1,j}}$, $\vec{C_{i,p(j)}}$ and $\vec{C_{i+1,p(j)}}$ (see Fig 1F and Fig S3 for a detailed example). $\vec{LAF_c}$ is a submatrix of $\vec{LAF}$, containing the LAF measurements corresponding to the positions in $\vec{C_c}$. $\mathbb{P}(LAF_{i,j}|C_{i,j},\mu,\hat{T})$ is first computed for each copy number in $\vec{C_c}$ individually, which are then multiplied to compute $\mathbb{P}(\vec{LAF_c}|\vec{C_c},\mu,\hat{T})$. Starting from the first two LAF measurement positions, $\vec{C_c}$ is iteratively shifted across $\vec{C}$ one position at a time. $\mathbb{P}(\vec{LAF}|\vec{C},\mu,\hat{T})$ is calculated by taking the product of all $\mathbb{P}(\vec{LAF_c}|\vec{C_c},\mu,\hat{T})$.

Computing $\mathbb{P}(\vec{C}|\ \hat{T})$

Next, we aim to assign a probability to observing a sequence of copy numbers $\vec{C_j}$ in a tumor subclone j given $\hat{T}$. We note that the alleles are more informative for evolutionary distance than the copy numbers (see Fig S7 and Supplementary Results). For instance, if the copy number is 2 in two subclones, we may conclude that the subclones are the same at this position. However, the underlying alleles could be AB and BB, in which case the evolutionary distance is nonzero.

To incorporate the allelic evolutionary distance in the calculation of $\mathbb{P}(\vec{C}|\hat{T})$, we can sum the probability of all alleles that can be generated for a $\vec{C_{i,j}}$ as:

$$\mathbb{P}(C_{i,j}|\hat{T}) = \sum_{q \subset Q} \mathbb{P}(q|\hat{T}) \tag{5}$$

However, from computing $\mathbb{P}(\vec{LAF}|\vec{C},\mu,\hat{T})$ we already know that one element in Q is much more likely than others given our LAF measurements. Thus, we reason that it is possible to approximate $\mathbb{P}(C_{i,j}|\hat{T})$ with the probability of the most likely alleles.

To compute $\mathbb{P}(\vec{C}|\ \hat{T})$, we first obtain the most likely alleles corresponding to $\vec{C_c}$ (described in Section "Deriving the most likely $\vec{A}$ from a combination of $\vec{C}$ and μ"). For these alleles, the Finite State Transducer (FST) shown in Fig 1E is used to compute the event distance that incorporates the horizontal dependency. The FST is used in the MEDICC algorithm for a similar purpose [13]. In the FST, a distance of one is counted for every loss or gain of an allele. In addition, no penalty is given when alleles at adjacent AF measurement positions are affected by the same event. $\mathbb{P}(\vec{C}|\ \hat{T})$ is calculated as the product of the event distance computed for each $\vec{C_c}$ using the FST. Since $\mathbb{P}(\vec{LAF}|\vec{C},\mu,\hat{T})$ and the event distance are inversely proportional, $\mathbb{P}(\vec{C}|\ \hat{T})$ is computed as the reciprocal of the total event distance for $\vec{C}$. Examples of this step are illustrated in Figs 1G and S3.

Maximizing $\mathbb{P}(\vec{C}, \mu | L\vec{A}F, \hat{T})$

Finally, the C-step is completed by inferring a combination of $\vec{C}$ and μ for which $\mathbb{P}(\vec{C}, \mu | L\vec{A}F, \hat{T})$ is maximized. To achieve this, we exhaustively evaluate the values of μ between 0 and 1 in steps of 0.01. For every μ, we vary each copy number in $\vec{C}_c$ from a predefined kmin to kmax and select the copy numbers that maximize $\mathbb{P}(L\vec{A}F_c | \vec{C}_c, \mu, \hat{T})\mathbb{P}(\vec{C}_c | \hat{T})$. $\mathbb{P}(\vec{C}, \mu | L\vec{A}F, \hat{T})$ is computed by taking the product of every $\mathbb{P}(L\vec{A}F_c | \vec{C}_c, \mu, \hat{T})\mathbb{P}(\vec{C}_c | \hat{T})$. The $\vec{C}$ and μ that overall maximize $\mathbb{P}(\vec{C}, \mu | L\vec{A}F, \hat{T})$ are selected as the optimal solution. A more detailed example of how $\mathbb{P}(\vec{C}, \mu | L\vec{A}F, \hat{T})$ is computed for one $\hat{C}_c$ is provided in Fig S3.

Deriving the most likely $\vec{A}$ from a combination of $\vec{C}$ and μ

To derive the alleles most likely corresponding to a LAF measurement, we define a threshold at the average value between each adjacent LAF measurement in $\mathbb{P}(L\vec{A}F | \vec{C}, \mu, \hat{T})$ (Fig 1D). We note that our model is unable to differentiate between the alleles AA and BB. As a result of the low abundance of proximate measurements generated with targeted sequencing, it is not possible to accurately phase alleles. Thus, when computing the horizontal dependency, there is no guarantee that allele A at position *i* is on the same haplotype as allele A at position *i+1*. Therefore, the method will always select the combination with the highest number of B alleles in such ambiguous scenarios.

T-step

Reconstructing *T*

To reconstruct the evolutionary tree T of sampled subclones using the inferred alleles (see Fig 2A for an example of T), we assume that the optimal tree has a minimum event distance between all subclones in the tumor, and thus corresponds to the minimum spanning arborescence (MSA) [13]. Sample by sample distance matrices are generated to describe to relationship between each pair of subclones. The distance matrix D_A (Fig 2B) is constructed by calculating the allelic event distance between all combinations of subclones using the FST (Fig 1E). Distance matrix D_S describes the distances based on somatic SNVs, and initially only contains a value of 1 to indicate that a parental relationship is possible. The values in both matrices may be penalized as discussed below. As the distances based on alleles and somatic SNVs must both agree on a relation between subclones, matrices D_A and D_S are multiplied to generate the final distance matrix D_F. This final distance matrix is used as input to Edmonds' algorithm, which infers an MSA [36].

The inferred alleles and the measured somatic SNVs provide additional information that we can use to restrict or resolve the relations between subclones in the tree.

Restricting and penalizing $\hat{T}$ based on LOH - Edges in $\hat{T}$ can be restricted based on regions with loss of heterozygosity (LOH), as re-gaining lost alleles is highly unlikely (Fig 2C). By default, we consider LOH to be present in a subclone when at least 10 consecutive LAF measurements are smaller than 0.3, and either of the parental alleles has been estimated as lost in $\hat{\vec{A}}$. Both settings can be changed by the user if necessary. In Fig 2D, an example is shown where the LAF measurements are not smaller than 0.3. In this scenario, we cannot confidently decide whether the region shows LOH and that the percentage of normal admixture is high, or if $\hat{\vec{A}}$ is incorrect. Thus, rather than restricting

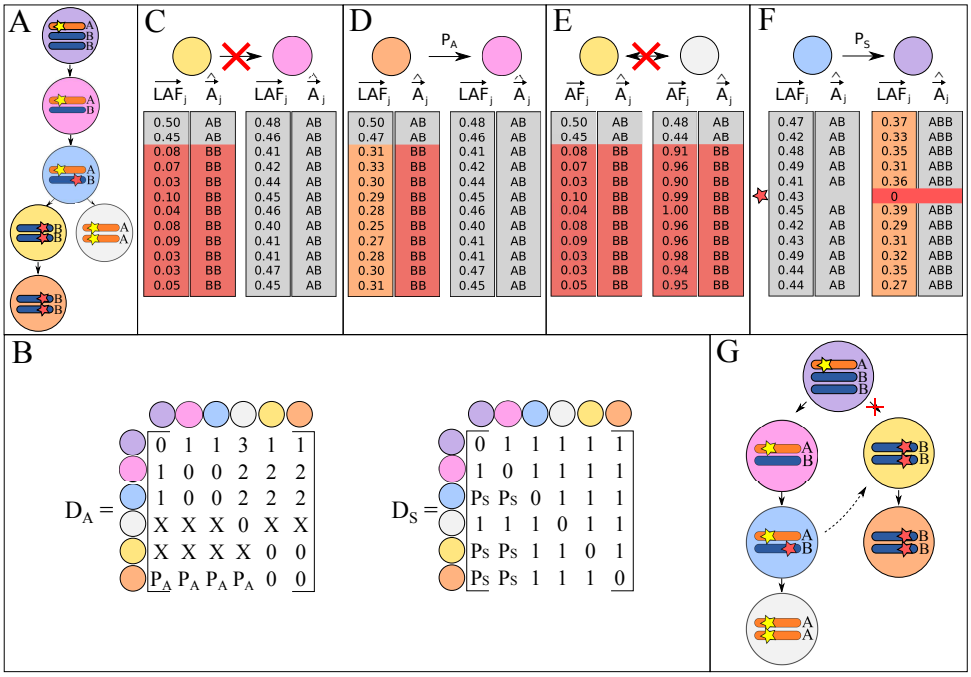


Fig. 2. (A) Example of the true T for 6 hypothetical subclones. (B) Distance matrices reconstructed from the event distance based on alleles (D_A) and somatic SNVs (D_S). 'X' indicates that a subclone cannot be the parent of another subclone. (C) and (D) Edges can be restricted or penalized based on LOH. Each row in the matrix represents a measurement position on the genome. The measured LAF and the $\hat{A}$ inferred by TargetClone are shown in separate columns. A grey color represents a balanced situation, orange allelic imbalance, and red LOH. The first two measurements are not shown in the tree in panel (A). (E) If different parental alleles are lost, edges can be restricted. The ground truth alleles are AA in the grey subclone, but TargetClone will report the alleles as BB. (F) Edges can be penalized if the loss of somatic SNVs is unlikely. (G) Example of the MSA for the subclones shown in (A). The red cross indicates that an edge in the MSA is removed when resolving the ISA. The dashed line indicates a newly added edge after resolving the ISA.

an edge between the subclones, we add a penalty P_A to the current value in D_A.

Restricting $\hat{T}$ based on the loss of different parental alleles - Relations between subclones can also be restricted based on AF measurements. If two subclones contain LOH and have lost a different parental allele, the first subclone cannot be the parent of the second subclone and vice versa (Fig 2E). Although TargetClone cannot distinguish between the parental alleles, we consider a different parental allele to be lost when the AF is lower or higher than 0.1 and 0.9, respectively. These default values may be changed by the user.

Restricting $\hat{T}$ based on somatic SNVs - The edges between subclones can also be restricted based on the measured somatic SNVs. One assumption is that somatic SNVs are typically not lost, unless the allele that these are present on is also lost. If no evidence is present of a lost allele (Fig. 2F), we assign a penalty P_S to these types of relations.

Resolving the ISA by editing the MSA - It often occurs that a MSA is obtained that violates the ISA (Fig 2G). Based on the minimum distance assumption, we reason that

it is possible to use the MSA as a starting point, and perform edit operations until the ISA is no longer violated. To this end, under the assumption that subclones should differ minimally from their parents, we expect that the edge in which the most somatic SNVs are introduced is most likely spurious. In case of a tie, a random edge is selected from the spurious edges. Our method iteratively removes the selected edge from the tree and reruns Edmonds' algorithm on all remaining possible edges between all subclones to infer a new tree until the ISA is resolved. By default, 50 updated trees are generated from the starting MSA, from which the tree with the lowest allelic distance between all subclones is selected as the final solution. 50 trees are explored to prevent the method from getting stuck in a local maximum and thus increases the likelihood that the method generates the same tree for each run.

There are situations in which the ISA may not hold, for example in scenarios where somatic SNVs are drivers of tumor evolution [37], and are therefore expected to independently recur in independent subclones. For this reason, if the ISA cannot be resolved, the edited tree with the fewest violations of the ISA and lowest total distance will be reported. The total distance is computed by taking the sum of all edge weights in the tree, which are obtained from the final distance matrix D_F. Furthermore, we allow the user to select somatic SNVs to be excluded from analysis with TargetClone. Furthermore, the final top 10 trees are visualized using the Bokeh plotting library [38], as described in the Supplementary Methods.

Simulation data

Generation of simulation data

Starting from a healthy, diploid cell, we formed subclones with new somatic SNVs and CNVs for 4 rounds (see Supplementary Methods and Fig S4 A for details on how the simulated data is created). On average, 5 samples are generated, including the healthy cell. The relations between the subclones and precursors decide the ground truth T. All generated subclones and precursors were sampled, which were assigned the same tumor fraction. Selecting the same tumor fraction allows us to additionally test what the effect is of each tumor fraction individually on the performance. In total, per sample, 500 AF/LAF and 50 somatic SNV measurements were generated based on the simulated somatic SNV and CNV profiles to model targeted sequencing data. These measurements were assigned randomly to each chromosome arm, but each chromosome arm on average has an equal number of SNPs.

In our TGCC dataset, we assumed that our sequencing noise is Gaussian distributed, and estimated the standard deviation to be 0.02 in our reference samples. Thus, we selected noise levels of 0.005, 0.01, 0.015, 0.02, 0.025 and 0.03 to represent realistic levels of noise, and 0, 0.04, 0.06, 0.08 and 0.1 representing more extreme sequencing noise levels to test the limits of the method. By default, TargetClone uses a diploid precursor in the initial tree $\hat{T}1$. In Section "TargetClone yields high-quality trees", we also explore the effect on the results if a random precursor ploidy is used.

All results on simulated data discussed in the main text refers to the data generated as described in this section. In addition, we also generated a more realistic simulation dataset closely modelling TGCC data. The generation of this data and related results are discussed in the Supplementary Data.

Computing the error on the simulation data

E_C is the error of $\hat{\vec{C}}$, which is computed as the absolute distance with respect to the true $\vec{C}$, which is normalized for the size of $\vec{C}$. The error in $\hat{\vec{A}}$, E_A, is defined as the average event distance between $\vec{A}$ and $\hat{\vec{A}}$ across all positions. The horizontal dependency is not taken into account in the calculation of the error, as we wish to score the error at each position in $\hat{\vec{A}}$ individually. E_μ, which is the error of $\hat{\mu}$, is computed as the mean absolute error with respect to μ. To test how well ancestry relationships are reconstructed in our trees, we investigated how often parent-child relations were inferred incorrectly. For each pair of samples, we computed how often a parent-child relationship was absent in the inferred tree (false negative) and we also computed how often parent-child relationships were present in the inferred tree, but not in the ground truth tree (false positive). The total tree error, E_T, is calculated as the sum of the number of false positives and false negatives, which is normalized by the total number of sample pairs. The error calculation formulas are provided in the Supplementary Methods.

Results

Simulation data results

To test TargetClone on realistic data for which the ground truth is known, we generated 101 simulation datasets as described in the methods section. Fig 3A-D shows the error of inferring $\vec{C}$, $\vec{A}$, μ and T across the simulations as a function of sequencing noise. The grey shaded areas indicate the mean of the error and 95% confidence interval obtained by running TargetClone on 101 simulation datasets with random data. In each random dataset, a different μ between 0 and 1 was selected. The same AF and somatic SNV measurement positions as in the non-random simulation datasets were selected. At each AF and somatic SNV measurement position, a random AF and somatic SNV measurement between 0 and 1 was selected. As a result, they provide a reference error rate based on the performance of the method by random chance.

TargetClone yields high-quality trees

The error profile of $\hat{\vec{C}}$ and $\hat{\vec{A}}$ reveals that the inference of copy numbers and alleles is highly accurate, in particular in the range of realistic sequencing noise levels. The error rate increases as sequencing noise increases, ultimately reaching the error rate expected by random chance for very high noise levels. The inference of μ is more robust to sequencing noise, indicating that the LAF measurements are still sufficiently informative to estimate μ correctly despite the increase in noise level. Noteably, the error rate of predicting μ correctly by random chance has larger confidence intervals, which results from μ estimates always being in the range of 0.7 - 0.91 in each simulated dataset. Thus, since all μ between 0 and 1 are tested, the error decreases as the true μ of the dataset increases, particularly showing low error rates when the true μ lies within this range of estimated μ.

In Fig S5 we show that re-running TargetClone yields approximately the same results.

To assess the quality of the solution for different initializations, we repeated the optimization for random starting trees ($\hat{T}1$). In these random trees, the relationships between all subclones were selected randomly. For each subclone that was selected as a

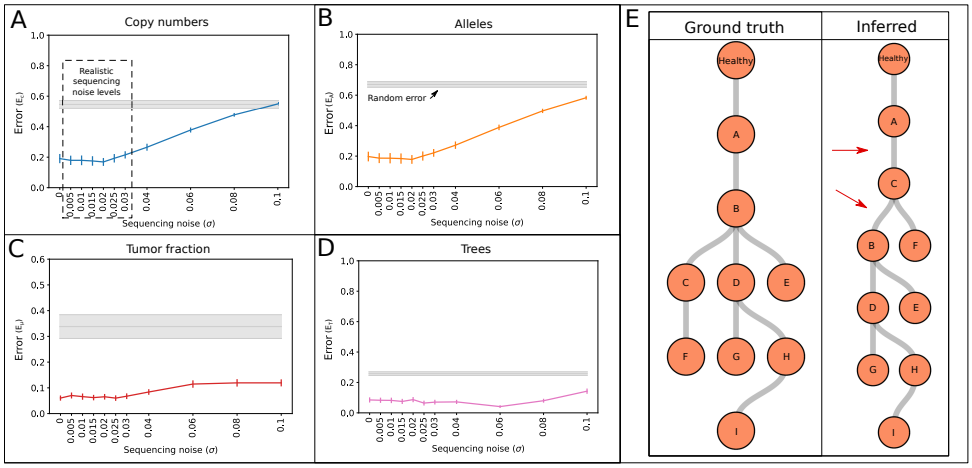


Fig. 3. (A-D) The error of inferring $\vec{C}$, $\vec{A}$, μ and T as a function of sequencing noise. For every noise level, the mean of the error and 95% confidence interval are reported across 101 simulated datasets, each with a unique μ between 0 and 1. The grey shaded areas represent the mean of the error and 95% confidence intervals in 101 simulated datasets where random AF and SNV measurements were selected. (E) Example of a simulated tree (ground truth) compared to the tree inferred by TargetClone. The red arrows indicate incorrectly placed edges.

parent in the random tree, the ploidy of the alleles were selected randomly, which are normally diploid. Fig S6 shows that very similar results are obtained, demonstrating robustness for the initialization of the optimization.

In Fig S7 and the Supplementary Results, we show that combining alleles and somatic SNVs, together with resolving the ISA, yields the largest benefit in reconstructing the trees as compared to when the trees are reconstructed with alleles, copy numbers or somatic SNVs individually. We additionally show that the number and distribution of SNP measurements and the number of measured SNVs does not significantly affect the quality of the inferred copy numbers, alleles and tumor fraction (Supplementary Results, Fig S8 and Fig S9).

Fig 3E shows an inferred tree with two differences with respect to the ground truth tree. Relations B-C and B-F are missed in the inferred tree (false negatives), and relations C-B, C-E, C-D, C-G, C-H and C-I are introduced (false positives). The total number of sample pairs in this tree is 45, and thus the error rate of this tree would be $8/45 = 0.18$. For realistic noise levels, the mean tree error obtained by TargetClone is approximately 0.1. (Fig 3D, see Fig S10 for a figure showing the false positive and false negative rates independently). Clearly, trees with so few errors are useful to investigate subclonal development and yield similar conclusions, despite the few differences with respect to the ground truth.

Tumor fraction is a determinant of error rate

Fig 4A and Fig S11 show that robust performance is measured at realistic and common tumor fractions in microdissected samples [39–41]. For lower tumor fractions, a higher error rate for $\hat{\vec{C}}$ and $\hat{\vec{A}}$ is obtained than for high tumor fractions. Thus, a high amount

of healthy cell contamination, which pushes the LAF measurements towards 0.5, obfuscates information about the tumor subclone. Furthermore, the estimation of T is more accurate at realistic tumor fractions. In short, obtaining high sample tumor fractions benefits subclonal reconstruction accuracy, further justifying the advantage of microdissections.

Ambiguous alleles can be correctly resolved

Many combinations of alleles and tumor fraction give rise to the same LAF. For example, both allele combinations AABB and AB for a μ of 0.5 give rise to a LAF measurement of 0.5. Thus, the exact allele at such a position is impossible to derive based on the LAF measurement of that position alone, and hence is considered ambiguous. In our simulation data, for which the ground truth alleles are known, on average 75% of simulated alleles are ambiguous (Fig 4B).

To investigate the effect of these ambiguities, we aimed to demonstrate how well our method is able to resolve the correct allele. Interestingly, TargetClone is able to infer the correct alleles for around 80% of these ambiguous positions. In part this is due to the assumption of vertical dependency, which ensures alleles in $\vec{A}$ are chosen that minimize the event distance to its parental subclone. To investigate the importance of the presence of the vertical dependency in a dataset for resolving ambiguities, we computed how often the allelic event distance between a subclone and its parent is larger than the distance to any other subclone in a tree. We correlated these values with the number of unresolved ambiguities in the same subclones, and found a Pearson correlation coefficient of 0.23. Thus, we conclude that the ability of TargetClone to resolve ambiguities is not significantly affected by cases where the vertical dependency between the subclones is not as strong.

Second, LOH regions are informative of μ, and as a result greatly restrict the number of possible alleles. For example, a LAF of approximately 0.33 can be measured in a sample with a tumor subclone with alleles ABB or ABBB at one position with tumor fractions of 0.9 and 0.5, respectively. However, if LOH is present at another position, where a LAF of for example 0.09 is measured, the ambiguity is resolved, as this LAF measurement cannot be obtained with a tumor fraction of 0.5 at realistic sequencing noise levels.

It is also important to note that errors in $\hat{\vec{A}}$ resulting from measurement ambiguities may not necessarily negatively affect $\hat{T}$. For example, if the measured LAF is 0.5, it may be explained by multiple combinations of $\vec{A}$ and μ, such as AABB or AB with a μ of 0.5. However, the event distance between two subclones does not change if the alleles are inferred to be AB in both subclones instead of AABB, and thus, no effect is observed on $\hat{T}$ even though an error is made in $\hat{\vec{A}}$. In conclusion, we showed that the assumptions made in our model are sufficient to resolve measurement ambiguities.

TargetClone can reconstruct trees for polyclonal samples

To investigate the effect of multiple co-existing subclones in a sample on the performance of TargetClone, we generated additional simulation datasets with a sequencing noise level of 0.02. The μ of these datasets was fixed at 0.9. As is shown in Fig 4A, the error rates of the method are low with relatively small confidence intervals at this μ, thus allowing us to test the influence of polyclonality at a realistic μ that itself does not largely

influence the results. Each simulated sample consists of one major subclone (at least 50% of the total tumor content), and increasing levels of contamination from random other subclones from the same tumor. We observe that the inference of $\vec{C}$, $\vec{A}$, μ and T is robust to increasing number of subclones (Fig 4C and Fig S12). For T, the error rate at a contamination level between 40 and 50% is as low in samples containing 5 subclones (4 minor subclones contaminating around 10%) as in samples containing 2 subclones (major and minor subclone both present in around 50%). Thus, reducing the total level of contaminating minor subclones yields higher performance improvement than reducing the number of contaminating subclones, which is consistent with our assumption that samples require one major tumor subclone. It has been shown that in practice, microdissected samples can most often indeed contain one major subclone, with relatively small contamination of minor subclones [23].

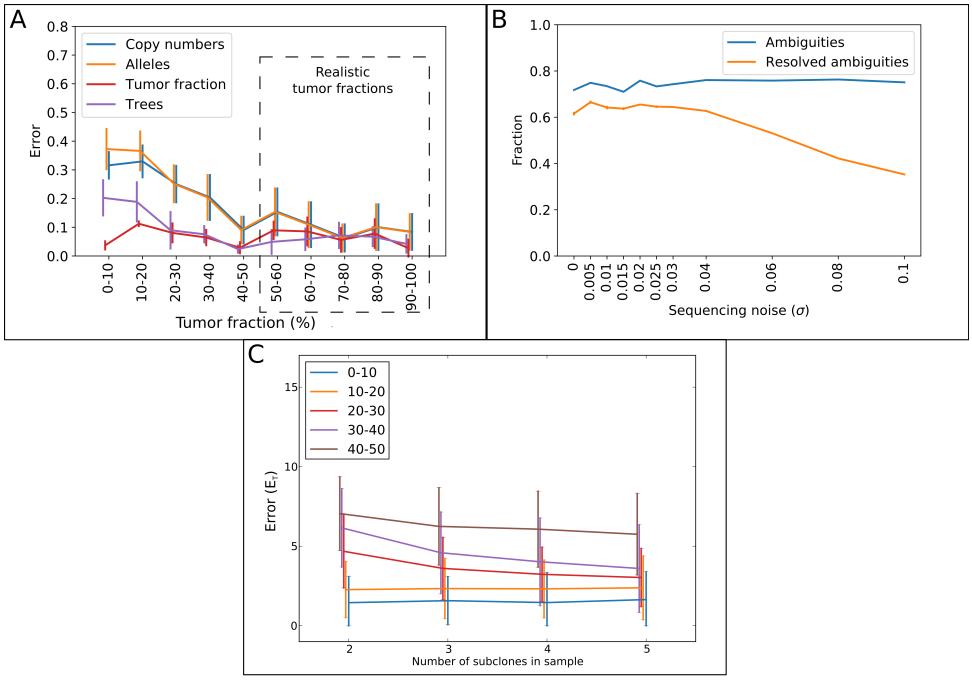


Fig. 4. (A) Mean of the error rates and 95% confidence intervals as a function of different tumor fractions at a sequencing noise level of 0.02. Every μ was tested once. (B) The blue line shows the average fraction of ambiguous alleles that are present in 101 simulated datasets, each with a different tumor fraction between 0 and 1. The orange line indicates the mean and 95% confidence intervals of resolved ambiguities, normalized by the size of $\hat{\vec{A}}$. (C) Mean of the tree reconstruction error rates and 95% confidence intervals as a function of the number of subclones in the sample. A total of 100 simulations were performed for each number of subclones, for each of which a noise level of 0.02 and μ of 0.9 was selected. Each line shows the total percentage of the contaminating minor subclones in each sample. Every contamination percentage within the shown range was tested once.

Real data results

We applied TargetClone to samples from 4 patients with TGCC (NS) with intrinsic resistance to chemotherapy. Multiple histological components were microdissected from each tumor (Fig S2), which were subjected to targeted sequencing [23]. In total, each patient has 9, 6, 18 and 10 samples, with 15, 43, 32 and 31 measured somatic SNVs, and 427, 420, 435 and 407 AF measurements (in patient T6107, T6108, T3209 and T1382, respectively).

The sequencing depth is 1000x on average. Since no ground truth is known for the development of these specific tumors, the results are compared to knowledge previously described in literature (Fig 5A). In summary, TGCC are expected to start development from a tetraploid precursor GCNIS (referred to as CIS in sample names). GCNIS can further develop into NS, which may consist of multiple histological components, including Embryonal Carcinoma (EC), Yolk Sac Tumor (YST), Teratoma (TE) and Embryonal Bodies (EB) [33, 42, 43]. It has been shown that TE and YST can only develop from EC [33, 44, 45].

Based on this knowledge, we defined that in the initial tree $\hat{T}1$, the parent of every subclone is a tetraploid cell, rather than a healthy, diploid cell. Fig 5 shows the inferred subclonal evolution tree for 2 patients, T6107 (Fig 5B) and T618 (Fig 5C). The trees reconstructed for the other 2 patients are shown in Fig S13A (T3209) and Fig S13B (T1382).

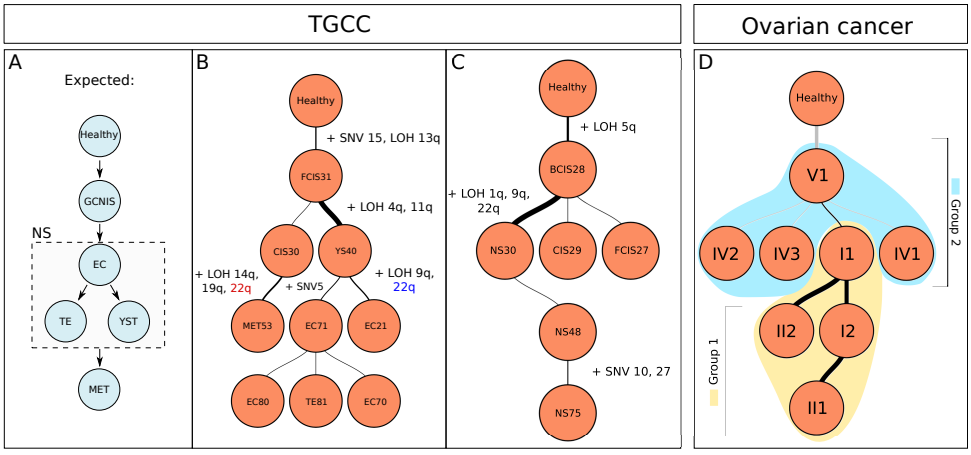


Fig. 5. (A) Expected development of TGCC based on knowledge described in literature. (B) Tree reconstructed by TargetClone for T6107. (C) Tree reconstructed by TargetClone for T618. A few events have been annotated to show the relations between samples. In **(B)**, LOH at chromosome 22q is colored in blue and red to indicate that a different parental allele has been lost. A thicker line indicates that a larger number of events is introduced in the subclone. (D) Tree reconstructed by TargetClone for P1 of the ovarian cancer dataset. The two sample groups are placed in two clusters, as highlighted in yellow and blue. A description of how the trees are visualized can be found in the Supplementary Methods.

On average, the trees for the real data were reconstructed in 30 minutes on 1 CPU core with 12GB of memory.

Case 1: T6107
Fig 5B shows that the predicted evolution tree of T6107 closely resembles the predefined expectations in Fig 5A. Interestingly, samples MET53 and EC21 are correctly placed in different branches. Both samples contain LOH at chromosome 22q, but from the AF it becomes clear that a different parental allele has been lost and thus there exists no direct relation between these samples. Sample MET53 is predicted to have formed from the early precursor CIS30. Sample MET53 lacks all somatic SNVs that have been measured in samples other than CIS30 and FCIS31, and contains a unique pattern of LOH.

The placement of sample YS40 does not correspond to the expectations, as YST can only originate from EC. Nevertheless, YS40 lacks one somatic SNV compared to the EC and TE samples, and thus the ISA cannot be resolved if YS40 is placed elsewhere. As an explanation, it is likely that an unsampled EC subclone existed after FCIS31, which gave rise to YS40, EC21 and the other EC and TE samples.

Case 2: T618
Fig 5C shows the inferred tree for patient T618. CIS is expected to develop into BCIS, which in turn develops into FCIS. FCIS can then develop further into the histological components of NS. From the data, we note an indication that a different parental allele may have been lost at chromosomes 11 and 22 in BCIS28 and FCIS27 and the primary tumor (NS). Thus, it is likely that an unsampled precursor exists that branched into CIS29, FCIS27 and into BCIS28, which then further developed into NS. In our result, sample CIS29 is instead predicted to develop from BCIS28 for two reasons. First of all, LOH is not detected by the model on chromosomes 11 and 22 in BCIS28 and FCIS27 as no 10 consecutive measurements support that LOH. Finally, CIS29 contains additional somatic SNVs that have not been measured in FCIS27 and BCIS28. The primary tumor (NS) has acquired additional mutations, and independent runs of the primary tumor sample (NS48, NS30, NS75) are placed at the bottom of the tree as expected.

The choice of precursor ploidy influences the quality of $\hat{T}$
No proof yet exists for the assumption that TGCC are initiated by genome duplication. To further investigate this question, we also reconstructed evolutionary trees for our TGCC cases with an assumed diploid precursor (Fig S14). The reconstructed tree for T3209 does not follow the biological expectations very well, as sample TE86 cannot be the precursor of EC samples. The total distance between all subclones is higher in the trees generated with a diploid precursor (294, 1054, 3473, 1213 with a diploid precursor and 227, 657, 578, 943 with tetraploid precursor in T618, T6107, T1382 and T3209, respectively). Although the tree for T1382 could not be reliably reconstructed due to high numbers of unsampled subclones and high levels of sequencing noise, and for T618 only a limited number of samples was sequenced, more support is obtained for the assumption that TGCC develop after a duplication of the diploid genome. Although no hard conclusions about precursor ploidy can be drawn from this limited set of samples, the observation that higher distances are obtained and that biological assumptions can be violated when a different initial ploidy is selected, highlights the importance of choosing the correct precursor ploidy. If the ploidy of the precursor is not known, we recommend selecting the ploidy for which the minimum total distance between all subclones in the final tree is reported.

A comparison of TargetClone to existing methods on targeted sequencing data

Finally, we aimed to determine how TargetClone compares to existing tools to reconstruct subclonal evolution trees on targeted sequencing data with microdissected samples. This comparison is challenging, as no method exists that is specifically designed to work with targeted sequencing data from microdissected samples. For this reason, we performed the comparisons under the assumption that one tumor subclone is present per sample. In our comparison we included PhyloWGS, which is currently the only method that combines SNVs and CNVs to infer evolutionary trees (see Fig S1), thus making it the most suitable method to compare with TargetClone. Second, we selected the SNV-only method LICHeE, which infers trees from cellular prevalences estimated with PyClone [46]. Third, we ran LICHeE directly on VAFs to demonstrate the effect of including cellular prevalences. Details on the settings of these methods are described in the Supplementary Methods.

The trees inferred by PhyloWGS, PyClone + LICHeE and LICHeE are provided in Figs S19-S21. Inspection of these trees (described in detail in the Supplementary Results) reveals that none of these trees match with the established knowledge on TGCC development. PhyloWGS appears to miss many subclones and LICHeE fails to detect important relations between subclones that are apparent from LOH patterns. Notably, all of the relations missed by PhyloWGS, PyClone and LICHeE were captured by TargetClone, with the exception of T1382, for which we cannot make a clear statement about the quality of the inferred tree due to the large number of unsampled subclones. Thus, we conclude that the analysis of targeted sequencing data is a difficult task that is not well dealt with by existing methodology. TargetClone, which is tailored to deal with targeted sequencing data, does provide insightful trees containing evolutionary relations that are missed by the currently available tools. These findings are supported by our comparison of TargetClone with existing methods on simulated targeted sequencing data, which is discussed in the Supplementary Results.

TargetClone applied to an ovarian cancer dataset

To determine how well TargetClone performs on another tumor type, we applied it to 8 samples taken from physically separated tumor sites in the abdomen of an ovarian cancer patient [34]. Although these samples were not microdissected, it is shown in the original paper that there exist two sample groups with independent clusters of mutations, and a number of samples contain private mutations with VAF > 0.1. Based on these observations, we expect that the topographic sampling sufficiently reduces heterogeneity to major clones, thus providing an additional test case for TargetClone.

In total, 58 somatic SNVs were measured with targeted sequencing and the AF was measured at approximately 300000 SNP positions using a SNP array. It was previously observed that sample group 1 (I1, I2, II1, II2) and 2 (IV1, IV2, IV3 and V1) contain two clusters of mutations that are mutually exclusive, and we thus expect TargetClone to identify that these groups to have independent origins. Noteably, sample group 2 shares a number of mutations with group 1. However, the low allele frequencies of these mutations point to likely contamination with other subclones.

TargetClone reconstructs a tree in which both groups are clustered together, matching our expectations (Fig 5C). In conclusion, TargetClone provides useful insight into the

development of this tumor, even though the data consists of non-microdissected heterogeneous samples.

Comparing TargetClone with existing whole genome sequencing-based methods
Finally, we aimed to determine the benefits of running TargetClone on targeted sequencing data instead of using existing tools applied to WGS data. To do so, we compared the results of TargetClone on SNP array and targeted sequencing data (Fig 5C) with the result obtained by PhyloWGS, PyClone coupled with LICHeE (Fig S25), and LICHeE with VAFs (Fig S26) on WGS data of our ovarian cancer dataset.

PhyloWGS could not infer a tree. The trees reported by PyClone coupled with LICHeE and LICHeE alone do not capture the relationships between the two sample groups with mutual exclusive mutations (details in the Supplementary Results). These poor results are most likely explained by the low read depth (3X on average) of our WGS dataset. Taken together, we have shown that running TargetClone on targeted sequencing data does not miss information that is captured by applying existing methods on WGS data.

Discussion

In this article, we described TargetClone, a novel method to infer copy numbers, alleles, the fraction and subclonal evolution trees of tumors from SNP AF and somatic SNVs measured in microdissected samples. We demonstrated on simulation data that our method obtains low error rates for inferring $\vec{C}$, $\vec{A}$, μ and T at realistic levels of sequencing noise and realistic sample tumor fractions. Furthermore, we show that at approximately 80% of ambiguous LAF measurements the correct alleles are estimated. Existing algorithms always rely on read depth information, either by requiring that somatic SNVs are located in copy number-neutral regions, or by directly using CNVs. We have now demonstrated that in samples that contain at least one major subclone, a combination of somatic SNVs and AFs can be sufficient to accurately reconstruct copy numbers, alleles, fractions and evolutionary trees of tumors. These findings suggest that it is possible to obtain a good insight into subclonal tumor evolution even if read depth information is noisy and biased.

A current limitation of our approach is the assumption that purified samples contain only one tumor subclone. We showed that, in practice, TargetClone is not markedly affected by samples containing more than one subclone, as it still produces trees with few errors up until on average 20% of contamination with minor subclones. Although it has been shown that it is possible to obtain samples with at least one major subclone and limited minor contamination [23], it may not always be known beforehand what the total percentage of contamination in a sample is. In the future, single-cell sequencing may mitigate this limitation.

We also note that there are some limitations to the use of the FST. In short, the FST does not model biological constraints, allowing for example the re-gain of alleles when inferring the most likely alleles in a subclone. To overcome this, our model limits relations between subclones when inferring T if there is evidence in $\hat{\vec{A}}$ that alleles would require to be re-gained. A potential alternative would be to adapt the FST to include restrictions based on biological constraints, removing the need for ad-hoc corrections. However, we argue that enforcing such restrictions at an early stage in the model would

reduce the potential to estimate $\vec{A}$ correctly if many subclones were unsampled. Since the model infers alleles that minimizes the event distance, in such scenarios the inferred alleles will be more similar between subclones, misrepresenting the actual underlying allelic composition.

TargetClone currently does not scale to whole exome sequencing data, as our method infers $\vec{C}$ and $\vec{A}$ for every SNP individually. Runtimes can be reduced by a pre-segmentation of SNPs into regions with equal AF. Furthermore, resolving the ISA will become more difficult when a higher number of, potentially noisy, somatic SNVs are measured. We therefore recommend to either exclude somatic SNVs with low confidence and quality from reconstructing the ISA, which is provided as an option in TargetClone, or cluster the somatic SNVs into groups of somatic SNVs that are shared or absent across samples to reduce the influence of noise.

We employed TargetClone on four TGCC cases and one ovarian cancer case to study their subclonal evolution. We found that the inferred trees are mostly consistent with our expectations of the development of these tumors. Thus, the reconstructed trees are helpful to study relations between tumor subclones, which can assist in gaining insight into development and progression of the tumor.

Acknowledgments
M.N. would like to thank the Delft Bioinformatics Lab and Berend Snel for the helpful discussions, and Wigard P. Kloosterman for the help with processing and analyzing the ovarian cancer dataset.

References

[1] N. McGranahan and C. Swanton, *Clonal Heterogeneity and Tumor Evolution: Past, Present, and the Future,* Cell **168**, 613 (2017).

[2] P. Nowell, *The clonal evolution of tumor cell populations,* Science **194**, 23 (1976).

[3] R. Rosenthal, N. McGranahan, J. Herrero, and C. Swanton, *Deciphering Genetic Intratumor Heterogeneity and Its Impact on Cancer Evolution,* Annual Review of Cancer Biology **1**, 223 (2017).

[4] P. L. Bedard, A. R. Hansen, M. J. Ratain, and L. L. Siu, *Tumour heterogeneity in the clinic,* Nature **501**, 355 (2013).

[5] M. R. Junttila and F. J. de Sauvage, *Influence of tumour micro-environment heterogeneity on therapeutic response,* Nature **501**, 346 (2013).

[6] V. Popic, R. Salari, I. Hajirasouliha, D. Kashef-Haghighi, R. B. West, and S. Batzoglou, *Fast and scalable inference of multi-sample cancer lineages,* Genome Biology **16**, 91 (2015).

[7] W. Jiao, S. Vembu, A. G. Deshwar, L. Stein, and Q. Morris, *Inferring clonal evolution of tumors from single nucleotide somatic mutations.* BMC bioinformatics **15**, 35 (2014), arXiv:1210.3384 .

[8] F. Strino, F. Parisi, M. Micsinai, and Y. Kluger, *TrAp: a tree approach for fingerprinting subclonal tumor composition,* Nucleic Acids Research **41**, e165 (2013).

[9] M. El-Kebir, L. Oesper, H. Acheson-Field, and B. J. Raphael, *Reconstruction of clonal trees and tumor composition from multi-sample sequencing data,* Bioinformatics **31**, i62 (2015).

[10] S. Zaccaria, M. El-Kebir, G. W. Klau, and B. J. Raphael, *The Copy-Number Tree Mixture Deconvolution Problem and Applications to Multi-sample Bulk Sequencing Tumor Data,* (2017) pp. 318–335.

[11] L. Oesper, A. Mahmoody, and B. J. Raphael, *Inferring intra-tumor heterogeneity from high-throughput DNA sequencing data,* Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) **7821 LNBI**, 171 (2013).

[12] G. Ha, A. Roth, J. Khattra, J. Ho, D. Yap, L. M. Prentice, N. Melnyk, A. McPherson, A. Bashashati, E. Laks, J. Biele, J. Ding, A. Le, J. Rosner, K. Shumansky, M. a. Marra, C. B. Gilks, D. G. Huntsman, J. N. McAlpine, S. Aparicio, and S. P. Shah, *TITAN: inference of copy number architectures in clonal cell populations from tumor whole-genome sequence data.* Genome research **24**, 1881 (2014).

[13] R. F. Schwarz, A. Trinh, B. Sipos, J. D. Brenton, N. Goldman, and F. Markowetz, *Phylogenetic Quantification of Intra-tumour Heterogeneity,* PLoS Computational Biology **10**, e1003535 (2014).

[14] Z. Yu, A. Li, and M. Wang, *CloneCNA: detecting subclonal somatic copy number alterations in heterogeneous tumor samples from whole-exome sequencing data,* BMC Bioinformatics **17**, 310 (2016).

[15] Z. Yu, A. Li, and M. Wang, *CLImAT-HET: detecting subclonal copy number alterations and loss of heterozygosity in heterogeneous tumor samples from whole-genome sequencing data,* BMC Medical Genomics **10**, 15 (2017).

[16] A. G. Deshwar, S. Vembu, C. K. Yung, G. Jang, L. Stein, and Q. Morris, *PhyloWGS: Reconstructing subclonal composition and evolution from whole-genome sequencing of tumors,* Genome Biology **16**, 35 (2015).

[17] K. C. Amarasinghe, J. Li, S. M. Hunter, G. L. Ryland, P. A. Cowin, I. G. Campbell, and S. K. Halgamuge, *Inferring copy number and genotype in tumour exome data,* BMC Genomics **15**, 732 (2014).

[18] C. a. Miller, B. S. White, N. D. Dees, M. Griffith, J. S. Welch, O. L. Griffith, R. Vij, M. H. Tomasson, T. a. Graubert, M. J. Walter, M. J. Ellis, W. Schierding, J. F. DiPersio, T. J. Ley, E. R. Mardis, R. K. Wilson, and L. Ding, *SciClone: Inferring Clonal Architecture and Tracking the Spatial and Temporal Patterns of Tumor Evolution,* PLoS Computational Biology **10**, e1003665 (2014).

[19] N. E. Potter, L. Ermini, E. Papaemmanuil, G. Cazzaniga, G. Vijayaraghavan, I. Titley, A. Ford, P. Campbell, L. Kearney, and M. Greaves, *Single-Cell mutational profiling and clonal phylogeny in cancer,* Genome Research **23**, 2115 (2013).

[20] N. E. Navin, *The first five years of single-cell cancer genomics and beyond,* Genome Research **25**, 1499 (2015).

[21] V. Espina, M. Heiby, M. Pierobon, and L. a. Liotta, *Laser capture microdissection technology.* Expert review of molecular diagnostics **7**, 647 (2007).

[22] M. R. Emmert-Buck, R. F. Bonner, P. D. Smith, R. F. Chuaqui, Z. Zhuang, S. R. Goldstein, R. A. Weiss, and L. A. Liotta, *Laser capture microdissection.* Science (New York, N.Y.) **274**, 998 (1996).

[23] L. C. Dorssers, A. J. Gillis, H. Stoop, R. van Marion, M. M. Nieboer, J. van Riet, H. J. van de Werken, J. W. Oosterhuis, J. de Ridder, and L. H. Looijenga, *Molecular heterogeneity and early metastatic clone selection in testicular germ cell cancer development,* bioRxiv (2018), 10.1101/385807, https://www.biorxiv.org/content/early/2018/08/08/385807.full.pdf .

[24] T. Kader, D. L. Goode, S. Q. Wong, J. Connaughton, S. M. Rowley, L. Devereux, D. Byrne, S. B. Fox, G. Mir Arnau, R. W. Tothill, I. G. Campbell, and K. L. Gorringe, *Copy number analysis by low coverage whole genome sequencing using ultra low-input DNA from formalin-fixed paraffin embedded tumor tissue,* Genome Medicine **8**, 121 (2016).

[25] Y. Chen, L. Zhao, Y. Wang, M. Cao, V. Gelowani, M. Xu, S. A. Agrawal, Y. Li, S. P. Daiger, R. Gibbs, F. Wang, and R. Chen, *SeqCNV: a novel method for identification of copy number variations in targeted next-generation sequencing data,* BMC Bioinformatics **18**, 147 (2017).

[26] J. Li, R. Lupat, K. C. Amarasinghe, E. R. Thompson, M. A. Doyle, G. L. Ryland, R. W. Tothill, S. K. Halgamuge, I. G. Campbell, and K. L. Gorringe, *CONTRA: copy number analysis for targeted resequencing,* Bioinformatics **28**, 1307 (2012).

[27] V. Boeva, T. Popova, K. Bleakley, P. Chiche, J. Cappo, G. Schleiermacher, I. Janoueix-Lerosey, O. Delattre, and E. Barillot, *Control-FREEC: a tool for assessing copy number and allelic content using next-generation sequencing data,* Bioinformatics **28**, 423 (2012).

[28] M. Fromer, J. L. Moran, K. Chambert, E. Banks, S. E. Bergen, D. M. Ruderfer, R. E. Handsaker, S. A. McCarroll, M. C. O'Donovan, M. J. Owen, G. Kirov, P. F. Sullivan, C. M. Hultman, P. Sklar, and S. M. Purcell, *Discovery and Statistical Genotyping of Copy-Number Variation from Whole-Exome Sequencing Depth,* The American Journal of Human Genetics **91**, 597 (2012).

[29] J.-Y. Nam, N. K. D. Kim, S. C. Kim, J.-G. Joung, R. Xi, S. Lee, P. J. Park, and W.-Y. Park, *Evaluation of somatic copy number estimation tools for whole-exome sequencing data,* Briefings in Bioinformatics **17**, 185 (2016).

[30] A. Magi, L. Tattini, I. Cifola, R. D'Aurizio, M. Benelli, E. Mangano, C. Battaglia, E. Bonora, A. Kurg, M. Seri, P. Magini, B. Giusti, G. Romeo, T. Pippucci, G. D. Bellis, R. Abbate, and G. F. Gensini, *EXCAVATOR: detecting copy number variants from whole-exome sequencing data,* Genome Biology **14**, R120 (2013).

[31] D. C. Koboldt, Q. Zhang, D. E. Larson, D. Shen, M. D. McLellan, L. Lin, C. A. Miller, E. R. Mardis, L. Ding, and R. K. Wilson, *VarScan 2: Somatic mutation and copy number alteration discovery in cancer by exome sequencing,* Genome Research **22**, 568 (2012).

[32] N. E. Navin and J. Hicks, *Tracing the tumor lineage,* Molecular Oncology **4**, 267 (2010).

[33] J. W. Oosterhuis and L. H. J. Looijenga, *Testicular germ-cell tumours in a broader perspective*. Nature reviews. Cancer **5**, 210 (2005).

[34] M. Hoogstraat, M. S. de Pagter, G. A. Cirkel, M. J. van Roosmalen, T. T. Harkins, K. Duran, J. Kreeftmeijer, I. Renkens, P. O. Witteveen, C. C. Lee, I. J. Nijman, T. Guy, R. van 't Slot, T. N. Jonges, M. P. Lolkema, M. J. Koudijs, R. P. Zweemer, E. E. Voest, E. Cuppen, and W. P. Kloosterman, *Genomic and transcriptomic plasticity in treatment-naive ovarian cancer,* Genome Research **24**, 200 (2014).

[35] I. Hajirasouliha, A. Mahmoody, and B. J. Raphael, *A combinatorial approach for analyzing intra-tumor heterogeneity from high-throughput sequencing data.* Bioinformatics (Oxford, England) **30**, i78 (2014).

[36] J. Edmonds, *Optimum branchings,* Journal of Research of the National Bureau of Standards Section B Mathematics and Mathematical Physics **71B**, 233 (1967).

[37] J. Kuipers, K. Jahn, B. J. Raphael, and N. Beerenwinkel, *A statistical test on single-cell data reveals widespread recurrent mutations in tumor evolution,* bioRxiv (2016), 10.1101/094722, http://www.biorxiv.org/content/early/2016/12/16/094722.full.pdf .

[38] Bokeh Development Team, *Bokeh: Python library for interactive visualization* (2014).

[39] G. Deng, *BRAF Mutation Is Frequently Present in Sporadic Colorectal Cancer with Methylated hMLH1, But Not in Hereditary Nonpolyposis Colorectal Cancer,* Clinical Cancer Research **10**, 191 (2004).

[40] S. Lassmann, C. Kreutz, A. Schoepflin, U. Hopt, J. Timmer, and M. Werner, *A novel approach for reliable microarray analysis of microdissected tumor cells from formalin-fixed and paraffin-embedded colorectal cancer resection specimens,* Journal of Molecular Medicine **87**, 211 (2009).

[41] W. Xu, Y. Chen, W. He, Z. Fu, T. Pan, H. He, J. Yu, Q. Wei, S. Zheng, and S. Zhang, *Protein fingerprint of colorectal cancer, adenomatous polyps, and normal mucosa using ProteinChip analysis on laser capture microdissected cells.* Discovery medicine **17**, 223 (2014).

[42] L. Boublikova, T. Buchler, J. Stary, J. Abrahamova, and J. Trka, *Molecular biology of testicular germ cell tumors: Unique features awaiting clinical application,* Critical Reviews in Oncology/Hematology **89**, 366 (2014).

[43] M. a. Rijlaarsdam, D. M. J. Tax, A. J. M. Gillis, L. C. J. Dorssers, D. C. Koestler, J. de Ridder, and L. H. J. Looijenga, *Genome Wide DNA Methylation Profiles Provide Clues to the Origin and Pathogenesis of Germ Cell Tumors,* Plos One **10**, e0122146 (2015).

[44] J. K. Killian, L. C. Dorssers, B. Trabert, A. J. Gillis, M. B. Cook, Y. Wang, J. J. Waterfall, H. Stevenson, W. I. Smith, N. Noyes, *et al., Imprints and dppa3 are bypassed during pluripotency-and differentiation-coupled methylation reprogramming in testicular germ cell tumors,* Genome research **26**, 1490 (2016).

[45] C. M. Spiller and J. Bowles, *Germ cell neoplasia in situ: The precursor cell for invasive germ cell tumors of the testis,* The International Journal of Biochemistry & Cell Biology **86**, 22 (2017).

[46] A. Roth, J. Khattra, D. Yap, A. Wan, E. Laks, J. Biele, G. Ha, S. Aparicio, A. Bouchard-Côté, and S. P. Shah, *PyClone: statistical inference of clonal population structure in cancer,* Nature Methods **11**, 396 (2014).

[47] C. A. Miller, O. Hampton, C. Coarfa, and A. Milosavljevic, *ReadDepth: A Parallel R Package for Detecting Copy Number Alterations from Short Sequencing Reads,* PLoS ONE **6**, e16327 (2011).

[48] A. Soylev, C. Kockan, F. Hormozdiari, and C. Alkan, *Toolkit for automated and rapid discovery of structural variants,* Methods **129**, 3 (2017).

[49] E. Talevich, A. H. Shain, T. Botton, and B. C. Bastian, *CNVkit: Genome-Wide Copy Number Detection and Visualization from Targeted DNA Sequencing,* PLOS Computational Biology **12**, e1004873 (2016).

[50] P. Van Loo, S. H. Nordgard, O. C. Lingjaerde, H. G. Russnes, I. H. Rye, W. Sun, V. J. Weigman, P. Marynen, A. Zetterberg, B. Naume, C. M. Perou, A.-L. Borresen-Dale, and V. N. Kristensen, *Allele-specific copy number analysis of tumors,* Proceedings of the National Academy of Sciences **107**, 16910 (2010).

[51] L. Oesper, G. Satas, and B. J. Raphael, *Quantifying tumor heterogeneity in whole-genome and whole-exome sequencing data,* Bioinformatics **30**, 3532 (2014).

[52] H. Li, *A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data,* Bioinformatics **27**, 2987 (2011), arXiv:1203.6372 .

[53] *Picard,* (http://broadinstitute.github.io/picard/).

[54] C. T. Saunders, W. S. W. Wong, S. Swamy, J. Becq, L. J. Murray, and R. K. Cheetham, *Strelka: accurate somatic small-variant calling from sequenced tumor–normal sample pairs,* Bioinformatics **28**, 1811 (2012).

Supplementary Data
Supplementary Methods

Derivations

The goal of TargetClone is to infer the most likely tree T given AF and SNV measurements:

$$\arg\max_{T} \mathbb{P}(T|\vec{AF}, \vec{SNV}) \tag{6}$$

We infer the best $\vec{C}$ and μ by maximizing:

$$\arg\max_{\vec{C},\mu} \mathbb{P}(\vec{C}, \mu|\vec{LAF}, \hat{T}) \tag{7}$$

Applying Bayes' rule, we can write:

$$\mathbb{P}(\vec{C}, \mu|\vec{LAF}, \hat{T}) = \frac{\mathbb{P}(\vec{LAF}|\vec{C}, \mu, \hat{T})\mathbb{P}(\vec{C}, \mu|\hat{T})}{\mathbb{P}(\vec{LAF}|\hat{T})} \tag{8}$$

$\vec{C}$ and μ are independent, so:

$$\mathbb{P}(\vec{C}, \mu|\vec{LAF}, \hat{T}) = \frac{\mathbb{P}(\vec{LAF}|\vec{C}, \mu, \hat{T})\mathbb{P}(\vec{C}|\hat{T})\mathbb{P}(\mu|\hat{T})}{\mathbb{P}(\vec{LAF}|\hat{T})} \tag{9}$$

μ does not depend on $\hat{T}$:

$$\mathbb{P}(\vec{C}, \mu|\vec{LAF}, \hat{T}) = \frac{\mathbb{P}(\vec{LAF}|\vec{C}, \mu, \hat{T})\mathbb{P}(\vec{C}|\hat{T})\mathbb{P}(\mu)}{\mathbb{P}(\vec{LAF}|\hat{T})} \tag{10}$$

As initially every μ and topology of $\hat{T}$ is equally likely, we do not need to compute the probability of observing μ:

$$\mathbb{P}(\vec{C}, \mu|\vec{LAF}, \hat{T}) \propto \frac{\mathbb{P}(\vec{LAF}|\vec{C}, \mu, \hat{T})\mathbb{P}(\vec{C}|\hat{T})}{\mathbb{P}(\vec{LAF}|\hat{T})} \tag{11}$$

Finally, the probability of $\mathbb{P}(\vec{LAF}|\hat{T})$ is constant with respect to $\vec{C}$ and μ, and we can thus omit the denominator:

$$\mathbb{P}(\vec{C}, \mu|\vec{LAF}, \hat{T}) \propto \mathbb{P}(\vec{LAF}|\vec{C}, \mu, \hat{T})\mathbb{P}(\vec{C}|\hat{T}) \tag{12}$$

Computing LAF

The LAF measured at position i in subclone j is computed for every element in Q as:

$$LAF_{i,j} = \frac{\min(((1-\mu)L_{Ahi} + \mu L_{Ati}), ((1-\mu)L_{Bhi} + \mu L_{Bti}))}{((1-\mu)L_{Ahi} + \mu L_{Ati}) + ((1-\mu)L_{Bhi} + \mu L_{Bti})} \quad (13)$$

where L_{Ahi}, L_{Bhi} and L_{Ati}, L_{Bti} are the total number of 'A' and 'B' alleles at position i in the healthy cell and the tumor subclone, respectively.

As the total number of 'A' and 'B' alleles are always equal to 1 in healthy cells, Eq 13 simplifies to:

$$LAF_{i,j} = \frac{\min(((1-\mu) + \mu L_{Ati}), ((1-\mu) + \mu L_{Bti}))}{((1-\mu) + \mu L_{Ati}) + ((1-\mu) + \mu L_{Bti})} \quad (14)$$

Gaussian Mixture Model

As our model is based on AF measurements, noise is introduced by variation in the read counts. For WGS data, it has been shown that the read depth can be accurately modeled using binomial and Poission distributions [47, 48]. As was shown by Hajirasouliha et al [35], the read depth can be approximated with a Gaussian distribution when the depth is larger than 1000x, which is typically the case for targeted sequencing data. Therefore, $\mathbb{P}(LAF_{i,j}|C_{i,j}, \mu, \hat{T})$ is modeled as a Gaussian mixture model:

$$\mathbb{P}(LAF_{i,j}|C_{i,j}, \mu, \hat{T}) = \sum_{n=1}^{N} \mathbb{P}(q_{i,j}^{n}) \mathcal{N}(\mu_n, \sigma) \quad (15)$$

where N is the total number of possible alleles that can result from $C_{i,j}$, $\mathbb{P}(q_{i,j}^{n})$ is computed using Eq 4 described in the main text. The means of the component μ_n are equal to the LAFs that can be generated from $C_{i,j}$. The noise, σ, is estimated from the LAF measurements in the normal samples of our real TGCC dataset. The interval of the distribution is limited between 0 and 0.5 to adequately model LAF measurements.

Visualizing $\hat{T}$

For every $\hat{T}$ inferred at every iteration of the algorithm, we sum the event distance between all subclones to obtain a total score. The inferred trees are divided into two groups based on if the ISA could be resolved or not. All trees are sorted within these groups based on the total score. To reduce the amount of information in the final output, only the top 10 trees are reported, where the trees with a resolved ISA are prioritized. The annotated events include gained and lost somatic SNVs, and gained LOH. LOH events are grouped per chromosome arm for clarity. The final visualized output is generated using the Bokeh plotting library.

Generating TGCC-based simulation data

In addition to a generic simulation dataset, we also generated a simulation dataset based on the expected development of TGCC (see Fig S4B). Starting from a diploid cell, the genome is doubled to form a tetraploid precursor. This precursor further develops into a malignant subclone by acquiring 10 whole chromosome losses, 10 chromosome arm losses and 20 somatic SNVs in the stated order. The affected chromosomes and chromosome arms are randomly selected without replacement. The somatic SNV positions

are selected from 36 predefined genomic locations defined based on real observations in our TGCC sequencing dataset. Finally, the malignant subclone acquires 6 copies of chromosome 12p, which is a hallmark of TGCC, and is not allowed to be lost in subsequent cell divisions. All chromosome (arm) losses, gains and somatic SNVs affect only one allele.

In the next step, the malignant precursor continues to divide and form a new subclone. Every subclone also has the capability to divide and form a child subclone for a total of 4 rounds. In each child 8 chromosome arms are gained, 3 chromosome arms are lost, and 2 somatic SNVs are introduced, in this order. If a loss affects an allele containing a somatic SNV, the somatic SNV is lost as well. A subclone is considered unviable if the last allele of a chromosome is lost. In these situations, the subclone is unable to divide further and is removed from the simulation. This process yields 8 subclones (including the healthy cell and precursors) on average.

All generated subclones and precursors are sampled. Every sample is assigned the same predefined tumor fraction. We tested tumor fractions between 0 and 1 in steps of 0.01. A tumor fraction of 0 was included to show the performance of the method on cases where the samples contain no tumor fraction, which may for example occur when it is unclear during sampling if a region contains tumor components or not. Similarly, a tumor fraction of 1 may not be realistic, but provides a reference for how well the method would perform in a perfect world scenario. In each sample, we generate AF and LAF measurements based on the CNV profiles and introduced somatic SNVs. The AF/LAF measurement positions are selected from the real TGCC data.

Computing the error on the simulation data

The error of $\hat{\vec{C}}$ compared to the true $\vec{C}$ across n measurement positions in m samples is computed as:

$$E_C = \frac{1}{nm} \sum_{i=1}^{n} \sum_{j=1}^{m} |C_{i,j} - \hat{C_{i,j}}| \tag{16}$$

To compute the error in the predicted alleles $\hat{\vec{A}}$, we make use of the event distance as explained in Section "Computing $\mathbb{P}(L\vec{A}F|\vec{C}, \mu, \hat{T})$" in the main text:

$$E_A = \frac{1}{nm} \sum_{i=1}^{n} \sum_{j=1}^{m} ED(A_{i,j}, \hat{A_{i,j}}) \tag{17}$$

Similarly, the error of $\hat{\mu}$ for m samples is computed as:

$$E_\mu = \frac{1}{m} \sum_{j=1}^{m} |\mu_j - \hat{\mu_j}| \tag{18}$$

The error of $\hat{T}$ is computed as the ratio of false positive and false negative sample pairs, as described in the main text.

Supplementary Results

Combining alleles and somatic SNVs improves performance

To test if the way in which TargetClone combines somatic SNVs and allele information improves performance, we compared our results to results obtained with distance matrices computed from other data types (Fig S7). In these tests, the MSA was inferred in one method iteration, without attempting to resolve the ISA. For copy numbers, the distance matrix was reconstructed by computing the absolute distance between the inferred copy numbers, incorporating the horizontal dependency, which is detailed in MEDICC. For the alleles, we computed the event distance using the FST shown in Fig 1E. For somatic SNVs, the distance between samples was set to 1 if a relation is possible, or a penalty was assigned if a relation is unlikely as is detailed in Section "Reconstructing T" (Fig 2B). Finally, a distance matrix was also generated by computing the Euclidean distance between the LAF measurements.

We observe that trees based on allele information exclusively are slightly better than those inferred from only copy number information. Both perform much better than trees inferred from somatic SNVs only, and yield a further improvement in error rate compared to trees inferred using the Euclidean distance computed directly on the LAF measurements. However, the error obtained by TargetClone is lower than any other method. This result demonstrates that the combination of distance metrics based on alleles and somatic SNVs, including resolving the ISA, improves tree reconstruction accuracy.

The number and distribution of measured SNPs does not significantly influence performance

To assess the effect of the number of measured SNPs, we ran additional simulations in which the number of SNPs was increased between 100 and 50000. As TargetClone was not designed to handle such a large number of SNPs, we segmented the run with 50000 SNPs to chromosome arms and assigned 1 SNP to each arm. A μ of 0.9 was selected for each dataset as the error is low at this tumor fraction in the simulation data and thus it is easier to determine the effect of the number of SNPs. We removed all somatic SNVs from the datasets to avoid potential bias from SNVs during tree reconstruction. Fig S8 shows that the number of SNPs does not have a signficant effect on the performance of TargetClone. Furthermore, as the SNPs are distributed randomly in each simulated dataset, we can observe from the small confidence intervals across the increasing number of SNPs that the effect of the distribution of SNPs is small. In addition, we show in the Results that the results of our generic simulations (with 500 randomly distrubted SNPs) differs minimally from our results obtained with a TGCC-based simulation (450 SNPs at fixed positions). Thus, we conclude that the effect of the number of distribution of SNPs on the performance of TargetClone is minimal.

The number of measured SNVs does not significantly influence performance

To test if TargetClone performs better with a specific number of measured SNVs, we performed additional simulations in which the number of measured SNVs was varied between 10 and 10000 SNVs. We selected a μ of 0.9 and a noise level of 0.02. We see no significant change in the performance of TargetClone as the number of input SNVs change (Fig S9), showing that TargetClone is robust to different numbers of somatic SNVs.

The TGCC-based simulations yield similar results to the generic simulations

In Fig S22, we show the mean error and 95% confidence intervals for inferring $\vec{C}$, $\vec{A}$, μ and T in our TGCC-based simulation dataset. The grey shared area shows the mean error and 95% confidence interval for a 100 simulation datasets where the LAF measurements and somatic SNV measurements were assigned a random value between 0 and 1. Interestingly, the error rates differ minimally from the error rates obtained in the generic simulation dataset. Furthermore, these results show that TargetClone is capable of accurately inferring tumor evolution despite the initial duplication to a tetraploid precursor, and the assumption of tetraploidy in the initial tree $\hat{T}1$.

A comparison of TargetClone to existing methods on targeted sequencing data

First, we corrected the read depth of our TGCC samples using the amplicon sequencing correction methods in CNVKit [49]. However, the segmentation immediately shows that even after such corrections, it remains difficult to properly detect copy numbers from our read depth data (Fig S15). In addition, the number of measurements is small and there exists a lot of variation between adjacent measurements, complicating finding a good segmentation.

We attempted to use ASCAT to estimate $\vec{C}$ and μ [50]. However, ASCAT failed to output copy numbers for all but 3 of our 42 samples (Fig S16). We then used THetA2 on the read depth corrected by CNVKit and the SNP AF to generate $\vec{C}$ and μ estimates assuming 1 tumor subclone per sample [51]. THetA failed on sample EC80 of T6107, NS30 and NS48 of T618 and EC22 of T1382. In Figs S17 and S18 we show a comparison of the copy number estimates between THetA and TargetClone for the most notable results. For sample EC70 of T6107, THeTA estimated a copy number of 0 for chromosomes 12, 15, 18, 21 and 22 with a normal contamination percentage of 0% (Fig S17B). However, with LAF measurements > 0.2 at all of these chromosomes, and somatic SNVs present with a VAF > 0.1 at chromosomes 12 and 15, it seems unlikely that the copy number is truly 0 (Fig S17C). A similar situation can be observed for sample TE74 of T3209. Here, THetA inferred a copy number of 2 everywhere but at chromosome 17 (Fig S18B). However, looking at the raw measurements (Fig S18C), we again do not see clear evidence for these results in the data. The copy numbers reported by TargetClone generally match the raw measurements profile more closely. Overall, the copy numbers are more closely distributed around a copy number of 4, which matches the assumption that TGCC develops by first duplicating to a tetraploid genome.

Although comparing μ is difficult as no ground truth is known, we see a few examples where μ is estimated to be 1 by THetA (Fig S16). For example, for sample EC70 of T6107, if the copy number were truly 0 at chromosome 12 with a μ of 1, the expected AF of the somatic SNVs would be 0, rather than 0.1 (Fig S17C).

The μ estimated by THetA were used as input to CNVKit to obtain estimates of major and minor copy numbers for the tumor component in each sample. We used these major and minor copy numbers together with somatic SNVs to run PhyloWGS and PyClone.

To make the comparison with PhyloWGS as equal as possible to TargetClone, we assumed that each sample contained 1 tumor subclone and set the cellular prevalence for each CNV equal to the sample μ estimated by THetA. All somatic SNVs that are not shared between all samples had to be excluded from analysis with PhyloWGS. We used

4 chains to run PhyloWGS and the default configuration for PyClone. When running LICHeE, the minimum present VAF was set to 0.0001, and the maximum absent VAF to 0.9999.

Samples for which THetA could not infer a $\vec{C}$ and μ (Fig S16) are not reported in the PhyloWGS and PyClone-based trees.

In Fig S19, we show the trees inferred by PhyloWGS. In short, these trees do not match our biological knowledge of TGCC development. The reported tree for T3209 is unlikely as it consists of 3 tumor subclones, while at least 5 different histological elements were sampled for this tumor. For T618, the expected development of FCIS from CIS through BCIS is not reflected in the tree. Additionally, many somatic SNVs are present in sample NS75 only, but the only reported subclone that is unique to NS75 gains only 1 somatic SNV. For T1382 and T6107, PhyloWGS was unable to infer a tree.

Fig S20 shows the results of coupling LICHeE with cellular prevalences estimated by PyClone. Interestingly, PyClone reports only 3, 3, 4 and 4 subclonal clusters for T3209, T6107, T618 and T1382, respectively, which is fewer than expected for T3209 and T6107 given that we sampled more histological elements than the reported number of subclonal clusters for each patient. For all patients, we see that important relations based on LOH are missed. For T1382, samples MET32, MET35, MET30 and MET32 share LOH on chromosome 6q, 11q and 14q, but this relationship is not visible in the tree. For T6107, LICHeE misses expected relations between the CIS/FCIS and EC samples.

In Fig S21, the results are shown of running LICHeE on VAFs. For T3209, the relation between all samples other than CIS32 and CIS73 based on LOH on chromosomes 4, 14, 15 and 22 is not captured. For T618, samples NS30, NS48 and NS75 are predicted by LICHeE to have independent origin, but these samples share LOH on chromosomes 9q and 22. All metastasis samples of T1382 are inferred to have originated independently from a healthy cell, while in reality these samples share at least 13 somatic SNVs.

A comparison of TargetClone to existing methods on simulated data

In addition to real data, we also tested how TargetClone compares to existing methods on simulated data where a ground truth is known. We ran these comparisons on simulated data based on our TGCC data, which resemble targeted sequencing data more directly than our generic simulations. As was shown in Section "The TGCC-based simulations yield similar results to the generic simulations", the results we obtained on this dataset minimally differ from the results on the generic simulations, and we thus reason that our TGCC-based simulation data is suitable enough to compare methods on.

Since our simulation data do contain ground truth copy number information, it is possible to compare with methods that rely on copy number information and/or somatic SNVs. As MEDICC, a copy number based method, also uses a FST, it is interesting to see how our model compares to this work. From the existing somatic SNV-based methods listed in Fig S1, 4 methods (PhyloSub, AncesTree, CITUP and LICHeE) are able to reconstruct trees from SNVs in multiple samples. As all of these methods are based on similar principles related to somatic SNVs and no previous study has compared the performance of all 4 methods, we limited our comparison to LICHeE.

To make the comparison fair, we generated a simulation dataset without sequencing noise and normal cell contamination, as the copy numbers used as input to MEDICC do

not contain noise and are also not affected by tumor content. For LICHeE, we considered somatic SNVs with a frequency < 0.001 as absent, and a frequency > 0.999 as germline. TargetClone was run with a precursor ploidy of 2.

In contrast to TargetClone, MEDICC and LICHeE reconstruct trees in which all samples are placed at leaf nodes, and thus comparing the resulting trees directly, e.g. by computing the edit distance, is not informative. To enable a meaningful comparison, we therefore converted each tree to a distance matrix and compared the ranked correlation of the distances with the distance matrix of the ground truth tree. For LICHeE, we computed the pairwise distance between samples as the sum of the edge weights across the shortest path between these samples. MEDICC provides a distance matrix based on the CNVs, which we used directly to obtain pairwise sample distances. For TargetClone, we used the distances obtained from distance matrix A_d (see Fig 2B)).

We argue that, if the distance between two samples in the ground truth tree is small, a small distance should also be observed in an inferred tree if the samples are placed correctly. We first ranked each pairwise combination of samples in the ground truth tree by their allelic distance (from matrix A_d). For each tested tool, the sample pairs were ordered according to this ranking, and the ranked distances were correlated with the ground truth (Fig S23). We see that the distances inferred by TargetClone correlate with the ground truth better than the trees reported by MEDICC and LICHeE. Noteably, in contrast to the ground truth, LICHeE and MEDICC often report a larger distance between the precursor, Germ Cell Neoplasia In Situ (GCNIS) and pre-GCNIS than between GCNIS and its child subclones, resulting in a negative correlation (see Fig S24 for an example). These results do not imply that MEDICC and LICHeE generally perform badly, but that these methods are less suitable to our specific case.

Comparing TargetClone to existing whole genome sequencing-based methods

To demonstrate the possible benefits of using TargetClone on targeted sequencing data instead of existing methods on WGS data, we used our ovarian cancer dataset, for which both data types are available. We downloaded the aligned BAM files (hg19) and first merged these per patient and filtered unmapped reads using samtools [52]. The sample names in the read groups were corrected using Picard tools [53]. The resulting BAM files were sorted and indexed using samtools. Notably, the reported average read depth of the samples is 3x. SNPs (no indels) were called using samtools mpileup coupled with bcftools call of samtools. These SNPs were filtered for minimum read depth of 30, maximum read depth of 100 and minimum RMS mapping quality of 20 using varFilter of vcfutils of samtools. We ran CNVKit in WGS mode coupled with THetA as described in Section "A comparison of TargetClone to existing methods on targeted sequencing data" to estimate major/minor copy numbers. THetA could not infer a $\vec{C}$ and μ for samples IV3 and V1.

We used Strelka to call somatic SNVs on autosomes and sex chromosomes [54]. These SNVs were filtered for lowEVS or lowDepth. As the runtime of PhyloWGS increases linearly with the number of somatic SNVs, we initially limited the number of SNVs in the PhyloWGS parser to 5000, which did not complete within reasonable time. Thus, we further reduced the number of SNVs to 1000. For PyClone and LICHeE we used the same settings as discussed in Section "A comparison of TargetClone to existing methods on

targeted sequencing data".

PhyloWGS was unable to infer a tree. Notably, PyClone identifies 3 subclonal clusters, but running LICHeE on the estimated cellular prevalences results in a tree in which all samples originate from the same precursor subclone (Fig S25). Running LICHeE on VAFs similarly results in a tree in which all samples directly originate from the germline sample (Fig S26).

We believe that a main reason why the existing WGS tools do not output expected results is related to the low read depth in our dataset. In conclusion, these results show that it may be beneficial to use targeted sequencing coupled with TargetClone in cases where the read depth of WGS analysis is low.

Supplementary Figures

Method	Author	Somatic SNVs	Read depth/ CNVs	Trees	Multi-sample
Clomial	*Zare et al, 2014*	y	n	n	y
Phylosub	*Jiao et al, 2014*	y	n	y	y
PyClone	*Roth et al, 2014*	y	n	n	y
SciClone	*Miller et al, 2014*	y	y	n	y
TrAp	*Strino et al, 2013*	y	n	y	n
AncesTree	*El-Kebir et al, 2015*	y	n	y	y
LICHeE	*Popic et al, 2015*	y	n	y	y
CITUP	*Malikic et al, 2015*	y	n	y	y
CTNMD	*Zaccaria et al, 2017*	n	y	y	y
PhyloWGS	*Deshwar et al, 2015*	y	y	y	y
CloneHD	*Fischer et al, 2014*	y	y	n	y
MEDICC	*Schwarz et al, 2014*	n	y	y	y
TuMult	*Letouzé et al, 2010*	n	y	y	y

Fig. S1. (A) Existing methods that can decompose subclones from mixed samples and/or reconstruct subclonal evolution trees. For each method, it is listed which data types are used and if trees are reconstructed or not. As this paper focuses on mixed samples, single-cell-based methods have been omitted from this overview.

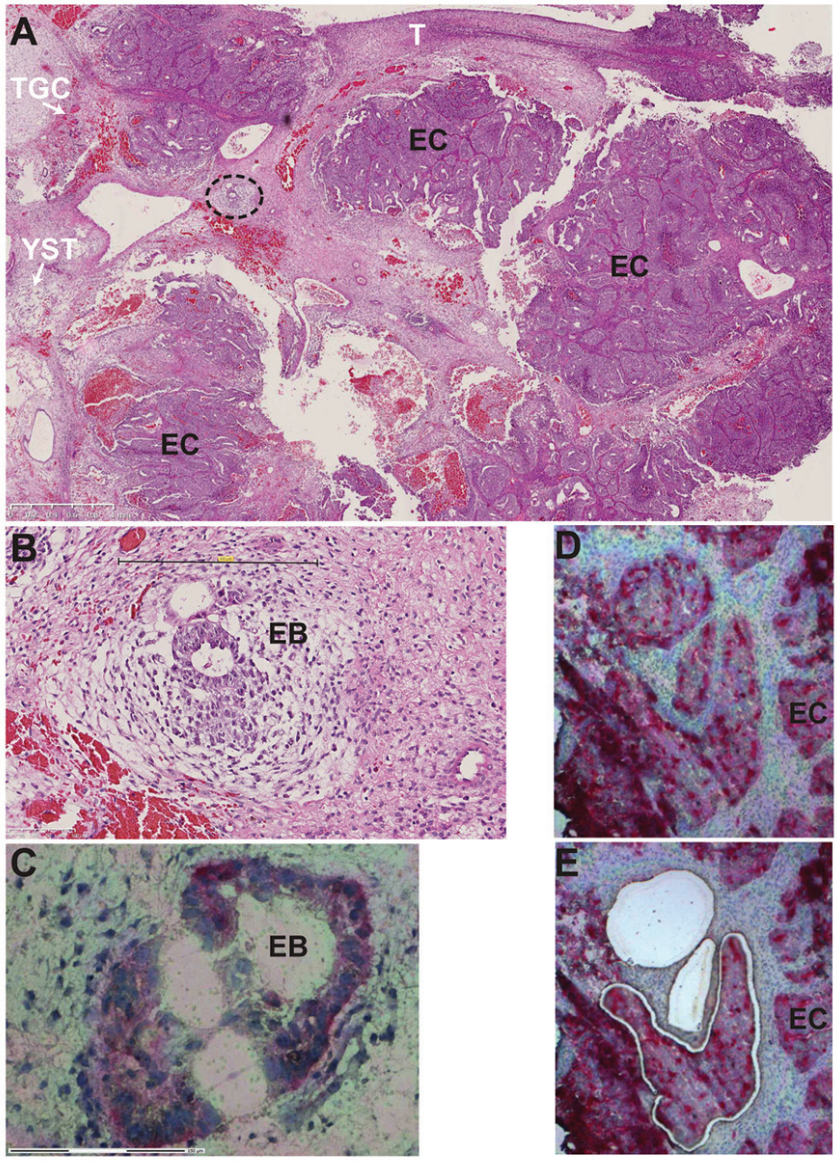


Fig. S2. Example of microdissections applied to our real data case of testicular germ cell cancer (nonseminoma) [23]. (A) H&E staining (original magnification x 2) of a section from T3209 showing the complexity of this primary testicular mixed germ cell tumor. The major tumor component in this section is solid and glandular embryonal carcinoma (EC), with in between highly vascular mesenchymal teratomatous tissue with scattered epithelial structures (T), small areas of yolk sac tumor (YST) and trophoblastic giant cells (TGC). Larger areas of teratoma and yolk sac tumor are present in adjacent sections of this case. A so-called embryoid body (EB), comparable to a day 10-human embryo, derived from a single embryonal carcinoma cell, is present in the encircled area, and shown at higher magnification in panel (B). Pictures taken from PALM-assisted purification of tumor cells from frozen tissue sections, visualized by direct alkaline phosphatase reactivity, are shown in panels (C) and (D) (before purification), and (E) (during purification).

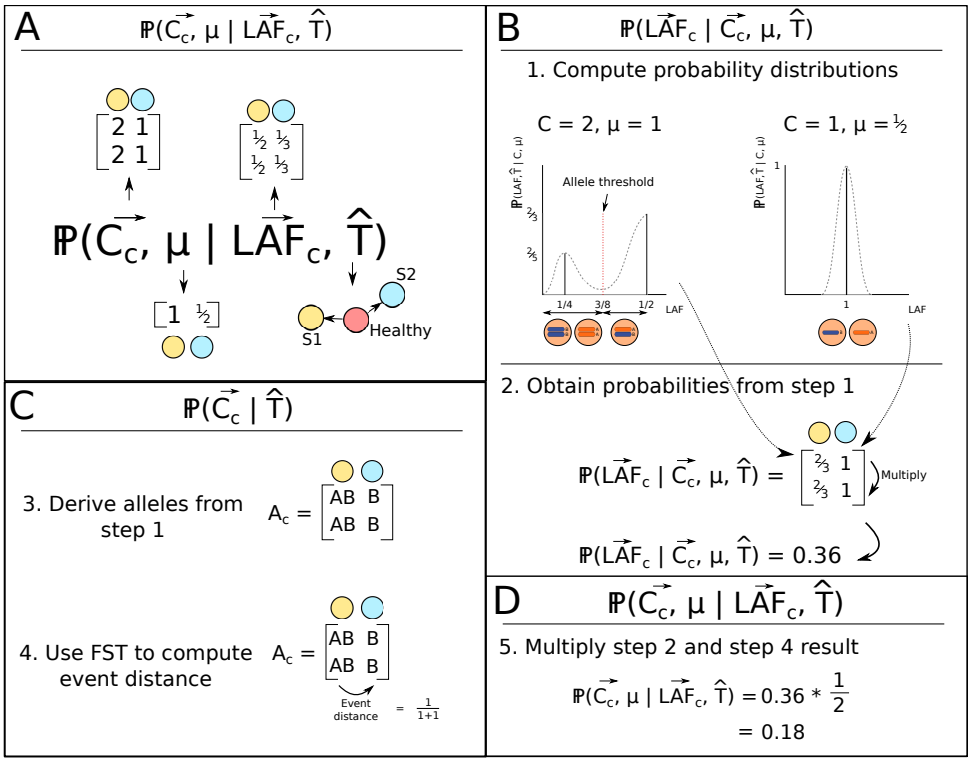


Fig. S3. Toy example calculation of $\mathbb{P}(\vec{C}, \mu | L\vec{A}F, \hat{T})$ for one $\vec{C}_c$, thus with two samples and two measurements. (A) We start with estimates of $\vec{C}_c$ and μ given the LAF measurements and an initial tree where the parent of each sample is diploid. (B) Computation of $\mathbb{P}(L\vec{A}F|\vec{C}, \mu, \hat{T})$ for one $\vec{C}_c$. In step 1, we compute the probability distribution for the current μ estimates, which are 1 and 0.5, and each copy number in $\vec{C}_c$, which are 2 and 1, respectively. An example of how the probabilities are computed is detailed in Fig. 1. In step 2, we obtain the actual probabilities that would be assigned to the measured LAF for these C in $\vec{C}_c$ and μ. All four values in $\vec{C}_c$ are multiplied to obtain the final probabilities. (C) Computation of $\mathbb{P}(\vec{C}|\hat{T})$ for one $\vec{C}_c$. In step 3, we use the known LAF measurements to derive from the probability distributions of step 1 what the alleles would be. In step 4, we compute the event distance based on the alleles corresponding to $\vec{C}_c$ as derived in step 3. Under the horizontal dependency assumption, the FST will compute an event distance of 1. The total probability is computed as 0.5. (D) In step 5, $\mathbb{P}(\vec{C}_c, \mu | L\vec{A}F_c, \hat{T})$ is computed by multiplying the probabilities obtained at step 2 and step 4.

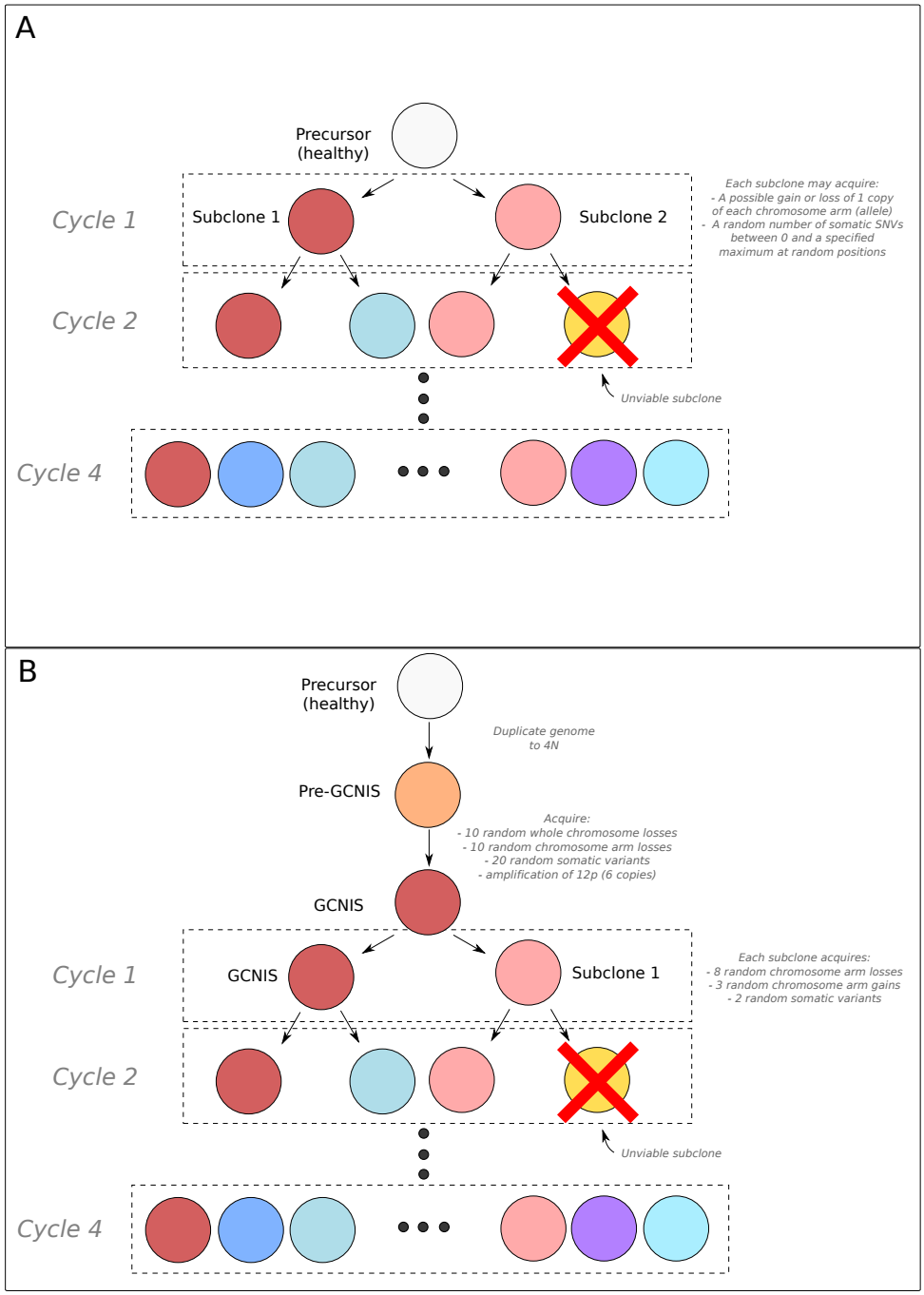


Fig. S4. Generation of simulation data for (A) the generic simulations and (B) the TGCC-based simulations. Unviable subclones are not allowed to continue through further cell divisions. The final remaining subclones at cycle 4 are sampled to generate input for TargetClone.

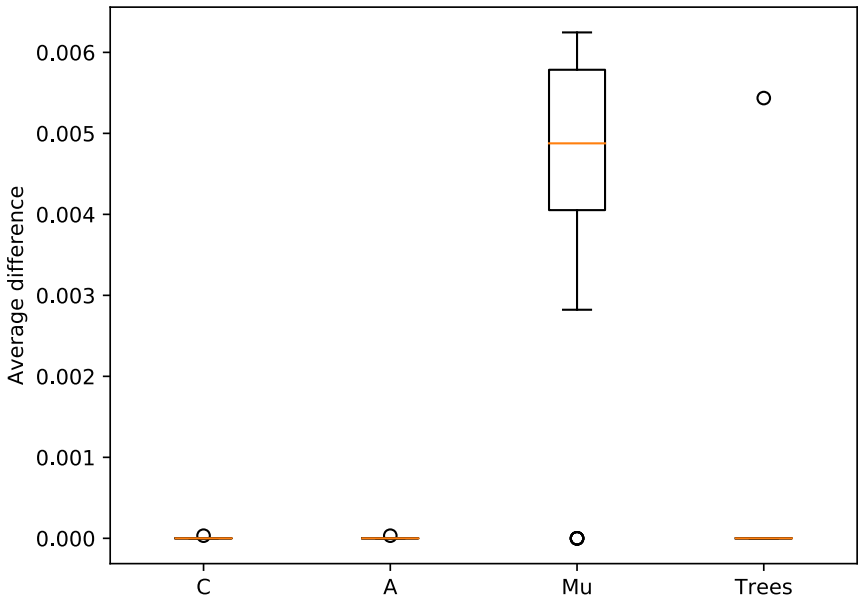


Fig. S5. Re-running TargetClone 100 times on the same simulated dataset gives approximately the same results. For each simulation re-run, we computed the difference to the error of all other re-runs, of which the average is reported in the figure. The tumor fractions differ more often between re-runs than $\vec{C}$, $\vec{A}$ and T, but the low average difference indicates that this happens in a minimum number of re-runs.

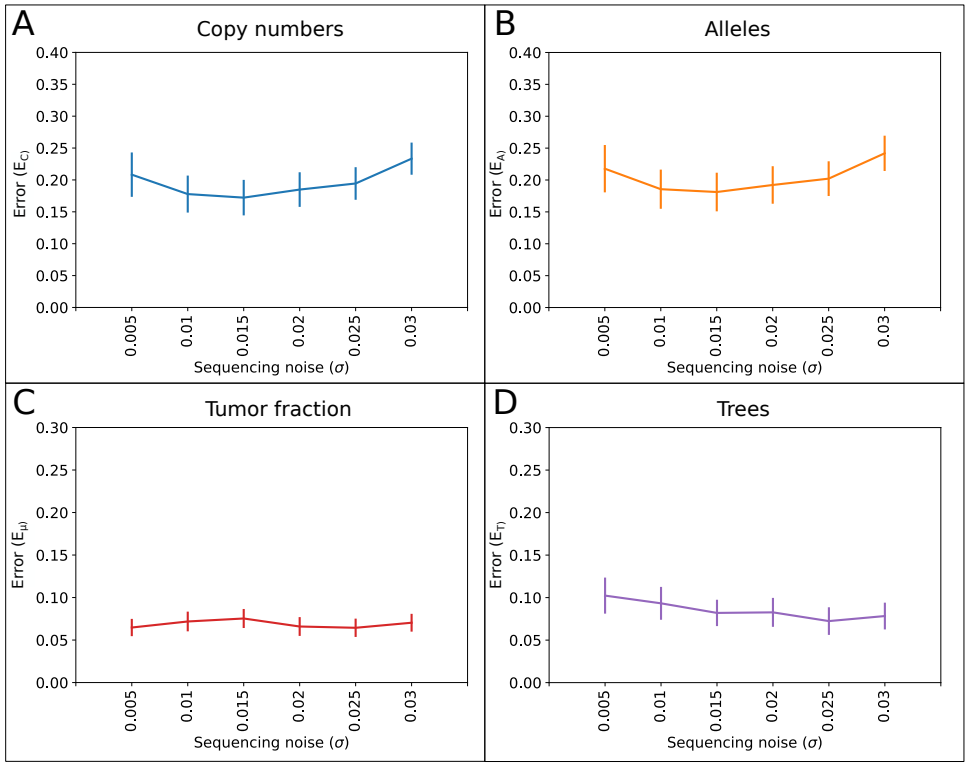


Fig. S6. Mean of the error and 95% confidence intervals for $\hat{\hat{C}}$, $\hat{\hat{A}}$, $\hat{\mu}$ and $\hat{T}$ in the simulated datasets where a random tree was used as $\hat{T}1$. Only realistic noise levels are shown. At every noise level, 101 simulated datasets were generated, each wth a unique μ between 0 and 1.

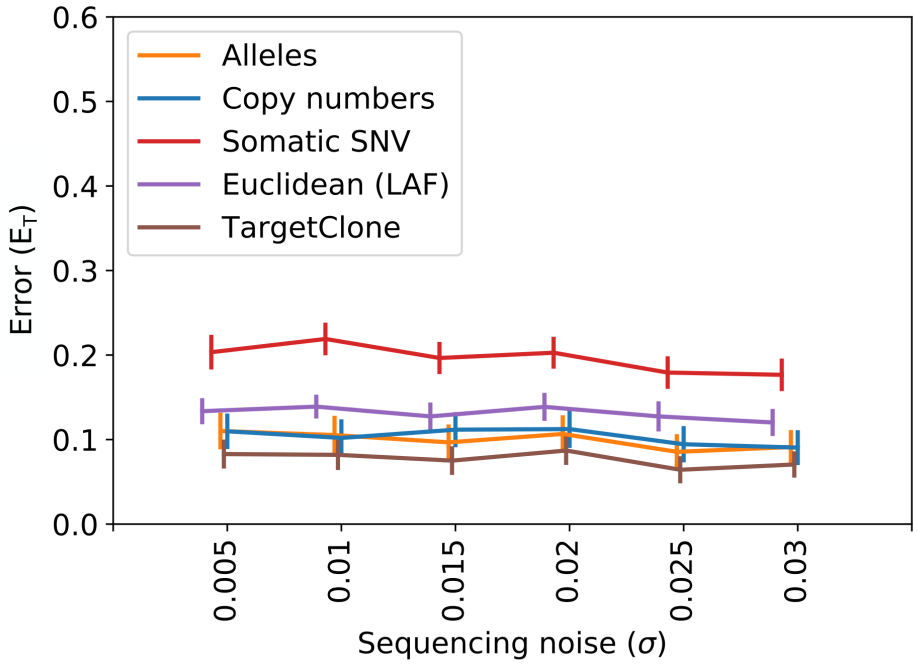


Fig. S7. The mean of the tree reconstruction error and 95% confidence intervals when different data types are used to reconstruct the distance matrices in comparison to the error obtained by TargetClone. A total of 101 simulated datasets were tested, each with a different μ between 0 and 1.

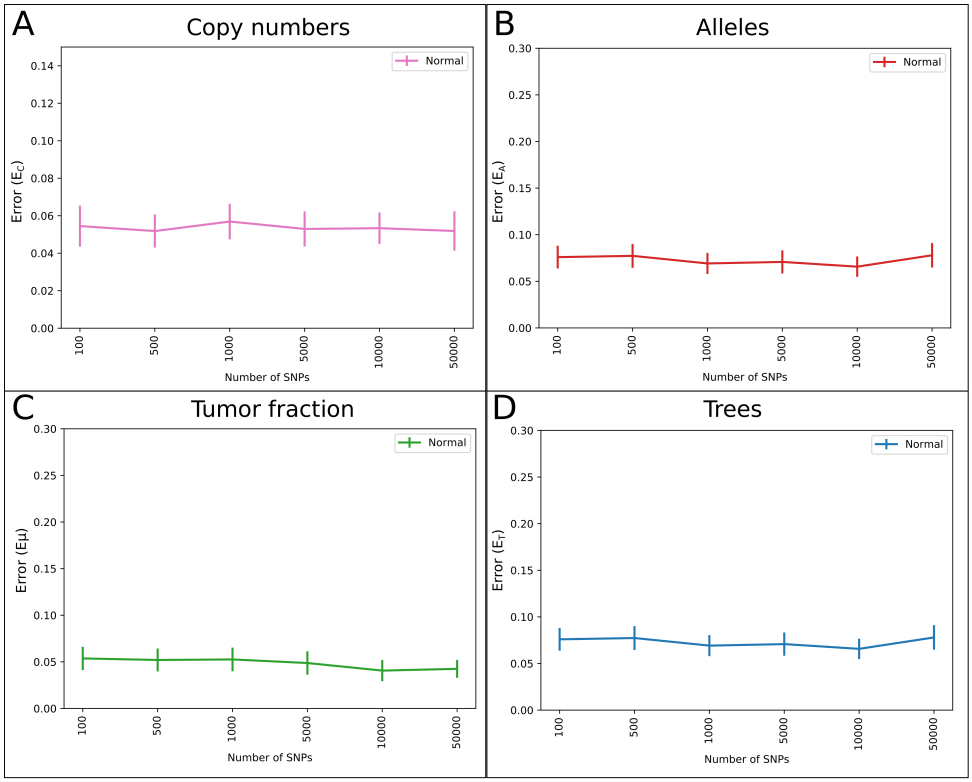


Fig. S8. Increase in the number of SNPs to show the effect of having fewer or more LAF measurements. For each number of SNPs 100 simulated datasets were generated with a noise level of 0.02 and a μ of 0.9. Because we measured the error rate with a μ of 0.9, T_e is significantly lower than T_e in Fig 3D.

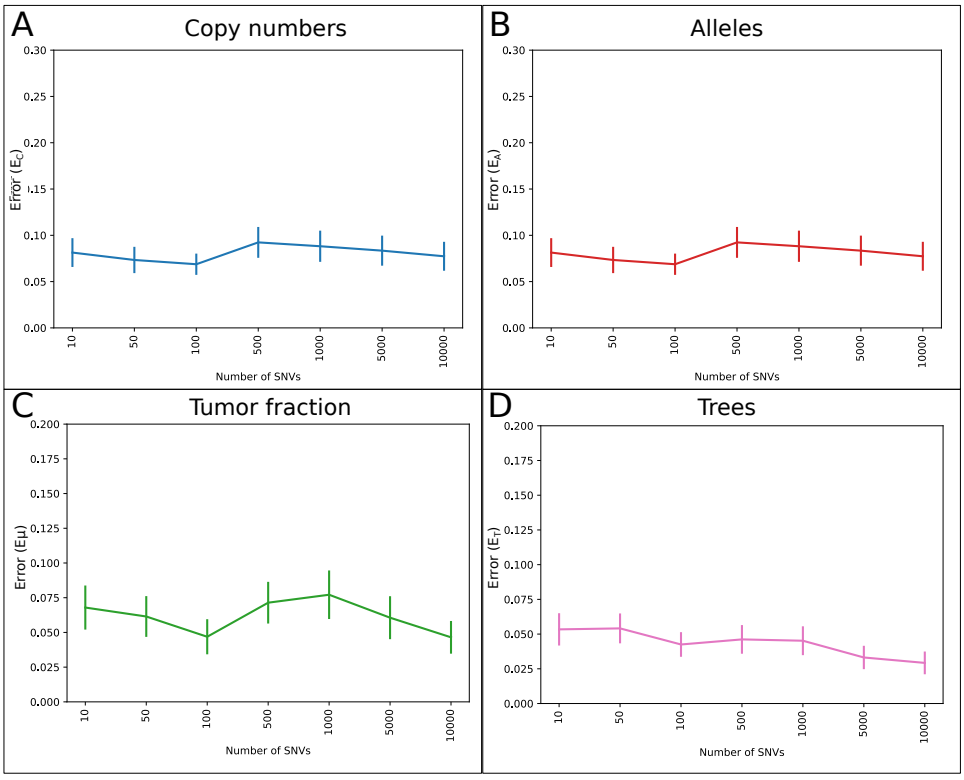


Fig. S9. The error rates obtained when the number of somatic SNV measurements are increased. For each number of SNVs, 100 simulated datasets were generated with a noise level of 0.02 and a μ of 0.9.

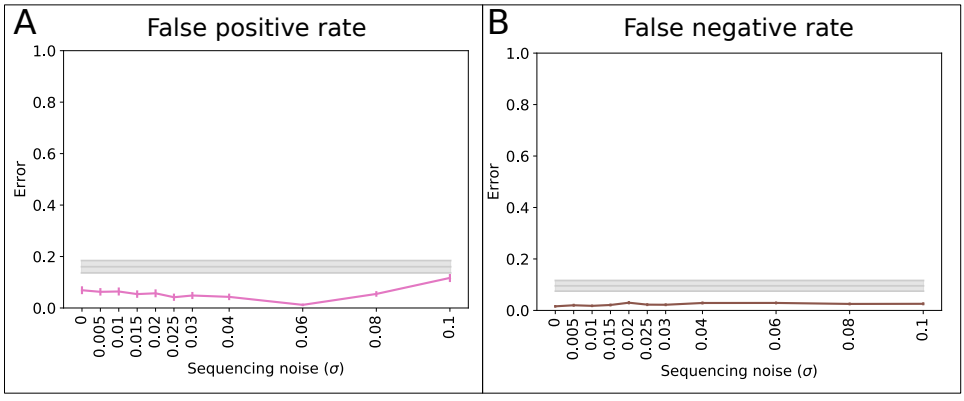


Fig. S10. The false positive and false negative rates for the trees inferred in our simulation data. The combined FPR and FNR is shown in Fig 3D.

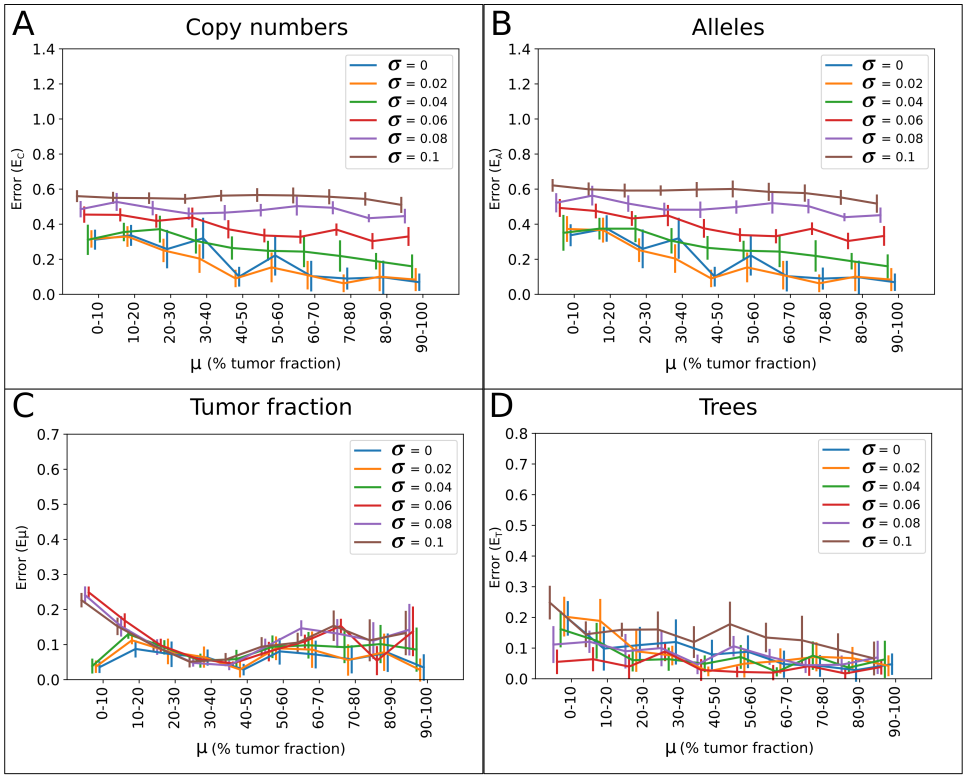


Fig. S11. Error rates for $\hat{C}$, $\hat{A}$, $\hat{\mu}$ and $\hat{T}$ as a function of μ in the simulated datasets. Every simulated dataset has one unique μ between 0 and 1. The mean of the error and 95% confidence intervals are reported in bins of μ. The noise levels are shown as separate lines. Not all tested noise levels are shown to improve visualization.

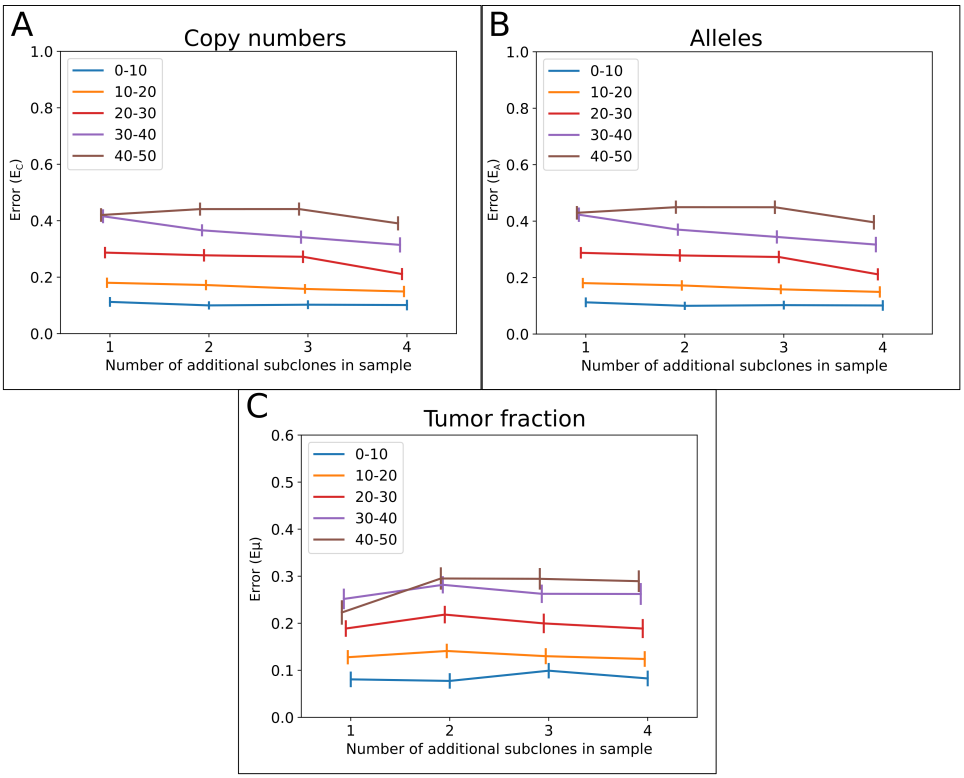


Fig. S12. The mean error and 95% confidence interval of $\hat{\hat{C}}$, $\hat{\hat{A}}$ and $\hat{\mu}$ as the number of subclones in the sample increases. Each line incidates the total percentage of contamination of the minor subclones in the sample. For each simulated dataset, a μ of 0.9 and a noise level of 0.02 was selected.

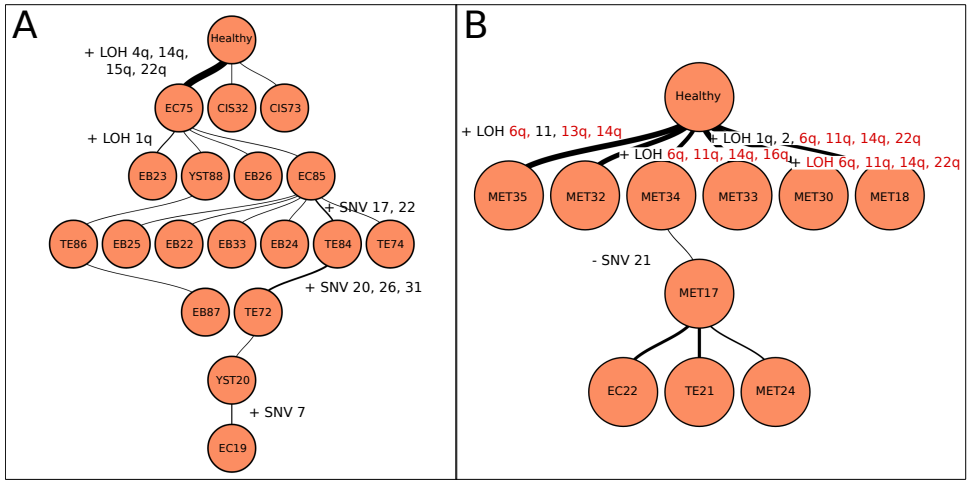


Fig. S13. Reconstructed trees for (A) T3209 and (B) T1382 when a tetraploid precursor is used. For T3209, we selected the second best reported tree, as the development of other histological components (other than CIS) from EC75 instead of TE86 matches biological expectation better. For T1382 the ISA could not be resolved and thus the MSA with the fewest ISA violations is reported. All events that are introduced multiple times independently are highlighted in red. A thicker line indicates that a higher number of events is gained in the subclone.

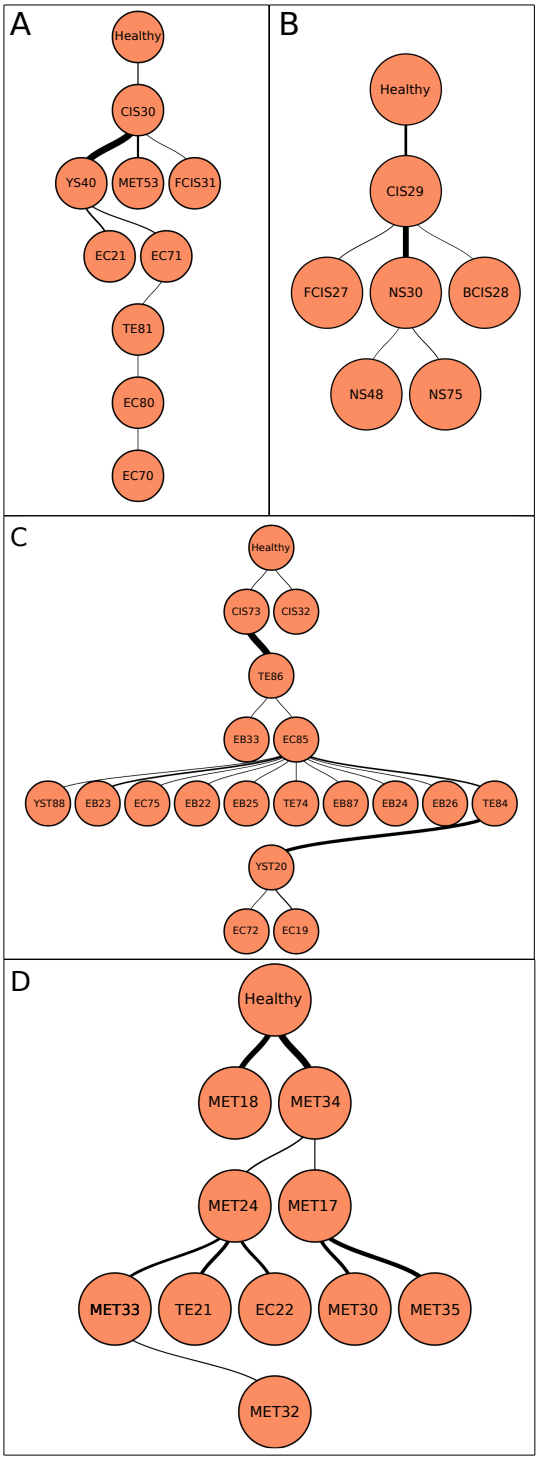


Fig. S14. Reconstructed trees for (A) T6107, (B) T618, (C) T3209 and (D) T1382 when a diploid precursor is used. For T1382 the ISA could not be resolved and thus the MSA with the fewest ISA violations is reported. A thicker line indicates that a higher number of events is gained in the subclone.

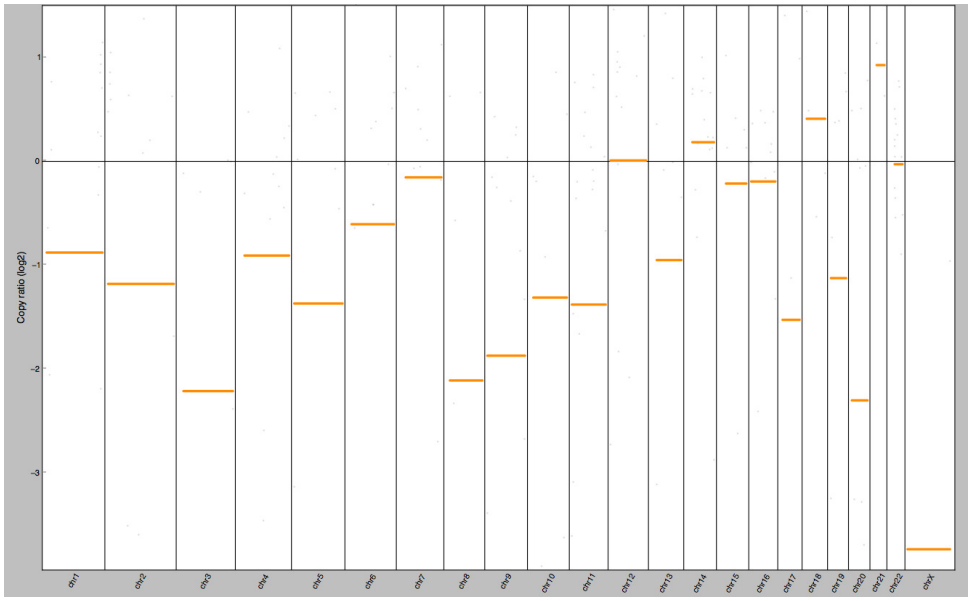


Fig. S15. Segmentation of the corrected read depth sample EC85 of T3209 by CNVKit.

T3209 A

Sample	μ (TargetClone)	μ (ASCAT)	μ (THetA)
CIS32	0.98		0.791
CIS73	0.90		1
EB22	0.96	0.75	0.994
EB23	0.76		0.976
EB24	0.94		0.983
EB25	1	0.92	0.923
EB26	0.98		0.985
EB33	0.97		0.746
EB87	0.57		0.999
EC19	0.94		1
EC75	0.91		1
EC85	0.92		0.81
TE72	0.87		1
TE74	0.98		1
TE84	0.76		0.774
TE86	0.51		0.92
YST20	0.85		0.942
YST88	0.99		0.597

T6107 B

Sample	μ (TargetClone)	μ (ASCAT)	μ (THetA)
EC70	0.45		1
EC21	0.41		0.902
CIS30	0.88		0.777
EC80	1		
TE81	0.83		0.893
EC71	0.21	0.85	1
YST40	0.88		0.806
FCIS31	0.96		0.971
MET53	0.58		0.995

T618 C

Sample	μ (TargetClone)	μ (THetA)
CIS29	0.67	0.597
BCIS28	1	0.677
FCIS27	0.71	0.554
NS30	0.62	
NS48	0.87	
NS75	0.72	0.647

T1382 D

Sample	μ (TargetClone)	μ (THetA)
EC22	0.41	
TE21	0.83	0.993
MET17	0.86	0.989
MET18	0.83	0.995
MET24	0.87	0.967
MET30	0.66	0.894
MET32	0.53	0.955
MET33	0.5	0.815
MET35	1	0.924

Fig. S16. Comparison of μ estimates of TargetClone to ASCAT and THetA for (A) T3209, (B) T6107, (C) T618 and (D) T1382.

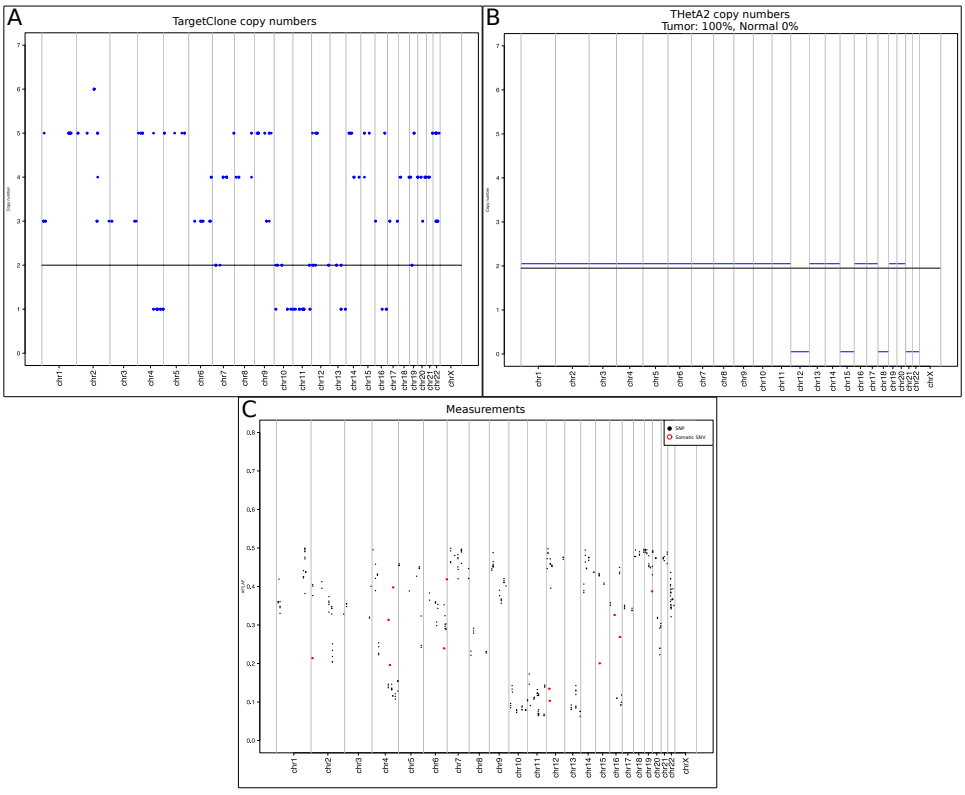


Fig. S17. Comparison of $\vec{C}$ estimates of (A) TargetClone to (B) THetA for sample EC70 of T6107. The SNP (AF) and somatic SNV (VAF) measurements of this sample are shown in (C).

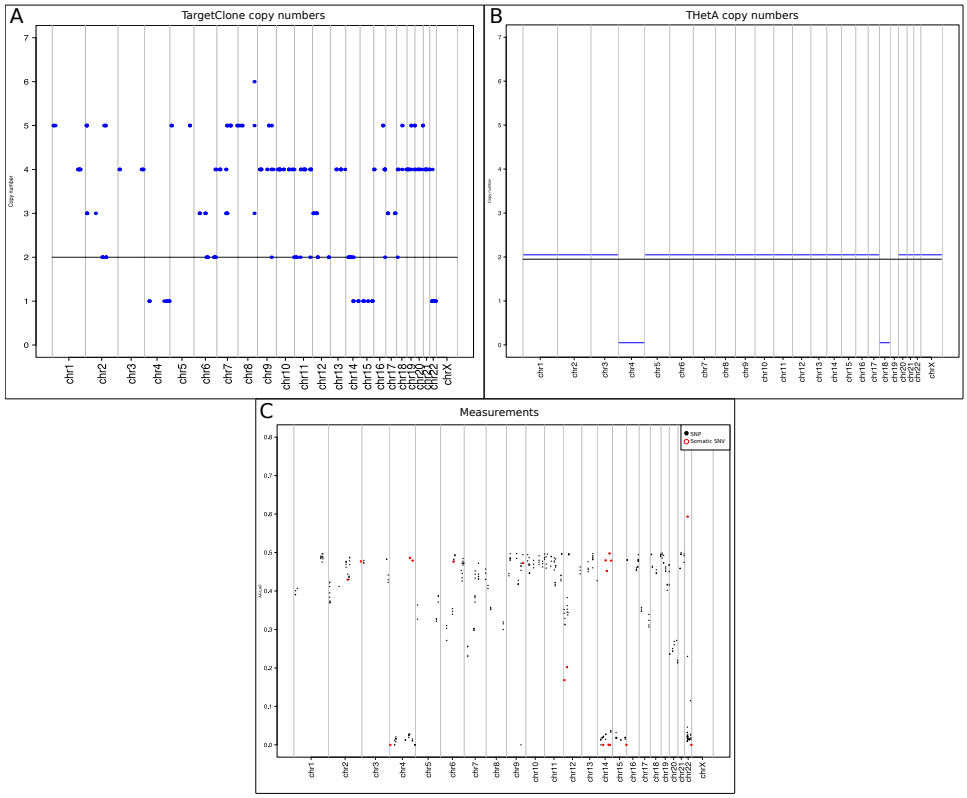


Fig. S18. Comparison of $\vec{C}$ estimates of (A) TargetClone to (B) THetA for sample TE74 of T3209. The SNP (AF) and somatic SNV (VAF) measurements of this sample are shown in (C). In (B), THetA estimated a copy number of 4301 for chromosome 19, which was left out of this figure.

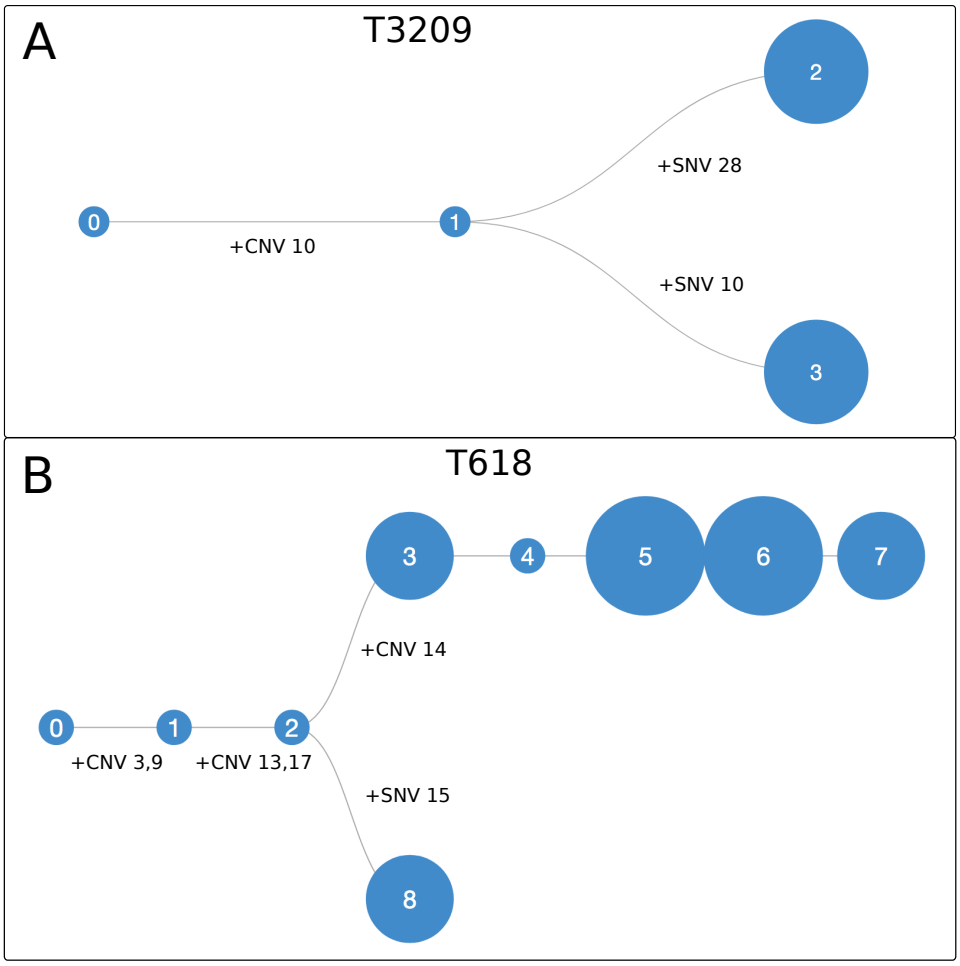


Fig. S19. Trees inferred by PhyloWGS for (A) T3209 (B) T618. The most interesting events are annotated in the trees. The order of the somatic SNVs is equal to the order of the somatic SNVs in the original input file and thus corresponds to the events annotated in the trees generated by TargetClone. (A) Samples in subclones: each subclone is present in every sample. (B) Samples in subclones: 1: all samples, 2, 3 and 8: CIS29, FCIS27, NS75. 4, 5 and 6: FCIS27. 7: NS75.

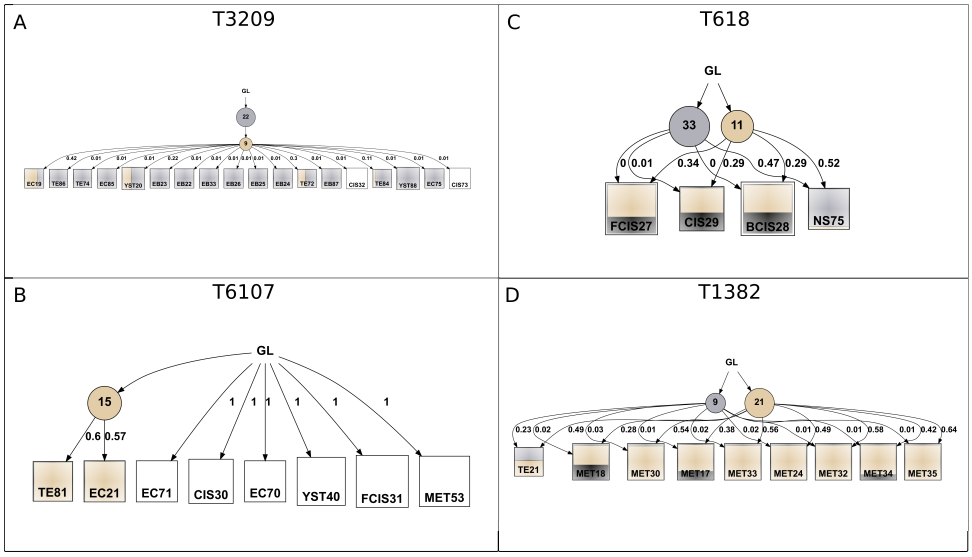


Fig. S20. Trees inferred by LICHeE using the cellular prevalences inferred by PyClone for (A) T3209, (B) T6107, (C) T618 and (D) T1382.

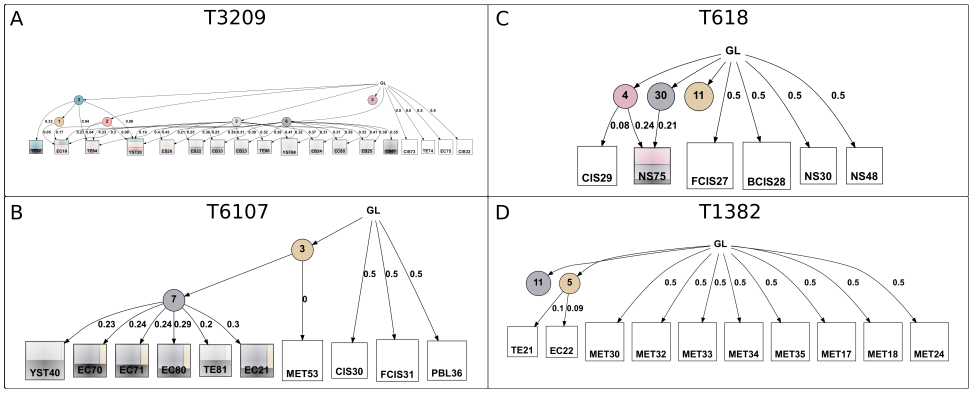


Fig. S21. Trees inferred by LICHeE using the VAF of somatic SNVs for (A) T3209, (B) T6107, (C) T618 and (D) T1382.

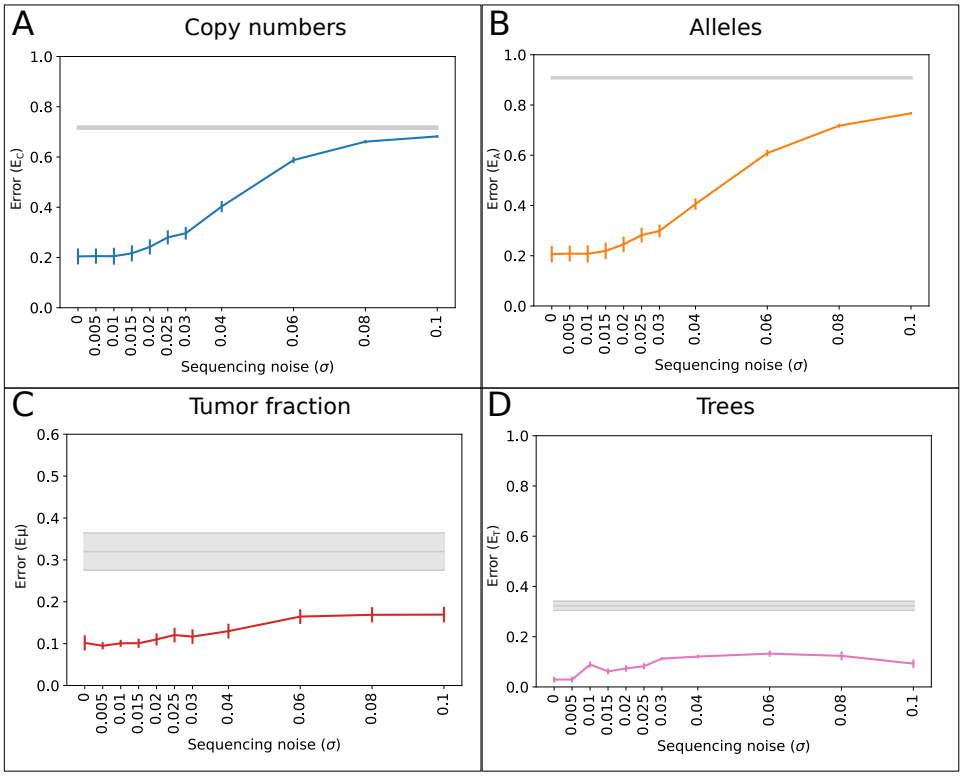


Fig. S22. Mean of the inference error and 95% confidence intervals on the TGCC-based simulations for (A) copy numbers (B) alleles (C) tumor fraction and (D) trees.

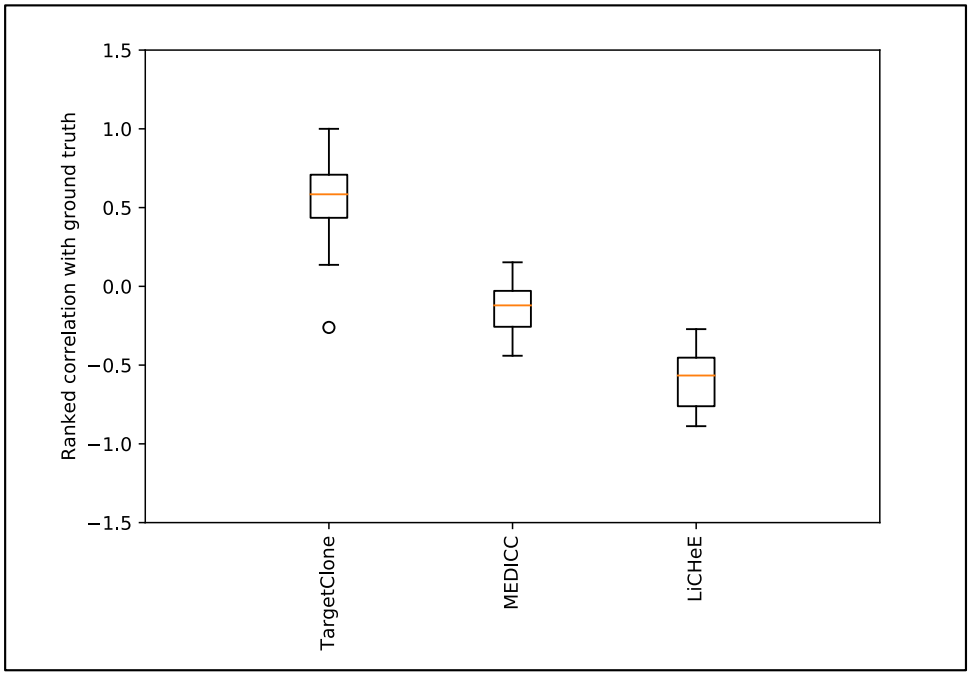


Fig. S23. Schematic representation of trees reconstructed by TargetClone, LICHeE and MEDICC for one simulation dataset. The precursor node indicates the 4N precursor. The numbers on the edges represent the estimated distances between the nodes. The tree reconstructed by LiCHeE correlates negatively with the ground truth as the distances between the precursor and pre-GCNIS nodes are larger than the distances to subclones A,B and C, whereas the ground truth distances are the opposite. A similar pattern is observed for MEDICC.

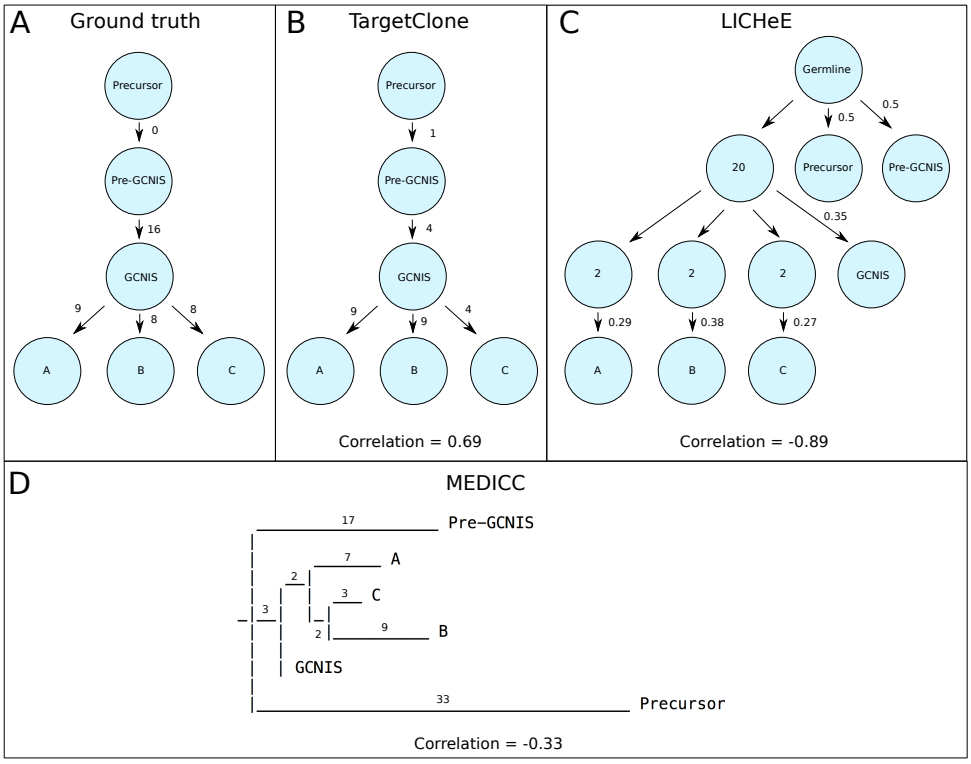


Fig. S24. Correlation of the distance matrices produced by TargetClone, MEDICC and LICHeE with the ranked ground truth distances.

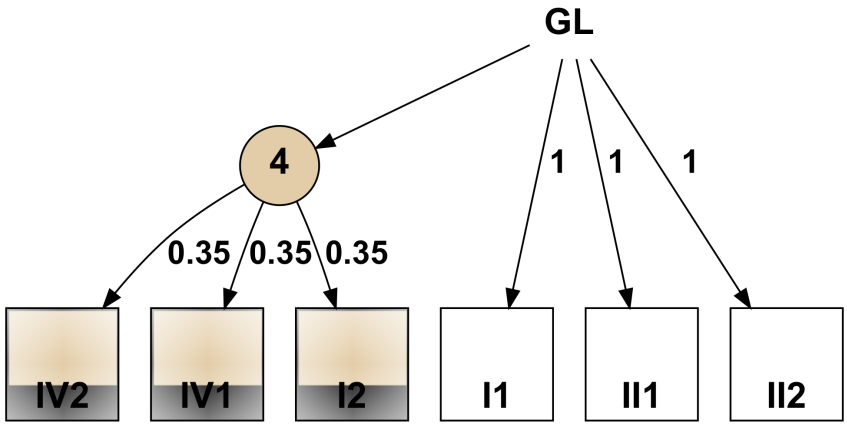


Fig. S25. Tree inferred by coupling PyClone with LICHeE for our ovarian dataset.

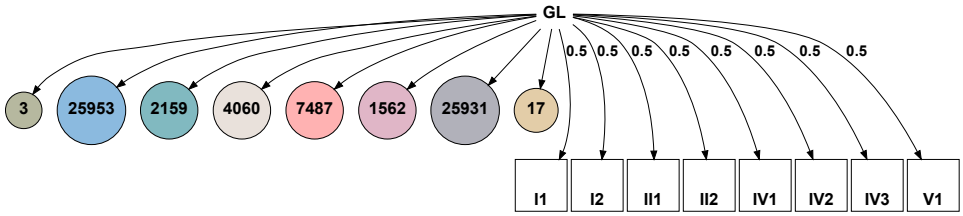


Fig. S26. Tree inferred by LICHeE for our ovarian dataset.

3

svMIL: Predicting the pathogenic effect of TAD boundary-disrupting somatic structural variants through multiple instance learning

Marleen M. Nieboer, Jeroen de Ridder

Adapted from Bioinformatics 36, i692–i699 (2020). https://doi.org/10.1093/bioinformatics/btaa802

Abstract

Motivation: Despite the fact that structural variants (SVs) play an important role in cancer, methods to predict their effect, especially for SVs in non-coding regions, are lacking, leaving them often overlooked in the clinic. Non-coding SVs may disrupt the boundaries of Topologically Associated Domains (TADs), thereby affecting interactions between genes and regulatory elements such as enhancers. However, it is not known when such alterations are pathogenic. Although machine learning techniques are a promising solution to answer this question, representing the large number of interactions that an SV can disrupt in a single feature matrix is not trivial.

Results: We introduce svMIL: a method to predict pathogenic TAD boundary-disrupting SV effects based on multiple instance learning, which circumvents the need for a traditional feature matrix by grouping SVs into bags that can contain any number of disruptions. We demonstrate that svMIL can predict SV pathogenicity, measured through same-sample gene expression aberration, for various cancer types. In addition, our approach reveals that somatic pathogenic SVs alter different regulatory interactions than somatic non-pathogenic SVs and germline SVs.

Availability: All code for svMIL is publicly available on GitHub:
https://github.com/UMCUGenetics/svMIL

Introduction

Pan-cancer genome sequencing projects, such the TCGA and PCAWG, have yielded unprecedented insights into the catalogue of somatic mutations in cancer genomes. Results from these efforts revealed that, on average, cancer genomes contain between four and five driver mutations [1]. The majority of these drivers are within the coding region of the genome. However, due to whole genome sequencing it is now clear that non-coding drivers also play an important role in cancer initiation and progression, although such driving non-coding events are scarcer than may be anticipated based on the sheer size of the non-coding genome [2].

In addition to single nucleotide variants (SNVs) and small insertions and deletions (indels), a typical cancer genome contains tens to several hundreds of somatic structural variants (SVs), which are broadly classified into simple SVs (e.g. deletions, duplications, inversions and translocations), and complex SVs (e.g. deletions flanked by insertions) [3]. While fewer in number than SNVs, due to their size, SVs affect many more bases and therefore can have consequential deleterious effects [4]. For instance, SVs may have increased impact on regulatory elements, genome architecture and the interplay between them.

One important mechanism through which non-coding SVs can exert pathogenic effects is by disrupting the boundaries between Topologically Associated Domains (TADs). TADs are regions in the genome wherein sequences physically interact with each other more frequently than with sequences outside the domain [5]. As a result, TADs are important architectural features that constrain 3D regulatory interactions of enhancers to the genes within the TAD. Disruptions of TADs and/or their boundaries, e.g. through SVs, can lead to de novo promoter-enhancer interactions resulting in aberrant expression patterns. This mechanism has been shown to play a role in causing different pathogenic

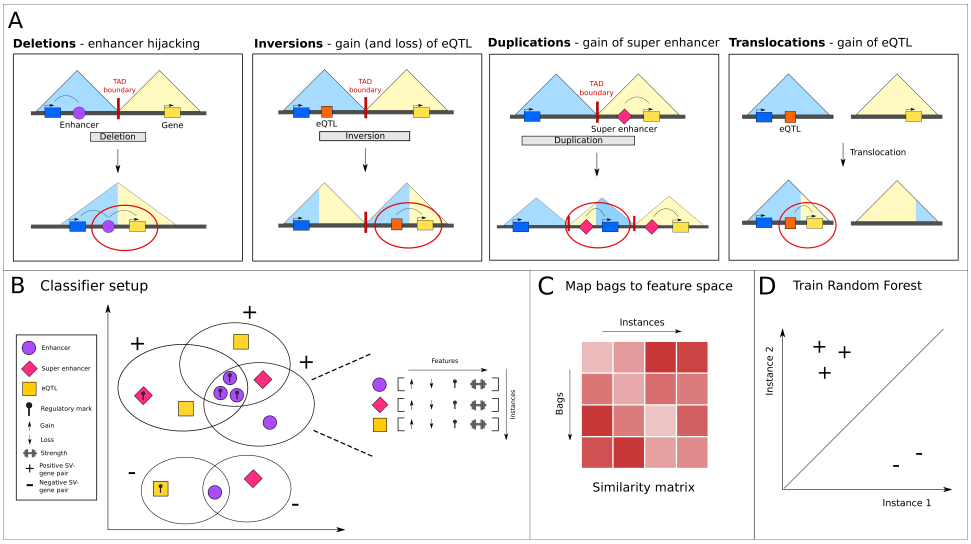


Fig. 1. Overview of the steps in the svMIL method. (A) Rules applied by our model to link non-coding SVs to their effect on genes, and some biological examples of how these effects could be caused. (B) Each SV-gene pair is a bag, which contains instances representing regulatory elements. Each instance has its own feature vector. The number of features is the same between each instance, but each bag can have a different number and different types of instances. In this example, positive bags are identified by shared affected enhancers with a specific regulatory mark. (C) In the MILES approach, bags are mapped to a feature space by constructing a bag-to-instance similarity matrix. Positive bags will have smaller distances to positive instances than to negative instances, which (D) allows for a separation in feature space using a standard classifier.

congenital phenotypes [6–10], but likely also plays a role in cancer [11–15]. For this reason, studying how and which genes are affected by non-coding SVs is important to complete the catalogue of cancer driver genes.

Despite the clear impact of non-coding SVs on genome architecture, there are no comprehensive tools available to prioritize which SVs are likely to contribute to cancer, and which ones are bystander variants. This is in stark contrast to methods that predict the effect or deleteriousness of (non-)coding SNVs and indels. Currently, VEP is the only tool with support for SVs, but cannot assign a score for all non-coding SVs [16]. SVScore was specifically designed to determine the effect of SVs by summarizing the pathogenicity scores of individual SNVs inside the SV [17]. However, this approach does not model the full spectrum of mechanisms by which SVs can cause cancer, such as through the disruption of TAD boundaries. The TAD fusion score method scores SVs by their effect on the 3D genome structure, but is limited to deletions [18]. Thus, there is currently a gap in interpreting the effects of non-coding SVs in the genome. Moreover, it was recently shown that it is not unusual for 60% of all TADs to be affected by SVs in cancer cells [19]. When simply counting how many genes are in these affected TADs, the regulation of as much as 20 genes could potentially be disrupted per TAD (Fig S1). Altogether, these factors make it very difficult to identify pathogenic SVs and further underscore the need for tools that aid in distinguishing pathogenic from non-pathogenic (bystander) SVs.

In the past, the identification of pathogenic SNVs and indels has been successfully solved with machine learning models [20–22]. However, one challenge in machine learning is defining the features that would distinguish pathogenic from non-pathogenic SVs. As the number and types of regulatory elements that are affected can differ per SV, it becomes problematic to design a rich representation of SV effects that fits in a traditional feature matrix. One approach that is particularly useful in these scenarios is known as Multiple Instance Learning (MIL). MIL is commonly explained using the analogy of a set of keychains and a door that is opened by one specific key [23]. The challenge is to distinguish between keychains that contain at least one key that opens the door (positive keychains or 'bags' in MIL terminology), from keychains that do not open the door (negative keychains or bags), without knowing which key opens the door. A keychain may contain a variable number of keys (called instances in MIL terminology) and therefore cannot be easily described in a regular feature representation. The keys, on the other hand, can be described in terms of a feature representation, such as the shape of the key and the length of the key. Several MIL classifiers have been proposed that aim to identify the feature description of so-called concepts, i.e. the key that opens the door, or that map the bags to a new feature space wherein regular classifiers can be applied, thus solving the classification problem [24–27].

Here we note that the prediction of pathogenic SVs follows a similar structure and can therefore be formulated as a MIL problem. We consider a combination of an SV and its putative target gene as a bag. Each bag contains any number of regulatory elements (the instances) which are either gained (e.g. by removal of insulating TAD boundaries) or lost (e.g. by inverting the element itself outside of the TAD). Annotations such as histone marks and chromatin states are then used as features to describe the instances.

A second challenge is defining meaningful labels, i.e. determining which bags are pathogenic (positive) and non-pathogenic (negative). A ground truth set of pathogenic

somatic SVs is not readily available. In a recent study from the PCAWG consortium, recurrence was used as a measure for pathogenicity [2]. However, as the number of significantly recurrent SVs is low, even across cancer types, using this metric on a per-SV basis is not useful. Instead, we leveraged a high-quality breast cancer dataset, generated by the Hartwig Medical Foundation (HMF), for which high-depth whole genome sequencing (> 90X) and RNA-sequencing was uniformly performed for all patients [28]. All somatic SV calls were provided by the HMF and were generated using standardized pipelines. We used these data to define positive bags as those SV-gene pairs for which the gene expression was significantly different in the sample with the SV compared to the samples without any SVs in the genomic vicinity.

In the remainder of this work, we demonstrate that svMIL can successfully separate pathogenic from non-pathogenic SVs (as defined by same-sample gene expression data), and validate this in 1 additional PCAWG cancer dataset. Furthermore, we explore the regulatory elements that are affected by the top-ranking SVs, and show that these are highly similar to our observations for known cancer genes.

Methods

Data

Whole-genome SV, SNV and CNV calls were obtained for 182 breast cancer patients from the Hartwig Medical Foundation (HMF). For 171 of these, RNA-seq data was also available. All data processing used hg19 as the reference genome. The RNA-seq data were processed using an in-house pipeline (https://github.com/UMCUGenetics/RNASeq). We excluded 9 patients that did not pass RNA-seq quality control. All 162 patients included in this work are listed in Table S1. Read counts were normalized using the Trimmed Mean of M-values (TMM) method in EdgeR [29]. For the PCAWG data, publicly available SV, SNV and CNV calls and pre-processed RNA-seq data were downloaded for 70 ovarian cancer samples from the ICGC data portal. Germline SVs were obtained from gnomAD and were randomly subsampled to match the number of SVs in the HMF breast cancer cohort (73293, 56430 deletions, 16607 duplications, 256 inversions) [30]. Because deletions are overrepresented in germline SVs, we did not select for SV type to retain the original distribution of SVs, as these could hold information about how often, and which, TADs are disrupted. Regulatory data were downloaded for the respective healthy tissue type where available, using data across cell types where stated otherwise (see also Tables S2-S3). TAD coordinates were obtained from the 3D genome browser [31]. The following data were used as regulatory elements. eQTLs were downloaded from GTEx [32]. JEME was used to obtain enhancers and target genes [33]. Promoters (across all cell types) were obtained from the Eukaryotic Promoter Database [34]. Super enhancers were obtained from dbSUPER [35]. CpG islands (across cell types) were obtained from the UCSC genome annotation database. Transcription factors (across cell types) were downloaded from ORegAnno [36]. ChromHMM states were obtained from Taberlay et al. [37]. We downloaded intrachromosomal Hi-C matrices at 5 kb resolution from Rao et al. [38], filtering out interactions occurring less than 6 times. Each side of a Hi-C interaction in this matrix was treated as a separate regulatory element of 5kb in size. Histone marks, CTCF sites, RNA pol II binding sites and DNAse I hypersensitivity sites were downloaded

from ENCODE [39]. A full list of all data sources and processing steps can be found in the Supplementary Data.

svMIL Model
The objective of our model is to rank input SVs from one or more patients by their likelihood to be involved in the development or progression of cancer. The model consists of 2 main steps:

Step 1: identifying the genes putatively disrupted by SVs: for each SV disrupting a TAD or TAD boundary, the genes are identified that could potentially be affected by re-wiring regulatory interactions between the affected TADs. This results in a list of candidate SV-gene pairs.
Step 2: learning characteristics of pathogenic SVs: for each candidate SV-gene pair, we use a machine learning approach to learn which re-wiring patterns alter gene expression and could therefore be indicative of pathogenic SVs.

The model enables us to assign a probability score to each SV reflecting its pathogenicity, which can be used as a metric to rank and classify SVs.

Step 1: identifying the genes putatively disrupted by SVs
Associating genes with regulatory elements
For every gene, we define a potential regulator set as all regulatory elements that are present within the same TAD as the gene. For eQTLs, enhancers and promoters, we further limit this set to the elements for which the respective gene was listed as a target.

Defining rules to link structural variants to putative target genes
SVs are linked to genes through the regulatory elements that they affect. More specifically, SVs cause gains and losses of regulatory elements depending on the type of SV. Therefore, we define a set of rules to determine how the potential regulator set of each gene in each patient is changed for different SV types, focusing on deletions, duplications, inversions and translocations (Fig 1A). We only include SVs that start and end in a TAD.

Deletions - If a deletion overlaps a TAD boundary, all genes in the TADs on either side of the deletion gain the regulatory elements from the TAD on the other side of the deletion.
Inversions - For inversions, genes lose regulatory elements that are inverted outside of the respective TAD, and gain elements that are inverted into the TAD. The genes residing in the inversion gain regulatory elements of the TAD that these genes are inverted into.
Duplications - If a duplication crosses a TAD boundary, it will generate a new TAD. Within this new TAD, genes in the duplication on the one side of the TAD boundary will be brought into contact with regulatory elements in the duplication on the other side of the TAD boundary.
Translocations - For each translocation independently, a derivative TAD is constructed based on the SV orientation. The gains and losses in the regulator set of every gene inside this new TAD are then determined.

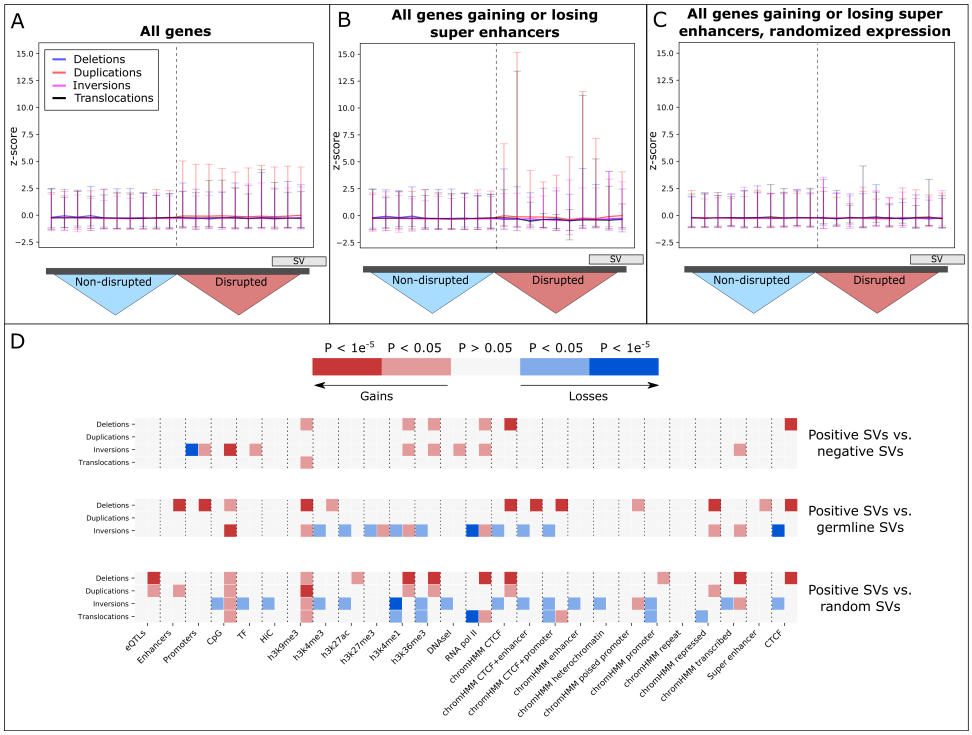


Fig. 2. (A) The z-scores of all genes in disrupted TADs (right half) compared to the adjacent, non-disrupted TADs (left half), shown for each SV type. The two sides of the disrupted TAD pair, and the adjacent TADs on the left and right, were collapsed. The trend indicates the median z-score in each bin. Error bars indicate the 95th and 5th percentiles. (B) The z-scores shown specifically for genes that gain or lose super enhancers, as identified using the rules. (C) The z-scores shown in the same TADs as in (B), but when the gene expression is randomly shuffled. (D) Gains and losses of regulatory information that are significantly different between the positive set of SVs (effect on gene expression), negative set (no effect of gene expression), germline SVs, and when the positions of the positive SVs are shuffled randomly.

Applying these rules results in a list of candidate SV-gene pairs and the associated gained and lost regulatory elements that are the result of the SV for this gene. To ensure that we are only looking at non-coding effects on genes, we removed SV-gene pairs for which the gene is overlapped (minimum 1 bp) by another SV, SNV, or CNV in the same patient sample, as it is assumed that such coding events are the likely cause of any aberrant gene expression. As the affected genes can be overlapped by the respective SV itself, we make an exception for duplications and inversions (Fig 1A). For inversions, we do not remove genes that are overlapped by the inversion when these are not affected by any other mutation. For duplications, we similarly keep the genes that are only overlapped by the duplication. As duplications often coincide with copy number (CN) amplifications (CN > 2.3), we allow genes to be affected by such events only if these are also linked to a duplication by the rules.

Determining genes with altered expression

For each SV-gene pair, identified using the rules, we computed a z-score by comparing the expression of the gene in the patient with the SV to a null-distribution constructed based on all other patients (one-sample t-test). To ensure that the expression is changed by the non-coding SV specifically, we constructed the null-distribution only from patients that do not have an SNV, SV or CNV overlapping the gene. In addition, we removed potential non-coding effects from the null-distribution by excluding genes in patients where an SV disrupts the boundary of the TAD in which the gene is located. The z-scores are used as a measure of SV effect on the respective gene.

Step 2: machine learning to learn the characteristics of pathogenic SVs

To learn which SV-gene pair is likely pathogenic, and investigate if the gain or loss of specific regulatory elements are predictive for this, we trained a MIL model (Fig 1B).

Defining bags and instances

Every SV-gene pair is considered a bag containing regulatory elements, the instances. The instances in each bag are defined as the regulatory elements lost or gained by the gene affected by the SV, as determined by the rules. We used a combination of eQTLs, enhancers and super enhancers as instances.

Instance features

Every instance is described with a feature vector combining three layers of features (Fig 1B). The first layer consists of two features, indicating if the regulatory element is gained or lost. The second layer contains annotations of the region in which the regulatory element is located. This includes histone marks (H3K9me3, H3K4me3, H3K27ac, H3K27me3, H3K4me1, H3K36me3), chromHMM states (CTCF, CTCF+enhancer, CTCF+promoter, enhancer, promoter, poised promoter, heterochromatin, repeat, repressed, transcribed), transcription factor binding profiles (DNASeI hypersensitivity sites, RNA polymerase II, CTCF, transcription factor binding sites), CpG islands and Hi-C interacting regions. The feature vector contains a 0 or 1 depending on if the regulatory element overlaps any of these annotations (minimum 1 bp). Finally, we added a third layer to indicate the strength of the annotations. For enhancers, the prediction confidence score was used.

For all histone marks, RNA polymerase II and CTCF, the peak intensity was used as an indicator of strength. This information was not available for any of the other annotations, and was thus left out. All features were normalized between 0 and 1.

Bag labels
We used the z-score of gene expression compared to unaffected TADs as a proxy for pathogenicity. Bags for SV-gene pairs with a z-score larger than 1.5 or smaller than -1.5 were labeled positive, and negative otherwise. To obtain equal class sizes, we randomly subsampled the negative bags to the same number of positive bags.

Multiple Instance Learning model
To obtain a final classifier, we used the MILES approach [27]. One important feature of MILES is that it allows reconstructing which features of gained or lost regulatory elements are associated with positive SV-gene pairs. In MILES, the general idea is to map the bags to a feature space in which a standard classifier can be trained (Fig 1C). This feature space is constructed by computing a similarity score between every bag and every instance. Positive bags are expected to be more similar to instances of other positive bags, but dissimilar to negative instances, therefore creating a separation in feature space (Fig 1D). As all regulatory elements in each bag could equally contribute to the effect on gene expression, we compute the similarity score by first averaging the features of all instances in each bag, and then computing the absolute distance to all other instances of the other bags (collective assumption) [40]. A random forest is trained on the matrix of similarity scores to classify the SV-gene pairs. As the similarity matrix represents distances between bags as objects to instances as features, we used the random forest feature importance to rank individual instances.

Performance evaluation using cross-validation
We evaluate the model performance using 3 cross-validation (CV) approaches. The first is a bag-based CV, in which bags are randomly distributed into the training and test set across 10 folds. To mimic a clinical setting, we use a leave-one-patient-out CV, in which all bags of one patient are held out in each fold. Lastly, we use a leave-one-chromosome-out CV, in which each chromosome is left out in every fold, to test for spatial correlation between SV-gene pairs. For each approach, in every fold one similarity matrix is constructed for the training bags, and one for the test bags, for which the similarity is computed to the instances of the training bags. Folds are stratified by randomly subsampling the number of negative bags to match the number of positive bags. The classifier was optimized using a random parameter search, applying 3-fold CV directly on the similarity matrix to reduce computational time.

Results

Altered gene expression is only visible in TADs disrupted by SVs
To show if SVs can affect gene expression by disrupting TADs and TAD boundaries, we compared the expression of genes in disrupted to non-disrupted TADs in the breast cancer patients. We define the TADs that an SV ends in on the left and right side as a 'TAD pair'. We computed the z-score of gene expression inside these TAD pairs to all patients

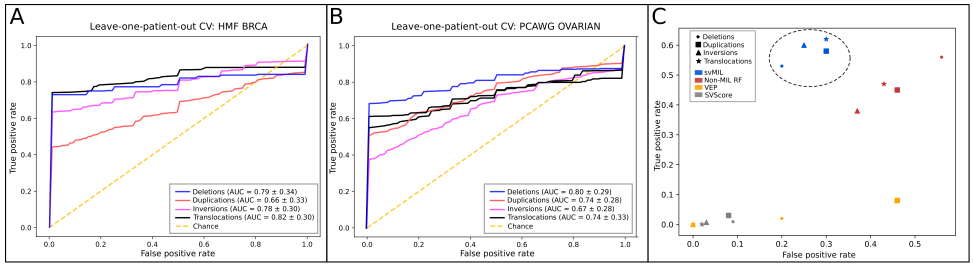


Fig. 3. ROC curves based on leave-one-patient out CV for the models trained on each SV type for (A) the HMF BRCA dataset and (B) the PCAWG ovarian cancer dataset. (C) The TPR and FPR of svMIL compared to 3 other methods. svMIL is highlighted by the dotted oval.

in which this pair is not disrupted by an SV, and filtered out mutated genes (see Methods). As a negative control, we repeated this procedure for the TADs immediately to the left and right of each pair, keeping only adjacent TADs that are not disrupted by SVs in the respective patient. To account for varying TAD size, each TAD was divided into 10 bins. A minor increase in z-score is visible in the affected TADs, in particular for duplications and translocations (Fig 2A). To determine if the gene expression is altered specifically by SVs and not due to effects such as methylation, we focused on all genes gaining or losing super enhancers as identified using the rules. Here also an overall increase in gene expression is visible (Fig 2B), which is not observed when randomly shuffling the expression (Fig 2C). Thus, we see that SVs are able to alter gene expression by re-shaping interactions between genes and regulatory elements.

Somatic SVs affect a unique combination of regulatory elements

To determine if SVs alter gene expression by affecting specific classes of regulatory elements, we compared their gain and loss patterns to those of SVs that do not affect expression (Fig 2D, positive vs negative SVs). Overall, we observe significant differences for each SV type ($P < 0.05$, χ^2 test with Bonferroni correction). These patterns are not observed when comparing to a case where the somatic SVs are assigned random positions (maintaining their size, type and chromosome), indicating that we do not find these differences by random chance (Fig 2D, positive vs random SVs). Furthermore, the different gains and losses of regulatory elements observed when comparing to germline SVs suggest that somatic SVs occur at different genomic positions with different effects on regulatory elements (Fig 2D, positive vs germline SVs). Taken together, these findings indicate that disrupted interactions with regulatory elements contain information on the pathogenicity of somatic SVs.

svMIL can successfully identify pathogenic SVs in various cancer types

To determine if re-wiring patterns are informative predictors of pathogenicity, we trained MIL models with a random forest classifier to separate SV-gene pairs with large effects on gene expression from SV-gene pairs with no effects (Methods). A model was trained on each SV type individually, using the same number of bags in each class (deletions: 168, duplications: 906, inversions: 338 and translocations: 133).

To mimic a clinical setting, in which prioritization of the SVs in the cancer genome of a new and unseen patient is required, we performed leave-one-patient-out CV. We find that svMIL can successfully identify pathogenic SVs in unseen patients with AUCs of 0.79, 0.66, 0.78 and 0.82 for deletions, duplications, inversions and translocations, respectively (Fig 3A). Notably, such classification performance is not observed when the SV positions are shuffled randomly (Fig S2A). Thus, our method is suitable to predict pathogenic SVs in clinical settings.

As a note of warning, bag-based CV, which is the classical CV strategy, yields much higher performances (Fig S2B). However, we observed that patients frequently have multiple spatially clustered SVs affecting the same gene, causing gains and losses of the same instances. Since these pairs are randomly distributed across the training and test set in each fold, this may cause some information leakage and biased CV results. Indeed, when validating our model in a per-chromosome CV setting, similar results were obtained to the leave-one-patient-out CV (Fig S2C). A similar issue could arise in the leave-one-patient-out CV if many SVs are shared between multiple patients. Therefore, caution is advised when interpreting results of bag-based CV or leave-one-patient-out CV settings in these situations.

Furthermore, we realize that our p-value threshold for eQTLs of $P < 5e-8$ is especially stringent, and that lowering the threshold to $P < 0.05$ improves the AUC for deletions to 0.88. However, due to the sheer increase in the number of eQTLs, lowering the threshold increases the run time to up to 24 hours for deletions alone, and is therefore not realistic to run in routine analysis. Nevertheless, these results reveal potential for further improvement in predictive ability.

Finally, we aimed to demonstrate the performance in other cancers. To this end, we retrained svMIL on ovarian cancer samples from the PCAWG dataset. Notably, the number of bags is slightly larger than for our breast cancer dataset (deletions: 256, duplications: 1009, inversions: 818, translocations: 229). Nevertheless, the ROC curves demonstrate that also for ovarian cancer, similar performances are obtained (Fig 3B), indicating that our method is applicable to various cancer types.

svMIL outperforms the state-of-the-art methods

Next, we aimed to show how our method compares to the state-of-the-art. We compared the true positive rate (TPR) and false positive rate (FPR) of our method to VEP, SVScore and a random forest classifier without MIL (non-MIL-RF). For VEP, SVs with a 'moderate' or 'high' impact score were considered pathogenic. For SVScore, pathogenic SVs were selected using scores above the 90th percentile (for each SV type separately). For the non-MIL-RF, we used gains and losses of regulatory elements of each SV-gene pair as binary features. If multiple regulatory elements of a type are affected, the feature value was capped at 1. We used an operating point of 0.5 for both svMIL and non-MIL-RF and tested performance using leave-one-patient-out CV. For each method, an SV is considered positive if it is part of a positive SV-gene pair.

Overall, we see that svMIL obtains higher TPR than the other methods when predicting SV pathogenicity in unseen patients (Fig 3C). Non-MIL-RF also scores high TPR, but at a cost of increased FPR compared to svMIL, showing the benefit of identifying pathogenic SVs using a MIL-based approach. VEP and SVScore both score low FPR, but

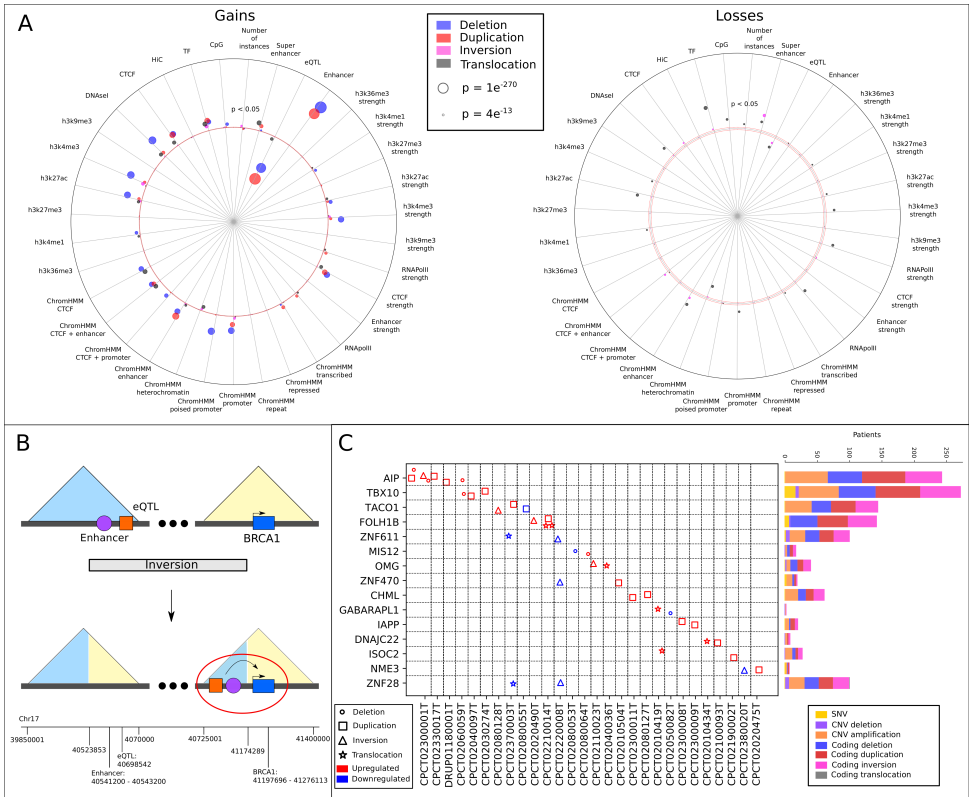


Fig. 4. (A) Gains and losses in the top 100 instances per SV type. Values above the red dotted lines are significant ($p < 0.05$, Bonferroni) with a z-score larger than 0 (more than in 100 randomly selected instances), values below the red dotted line with a z-score smaller than 0 (less than in 100 randomly selected instances). Everything within the red dotted lines is not significant. (B) Inversion bringing enhancers and eQTLs from another TAD close to BRCA1, resulting in overexpression. For simplicity, only 1 enhancer and 1 eQTL is shown. (C) Top 15 genes most recurrently affected across patients by SVs within the top 100 instances of each SV type. Only patients with positive SV-gene pairs are included in the analysis. The bars show the number of patients with a coding mutation overlapping that same gene. Some patients have multiple mutations.

also never have TPR above 0.1. As both methods do not include TAD disruptions or gene expression in their predictions, these results show that performance can be improved if such additional data layers are integrated.

SVs frequently re-wire active (super) enhancers in open chromatin regions

Using our trained MIL classifiers, we investigated which regulatory elements are most commonly associated with pathogenic SVs. We obtained a ranked list of instances for each MIL model from the random forest feature importance scores. As a threshold, we focused on the top 100 instances of each model, which retains the majority of information for each SV type (Fig S3). Overall, we find that the affected regulatory elements are significantly associated with open chromatin regions and histone marks enriched

in active promoters and enhancers (compared to 100 randomly sampled instances, $p < 0.05$, Bonferroni) (Fig 4A). Each SV type affects a unique set of regulatory elements. For instance, deletions and duplications frequently cause gains of active enhancers, while inversions and translocations instead more often result in gains and losses of super enhancers, revealing a possible different mechanism by which target genes are affected. We furthermore note that the majority of top-ranked instances are gains rather than losses (89/100 and 58/100 gains for inversions and translocations, respectively), indicating that gains of regulatory elements could be a preferred method to affect gene expression in breast cancer.

To determine if the gains and losses could potentially be causal for cancer, we compared the instances within the ranked top 100 that affect known cancer genes (COSMIC genes; 5, 5, 1 and 4 out of 100 instances for deletions, duplications, inversions and translocations, respectively) to those affecting non-COSMIC genes (Fig S4). Overall, we see that COSMIC genes and non-COSMIC genes show similar patterns, indicating a potential similar effect on genes. Taken together, these findings suggest that our ranking can successfully identify pathogenic mechanisms of SVs and bona fide target genes of SVs.

svMIL can identify driver genes in unseen patients

Motivated by the similarities in gained and lost regulatory elements between known cancer genes and other genes, we investigated if our method can identify potential novel driver genes. To this end, we employed leave-one-patient-out CV, specifically for the patients with an SV linked to a known COSMIC gene. This enabled us to determine how often a known COSMIC gene is correctly classified as pathogenic, using a classifier trained on data that has never seen this SV-gene pair. In total, 34 (out of 64) SV-COSMIC gene pairs were predicted correctly, which are significantly more COSMIC genes than expected by random chance (P = 3.86e-172, t-test, compared to 100 iterations of leave-one-patient-out CV where random SV-gene pairs were assigned to the bags after the classification step). For none of these COSMIC genes there was any other evidence, such as SNVs or indels, that could have disrupted it, which means it would not have been identified otherwise. Notably, 4 of these SV-gene pairs identified genes specific to breast cancer (1 inversion targeting BRCA1, and 3 duplications targeting ERBB2 in the same patient). Of particular interest is the inversion affecting BRCA1, bringing 3 enhancers and 46 eQTLs in close proximity to the gene (Fig 4B). High BRCA1 expression has previously been linked to worse prognosis in several cancers, including breast cancer, and could therefore be an interesting finding for selecting treatment [41, 42]. In addition, BRCA1 is upregulated in 11.4% of breast cancer patients (CGC), indicating that this SV could be an alternative pathway to upregulate the gene [43]. Altogether, these results demonstrate the importance of incorporating non-coding SVs in the analysis of whole cancer genome sequencing data of patients.

In addition, we investigated if genes linked to high-ranking SVs are recurrently affected ($z > 1.5$ or $z < -1.5$) across different patients by other non-coding SVs, and could therefore be putative cancer drivers. We obtained the genes from the top 100 ranked instances of all SV types, and report the top 15 most recurrently mutated in Fig 4C. Some patients have multiple non-coding SVs targeting the same gene, indicative of selective

pressure. Most genes affected by deletions and duplications are only upregulated in their respective patients, while we see both up- and downregulation for inversions and translocations, showing that the majority of SVs can indeed be responsible for the gene expression change. Moreover, we identify a large number of SNVs, CNVs and coding SVs targeting the same gene in other patients, showing that non-coding SVs could be an alternative route to affect important genes in the cancer genome.

Finally, we note that the number of patients in which the top 15 genes are recurrently affected is not significant ($P < 0.05$, Bonferroni, compared to 100 distributions of randomly sampled positive SV-gene pairs) and appears to be much smaller than for other mutation types, which has also been reported previously [2]. In addition, we identify different recurrently affected genes if we do not filter by the top-ranked instances (Fig S5), supporting the importance of applying machine learning-based methods to identify pathogenic SVs, and including gene expression into these models.

Discussion

In this work, we described svMIL, a novel MIL-based method to rank SVs for their likelihood to be pathogenic based on altering interactions between genes and regulatory elements and thereby disrupting gene expression. Not all genes in disrupted TADs show affected expression levels. However, the genes that are affected can be pinpointed by looking at gains and losses of regulatory elements caused by SVs. Using our MIL model, we can now utilize this information to identify pathogenic SVs in various cancer types. We demonstrated, by mimicking a clinical setting using leave-one-patient-out CV, that svMIL can successfully identify pathogenic SVs in unseen patients. Within these unseen patients, we identified SVs affecting known cancer genes that were not disrupted by any other mutations, such as SNVs, thus revealing the importance of also studying non-coding SVs in clinical analyses. As such methods to predict the effect of non-coding SVs are currently lacking, our model provides an opportunity to study existing and future SV datasets in more detail.

We showed that top-ranking pathogenic SVs frequently affect active (super) enhancers in open chromatin regions. Furthermore, regulatory elements are more frequently gained than lost in the breast cancer samples, showing a possible preference to upregulate genes in these patients. Many genes disrupted by top-ranking pathogenic SVs are recurrently affected by non-coding SVs across multiple patients, with frequent evidence of other mutations, such as CNVs, disrupting the same genes. These findings show that non-coding SVs could be an important alternative strategy to disrupt genes that are important in cancer cells. As these same genes could not be identified independent of the ranking by pathogenicity produced by svMIL, it highlights the benefit of utilizing gene expression information and MIL models to identify pathogenic SVs.

As more and more data are becoming available for cancer patients, our ability to predict pathogenicity will continue to improve. For example, the absence of methylation and 3D folding data for each patient limits our ability to perfectly label which genes are truly affected by the SV only. Additionally, an increase in cell-type specific and validated regulatory elements will make it possible to further fine-tune predictive models.

Nevertheless, our presented model can aid in understanding the consequences of SVs in more detail, allowing the generation of new hypotheses about the role of SVs in

cancer.

Data availability
The WGS breast cancer data were requested from the Hartwig Medical Foundation and provided under data request DR-066.

References

[1] *Pan-cancer analysis of whole genomes,* Nature **578**, 82 (2020).

[2] E. Rheinbay, M. M. Nielsen, F. Abascal, J. A. Wala, O. Shapira, G. Tiao, H. Hornshøj, J. M. Hess, R. I. Juul, Z. Lin, L. Feuerbach, R. Sabarinathan, T. Madsen, J. Kim, L. Mularoni, S. Shuai, A. Lanzós, C. Herrmann, Y. E. Maruvka, C. Shen, S. B. Amin, P. Bandopadhayay, J. Bertl, K. A. Boroevich, J. Busanovich, J. Carlevaro-Fita, D. Chakravarty, C. W. Y. Chan, D. Craft, P. Dhingra, K. Diamanti, N. A. Fonseca, A. Gonzalez-Perez, Q. Guo, M. P. Hamilton, N. J. Haradhvala, C. Hong, K. Isaev, T. A. Johnson, M. Juul, A. Kahles, A. Kahraman, Y. Kim, J. Komorowski, K. Kumar, S. Kumar, D. Lee, K.-V. Lehmann, Y. Li, E. M. Liu, L. Lochovsky, K. Park, O. Pich, N. D. Roberts, G. Saksena, S. E. Schumacher, N. Sidiropoulos, L. Sieverling, N. Sinnott-Armstrong, C. Stewart, D. Tamborero, J. M. C. Tubio, H. M. Umer, L. Uuskü-la-Reimand, C. Wadelius, L. Wadi, X. Yao, C.-Z. Zhang, J. Zhang, J. E. Haber, A. Hobolth, M. Imielinski, M. Kellis, M. S. Lawrence, C. von Mering, H. Nakagawa, B. J. Raphael, M. A. Rubin, C. Sander, L. D. Stein, J. M. Stuart, T. Tsunoda, D. A. Wheeler, R. Johnson, J. Reimand, M. Gerstein, E. Khurana, P. J. Campbell, N. López-Bigas, J. Weischenfeldt, R. Beroukhim, I. Martincorena, J. S. Pedersen, and G. Getz, *Analyses of non-coding somatic drivers in 2,658 cancer whole genomes,* Nature **578**, 102 (2020).

[3] Li *et al.*, *Patterns of somatic structural variation in human cancer genomes,* Nature **578**, 112 (2020).

[4] Sudmant *et al.*, *An integrated map of structural variation in 2,504 human genomes,* Nature **526**, 75 (2015).

[5] J. R. Dixon, S. Selvaraj, F. Yue, A. Kim, Y. Li, Y. Shen, M. Hu, J. S. Liu, and B. Ren, *Topological domains in mammalian genomes identified by analysis of chromatin interactions,* Nature **485**, 376 (2012).

[6] E. Giorgio, D. Robyr, M. Spielmann, E. Ferrero, E. Di Gregorio, D. Imperiale, G. Vaula, G. Stamoulis, F. Santoni, C. Atzori, L. Gasparini, D. Ferrera, C. Canale, M. Guipponi, L. A. Pennacchio, S. E. Antonarakis, A. Brussino, and A. Brusco, *A large genomic deletion leads to enhancer adoption by the lamin B1 gene: a second path to autosomal dominant adult-onset demyelinating leukodystrophy (ADLD),* Human Molecular Genetics **24**, 3143 (2015).

[7] C. Redin, H. Brand, R. L. Collins, T. Kammin, E. Mitchell, J. C. Hodge, C. Hanscom, V. Pillalamarri, C. M. Seabra, M.-A. Abbott, O. A. Abdul-Rahman, E. Aberg, R. Adley, S. L. Alcaraz-Estrada, F. S. Alkuraya, Y. An, M.-A. Anderson, C. Antolik, K. Anyane-Yeboa, J. F. Atkin, T. Bartell, J. A. Bernstein, E. Beyer, I. Blumenthal, E. M. H. F.

Bongers, E. H. Brilstra, C. W. Brown, H. T. Brüggenwirth, B. Callewaert, C. Chiang, K. Corning, H. Cox, E. Cuppen, B. B. Currall, T. Cushing, D. David, M. A. Deardorff, A. Dheedene, M. D'Hooghe, B. B. A. de Vries, D. L. Earl, H. L. Ferguson, H. Fisher, D. R. FitzPatrick, P. Gerrol, D. Giachino, J. T. Glessner, T. Gliem, M. Grady, B. H. Graham, C. Griffis, K. W. Gripp, A. L. Gropman, A. Hanson-Kahn, D. J. Harris, M. A. Hayden, R. Hill, R. Hochstenbach, J. D. Hoffman, R. J. Hopkin, M. W. Hubshman, A. M. Innes, M. Irons, M. Irving, J. C. Jacobsen, S. Janssens, T. Jewett, J. P. Johnson, M. C. Jongmans, S. G. Kahler, D. A. Koolen, J. Korzelius, P. M. Kroisel, Y. Lacassie, W. Lawless, E. Lemyre, K. Leppig, A. V. Levin, H. Li, H. Li, E. C. Liao, C. Lim, E. J. Lose, D. Lucente, M. J. Macera, P. Manavalan, G. Mandrile, C. L. Marcelis, L. Margolin, T. Mason, D. Masser-Frye, M. W. McClellan, C. J. Z. Mendoza, B. Menten, S. Middelkamp, L. R. Mikami, E. Moe, S. Mohammed, T. Mononen, M. E. Mortenson, G. Moya, A. W. Nieuwint, Z. Ordulu, S. Parkash, S. P. Pauker, S. Pereira, D. Perrin, K. Phelan, R. E. P. Aguilar, P. J. Poddighe, G. Pregno, S. Raskin, L. Reis, W. Rhead, D. Rita, I. Renkens, F. Roelens, J. Ruliera, P. Rump, S. L. P. Schilit, R. Shaheen, R. Sparkes, E. Spiegel, B. Stevens, M. R. Stone, J. Tagoe, J. V. Thakuria, B. W. van Bon, J. van de Kamp, I. van Der Burgt, T. van Essen, C. M. van Ravenswaaij-Arts, M. J. van Roosmalen, S. Vergult, C. M. L. Volker-Touw, D. P. Warburton, M. J. Waterman, S. Wiley, A. Wilson, M. d. l. C. A. Yerena-de Vega, R. T. Zori, B. Levy, H. G. Brunner, N. de Leeuw, W. P. Kloosterman, E. C. Thorland, C. C. Morton, J. F. Gusella, and M. E. Talkowski, *The genomic landscape of balanced cytogenetic abnormalities associated with human congenital anomalies,* Nature Genetics **49**, 36 (2017).

[8] M. Franke, D. M. Ibrahim, G. Andrey, W. Schwarzer, V. Heinrich, R. Schöpflin, K. Kraft, R. Kempfer, I. Jerković, W.-L. Chan, M. Spielmann, B. Timmermann, L. Wittler, I. Kurth, P. Cambiaso, O. Zuffardi, G. Houge, L. Lambie, F. Brancati, A. Pombo, M. Vingron, F. Spitz, and S. Mundlos, *Formation of new chromatin domains determines pathogenicity of genomic duplications,* Nature **538**, 265 (2016).

[9] D. G. Lupiáñez, K. Kraft, V. Heinrich, P. Krawitz, F. Brancati, E. Klopocki, D. Horn, H. Kayserili, J. M. Opitz, R. Laxova, F. Santos-Simarro, B. Gilbert-Dussardier, L. Wittler, M. Borschiwer, S. A. Haas, M. Osterwalder, M. Franke, B. Timmermann, J. Hecht, M. Spielmann, A. Visel, and S. Mundlos, *Disruptions of Topological Chromatin Domains Cause Pathogenic Rewiring of Gene-Enhancer Interactions,* Cell **161**, 1012 (2015).

[10] Zhang *et al.*, *Local and global chromatin interactions are altered by large genomic deletions associated with human brain development,* Nature Communications **9**, 5356 (2018).

[11] D. Hnisz, A. S. Weintraub, D. S. Day, A.-L. Valton, R. O. Bak, C. H. Li, J. Goldmann, B. R. Lajoie, Z. P. Fan, A. A. Sigova, J. Reddy, D. Borges-Rivera, T. I. Lee, R. Jaenisch, M. H. Porteus, J. Dekker, and R. A. Young, *Activation of proto-oncogenes by disruption of chromosome neighborhoods,* Science **351**, 1454 (2016).

[12] J. Weischenfeldt, T. Dubash, A. P. Drainas, B. R. Mardin, Y. Chen, A. M. Stütz, S. M. Waszak, G. Bosco, A. R. Halvorsen, B. Raeder, T. Efthymiopoulos, S. Erkek, C. Siegl,

H. Brenner, O. T. Brustugun, S. M. Dieter, P. A. Northcott, I. Petersen, S. M. Pfister, M. Schneider, S. K. Solberg, E. Thunissen, W. Weichert, T. Zichner, R. Thomas, M. Peifer, A. Helland, C. R. Ball, M. Jechlinger, R. Sotillo, H. Glimm, and J. O. Korbel, *Pan-cancer analysis of somatic copy-number alterations implicates IRS4 and IGF2 in enhancer hijacking,* Nature Genetics **49**, 65 (2017).

[13] J. R. Dixon, J. Xu, V. Dileep, Y. Zhan, F. Song, V. T. Le, G. G. Yardımcı, A. Chakraborty, D. V. Bann, Y. Wang, R. Clark, L. Zhang, H. Yang, T. Liu, S. Iyyanki, L. An, C. Pool, T. Sasaki, J. C. Rivera-Mulia, H. Ozadam, B. R. Lajoie, R. Kaul, M. Buckley, K. Lee, M. Diegel, D. Pezic, C. Ernst, S. Hadjur, D. T. Odom, J. A. Stamatoyannopoulos, J. R. Broach, R. C. Hardison, F. Ay, W. S. Noble, J. Dekker, D. M. Gilbert, and F. Yue, *Integrative detection and analysis of structural variation in cancer genomes,* Nature Genetics **50**, 1388 (2018).

[14] A.-L. Valton and J. Dekker, *TAD disruption as oncogenic driver,* Current Opinion in Genetics and Development **36**, 34 (2016).

[15] K. C. Akdemir, V. T. Le, S. Chandran, Y. Li, R. G. Verhaak, R. Beroukhim, P. J. Campbell, L. Chin, J. R. Dixon, and P. A. Futreal, *Disruption of chromatin folding domains by somatic genomic rearrangements in human cancer,* Nature Genetics **52**, 294 (2020).

[16] W. McLaren, L. Gil, S. E. Hunt, H. S. Riat, G. R. S. Ritchie, A. Thormann, P. Flicek, and F. Cunningham, *The Ensembl Variant Effect Predictor,* Genome Biology **17**, 122 (2016).

[17] Ganel *et al.*, *SVScore: an impact prediction tool for structural variation,* Bioinformatics , btw789 (2016).

[18] Huynh *et al.*, *TAD fusion score: discovery and ranking the contribution of deletions to genome structure,* Genome Biology **20**, 60 (2019).

[19] Y. Zhang, L. Yang, M. Kucherlapati, F. Chen, A. Hadjipanayis, A. Pantazi, C. A. Bristow, E. A. Lee, H. S. Mahadeshwar, J. Tang, J. Zhang, S. Seth, S. Lee, X. Ren, X. Song, H. Sun, J. Seidman, L. J. Luquette, R. Xi, L. Chin, A. Protopopov, W. Li, P. J. Park, R. Kucherlapati, and C. J. Creighton, *A Pan-Cancer Compendium of Genes Deregulated by Somatic Genomic Rearrangement across More Than 1,400 Cases,* Cell Reports **24**, 515 (2018).

[20] J. Zhou and O. G. Troyanskaya, *Predicting effects of noncoding variants with deep learning–based sequence model,* Nature Methods **12**, 931 (2015).

[21] Kircher *et al.*, *A general framework for estimating the relative pathogenicity of human genetic variants,* Nature Genetics **46**, 310 (2014).

[22] M. F. Rogers, H. A. Shihab, M. Mort, D. N. Cooper, T. R. Gaunt, and C. Campbell, *FATHMM-XF: accurate prediction of pathogenic point mutations via extended features,* Bioinformatics **34**, 511 (2018).

[23] Dietterich *et al.*, *Solving the multiple instance problem with axis-parallel rectangles,* Artificial Intelligence **89**, 31 (1997).

[24] Zhou *et al.*, *Promoter prediction based on a multiple instance learning scheme,* in *Proceedings of the First ACM International Conference on Bioinformatics and Computational Biology - BCB '10* (ACM Press, New York, New York, USA, 2010) p. 295.

[25] Panwar *et al.*, *Genome-Wide Functional Annotation of Human Protein-Coding Splice Variants Using Multiple Instance Learning,* Journal of Proteome Research **15**, 1747 (2016).

[26] Bandyopadhyay *et al.*, *MBSTAR: multiple instance learning for predicting specific functional binding sites in microRNA targets,* Scientific Reports **5**, 8004 (2015).

[27] Chen *et al.*, *MILES: Multiple-Instance Learning via Embedded Instance Selection,* IEEE Transactions on Pattern Analysis and Machine Intelligence **28**, 1931 (2006).

[28] P. Priestley, J. Baber, M. P. Lolkema, N. Steeghs, E. de Bruijn, C. Shale, K. Duyvesteyn, S. Haidari, A. van Hoeck, W. Onstenk, P. Roepman, M. Voda, H. J. Bloemendal, V. C. G. Tjan-Heijnen, C. M. L. van Herpen, M. Labots, P. O. Witteveen, E. E. Smit, S. Sleijfer, E. E. Voest, and E. Cuppen, *Pan-cancer whole-genome analyses of metastatic solid tumours,* Nature **575**, 210 (2019).

[29] Robinson *et al.*, *A scaling normalization method for differential expression analysis of RNA-seq data,* Genome Biology **11**, R25 (2010).

[30] Collins *et al.*, *A structural variation reference for medical and population genetics,* Nature **581**, 444 (2020).

[31] Wang *et al.*, *The 3D Genome Browser: a web-based browser for visualizing 3D genome organization and long-range chromatin interactions,* Genome Biology **19**, 151 (2018).

[32] Aguet *et al.*, *Genetic effects on gene expression across human tissues,* Nature **550**, 204 (2017).

[33] Q. Cao, C. Anyansi, X. Hu, L. Xu, L. Xiong, W. Tang, M. T. S. Mok, C. Cheng, X. Fan, M. Gerstein, A. S. L. Cheng, and K. Y. Yip, *Reconstruction of enhancer–target networks in 935 samples of human primary cells, tissues and cell lines,* Nature Genetics **49**, 1428 (2017).

[34] Dreos *et al.*, *The eukaryotic promoter database in its 30th year: focus on non-vertebrate organisms,* Nucleic Acids Research **45**, D51 (2017).

[35] Khan *et al.*, *dbSUPER: a database of super-enhancers in mouse and human genome,* Nucleic Acids Research **44**, D164 (2016).

[36] Lesurf *et al.*, *ORegAnno 3.0: a community-driven resource for curated regulatory annotation,* Nucleic Acids Research **44**, D126 (2016).

[37] Taberlay *et al.*, *Reconfiguration of nucleosome-depleted regions at distal regulatory elements accompanies DNA methylation of enhancers and insulators in cancer,* Genome Research **24**, 1421 (2014).

[38] Rao *et al.*, *A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping,* Cell **159**, 1665 (2014).

[39] Dunham *et al.*, *An integrated encyclopedia of DNA elements in the human genome,* Nature **489**, 57 (2012).

[40] Carbonneau *et al.*, *Multiple instance learning: A survey of problem characteristics and applications,* Pattern Recognition **77**, 329 (2018).

[41] Wang *et al.*, *BRCA1 and BRCA2 expression patterns and prognostic significance in digestive system cancers,* Human Pathology **71**, 135 (2018).

[42] He *et al.*, *MiR-218 regulates cisplatin chemosensitivity in breast cancer by targeting BRCA1,* Tumor Biology **36**, 2065 (2015).

[43] Tate *et al.*, *COSMIC: the Catalogue Of Somatic Mutations In Cancer,* Nucleic Acids Research **47**, D941 (2019).

Supplementary Data

Data collection and processing

Patient data and genomic information

All 162 patients from the HMF dataset that were included in this work are listed in Table S1. These patients have 73293 SVs in total (24973 deletions, 14289 duplications, 18255, inversions, 15776 translocations).

From the PCAWG data SV (final_consensus_sv_bedpe_passonly.icgc.public), SNV (final_consensus_passonly.snv_mnv_indel.icgc.public.maf) and CNV calls (all_samples.consensus_CN.by_gene.170214) and RNA-seq data (tophat_star_fpkm_uq.v2_aliquot_gl.tsv) were downloaded from the ICGC data portal (https://dcc.icgc.org/releases/PCAWG/). All data for ovarian samples (Ovarian Cancer - AU) were extracted from these files. The 70 ovarian cancer samples have 18603 SVs in total (5024 deletions, 6751 duplications, 3055 inversions and 3773 translocations).

Supplementary figures

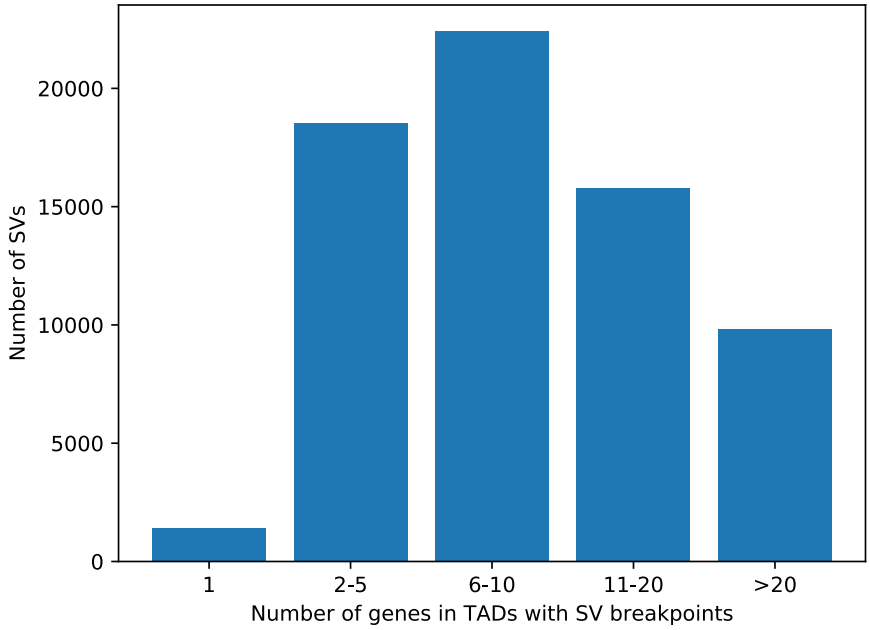


Fig. S1. Number of genes that can potentially be disrupted by SVs overlapping TAD boundaries.

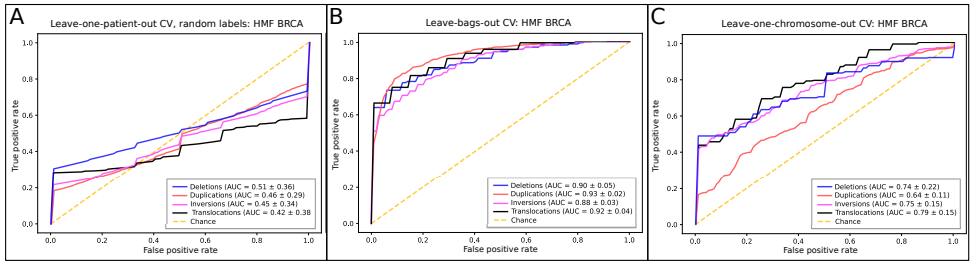


Fig. S2. ROC curves for the models trained on each SV type for the HMF BRCA dataset based on (A) leave-one-patient-out CV with randomized bag labels, (B) leave-bags-out CV and (C) leave-one-chromosome out CV.

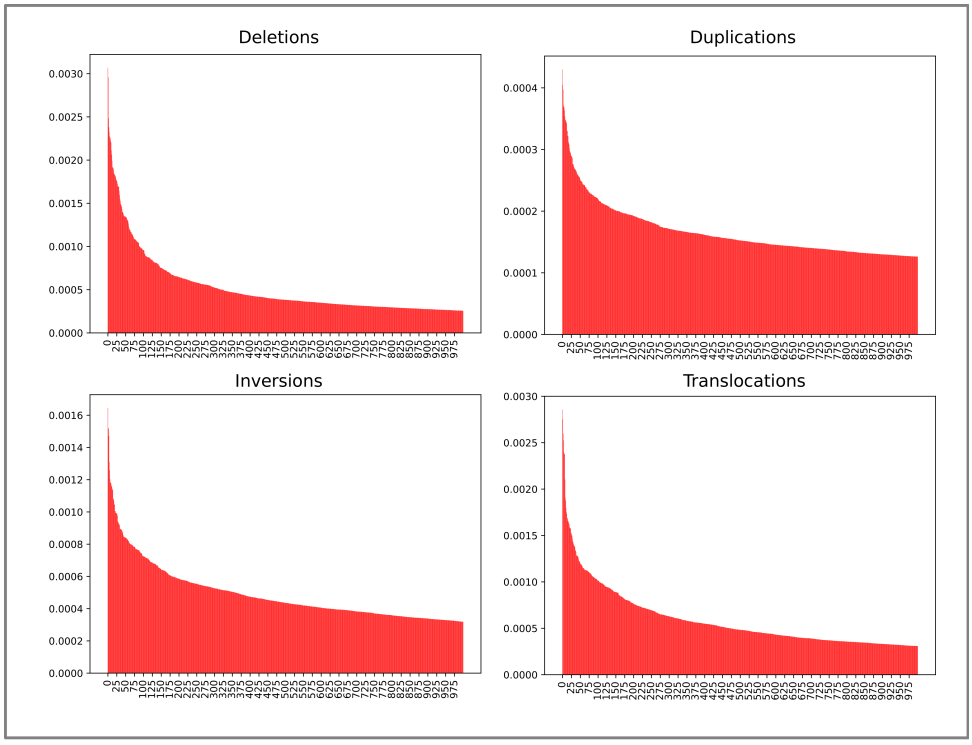


Fig. S3. Random forest feature importances of the classifiers for each SV type.

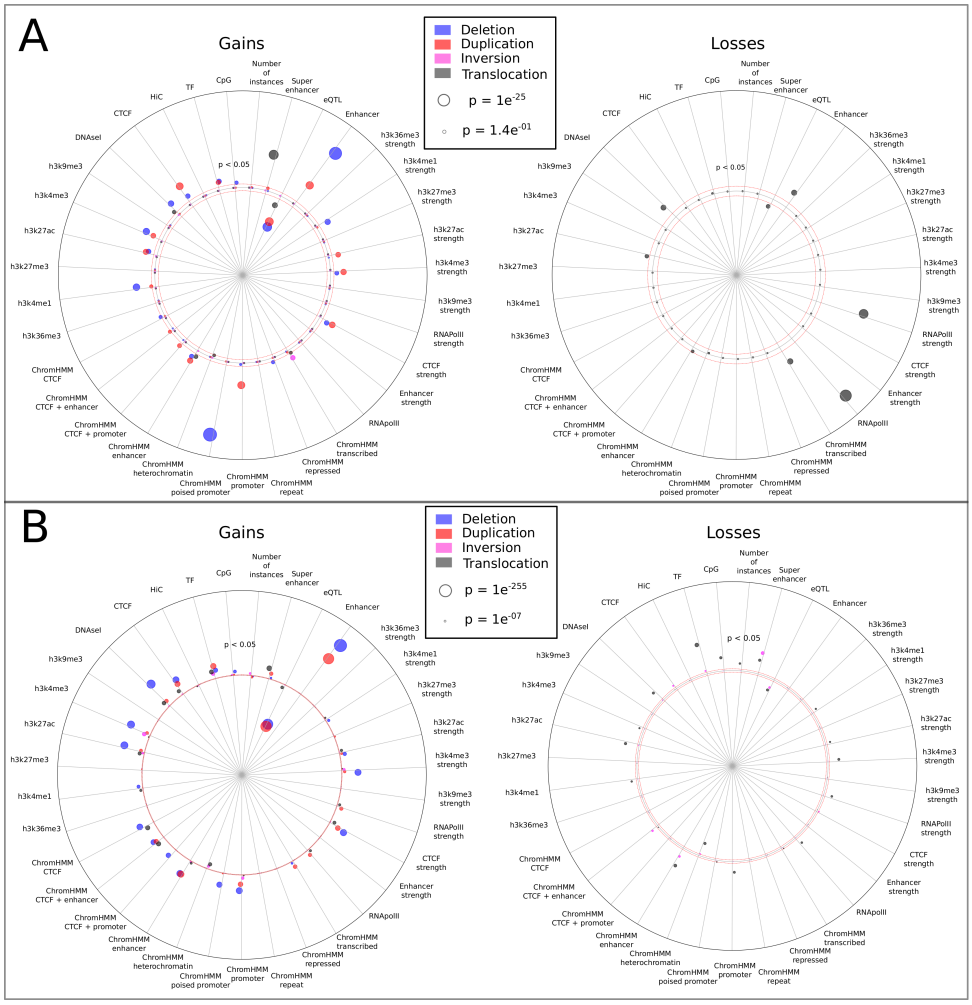


Fig. S4. Gains and losses of regulatory elements among the top 100 instances specific for (A) COSMIC genes and (B) non-COSMIC genes. Instances within the top 100 of inversions only contain gains, and no losses, for COSMIC genes.

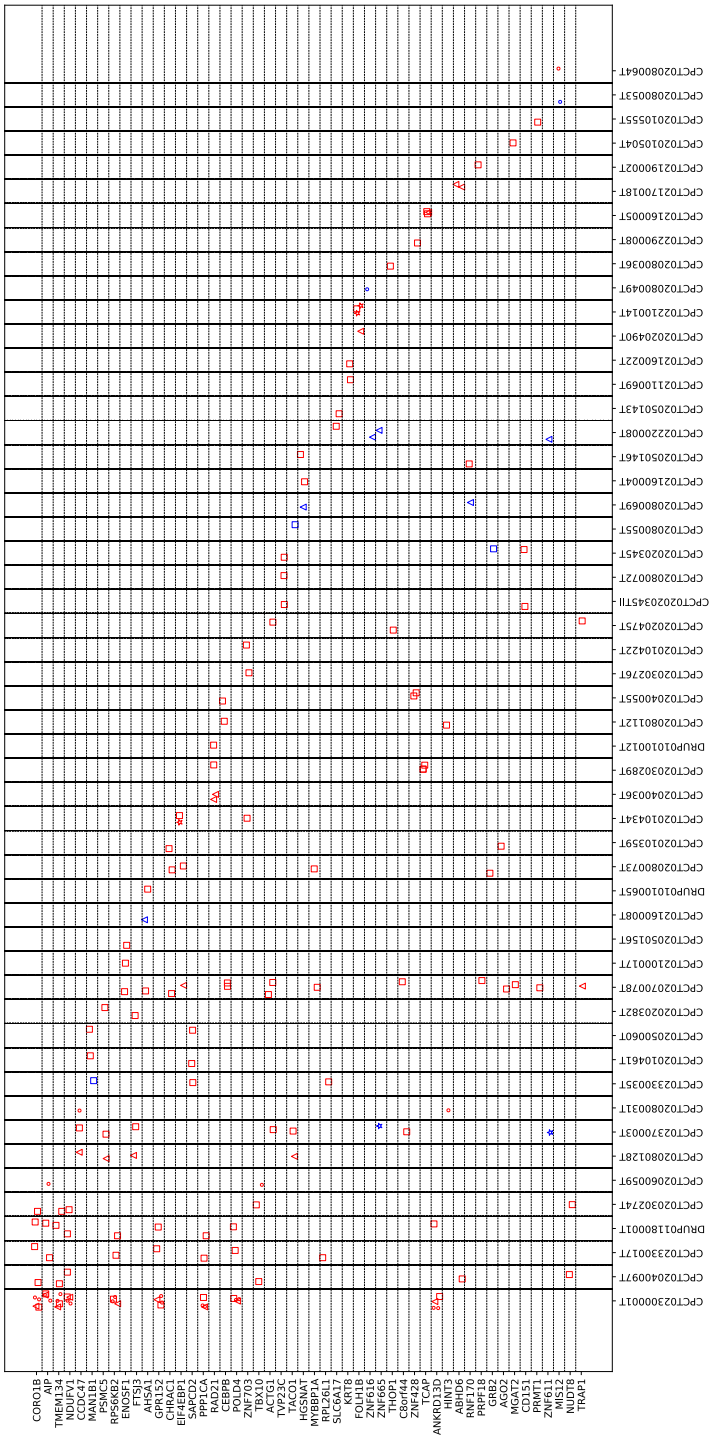


Fig. S5. Top 50 most recurrent genes across patients, selected for patients with positive SV-gene pairs.

Table S1: Specification of all patients in the HMF dataset and if these were included in this work.

Patient	Included?
CPCT02020436T	yes
CPCT02110069T	yes
CPCT02010434T	yes
CPCT02360011T	yes
CPCT02010523T	yes
CPCT02020322T	yes
CPCT02020347T	yes
CPCT02020493TII	no, missing RNA-seq data
CPCT02020345TII	yes
CPCT02160009T	yes
CPCT02100021T	yes
CPCT02050209T	yes
CPCT02210014T	yes
CPCT02180007T	yes
CPCT02150014T	yes
CPCT02370003T	yes
CPCT02300011T	yes
CPCT02010504T	yes
CPCT02220008T	yes
CPCT02080031T	yes
CPCT02010419T	yes
CPCT02170014T	yes
CPCT02030289TIII	no, missing RNA-seq data
CPCT02080125T	yes
CPCT02080069T	yes
CPCT02040035T	yes
CPCT02020508T	yes
CPCT02100067TII	no, missing RNA-seq data
CPCT02080022T	yes
CPCT02040031T	yes
CPCT02330067T	yes
CPCT02100048T	yes
CPCT02190002T	yes
CPCT02080076T	yes
CPCT02040097T	yes
CPCT02100027T	yes
CPCT02040071T	no, missing RNA-seq
CPCT02080053T	yes
CPCT02390001T	no, failed RNA-seq quality control
CPCT02170018T	yes
CPCT02160012T	yes
CPCT02020369T	yes
CPCT02020666T	no, failed RNA-seq quality control
CPCT02160004T	no, missing RNA-seq
CPCT02080127T	yes
CPCT02100023T	yes
CPCT02080019T	yes
CPCT02330017T	yes
CPCT02330035T	yes
CPCT02160013T	yes
CPCT02050160T	yes
CPCT02210010T	yes
CPCT02010520T	yes
CPCT02080036T	yes
CPCT02070078T	yes
CPCT02020514T	yes
CPCT02020344T	yes
DRUP01180001T	yes
CPCT02060075T	yes
CPCT02080055T	yes
CPCT02080029T	yes

Patient	Included?
CPCT02080112T	yes
CPCT02100093T	yes
CPCT02300005T	yes
CPCT02160022T	yes
CPCT02080029TII	no, missing RNA-seq data
CPCT02080073T	yes
CPCT02210004T	yes
CPCT02080025T	yes
DRUP01010037T	yes
CPCT02080063T	yes
CPCT02080016T	yes
CPCT02020385T	yes
CPCT02030264T	yes
CPCT02040055T	yes
CPCT02050143T	yes
CPCT02030276T	yes
CPCT02010419TII	no, missing RNA-seq data
CPCT02100017T	yes
CPCT02010555T	yes
CPCT02080128T	yes
CPCT02110004T	no, failed RNA-seq quality control
CPCT02100037T	yes
CPCT02110023T	yes
CPCT02100067T	yes
CPCT02030271T	yes
CPCT02080072T	yes
DRUP01010065T	yes
CPCT02160005T	yes
CPCT02060070T	yes
CPCT02330032T	yes
CPCT02160001T	yes
CPCT02080064T	yes
CPCT02300008T	yes
CPCT02010433T	yes
CPCT02080067T	yes
CPCT02100049T	yes
CPCT02020407T	yes
CPCT02020478T	yes
CPCT02110020T	yes
CPCT02010382TII	yes
CPCT02100024T	yes
CPCT02010447T	yes
CPCT02100075T	yes
CPCT02030289T	yes
CPCT02020341T	yes
CPCT02100043T	yes
CPCT02100035T	yes
CPCT02020345T	yes
CPCT02230002T	yes
CPCT02190009T	yes
CPCT02040078T	yes
DRUP01110005T	yes
CPCT02160018T	yes
CPCT02050156T	yes
CPCT02160014T	no, failed RNA-seq quality control
CPCT02010461T	yes
CPCT02010508T	yes
CPCT02020382T	yes

Patient	Included?
CPCT02060011T	yes
CPCT02100020T	yes
CPCT02100105T	yes
CPCT02040070T	yes
CPCT02030265T	yes
CPCT02030274T	yes
CPCT02010468T	yes
CPCT02020493T	yes
CPCT02180028T	yes
CPCT02080039T	yes
CPCT02080027T	yes
CPCT02050053T	yes
CPCT02050060T	yes
DRUP01010012T	yes
CPCT02240003T	yes
CPCT02050096T	yes
CPCT02010359T	yes
CPCT02290008T	yes
CPCT02300001T	yes
CPCT02100011T	yes
CPCT02380020T	yes
CPCT02040069T	yes
CPCT02100029T	yes
CPCT02080070TII	no, missing RNA-seq data
CPCT02030305T	yes
CPCT02390001TII	no, missing RNA-seq data
CPCT02020490T	yes
CPCT02100066T	yes
CPCT02020475T	yes
CPCT02300014T	yes
CPCT02050127T	yes
CPCT02080047T	yes
CPCT02060059T	yes
CPCT02160008T	yes
CPCT02050146T	yes
CPCT02040036T	yes
CPCT02010359TII	no, missing RNA-seq data
CPCT02050074T	yes
CPCT02020371T	yes
CPCT02010422T	yes
CPCT02010401T	yes
CPCT02240001T	yes
CPCT02050337T	yes
CPCT02050157T	yes
CPCT02080122T	yes
CPCT02080106T	yes
CPCT02010351T	yes
CPCT02080049T	yes
CPCT02300009T	yes
CPCT02100050T	no, failed RNA-seq quality control
CPCT02050071T	yes
CPCT02080070T	yes
DRUP01020002T	yes
CPCT02100132T	no, failed RNA-seq quality control
CPCT02330027T	yes
CPCT02010528T	no, failed RNA-seq quality control
CPCT02050082T	yes
CPCT02030265TII	no, missing RNA-seq data
CPCT02080060T	yes
CPCT02050138T	yes
CPCT02160015T	no, failed RNA-seq quality control
CPCT02040071TII	no, failed RNA-seq quality control

Table S2: Regulatory elements and sources used for the HMF BRCA dataset and for the germline SVs.

Dataset	Source	Cell type	Processing steps
eQTLs	GTEx v7	Breast	Only accepted p-values < 5 * 10e-8
Enhancers	JEME, elastic net	HMEC	
Promoters	Eukaryotic Promoter Database	All cell types	Selected all promoters containing either a TATA box, initiator motif, CCAAT box or GC box
CpG islands	UCSC genome annotation database	All cell types	
Transcription factors	ORegAnno, version 19-01-2016	All cell types	
H3K27me3	ENCODE - ENCFF291WFP	HMEC	
H3K36me3	ENCODE - ENCFF906MJM	HMEC	
H3K9me3	ENCODE - ENCFF065FJK	HMEC	
H3K4me1	ENCODE - ENCFF336DDM	HMEC	
H3K27ac	ENCODE - ENCFF154XFN	HMEC	
H3K4me3	ENCODE - ENCFF065TIH	HMEC	
DNAse I hypersensitivity sites	ENCODE - ENCFF301VRH	HMEC	
RNA pol II sites	ENCODE - ENCFF433ZKP	HMEC	
CTCF sites	ENCODE - ENCFF288RFS	HMEC	
ChromHMM states	GSE57498	HMEC	
Hi-C interactions	GSE63525 - intrachromosomal contact matrices	HMEC	Interactions observed less than 6 times were removed. Each side of the interaction is considered a separate regulatory element of 5kb.
Super enhancers	dbSUPER	HMEC	
TADs	3D Genome Browser	HMEC	

Table S3: regulatory elements and sources used for the PCAWG ovarian dataset. For regulatory elements not specified in this table, we used the same source as listed in Table S2.

Dataset	Source	Cell type	Processing steps
eQTLs	GTEx v7	Ovarian	Only accepted p-values < 5 * 10e-8
Enhancers	JEME, elastic net	Ovarian	
H3K27me3	ENCODE - ENCFF712UCB	Ovarian	
H3K36me3	ENCODE - ENCFF302DXB	Ovarian	
H3K9me3	ENCODE - ENCFF717WXC	Ovarian	
H3K4me1	ENCODE - ENCFF917PWI	Ovarian	
H3K27ac	ENCODE - ENCFF657AUA	Ovarian	
H3K4me3	ENCODE - ENCFF320JHG	Ovarian	
DNAse I hypersensitivity sites	ENCODE - ENCFF883WWT	Ovarian	
RNA pol II sites	ENCODE - ENCFF570SMG	Ovarian	
CTCF sites	ENCODE - ENCFF522DLJ	Ovarian	
Super enhancers	dbSUPER	Ovarian	
TADs	3D Genome Browser	Ovarian	Liftover from hg38 using UCSC liftover tool

4

Predicting pathogenic non-coding SVs disrupting the 3D genome in 1,646 whole cancer genomes using multiple instance learning

Marleen M. Nieboer, Luan Nguyen, Jeroen de Ridder

Accepted for publication in Nature Scientific Reports

Abstract

Over the past years, large consortia have been established to fuel the sequencing of whole genomes of many cancer patients. Despite the increased abundance in tools to study the impact of SNVs, non-coding SVs have been largely ignored in these data. Here, we introduce svMIL2, an improved version of our Multiple Instance Learning-based method to study the effect of somatic non-coding SVs disrupting boundaries of TADs and CTCF loops in 1646 cancer genomes. We demonstrate that svMIL2 predicts pathogenic non-coding SVs with an average AUC of 0.86 across 12 cancer types, and identifies non-coding SVs affecting well-known driver genes. The disruption of active (super) enhancers in open chromatin regions appears to be a common mechanism by which non-coding SVs exert their pathogenicity. Finally, our results reveal that the contribution of pathogenic non-coding SVs as opposed to driver SNVs may highly vary between cancers, with notably high numbers of genes being disrupted by pathogenic non-coding SVs in ovarian and pancreatic cancer. Taken together, our machine learning method offers a potent way to prioritize putatively pathogenic non-coding SVs and leverage non-coding SVs to identify driver genes. Moreover, our analysis of 1646 cancer genomes demonstrates the importance of including non-coding SVs in cancer diagnostics.

Introduction

On average, cancer develops through the accumulation of 4-5 driver mutations[1]. The implications of characterizing these mutations per cancer genome for developing novel anti-cancer therapies are undoubtedly large. Over the recent years, efforts such as the Cancer Gene Census (CGC) have been set up to catalogue all known genes that have been implicated by cancer-driving mutations[2]. Furthermore, a myriad of computational algorithms have been designed to predict the pathogenicity of mutations[3–10]. However, until now the majority of these studies have focused on mutations occurring in the coding part of the genome, while it is becoming increasingly clear that non-coding mutations may also drive cancer initiation and progression[11].

Elucidating the pathogenic effect of non-coding single-nucleotide variants (SNVs) is under very active study[12–16], and despite the fact that this is a challenging computational task, prediction results have been gradually improving. Relatively straightforward approaches are based on burden testing[17, 18], wherein elevated mutation densities point to mutations that are under positive selective pressure. However, these statistics-based approaches are not suitable for mutations with low recurrence across cancer patients, which is typically true for non-coding structural variants (SVs), as was recently demonstrated in a Pan-Cancer Analysis of Whole Genomes (PCAWG) study[19]. More recent work therefore focuses on using machine learning to identify patterns in genomic features overlapping and surrounding the SNVs, such as enhancers, histone modifications or transcription factor binding information[12, 13]. Despite this progress, almost no methods exist that allow identification of likely pathogenic non-coding SVs. This is counterintuitive, as the impact of somatic SVs (e.g. insertions, deletions, duplications, inversions and translocations) in terms of the number of affected bases far surpasses that of somatic SNVs. For this reason, elucidating the role of non-coding SVs is important for understanding cancer development and may prove to be indispensable for whole

genome sequencing (WGS)-based patient reporting.

Although in many cases the exact mechanism through which non-coding SVs cause cancer remains unclear, recent studies have shown that non-coding SVs may exert a pathogenic effect by disrupting the boundaries of Topologically Associated Domains (TADs). TADs are structures in the 3D genome in which DNA interacts more frequently with each other than with DNA outside of the TAD[20]. TADs are separated by boundaries across which interactions are much scarcer. Together, these structures maintain interactions between genes and regulatory elements such as enhancers. TADs are believed to be the result of a process called loop extrusion, in which DNA is pulled through a ring of cohesin until it is blocked by CCCTC-binding factor (CTCF)[21]. This theory is supported by the observation that convergent CTCF motifs were found to be enriched at the boundaries of TADs[22]. Non-coding SVs were found to be capable of causing congenital abnormalities[23–27] and cancer[28–32] by disrupting TAD boundaries and thereby enabling novel interactions to form between genes and regulatory elements. However, methods that exploit this principle for somatic SV prioritization or classification have only recently been introduced and remain scarce[33, 34].

While there are sufficient indications that disrupting TAD boundaries can be pathogenic, less is known about the role of disrupting CTCF-mediated chromatin loops that are formed inside of TADs. Previous work suggests that somatic SNVs can affect the binding sites of CTCF and thereby have cancer-driving potential[35]. On the other hand, it was found that not all CTCF loops disrupted by germline non-coding SVs equally contributed to the development of congenital phenotypes[36]. It therefore remains an open question whether somatic non-coding SVs exist that exert a pathogenic effect through CTCF loop disruption, but if they do it may be important to supplement non-coding SV prioritization information with CTCF loop data.

State-of-the art non-coding SNV prioritization algorithms are not straightforwardly applied to SVs. It is, for instance, much more difficult to define a suitable representation of the large number of interactions that may be altered by SVs. Moreover, no 'ground truth' labels on the pathogenicity of non-coding SVs are available that can be used for training. To this end, we previously proposed a Multiple Instance Learning (MIL)-based approach, called svMIL[34]. A common analogy to explain MIL is the problem of a number of keychains and a door that is opened by one specific key[37]. Without knowing beforehand which key opens the door, the goal is to distinguish the keychains containing at least one key that opens the door (positive keychains or 'bags') from keychains that do not open the door (negative keychains or 'bags'). As a keychain may contain a variable number of keys ('instances'), representing all keys in a single feature matrix is not trivial. Instead, in MIL, each key is individually described with features such as the length or shape of the key. The challenge for MIL-based classifiers is to separate positive bags (keychains) from negative bags (keychains) within the MIL feature space, which can for example be achieved by mapping the bags to a new feature space in which a regular classifier can be trained[38].

In svMIL, we formulated the prediction of pathogenic non-coding SVs as a MIL problem, wherein SV-gene pairs are considered as bags and the regulatory elements as instances (Fig 1a). Labels are obtained by leveraging patient matched gene expression data. Together, this representation enables identification of putatively pathogenic TAD

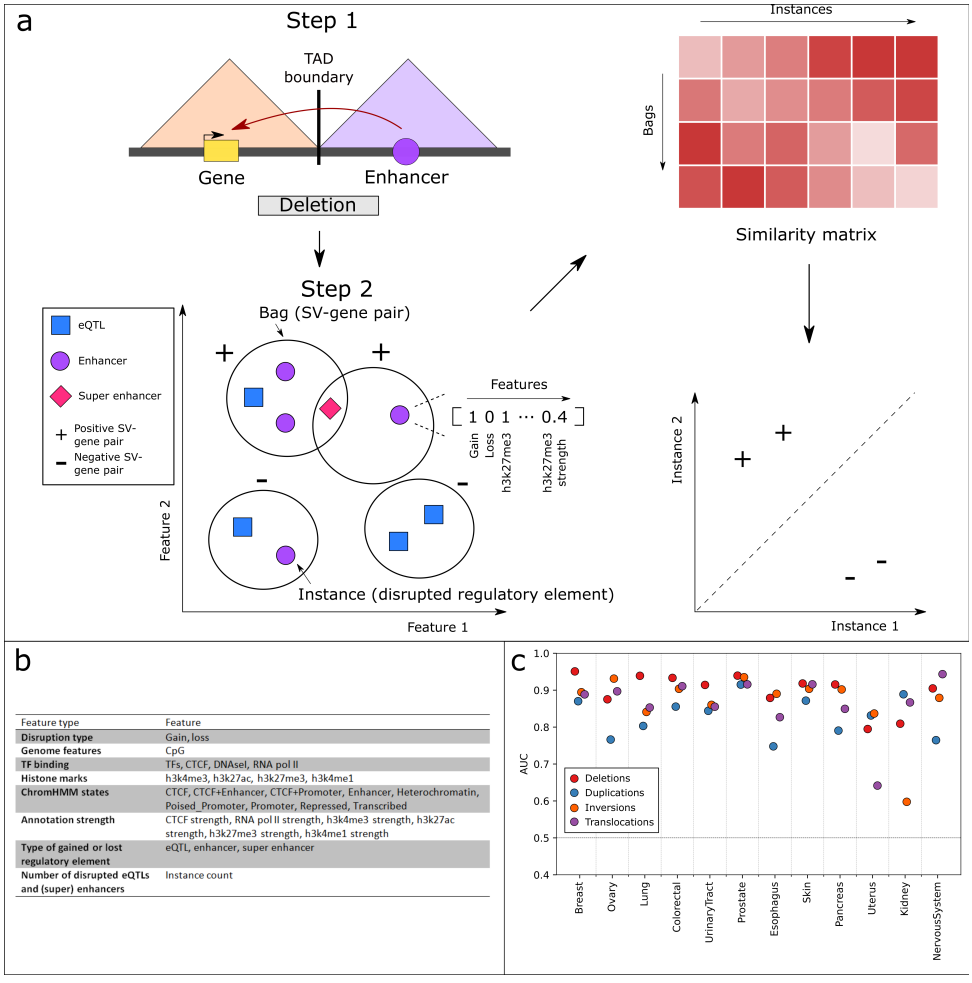
boundary-disrupting non-coding SVs by learning the characteristics of disrupted interactions between genes and regulatory elements. Here, we extend upon this framework and improve the svMIL algorithm, which was originally tested on a maximum of 162 breast cancer patients and 70 ovarian cancer patients, to scale to larger datasets. We additionally use feature selection to improve the AUC by around 0.1 to an average of 0.86 in 313 breast cancer patients. We apply the improved svMIL algorithm, svMIL2, to characterize pathogenic non-coding SVs across 12 cancer types. For this purpose, we leverage a high-quality pan-cancer dataset from the Hartwig Medical Foundation (HMF)[39], which consists of 1646 uniformly processed high-depth (>90x) metastatic tumor samples along with paired transcriptional profiling data. The availability of same-sample whole-genome sequencing (WGS) and RNA-sequencing data across many cancer types has already resulted in a number of novel studies[40–43], and likewise makes this dataset extremely suitable for this study.

In this work, we show that svMIL2 can confidently predict pathogenic TAD boundary-disrupting non-coding SV candidates across all cancer types, revealing that especially ovarian and pancreatic cancer appear to be more strongly driven by non-coding SVs than other cancers. Furthermore, non-coding SVs frequently disrupt active (super) enhancers in open chromatin regions uniformly across cancer types, which supports our previous findings in breast cancer[34]. Altogether, these findings indicate a common mechanism by which non-coding SVs may cause cancer.

Additionally, we explore the impact of non-coding SVs disrupting intra-TAD CTCF loops rather than TAD boundaries. Although we find that gene expression can be altered through mechanisms similar to TAD boundary disruptions in breast cancer, the frequency at which these events occur is low, confirming previous findings[36]. However, these initial results suggest that investigating the disruption of intra-TAD chromatin loops may be highly relevant in future studies to obtain a complete overview of cancer development and progression.

Multiple Instance Learning effectively predicts pathogenic non-coding SVs

svMIL predicts pathogenic TAD boundary-disrupting non-coding SVs in 2 steps: first predicting candidate pairs of somatic non-coding SVs and disrupted genes, and then applying machine learning to identify the pairs that are pathogenic (Fig 1a, see Methods for more details). In step 1, for every SV overlapping a TAD boundary, derivative TADs are constructed in which the disrupted interactions between genes and regulatory elements are modeled (Fig S1). Genes that gain or lose at least 1 regulatory element and the disrupting SV are considered a pair. In step 2, we learn pathogenic SV-gene pairs using a MIL model. Each SV-gene pair is defined as a bag containing the gained or lost eQTLs, enhancers and super enhancers as instances. Every instance is assigned a feature vector (Fig 1b) describing if the instance was gained or lost, which histone marks (h3k4me3, h3k27me3, h3k27ac, h3k4me1), chromatin states (CTCF, CTCF + enhancer, CTCF + promoter, promoter, poised promoter, heterochromatin, repressed, transcribed), transcription factor binding profiles (DNAse I hypersensitivity sites, RNA polymerase II, CTCF, transcription factor binding sites) and CpG islands it overlaps with, the peak intensity (used to indicate strength of the element) of these regulatory elements where available (histone marks, RNA polymerase II), the type of the regulatory element (eQTL, enhancer



Feature type	Feature
Disruption type	Gain, loss
Genome features	CpG
TF binding	TFs, CTCF, DNAseI, RNA pol II
Histone marks	h3k4me3, h3k27ac, h3k27me3, h3k4me1
ChromHMM states	CTCF, CTCF+Enhancer, CTCF+Promoter, Enhancer, Heterochromatin, Poised_Promoter, Promoter, Repressed, Transcribed
Annotation strength	CTCF strength, RNA pol II strength, h3k4me3 strength, h3k27ac strength, h3k27me3 strength, h3k4me1 strength
Type of gained or lost regulatory element	eQTL, enhancer, super enhancer
Number of disrupted eQTLs and (super) enhancers	Instance count

Fig. 1. Overview of the svMIL2 method and performance. (a) svMIL2 methodology. From disrupted TADs, pairs are identified between SVs and genes disrupted due to gained or lost regulatory elements. These SV-gene pairs are modeled as bags (keychain), in which the regulatory elements (eQTLs, enhancers or super enhancers) that the gene gained or lost due to the SV are instances (keys). Instances are described with features such as histone marks (see panel b). A similarity score is constructed between bags and instances by computing the absolute distance from the mean instance of each bag to all other instances. The resulting similarity matrix is used as input to a random forest model to classify bags. (b) All features used in the svMIL2 model to describe instances, grouped by feature category. (c) Performance in AUC of the svMIL2 model on 12 cancer types from the HMF dataset.

or super enhancer) and the number of regulatory elements disrupted by the SV (instance count) (see Table S1 for data sources).

To obtain a final classifier, we used the MILES approach with a random forest classifier[38]. In MILES, a feature space is created by computing a bag-to-instance similarity matrix by computing a distance between each bag to all instances, on which a regular classifier can then be trained. Positive bags are expected to have higher similarity to positive instances, but dissimilar to negative instances, resulting in a separation in feature space (Fig 1a). Here, an absolute distance is computed from the mean instance of each bag to all instances.

Bags are labeled positive if the z-score of the expression of the gene in an SV-gene pair to all other patients with no mutation affecting the gene (coding SNV, CNV, SV or non-coding SV) is larger than 1.5 or smaller than -1.5 (i.e. the SV led to altered expression of the paired gene), and negative otherwise.

Model performance is measured using leave-one-patient-out CV, mimicking a scenario in which an unseen patient comes into the clinic. In this CV setting, all SV-gene pairs of one patient are used as testing data, whereas the SV-gene pairs of all other patients are used as training data.

To improve svMIL, we include a rigorous feature selection approach to determine which features optimally benefit the classification result. To this end, we first explored the feature importance in the original model on the breast cancer samples, as this was the cancer type used to infer this model originally. We find that certain features have low variance across instances and do therefore not contribute to classification performance (Fig S2). By removing non-informative features and reducing noise in our instances (see Methods), we further enhance the ability of our previously described svMIL approach to predict pathogenic TAD boundary-disrupting non-coding SVs. Comparing the performance in a leave-one-patient-out CV setting of the original model to the updated model reveals that these improvements yield an increase in AUC of around 0.1 for all SV types except for duplications, which increases by 0.03 (Fig S3). Thus, the methodology of svMIL2 is highly effective at predicting pathogenic non-coding SV-gene pairs.

svMIL2 can accurately predict driver genes disrupted by non-coding SVs across cancer types

We applied svMIL2 to predict pathogenic non-coding SV-gene pairs in all 12 cancer types from HMF in a leave-one-patient-out CV setting and show that the AUC is consistently high, revealing that our method is also applicable to non-breast cancer data (Fig 1c), even in data with lower sample and SV counts (Fig S4a-b, Table S2). Out of 204 overlapping (100 bp) SVs within different patients, svMIL2 predicts 172 with the same label, showing that our method is robust.

Notably, lower performance is observed for translocations in uterus cancer and for inversions in kidney cancer, which is likely explained by a low sample count and low number of detected pathogenic SVs in these cancers (see Methods and Fig S4c, Table S2). Overall, differences in performance between SV types may be caused by the varying number of SVs of a certain type detected in each cancer.

To maximize the number of correctly identified pathogenic SV-gene pairs, the operating point of each model was individually optimized for the highest recall, requiring a

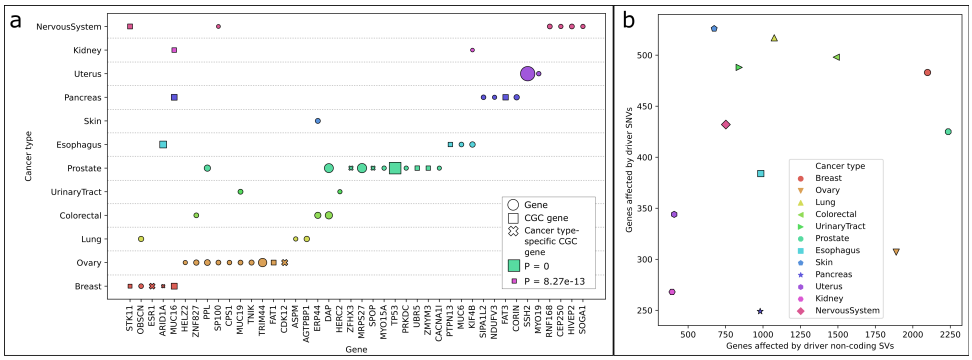


Fig. 2. Analysis of predicted pathogenic non-coding SV pairs. (a) Genes affected by pathogenic non-coding SVs as identified by svMIL2 with significant driver potential (showing top 50 most significant gene-cancer type pairs). To determine significant driver potential, random gene sets were sampled 10,000 times with the same size as the number of genes with candidate pathogenic non-coding SVs. A t-test was used to compute which of the candidate genes have more driver coding SNVs (snpEff moderate or high impact, filtered for consensus genes driven by SNVs from IntOGen) than expected by random chance. (b) Comparison of the number of genes affected by pathogenic non-coding SVs with the number of genes affected by driver SNVs reveals a preference for a different driving mechanism per cancer type.

minimum precision of 0.5. In total, 9261 candidate non-coding SV-affected driver genes were identified, ranging between on average 6-35 genes per patient depending on the cancer type (Fig S5a). 346 of the predicted genes are reported in the COSMIC CGC, of which 25 are also annotated to be specific for the respective cancer type (Fig S5b).

11 of the predicted genes have been previously reported as being affected by non-coding SVs, all of which result in significant changes to gene expression compared to non-mutated genes ($z > 1.5$ or $z < -1.5$, see Methods). Most notably, we identify a deletion (Fig S6a) and translocation (causing eQTL gains) affecting TP53 in prostate cancer, and an inversion (Fig S6b) and translocation causing ERBB2 to gain eQTLs and a (super) enhancer in ovarian cancer. These genes were reported to be driven by non-coding SVs in these cancer types previously[19]. PTEN (inversion causing gain of an enhancer, super enhancer and eQTL in ovarian cancer), BCL2 (deletion causing gain of an eQTL and enhancer, colorectal cancer), VMP1 (inversion causing gain of an enhancer, super enhancer and eQTL in pancreatic cancer) and LSAMP (translocation causing gain of eQTL in nervous system cancer) were also significant in the same study, albeit in different cancer types.

Other interesting findings include MYB, which is affected by an inversion leading to a (super) enhancer-hijacking event in a colorectal cancer patient, a phenomenon that has previously been observed to occur in ACC as a result of translocations[44]. We also identify a deletion causing GFI1 to gain an eQTL, enhancer and super enhancer in colorectal cancer and an inversion causing a gain of an eQTL in prostate cancer. Enhancer-hijacking was previously demonstated to lead to overexpression of GFI1 in medulloblastoma[45].

Activation of the proto-oncogene TAL1 was linked to recurrent deletions of a nearby TAD boundary in T-ALL[28], and we identify potential disruptions of this gene in esopha-

gus cancer (translocation causing gain of eQTL and enhancer) and uterus cancer (translocation causing gain of eQTL). In another study, mutations in the CTCF motif at a TAD boundary nearby NOTCH1 likely resulted in misregulation through novel gene-enhancer interactions[46]. svMIL2 identified an inversion in esophagus cancer causing the gene to gain an eQTL and potentially cause the upregulation of the gene. Finally, recurrently disrupted CTCF sites were observed near FOXC1 in esophagus, gastric and colon adenocarcinomas, and near BCL6 in hepatocellular carcinoma[31]. We identify a deletion causing FOXC1 to gain an eQTL and enhancer in pancreatic cancer, and a duplication resulting in a gain of an eQTL for BCL6 in colorectal cancer.

To validate if these predicted driver genes are significant findings, we determined how frequently they harbor predicted pathogenic SNVs. To this end, we defined the driver potential as the number of driver SNVs affecting the gene across patients within the respective cancer type according to snpEff (moderate or high impact). This list was further filtered for genes driven by SNVs from the IntOGen catalog[47]. Within each cancer type, significance of a gene is assessed by comparing the driver potential to the average driver potential in 10,000 randomly subsampled gene sets of the same size (t-test, Bonferroni corrected). This analysis reveals 112 genes disrupted by non-coding SVs with significant driver potential (Fig 2a, showing the top 50 most significant gene-cancer type combinations. The full list is provided in Table S3). 26 significant genes are also indicated as driver genes by the CGC, of which ESR1, ARID1A, CDK12, ZFHX3 and SPOP are known drivers in breast, ovarian and prostate cancer, respectively. Thus, our model can identify non-coding SVs affecting known driver genes in various cancer types in previously unseen patients.

The number of pathogenic non-coding SVs varies between cancer types

The highest number of pathogenic non-coding SVs is detected in breast, ovarian and prostate cancer, while only low numbers are identified in uterus and kidney cancer (Fig S4c, Table S2). Although the number of pathogenic non-coding SVs increases with the total number of SVs detected within a cancer type, uterus, nervous system and ovarian cancer have more pathogenic non-coding SVs relative to their total SV count (Fig S4d, Table S2). However, there does not appear to be a clear preference for specific SV types in any cancer type (Fig S7). To determine if certain cancer types may be largely driven by non-coding SVs, we plotted the number of genes affected by at least one predicted pathogenic non-coding SV to the genes with driver SNVs from snpEff and IntOGen as detailed above (Fig 2b). Ovarian and pancreatic cancer stand out as having relatively more pathogenic non-coding SVs than driver SNVs. As tumorigenesis is known to be driven by copy number alterations in these cancer types[48–50], these findings indicate that many of these events may exert driving effects through disrupting TAD boundaries.

Pathogenic non-coding SVs disrupt similar regulatory elements across cancer types

To determine if non-coding SVs exert pathogenicity through similar mechanisms across cancer types, we compared if gained and lost regulatory elements significantly differ between predicted pathogenic SVs and predicted non-pathogenic SVs. For each cancer type, the top 100 instances with highest feature importance were compared to 100 randomly selected instances from predicted non-pathogenic SVs (t-test, Bonferroni correc-

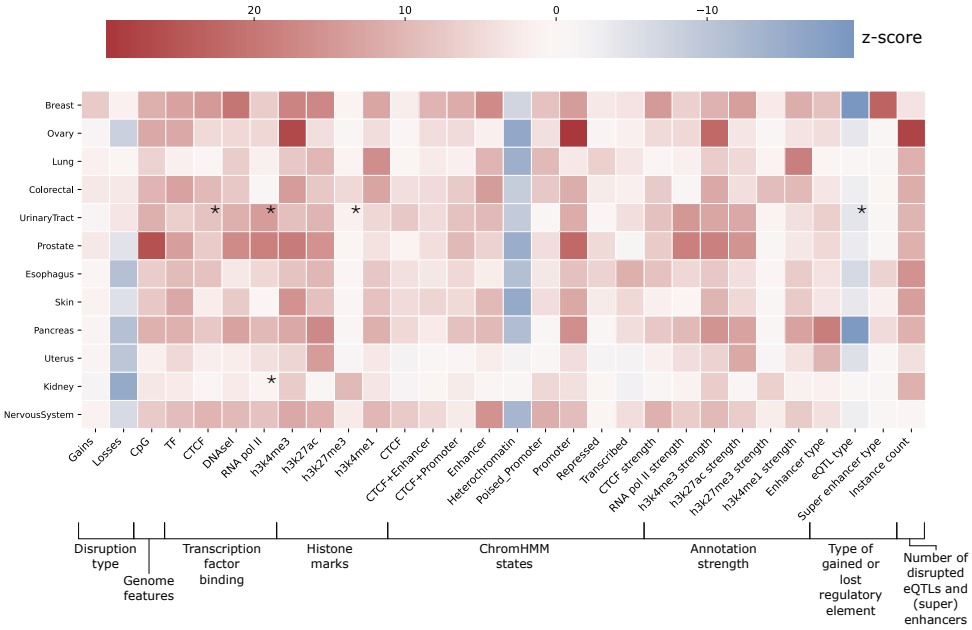


Fig. 3. Heatmap showing the instances observed more (red) or less (blue) frequently than expected by random chance in each cancer type. The colors represent the z-score. The asterisks indicate regulatory elements that were missing in a cancer type and for which GM12878 was used as default.

tion across all cancer types). Overall, we observe that highly similar regulatory elements are disrupted across cancer types (Fig 3). This is also visible if the affected regulatory elements are split into gains and losses (Fig S8). Interestingly, only breast cancer appears to be driven more by gains than losses of regulatory elements, which is not explained only by a higher number of deletions and duplications (Fig S7) and thus may represent a preferential mechanism to upregulate genes in this cancer type. For kidney and uterus, the overall lower significance is likely explained by a lower number of pathogenic SVs (Fig S4c, Table S2). Across cancer types, we notice a frequent disruption of enhancers and the active enhancer (h3k27ac) mark with high active signal strength (h3k27ac strength). For breast cancer, super enhancers are disrupted. Furthermore, lack of heterochromatin, repressed regions and h3k27me3 (marker of heterochromatin) is frequently observed, while more DNAseI hypersensitivity marks (accessible chromatin) are affected. In conclusion, these patterns indicate that pathogenic non-coding SVs appear to mostly alter active (super) enhancers in open chromatin regions, a mechanism which is recurrently observed across cancer types.

Tissue-specific regulatory elements are important for classifier performance

As regulatory data may not always be readily available for every tissue, we aimed to assess the impact of selecting less-than-optimal regulatory information on predictive performance. For every cancer type, we ran svMIL2 while swapping all regulatory data with all other cancer types and measured the effect on performance (see Methods). In ad-

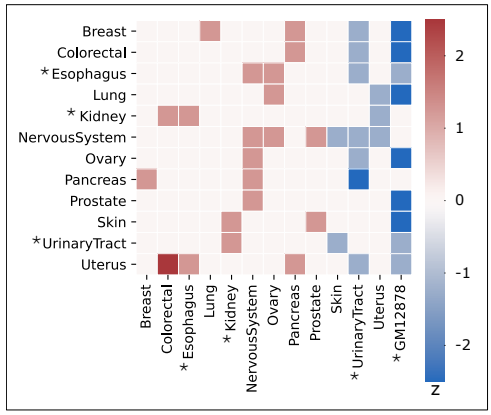


Fig. 4. The effect of swapping regulatory data between cancer types on model performance. The z-score is computed by comparing the total AUC difference in a swap across all SV types to the mean of performance differences from the original run to all other swaps, divided by the standard deviation of these differences. Higher z-scores thus mean that the performance is better with data from that tissue type relative to all other tested tissue types in the swap. For example, out of all swaps made, nervous system relatively performs best with data from nervous system, ovary and prostate, while the performance is worst with data from skin, urinary tract and uterus. The asterisks indicate cancer types with some missing data for which GM12878 was used.

dition, we compared the performance to a scenario where only data from GM12878 is used, which we use as a default when tissue-specific data is missing. Overall, it appears that the majority of swaps do not significantly alter performance, revealing the overlapping nature of regulatory information between tissue types (Fig 4), which has been noted previously[51]. Using regulatory data from GM12878 and urinary tract are typically poor choices that reduce predictive performance ($z < -1$). As urinary tract misses a lot of tissue-specific data and therefore already uses a lot of data from GM12878 in the original run, this reduction may not be surprising. On the contrary, certain swaps appear to improve performance ($z > 1$). These results may not be unexpected given that our samples consist of metastases, which may no longer necessarily completely represent the tissue of origin. However, as not all samples of our dataset within a cancer type metastasized to the same region, recommending an optimal alternative that will also be suitable for independent data is not trivial. Altogether, these findings are of particular importance for the choice of using GM12878 as a default in case of absent tissue-specific data. While the performance using GM12878 only in place of missing data is reasonable (see Fig 1c, where urinary tract, esophagus and kidney used GM12878 to replace missing data), the possibility of obtaining better AUC with the actual tissue type stresses the importance of generating regulatory datasets for each relevant tissue type.

Non-coding SVs alter gene expression by disrupting intra-TAD chromatin loops

Next, we aimed to determine if SVs disrupting intra-TAD chromatin loops may play a role in cancer. To this end, we ran svMIL2 using chromatin loops predicted by iTAD in place of TAD boundaries. As this software requires cohesin and CTCF peaks as input and these tracks are only available for breast, colorectal and lung cancer, our analysis

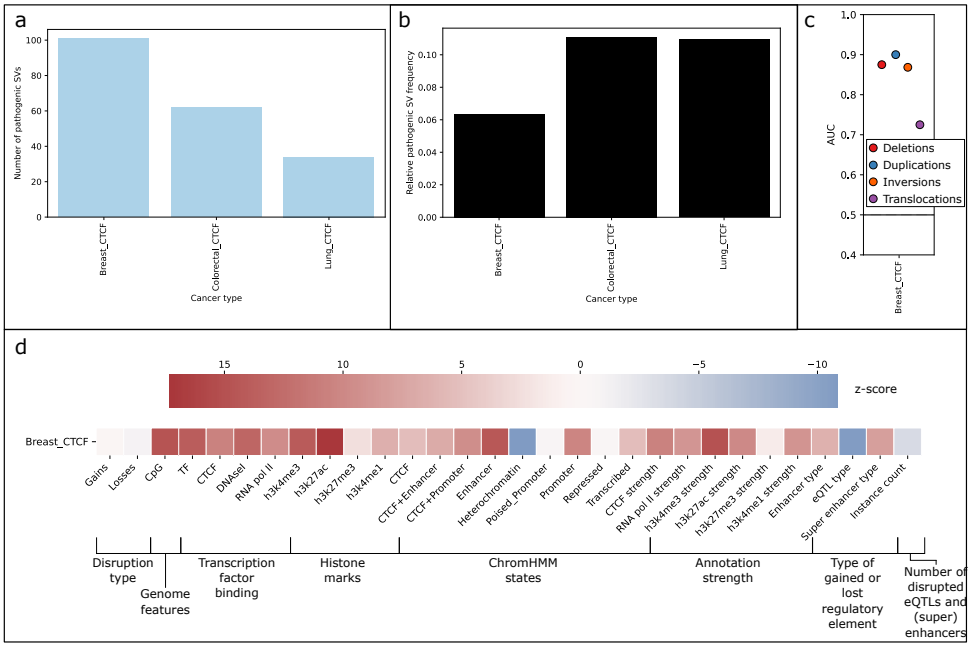


Fig. 5. Performance of svMIL2 when intra-chromatin CTCF loops are used in place of TADs. (a) Number of pathogenic SV-gene pairs identified to disrupt chromatin loops and (b) the percentage of the total SV-gene pairs these comprise. (c) AUC of svMIL2 when predicting pathogenic SVs using chromatin loops. (d) Regulatory elements more or less frequently disrupted than by random chance for SVs affecting chromatin loops. The colors represent the z-score.

is limited to those tissue types. By far, most chromatin loops were predicted in breast (breast: 22113, colorectal: 7522, lung: 9130). In contrast to the TAD-based scenario, the number of SV-gene pairs is far lower (54%, 74% and 83% less in breast, colorectal and lung cancer, respectively), with remarkably fewer pathogenic SV-gene pairs (breast: 101, colorectal: 62, lung: 34) (Fig 5a-b, Fig S4c, Table S2). Taken together, these findings reveal that pathogenic non-coding SVs are less likely to start and end within CTCF loops, but may still alter gene expression.

Due to low counts of candidate SV-gene pairs in colorectal and lung cancer, we could only reliably apply svMIL2 and obtain an AUC in breast cancer, where pathogenic SVs were predicted at high AUC for all SV types (Fig 5c). As the disruption of chromatin loops appears to also frequently result in gains of (super) enhancers in open chromatin regions (Fig 5d, Fig S9), the mechanism by which gene expression is altered is likely similar to that of TAD boundary disruptions.

Out of 94 predicted driver genes affected by SVs through CTCF loop disruption in breast cancer, 2 are reported as cancer-driving by the CGC. ZNF331 is affected by an inversion, while CHEK2, a well-known germline risk factor for breast cancer[52], is affected by translocations in 4 different patients.

In conclusion, we find evidence that non-coding SVs may be capable of altering gene expression in cancer by disrupting intra-TAD chromatin loops, but at a far lower frequency than by the disruption of TAD boundaries, confirming previous findings[36]. However, as our results are limited by the lack of available cohesin measurements across tissues and low sample counts, the importance of intra-TAD loops remains an important topic for future studies.

Discussion

In this work, we described an improved version of svMIL, svMIL2, to predict pathogenic TAD-boundary and CTCF-loop disrupting non-coding SVs from WGS cancer genomes with paired whole transcriptome sequencing data. We showed that svMIL2 can leverage these data to accurately predict pathogenic non-coding SVs across multiple cancer types. Across all cancer types, putative pathogenic non-coding SVs were predicted to disrupt 9261 genes, 346 of which are known cancer driver genes. Since all validation experiments are carried out through leave-one-patient-out CV, together with identifying non-coding SVs affecting known cancer drivers, these results demonstrate that our method is applicable to identify pathogenic non-coding SVs in a clinical setting where somatic variants of a newly diagnosed patients need to be prioritized. We also observe that the role of pathogenic non-coding SVs, as opposed to driver SNVs, varies between cancers. Despite these differences, non-coding SVs appear to similarly frequently disrupt active (super) enhancers in open chromatin regions in the majority of cancer types, pointing to common mechanisms by which TAD disruptions may be pathogenic. Taken together, these findings indicate that non-coding SVs play an important role in cancer and should be considered in WGS-based cancer diagnostics.

As opposed to the clear impact of disrupting TAD boundaries on the development of cancer, the effects of disrupting intra-TAD chromatin loops are not yet well understood. Using svMIL2, we were able to identify pathogenic non-coding SVs that alter expression of known cancer genes by disrupting CTCF loops in breast cancer. However, the num-

ber of candidate pathogenic SV-gene pairs resulting from CTCF loop disruptions is up to 10-fold lower than when only TADs are investigated. Therefore, SVs disrupting intra-TAD chromatin loops rather than TAD boundaries may seemingly be less pathogenic, which corresponds with previous experiments performed with germline SVs[36]. However, as we were only able to obtain cohesin and CTCF peak data for breast, colorectal and lung cancer, the actual relevance of chromatin loops may be underreported in this study. Nevertheless, these initial findings point to a potential involvement of disrupting CTCF loops in the development of cancer, and may be a highly interesting avenue for future studies.

While the majority of regulatory information is available in respective tissue types, we found that selecting the most suitable alternative for cases with missing data remains a difficult problem that potentially strongly affects classifier performance. As our dataset is comprised of metastatic cancer data, the reference tissue type may sometimes no longer be well-represented in the cancer at time of sampling, and thus selecting an optimal alternative tissue is not trivial. However, answering these questions will only really become possible once the missing regulatory data have been acquired in the respective tissues. Therefore, our results underscore the importance of completing the catalogue of celltype-specific regulatory information. Such data may also help create a better understanding of the role of SVs in the mitochondrial DNA (mtDNA). Common deletions have been identified in the mtDNA of especially gastric cancers[53], but the effect of such SVs on regulatory information is difficult to assess as mtDNA is often missing from regulatory datasets. While large-scale efforts to collect these data such as the ENCODE project[54] are still ongoing, other promising alternatives to acquire these data apply imputation from other cell types, which is performed by methods such as Avocado[51], ChromImpute[55] and PREDICTD[56]. However, as imputation with these methods is not yet possible for regulatory data in all tissue types, further research in this field is required.

Furthermore, our method could further benefit from improved SV calls. While our current dataset captures many SVs in the genome, adopting long-read sequencing techniques could improve detection of additional SVs in repetitive regions[57] and clear up potentially noisy calls. SVs obtained from longer reads can improve the training labels used in svMIL2, as expression of certain genes may be altered due to non-coding SVs but currently remain undetected due to missing calls. Label quality would also benefit from additional patient-matched datasets such as methylation data, which could be used to exclude genes that are deregulated due to methylation rather than non-coding SVs. However, such data is currently too costly to routinely generate for each patient. Similar labeling problems occur when genes are affected by variants of unknown significance or upstream pathway effects, which are difficult to account for. While methods such as DriverNet[58] or DawnRank[59] have been shown to improve driver prediction by integrating gene networks with SNV and CNV data, non-coding SVs have not yet been included in these studies. However, as the number of recurrent driver non-coding SVs is smaller than for SNVs or CNVs, as was shown previously[19, 34], the statistical validation applied will need to properly deal with the imbalance in contribution to the driver phenotype between the mutation types.

Although we demonstated that MIL is a suitable approach to identify pathogenic

non-coding SVs and previously showed the benefits of using MIL compared to a non-MIL random forest[34], alternative machine learning approaches may assist in learning about pathogenic non-coding SVs from a different perspective. For example, deep learning-based methods such as DeepSEA[12] and ExPecto[13] were recently used to prioritize non-coding SNVs by learning genomic features, such as chromatin states, of the region around the mutation. Such an approach could similarly be used to learn the characteristics of SV breakpoints, or disrupted TAD boundaries. These annotations on a smaller scale could teach us more about the local environment disrupted by non-coding SVs in detail, which is not straightforward with svMIL2.

WGS is rapidly becoming part of the routine diagnostic process of cancer centers. However, since the driving potential of non-coding SVs remains elusive, the vast majority of these costly WGS data remain underutilized. Our proposed svMIL2 model can accurately predict pathogenic non-coding SVs among the typically vast numbers of somatic SVs present in cancer genomes by learning from a combination of WGS, gene expression, TAD boundary and intra-TAD chromatin loop information. As more and more WGS datasets and epigenomics tracks will become available, it can be expected that these predictions will further improve. This will further enable the inclusion of non-coding SVs in WGS-based cancer diagnostic reporting.

Methods

Data

Pre-called whole-genome SV, CNV and SNV data and RNA-seq counts were obtained for 1944 cancer patients from the HMF, representing 29 cancer types in total. All variants were called using the HMF pipeline (https://github.com/hartwigmedical/pipeline), as detailed previously[60]. The RNA-seq data was processed using Isofox (https://github.com/hartwigmedical/hmftools/tree/master/isofox). The raw expression read counts were normalized across all patients using the Trimmed Mean of M-values (TMM) method. Cancer types with fewer than 20 samples or with uncertain or varying tissue origin were omitted from analysis, leaving 12 cancer types in total across 1,646 patients (breast: 313, ovary: 62, lung: 125, colon/rectum: 393, urinary tract: 118, prostate: 199, esophagus: 53, skin: 216, pancreas: 66, uterus: 26, kidney: 38, nervous system: 37).

For all data collection, hg19 was used as the reference genome. We downloaded CpG islands (across all cell types) from the UCSC genome annotation database. Transcription factors (across all cell types) were collected from the ORegAnno database[61]. ChromHMM states (HMEC) were obtained from Taberlay et al[62].

The following regulatory elements were downloaded for the tissue types closest matching the cancer type. A detailed overview of all regulatory data sources can be found in Table S1. eQTLs were downloaded from GTEx v7 (v8 for kidney, converted to hg19 using the UCSC liftover tool)[63]. Enhancers were obtained from JEME[64]. Super enhancers were collected from dbSUPER[65] and SEdb[66] (kidney, brain and prostate). TADs were downloaded from the 3D genome browser[67], using the UCSC liftover tool to convert from GRCh38 to hg19 for colorectal and ovary. CTCF, DNAse I, h3k4me3, h3k27me3, h3k27ac, h3k4me1 and RNA pol II peaks were downloaded from ENCODE[54].

For each cancer type, regulatory data was selected for the closest matching tissue of

origin. GM12878 was selected where tissue-specific regulatory data was missing, as this data type is available for all regulatory data and thus represents a typical baseline. The impact of selecting less-than-optimal tissue types is further explored in the Results and the procedure is detailed below.

svMIL2 model

svMIL2 follows 2 steps to identify pathogenic non-coding SVs: identifying genes putatively disrupted by TAD boundary-disrupting non-coding SVs, and using MIL to learn which of these SVs are pathogenic. For full details, please refer to the original svMIL publication[34].

In step 1, all genes are identified that are putatively affected by non-coding SVs disrupting boundaries of TADs (Fig S1). Only SVs that start and end within TADs are included, requiring at least 1 basepair overlap with the TAD. For each SV type, we determine which regulatory elements (eQTLs, enhancers and super enhancers) are disrupted by the SV. eQTLs have been previously shown to overlap with enhancers that regulate known cancer genes[68], and are therefore included to account for possibly undiscovered enhancers.

For deletions, all genes in the TAD on one side of the deletion will gain the regulatory elements on the other side of the deletion. Regulatory elements and genes that are overlapped by the deletion itself are not counted as these are not TAD-disrupting events.

For duplications, new TADs are created between the overlapped TAD boundary and the position where this overlapped boundary is re-inserted into the genome. Within this new TAD, genes overlapped by the duplication on one side of the TAD boundary will gain regulatory elements overlapped by the duplication on the other side of the TAD boundary. As no clear consensus exists about how many basepairs of a regulatory element need to be affected to disrupt its function, we require a minimum overlap of 1 basepair.

For inversions, genes lose regulatory elements that are inverted out of the TAD, and gain regulatory elements that are inverted into the TAD. Genes inside the inversion will gain regulatory elements of the TAD that these are inverted in to, and lose regulatory elements that were in the TAD it was inverted out of.

For translocations, we construct a derivative TAD based on the SV orientation in which the new positions of genes and regulatory elements are modeled. Genes gain and lose regulatory elements based on if these are introduced into or removed from the new TAD, respectively.

From these TAD disruptions, a list of SV-gene pairs is constructed containing the regulatory elements that the gene gained or lost as a result of the SV. All genes overlapped (1 basepair) by any coding mutation (SVs, SNVs or CNVs) are excluded to ensure that any effect on the gene is explained only by the non-coding SV. An exception is made for non-coding duplications and inversions, which may overlap the affected gene itself.

In step 2, a MIL model is trained to learn which gains and losses of regulatory elements are characteristic of pathogenic non-coding SVs. Every SV-gene pair is considered a bag, with the disrupted regulatory elements (eQTLs, enhancers and super enhancers) as instances. Each instance is described with a single feature vector. The first two features are binary, indicating if the regulatory element was gained or lost. The next set of features contain either a 0 or 1 depending on if the regulatory element over-

laps (minimum 1 bp) with any of the following annotations (Fig 1b): histone marks (h3k4me3, h3k27me3, h3k27ac, h3k4me1), chromHMM states (CTCF, CTCF+enhancer, CTCF+promoter, enhancer, promoter, poised promoter, heterochromatin, repressed, transcribed), transcription factor binding profiles (DNAseI hypersensitivity, RNA polymerase II, CTCF, transcription factor binding sites) and CpG islands. The third set of features uses the peak intensity of these annotations where available to indicate their strength (histone marks, RNA polymerase II, CTCF). Finally, binary features were used to indicate the type of the regulatory element (eQTL, enhancer, super enhancer) and the number of regulatory elements disrupted by this SV in total (instance count). All features were normalized between 0 and 1.

To label the bags (SV-gene pairs) as pathogenic or non-pathogenic, a z-score was computed from the gene expression to all patients without a disruption to the gene (e.g. coding SV, SNV, CNV or non-coding SV). Bags with z > 1.5 or z < -1.5 were labeled positive, and negative otherwise, which was determined to be the optimal threshold in the previous version of svMIL[34]. Negative bags were randomly subsampled to the number of positive bags to obtain class balance.

A final classifier was obtained by applying the MILES approach[38]. In MILES, a standard feature space is constructed by computing a similarity matrix between the bags and instances. Here, we computed the absolute distance from the mean instance of the bags to all instances. In this space, a random forest was trained to obtain a final classifier. A model was constructed for each SV type separately. All performances were measured using a leave-one-patient-out CV, which models a scenario in which an unseen patient would come into the clinic.

Using svMIL2

svMIL2 takes VCF files containing SVs per patient as input and generates SV-gene pairs based on TAD boundary disruption as detailed above. SV-gene pairs overlapped by coding SNVs, CNVs or SVs are filtered out. SNV files should be provided as VCF files per patient. For CNVs, a tab-delimited file is expected per patient containing the genes and their copy numbers. SV-gene pairs of which the gene has a copy number below 1.7 or above 2.3 are omitted from further analysis. Bags (SV-gene pairs) are labeled for MIL using normalized expression data as described above. To prioritize pathogenic SV-gene pairs, users can either run the MIL in a leave-one-patient-out CV setting, or train the model on one dataset and apply to another. A ranking can be obtained through the classifier probabilities assigned to each bag. A step-by-step tutorial for using svMIL2 is available on GitHub (see Data availability).

Feature selection to improve model performance

To improve the predictive performance of svMIL, we aimed to improve the quality of features through feature selection. Feature importance was assessed by computing the variance of a feature across all instances of the breast cancer samples (Fig S2). Certain features that were present in the original model (Hi-C, h3k9me3, h3k36me3, chromHMM repeat regions and enhancer, h3k9me3 and h3k36me3 strength) contained low variance (log(variance) < -10) and therefore did not contribute to the distinction between positive and negative instances, and were thus omitted.

Improving method accuracy by increasing the number of high-quality instances

To increase the number of informative instances in the model, the eQTL p-value stringency threshold was increased from 5e-8 to 0.05. To account for the resulting increased computational load, all eQTLs, histone marks, and transcription factor (TF) binding sites were binned using a 1kb sliding window.

To account for increased memory consumption resulting from a larger number of SV-gene pairs, bags of each SV type were randomly subsampled if their count exceeded 700, which did not significantly reduce performance on the breast cancer samples for all SV types but inversions, for which the AUC is lowered slightly (Fig S10).

Swapping regulatory elements between cancer types

The effect on performance when swapping regulatory data between cancer types was measured by computing the absolute difference in AUC between the original run and the swapped run, summed across the models for each SV type. A z-score was computed by comparing this summed difference to the mean and standard deviation of the summed differences of all swaps made for that cancer type. Thus, a higher z-score indicates a better performance with that tisue type relative to all other tested tissue types in the swap. For visualization purposes, z-scores were quantized to indicate non-significant effect ($-1 < z < 1$), significant effect ($-2 < z < -1$ and $1 < z < 2$), and highly significant effect ($z < -2$ and $z > 2$).

Running svMIL2 with CTCF loops instead of TAD boundaries

Intra-TAD chromatin loops were predicted using iTAD[69]. Due to the limited availability of cohesin peak data, predictions were limited to tissues for which both cohesin and CTCF peaks were available (breast, colorectal, lung). For cohesin, RAD21 TF ChIP-seq peaks were downloaded for MCF-7 (breast), HCT-116 (colorectal) and A549 (lung). For CTCF, the files listed in Table S1 were used. To predict pathogenic SV-gene pairs, svMIL2 was run using the predicted intra-TAD chromatin loops in place of TAD boundaries.

Data availability

All (processed) WGS and RNA-sequencing data were provided by the Hartwig Medical Foundation under data request DR-104. 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. All code and processed feature data is publicly available at https://github.com/UMCUGenetics/svMIL/. On GitHub a manual can be found reproducing all paper figures and running svMIL2 on a different dataset.

Acknowledgements

The authors thank Edwin Cuppen for his valuable scientific discussions and help with the HMF data.

Author contributions statement

Concept, study design - MN and JdR. Software implementation, data collection and analysis, drafting manuscript - MN. Reviewing manuscript - MN, LN, JdR.

Additional information

Competing interests

The authors declare no competing interests.

References

[1] *Pan-cancer analysis of whole genomes,* Nature **578**, 82 (2020).

[2] Z. Sondka, S. Bamford, C. G. Cole, S. A. Ward, I. Dunham, and S. A. Forbes, *The COSMIC Cancer Gene Census: describing genetic dysfunction across all human cancers,* Nature Reviews Cancer **18**, 696 (2018).

[3] P. Cingolani, A. Platts, L. L. Wang, M. Coon, T. Nguyen, L. Wang, S. J. Land, X. Lu, and D. M. Ruden, *A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff,* Fly **6**, 80 (2012).

[4] N.-L. Sim, P. Kumar, J. Hu, S. Henikoff, G. Schneider, and P. C. Ng, *SIFT web server: predicting effects of amino acid substitutions on proteins,* Nucleic Acids Research **40**, W452 (2012).

[5] I. A. Adzhubei, S. Schmidt, L. Peshkin, V. E. Ramensky, A. Gerasimova, P. Bork, A. S. Kondrashov, and S. R. Sunyaev, *A method and server for predicting damaging missense mutations,* Nature Methods **7**, 248 (2010).

[6] M. F. Rogers, H. A. Shihab, M. Mort, D. N. Cooper, T. R. Gaunt, and C. Campbell, *FATHMM-XF: accurate prediction of pathogenic point mutations via extended features,* Bioinformatics **34**, 511 (2018).

[7] P. Rentzsch, D. Witten, G. M. Cooper, J. Shendure, and M. Kircher, *CADD: predicting the deleteriousness of variants throughout the human genome,* Nucleic Acids Research **47**, D886 (2019).

[8] S. Flygare, E. J. Hernandez, L. Phan, B. Moore, M. Li, A. Fejes, H. Hu, K. Eilbeck, C. Huff, L. Jorde, M. G. Reese, and M. Yandell, *The VAAST Variant Prioritizer (VVP): ultrafast, easy to use whole genome variant prioritization tool,* BMC Bioinformatics **19**, 57 (2018).

[9] Ganel *et al.*, *SVScore: an impact prediction tool for structural variation,* Bioinformatics , btw789 (2016).

[10] D. Dahary, Y. Golan, Y. Mazor, O. Zelig, R. Barshir, M. Twik, T. Iny Stein, G. Rosner, R. Kariv, F. Chen, Q. Zhang, Y. Shen, M. Safran, D. Lancet, and S. Fishilevich, *Genome analysis and knowledge-driven variant interpretation with TGex,* BMC Medical Genomics **12**, 200 (2019).

[11] E. Khurana, Y. Fu, D. Chakravarty, F. Demichelis, M. A. Rubin, and M. Gerstein, *Role of non-coding sequence variants in cancer,* Nature Reviews Genetics **17**, 93 (2016).

[12] J. Zhou and O. G. Troyanskaya, *Predicting effects of noncoding variants with deep learning–based sequence model,* Nature Methods **12**, 931 (2015).

[13] J. Zhou, C. L. Theesfeld, K. Yao, K. M. Chen, A. K. Wong, and O. G. Troyanskaya, *Deep learning sequence-based ab initio prediction of variant effects on expression and disease risk,* Nature Genetics **50**, 1171 (2018).

[14] H. M. Umer, M. Cavalli, M. J. Dabrowski, K. Diamanti, M. Kruczyk, G. Pan, J. Komorowski, and C. Wadelius, *A Significant Regulatory Mutation Burden at a High-Affinity Position of the CTCF Motif in Gastrointestinal Cancers,* Human Mutation **37**, 904 (2016).

[15] L. Mularoni, R. Sabarinathan, J. Deu-Pons, A. Gonzalez-Perez, and N. López-Bigas, *OncodriveFML: a general framework to identify coding and non-coding regions with cancer driver mutations,* Genome Biology **17**, 128 (2016).

[16] H. Hornshøj, M. M. Nielsen, N. A. Sinnott-Armstrong, M. P. Świtnicki, M. Juul, T. Madsen, R. Sallari, M. Kellis, T. Ørntoft, A. Hobolth, and J. S. Pedersen, *Pan-cancer screen for mutations in non-coding elements with conservation and cancer specificity reveals correlations with expression and survival,* npj Genomic Medicine **3**, 1 (2018).

[17] N. D. Dees, Q. Zhang, C. Kandoth, M. C. Wendl, W. Schierding, D. C. Koboldt, T. B. Mooney, M. B. Callaway, D. Dooling, E. R. Mardis, R. K. Wilson, and L. Ding, *MuSiC: Identifying mutational significance in cancer genomes,* Genome Research **22**, 1589 (2012).

[18] D. Tamborero, A. Gonzalez-Perez, C. Perez-Llamas, J. Deu-Pons, C. Kandoth, J. Reimand, M. S. Lawrence, G. Getz, G. D. Bader, L. Ding, and N. Lopez-Bigas, *Comprehensive identification of mutational cancer driver genes across 12 tumor types,* Scientific Reports **3**, 2650 (2013).

[19] E. Rheinbay, M. M. Nielsen, F. Abascal, J. A. Wala, O. Shapira, G. Tiao, H. Hornshøj, J. M. Hess, R. I. Juul, Z. Lin, L. Feuerbach, R. Sabarinathan, T. Madsen, J. Kim, L. Mularoni, S. Shuai, A. Lanzós, C. Herrmann, Y. E. Maruvka, C. Shen, S. B. Amin, P. Bandopadhayay, J. Bertl, K. A. Boroevich, J. Busanovich, J. Carlevaro-Fita, D. Chakravarty, C. W. Y. Chan, D. Craft, P. Dhingra, K. Diamanti, N. A. Fonseca, A. Gonzalez-Perez, Q. Guo, M. P. Hamilton, N. J. Haradhvala, C. Hong, K. Isaev, T. A. Johnson, M. Juul, A. Kahles, A. Kahraman, Y. Kim, J. Komorowski, K. Kumar, S. Kumar, D. Lee, K.-V. Lehmann, Y. Li, E. M. Liu, L. Lochovsky, K. Park, O. Pich, N. D. Roberts, G. Saksena, S. E. Schumacher, N. Sidiropoulos, L. Sieverling, N. Sinnott-Armstrong, C. Stewart, D. Tamborero, J. M. C. Tubio, H. M. Umer, L. Uuskü-la-Reimand, C. Wadelius, L. Wadi, X. Yao, C.-Z. Zhang, J. Zhang, J. E. Haber, A. Hobolth, M. Imielinski, M. Kellis, M. S. Lawrence, C. von Mering, H. Nakagawa, B. J. Raphael,

M. A. Rubin, C. Sander, L. D. Stein, J. M. Stuart, T. Tsunoda, D. A. Wheeler, R. Johnson, J. Reimand, M. Gerstein, E. Khurana, P. J. Campbell, N. López-Bigas, J. Weischenfeldt, R. Beroukhim, I. Martincorena, J. S. Pedersen, and G. Getz, *Analyses of non-coding somatic drivers in 2,658 cancer whole genomes,* Nature **578**, 102 (2020).

[20] J. R. Dixon, S. Selvaraj, F. Yue, A. Kim, Y. Li, Y. Shen, M. Hu, J. S. Liu, and B. Ren, *Topological domains in mammalian genomes identified by analysis of chromatin interactions,* Nature **485**, 376 (2012).

[21] G. Fudenberg, M. Imakaev, C. Lu, A. Goloborodko, N. Abdennur, and L. A. Mirny, *Formation of Chromosomal Domains by Loop Extrusion,* Cell Reports **15**, 2038 (2016).

[22] A. L. Sanborn, S. S. P. Rao, S.-C. Huang, N. C. Durand, M. H. Huntley, A. I. Jewett, I. D. Bochkov, D. Chinnappan, A. Cutkosky, J. Li, K. P. Geeting, A. Gnirke, A. Melnikov, D. McKenna, E. K. Stamenova, E. S. Lander, and E. L. Aiden, *Chromatin extrusion explains key features of loop and domain formation in wild-type and engineered genomes,* Proceedings of the National Academy of Sciences **112**, E6456 (2015).

[23] E. Giorgio, D. Robyr, M. Spielmann, E. Ferrero, E. Di Gregorio, D. Imperiale, G. Vaula, G. Stamoulis, F. Santoni, C. Atzori, L. Gasparini, D. Ferrera, C. Canale, M. Guipponi, L. A. Pennacchio, S. E. Antonarakis, A. Brussino, and A. Brusco, *A large genomic deletion leads to enhancer adoption by the lamin B1 gene: a second path to autosomal dominant adult-onset demyelinating leukodystrophy (ADLD),* Human Molecular Genetics **24**, 3143 (2015).

[24] C. Redin, H. Brand, R. L. Collins, T. Kammin, E. Mitchell, J. C. Hodge, C. Hanscom, V. Pillalamarri, C. M. Seabra, M.-A. Abbott, O. A. Abdul-Rahman, E. Aberg, R. Adley, S. L. Alcaraz-Estrada, F. S. Alkuraya, Y. An, M.-A. Anderson, C. Antolik, K. Anyane-Yeboa, J. F. Atkin, T. Bartell, J. A. Bernstein, E. Beyer, I. Blumenthal, E. M. H. F. Bongers, E. H. Brilstra, C. W. Brown, H. T. Brüggenwirth, B. Callewaert, C. Chiang, K. Corning, H. Cox, E. Cuppen, B. B. Currall, T. Cushing, D. David, M. A. Deardorff, A. Dheedene, M. D'Hooghe, B. B. A. de Vries, D. L. Earl, H. L. Ferguson, H. Fisher, D. R. FitzPatrick, P. Gerrol, D. Giachino, J. T. Glessner, T. Gliem, M. Grady, B. H. Graham, C. Griffis, K. W. Gripp, A. L. Gropman, A. Hanson-Kahn, D. J. Harris, M. A. Hayden, R. Hill, R. Hochstenbach, J. D. Hoffman, R. J. Hopkin, M. W. Hubshman, A. M. Innes, M. Irons, M. Irving, J. C. Jacobsen, S. Janssens, T. Jewett, J. P. Johnson, M. C. Jongmans, S. G. Kahler, D. A. Koolen, J. Korzelius, P. M. Kroisel, Y. Lacassie, W. Lawless, E. Lemyre, K. Leppig, A. V. Levin, H. Li, H. Li, E. C. Liao, C. Lim, E. J. Lose, D. Lucente, M. J. Macera, P. Manavalan, G. Mandrile, C. L. Marcelis, L. Margolin, T. Mason, D. Masser-Frye, M. W. McClellan, C. J. Z. Mendoza, B. Menten, S. Middelkamp, L. R. Mikami, E. Moe, S. Mohammed, T. Mononen, M. E. Mortenson, G. Moya, A. W. Nieuwint, Z. Ordulu, S. Parkash, S. P. Pauker, S. Pereira, D. Perrin, K. Phelan, R. E. P. Aguilar, P. J. Poddighe, G. Pregno, S. Raskin, L. Reis, W. Rhead, D. Rita, I. Renkens, F. Roelens, J. Ruliera, P. Rump, S. L. P. Schilit, R. Shaheen, R. Sparkes, E. Spiegel, B. Stevens, M. R. Stone, J. Tagoe, J. V. Thakuria, B. W. van Bon, J. van de Kamp, I. van Der Burgt, T. van Essen, C. M. van Ravenswaaij-Arts, M. J. van Roosmalen, S. Vergult,

C. M. L. Volker-Touw, D. P. Warburton, M. J. Waterman, S. Wiley, A. Wilson, M. d. l. C. A. Yerena-de Vega, R. T. Zori, B. Levy, H. G. Brunner, N. de Leeuw, W. P. Kloosterman, E. C. Thorland, C. C. Morton, J. F. Gusella, and M. E. Talkowski, *The genomic landscape of balanced cytogenetic abnormalities associated with human congenital anomalies,* Nature Genetics **49**, 36 (2017).

[25] M. Franke, D. M. Ibrahim, G. Andrey, W. Schwarzer, V. Heinrich, R. Schöpflin, K. Kraft, R. Kempfer, I. Jerković, W.-L. Chan, M. Spielmann, B. Timmermann, L. Wittler, I. Kurth, P. Cambiaso, O. Zuffardi, G. Houge, L. Lambie, F. Brancati, A. Pombo, M. Vingron, F. Spitz, and S. Mundlos, *Formation of new chromatin domains determines pathogenicity of genomic duplications,* Nature **538**, 265 (2016).

[26] D. G. Lupiáñez, K. Kraft, V. Heinrich, P. Krawitz, F. Brancati, E. Klopocki, D. Horn, H. Kayserili, J. M. Opitz, R. Laxova, F. Santos-Simarro, B. Gilbert-Dussardier, L. Wittler, M. Borschiwer, S. A. Haas, M. Osterwalder, M. Franke, B. Timmermann, J. Hecht, M. Spielmann, A. Visel, and S. Mundlos, *Disruptions of Topological Chromatin Domains Cause Pathogenic Rewiring of Gene-Enhancer Interactions,* Cell **161**, 1012 (2015).

[27] Zhang *et al.*, *Local and global chromatin interactions are altered by large genomic deletions associated with human brain development,* Nature Communications **9**, 5356 (2018).

[28] D. Hnisz, A. S. Weintraub, D. S. Day, A.-L. Valton, R. O. Bak, C. H. Li, J. Goldmann, B. R. Lajoie, Z. P. Fan, A. A. Sigova, J. Reddy, D. Borges-Rivera, T. I. Lee, R. Jaenisch, M. H. Porteus, J. Dekker, and R. A. Young, *Activation of proto-oncogenes by disruption of chromosome neighborhoods,* Science **351**, 1454 (2016).

[29] J. Weischenfeldt, T. Dubash, A. P. Drainas, B. R. Mardin, Y. Chen, A. M. Stütz, S. M. Waszak, G. Bosco, A. R. Halvorsen, B. Raeder, T. Efthymiopoulos, S. Erkek, C. Siegl, H. Brenner, O. T. Brustugun, S. M. Dieter, P. A. Northcott, I. Petersen, S. M. Pfister, M. Schneider, S. K. Solberg, E. Thunissen, W. Weichert, T. Zichner, R. Thomas, M. Peifer, A. Helland, C. R. Ball, M. Jechlinger, R. Sotillo, H. Glimm, and J. O. Korbel, *Pan-cancer analysis of somatic copy-number alterations implicates IRS4 and IGF2 in enhancer hijacking,* Nature Genetics **49**, 65 (2017).

[30] A.-L. Valton and J. Dekker, *TAD disruption as oncogenic driver,* Current Opinion in Genetics and Development **36**, 34 (2016).

[31] K. C. Akdemir, V. T. Le, S. Chandran, Y. Li, R. G. Verhaak, R. Beroukhim, P. J. Campbell, L. Chin, J. R. Dixon, and P. A. Futreal, *Disruption of chromatin folding domains by somatic genomic rearrangements in human cancer,* Nature Genetics **52**, 294 (2020).

[32] J. R. Dixon, J. Xu, V. Dileep, Y. Zhan, F. Song, V. T. Le, G. G. Yardımcı, A. Chakraborty, D. V. Bann, Y. Wang, R. Clark, L. Zhang, H. Yang, T. Liu, S. Iyyanki, L. An, C. Pool, T. Sasaki, J. C. Rivera-Mulia, H. Ozadam, B. R. Lajoie, R. Kaul, M. Buckley, K. Lee, M. Diegel, D. Pezic, C. Ernst, S. Hadjur, D. T. Odom, J. A. Stamatoyannopoulos, J. R.

Broach, R. C. Hardison, F. Ay, W. S. Noble, J. Dekker, D. M. Gilbert, and F. Yue, *Integrative detection and analysis of structural variation in cancer genomes,* Nature Genetics **50**, 1388 (2018).

[33] Huynh *et al.*, *TAD fusion score: discovery and ranking the contribution of deletions to genome structure,* Genome Biology **20**, 60 (2019).

[34] M. M. Nieboer and J. de Ridder, *svMIL: predicting the pathogenic effect of TAD boundary-disrupting somatic structural variants through multiple instance learning,* Bioinformatics **36**, i692 (2020).

[35] E. M. Liu, A. Martinez-Fundichely, B. J. Diaz, B. Aronson, T. Cuykendall, M. MacKay, P. Dhingra, E. W. Wong, P. Chi, E. Apostolou, N. E. Sanjana, and E. Khurana, *Identification of Cancer Drivers at CTCF Insulators in 1,962 Whole Genomes,* Cell Systems **8**, 446 (2019).

[36] A. Despang, R. Schöpflin, M. Franke, S. Ali, I. Jerković, C. Paliou, W.-L. Chan, B. Timmermann, L. Wittler, M. Vingron, S. Mundlos, and D. M. Ibrahim, *Functional dissection of the Sox9–Kcnj2 locus identifies nonessential and instructive roles of TAD architecture,* Nature Genetics **51**, 1263 (2019).

[37] Dietterich *et al.*, *Solving the multiple instance problem with axis-parallel rectangles,* Artificial Intelligence **89**, 31 (1997).

[38] Chen *et al.*, *MILES: Multiple-Instance Learning via Embedded Instance Selection,* IEEE Transactions on Pattern Analysis and Machine Intelligence **28**, 1931 (2006).

[39] P. Priestley, J. Baber, M. P. Lolkema, N. Steeghs, E. de Bruijn, C. Shale, K. Duyvesteyn, S. Haidari, A. van Hoeck, W. Onstenk, P. Roepman, M. Voda, H. J. Bloemendal, V. C. G. Tjan-Heijnen, C. M. L. van Herpen, M. Labots, P. O. Witteveen, E. F. Smit, S. Sleijfer, E. E. Voest, and E. Cuppen, *Pan-cancer whole-genome analyses of metastatic solid tumours,* Nature **575**, 210 (2019).

[40] L. Angus, M. Smid, S. M. Wilting, J. van Riet, A. Van Hoeck, L. Nguyen, S. Nik-Zainal, T. G. Steenbruggen, V. C. G. Tjan-Heijnen, M. Labots, J. M. G. H. van Riel, H. J. Bloemendal, N. Steeghs, M. P. Lolkema, E. E. Voest, H. J. G. van de Werken, A. Jager, E. Cuppen, S. Sleijfer, and J. W. M. Martens, *The genomic landscape of metastatic breast cancer highlights changes in mutation and signature frequencies,* Nature Genetics **51**, 1450 (2019).

[41] A. Degasperi, T. D. Amarante, J. Czarnecki, S. Shooter, X. Zou, D. Glodzik, S. Morganella, A. S. Nanda, C. Badja, G. Koh, S. E. Momen, I. Georgakopoulos-Soares, J. M. L. Dias, J. Young, Y. Memari, H. Davies, and S. Nik-Zainal, *A practical framework and online tool for mutational signature analyses show intertissue variation and driver dependencies,* Nature Cancer **1**, 249 (2020).

[42] S. Christensen, B. Van der Roest, N. Besselink, R. Janssen, S. Boymans, J. W. M. Martens, M.-L. Yaspo, P. Priestley, E. Kuijk, E. Cuppen, and A. Van Hoeck, *5-Fluorouracil treatment induces characteristic T>G mutations in human cancer,* Nature Communications **10**, 4571 (2019).

[43] K. G. Samsom, S. Levy, L. M. Veenendaal, P. Roepman, L. L. Kodach, N. Steeghs, G. D. Valk, M. Wouter Dercksen, K. F. D. Kuhlmann, W. H. M. Verbeek, G. A. Meijer, M. E. T. Tesselaar, and J. G. Berg, *Driver mutations occur frequently in metastases of well-differentiated small intestine neuroendocrine tumours,* Histopathology , his.14252 (2020).

[44] Y. Drier, M. J. Cotton, K. E. Williamson, S. M. Gillespie, R. J. H. Ryan, M. J. Kluk, C. D. Carey, S. J. Rodig, L. M. Sholl, A. H. Afrogheh, W. C. Faquin, L. Queimado, J. Qi, M. J. Wick, A. K. El-Naggar, J. E. Bradner, C. A. Moskaluk, J. C. Aster, B. Knoechel, and B. E. Bernstein, *An oncogenic MYB feedback loop drives alternate cell fates in adenoid cystic carcinoma,* Nature Genetics **48**, 265 (2016).

[45] P. A. Northcott, C. Lee, T. Zichner, A. M. Stütz, S. Erkek, D. Kawauchi, D. J. H. Shih, V. Hovestadt, M. Zapatka, D. Sturm, D. T. W. Jones, M. Kool, M. Remke, F. M. G. Cavalli, S. Zuyderduyn, G. D. Bader, S. VandenBerg, L. A. Esparza, M. Ryzhova, W. Wang, A. Wittmann, S. Stark, L. Sieber, H. Seker-Cin, L. Linke, F. Kratochwil, N. Jäger, I. Buchhalter, C. D. Imbusch, G. Zipprich, B. Raeder, S. Schmidt, N. Diessl, S. Wolf, S. Wiemann, B. Brors, C. Lawerenz, J. Eils, H.-J. Warnatz, T. Risch, M.-L. Yaspo, U. D. Weber, C. C. Bartholomae, C. von Kalle, E. Turányi, P. Hauser, E. Sanden, A. Darabi, P. Siesjö, J. Sterba, K. Zitterbart, D. Sumerauer, P. van Sluis, R. Versteeg, R. Volckmann, J. Koster, M. U. Schuhmann, M. Ebinger, H. L. Grimes, G. W. Robinson, A. Gajjar, M. Mynarek, K. von Hoff, S. Rutkowski, T. Pietsch, W. Scheurlen, J. Felsberg, G. Reifenberger, A. E. Kulozik, A. von Deimling, O. Witt, R. Eils, R. J. Gilbertson, A. Korshunov, M. D. Taylor, P. Lichter, J. O. Korbel, R. J. Wechsler-Reya, and S. M. Pfister, *Enhancer hijacking activates GFI1 family oncogenes in medulloblastoma,* Nature **511**, 428 (2014).

[46] X. Ji, D. B. Dadon, B. E. Powell, Z. P. Fan, D. Borges-Rivera, S. Shachar, A. S. Weintraub, D. Hnisz, G. Pegoraro, T. I. Lee, T. Misteli, R. Jaenisch, and R. A. Young, *3D Chromosome Regulatory Landscape of Human Pluripotent Cells,* Cell Stem Cell **18**, 262 (2016).

[47] F. Martínez-Jiménez, F. Muiños, I. Sentís, J. Deu-Pons, I. Reyes-Salazar, C. Arnedo-Pac, L. Mularoni, O. Pich, J. Bonet, H. Kranas, A. Gonzalez-Perez, and N. Lopez-Bigas, *A compendium of mutational cancer driver genes,* Nature Reviews Cancer **20**, 555 (2020).

[48] A. Bhattacharya, R. D. Bense, C. G. Urzúa-Traslaviña, E. G. E. de Vries, M. A. T. M. van Vugt, and R. S. N. Fehrmann, *Transcriptional effects of copy number alterations in a large set of human cancers,* Nature Communications **11**, 715 (2020).

[49] D. J. McGrail, L. Federico, Y. Li, H. Dai, Y. Lu, G. B. Mills, S. Yi, S.-Y. Lin, and N. Sahni, *Multi-omics analysis reveals neoantigen-independent immune cell infiltration in copy-number driven cancers,* Nature Communications **9**, 1317 (2018).

[50] S. Lu, T. Ahmed, P. Du, and Y. Wang, *Genomic Variations in Pancreatic Cancer and Potential Opportunities for Development of New Approaches for Diagnosis and Treatment,* International Journal of Molecular Sciences **18**, 1201 (2017).

[51] J. Schreiber, T. Durham, J. Bilmes, and W. S. Noble, *Avocado: a multi-scale deep tensor factorization method learns a latent representation of the human epigenome,* Genome Biology **21**, 81 (2020).

[52] P. Apostolou and I. Papasotiriou, *Current perspectives on CHEK2 mutations in breast cancer,* Breast Cancer: Targets and Therapy **Volume 9**, 331 (2017).

[53] F. Ye, D. C. Samuels, T. Clark, and Y. Guo, *High-throughput sequencing in mitochondrial DNA research,* Mitochondrion **17**, 157 (2014).

[54] Dunham *et al.*, *An integrated encyclopedia of DNA elements in the human genome,* Nature **489**, 57 (2012).

[55] J. Ernst and M. Kellis, *Large-scale imputation of epigenomic datasets for systematic annotation of diverse human tissues,* Nature Biotechnology **33**, 364 (2015).

[56] T. J. Durham, M. W. Libbrecht, J. J. Howbert, J. Bilmes, and W. S. Noble, *PREDICTD PaRallel Epigenomics Data Imputation with Cloud-based Tensor Decomposition,* Nature Communications **9**, 1402 (2018).

[57] T. J. Treangen and S. L. Salzberg, *Repetitive DNA and next-generation sequencing: computational challenges and solutions,* Nature Reviews Genetics **13**, 36 (2012).

[58] A. Bashashati, G. Haffari, J. Ding, G. Ha, K. Lui, J. Rosner, D. G. Huntsman, C. Caldas, S. A. Aparicio, and S. P. Shah, *DriverNet: uncovering the impact of somatic driver mutations on transcriptional networks in cancer,* Genome Biology **13**, R124 (2012).

[59] J. P. Hou and J. Ma, *DawnRank: discovering personalized driver genes in cancer,* Genome Medicine **6**, 56 (2014).

[60] L. Nguyen, J. W. M. Martens, A. Van Hoeck, and E. Cuppen, *Pan-cancer landscape of homologous recombination deficiency,* Nature Communications **11**, 5584 (2020).

[61] Lesurf *et al.*, *ORegAnno 3.0: a community-driven resource for curated regulatory annotation,* Nucleic Acids Research **44**, D126 (2016).

[62] Taberlay *et al.*, *Reconfiguration of nucleosome-depleted regions at distal regulatory elements accompanies DNA methylation of enhancers and insulators in cancer,* Genome Research **24**, 1421 (2014).

[63] Aguet *et al.*, *Genetic effects on gene expression across human tissues,* Nature **550**, 204 (2017).

[64] Q. Cao, C. Anyansi, X. Hu, L. Xu, L. Xiong, W. Tang, M. T. S. Mok, C. Cheng, X. Fan, M. Gerstein, A. S. L. Cheng, and K. Y. Yip, *Reconstruction of enhancer–target networks in 935 samples of human primary cells, tissues and cell lines,* Nature Genetics **49**, 1428 (2017).

[65] Khan *et al.*, *dbSUPER: a database of super-enhancers in mouse and human genome,* Nucleic Acids Research **44**, D164 (2016).

[66] Y. Jiang, F. Qian, X. Bai, Y. Liu, Q. Wang, B. Ai, X. Han, S. Shi, J. Zhang, X. Li, Z. Tang, Q. Pan, Y. Wang, F. Wang, and C. Li, *SEdb: a comprehensive human super-enhancer database,* Nucleic Acids Research **47**, D235 (2019).

[67] Y. Wang, B. Zhang, L. Zhang, L. An, J. Xu, D. Li, M. N. K. Choudhary, Y. Li, M. Hu, R. Hardison, T. Wang, and F. Yue, *The 3D Genome Browser: a web-based browser for visualizing 3D genome organization and long-range chromatin interactions,* bioRxiv (2017).

[68] H. Chen *et al.*, *A Pan-Cancer Analysis of Enhancer Expression in Nearly 9000 Patient Samples,* Cell **173**, 386 (2018).

[69] B. J. Matthews and D. J. Waxman, *Computational prediction of CTCF/cohesin-based intra-TAD loops that insulate chromatin contacts and gene expression in mouse liver,* eLife **7** (2018), 10.7554/eLife.34077.

Supplementary Data

Table S1-S3 are available upon request.

Supplementary figures

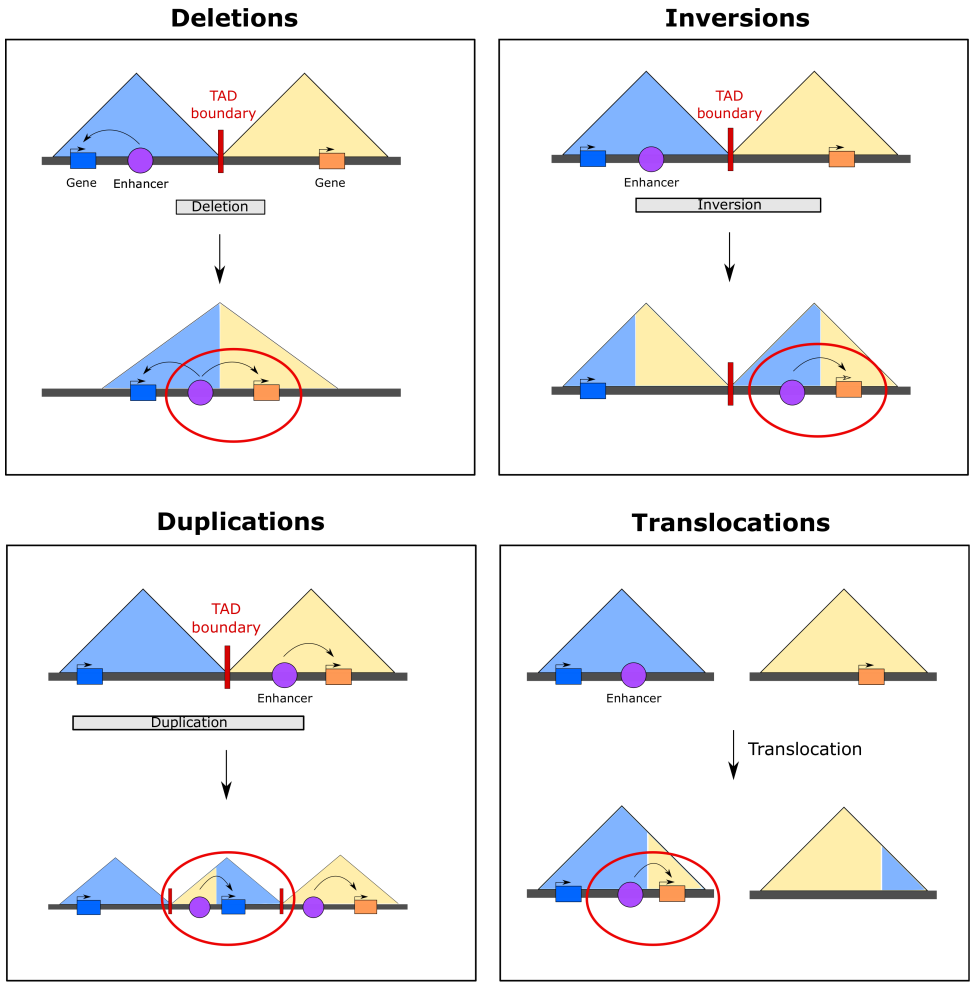


Fig. S1. Schematic illustration of disruptions of TAD boundaries by non-coding SVs are modeled in svMIL2. In each example, a gain of interaction with an enhancer is shown. For inversions, the gene in the left TAD also loses potential interactions with the enhancer.

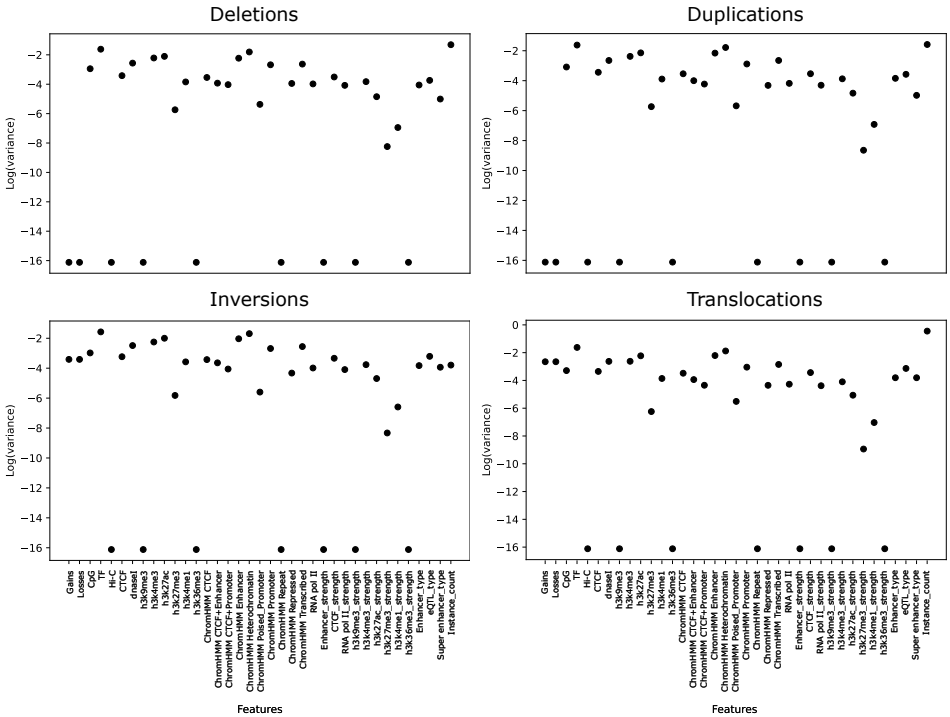


Fig. S2. Log of the variance across all instances. A margin of 0.00001 was added to variances of 0 to compute the log. Features with variances lower than -10 for all SV types were removed from the model.

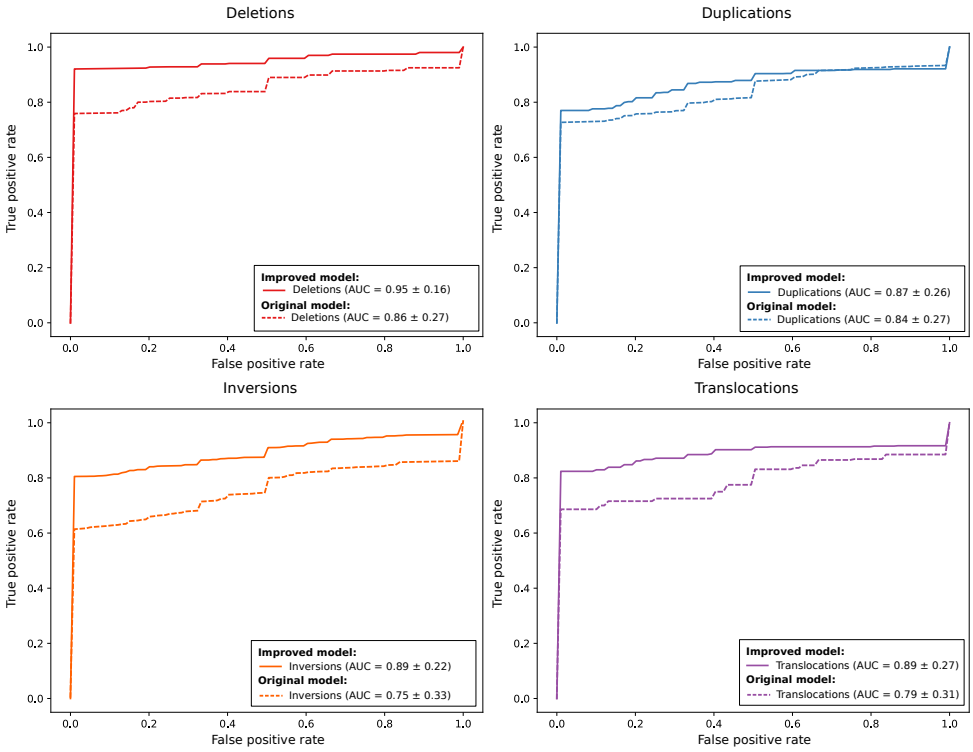


Fig. S3. Performance ROC curves of svMIL2 compared to the original svMIL on all breast cancer samples in a leave-one-patient-out CV setting per SV type.

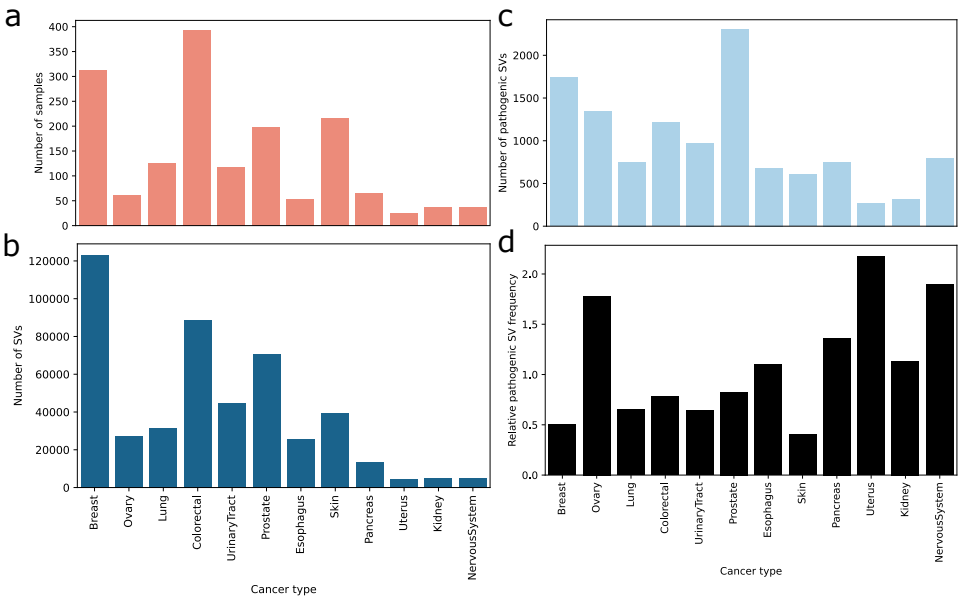


Fig. S4. Number of (a) samples, (b) SVs and (c) predicted pathogenic SVs in each cancer type. (d) Percentage of predicted pathogenic SVs compared to the total number of SVs across all samples in a cancer type. The numbers in this figure are also provided in Table S2.

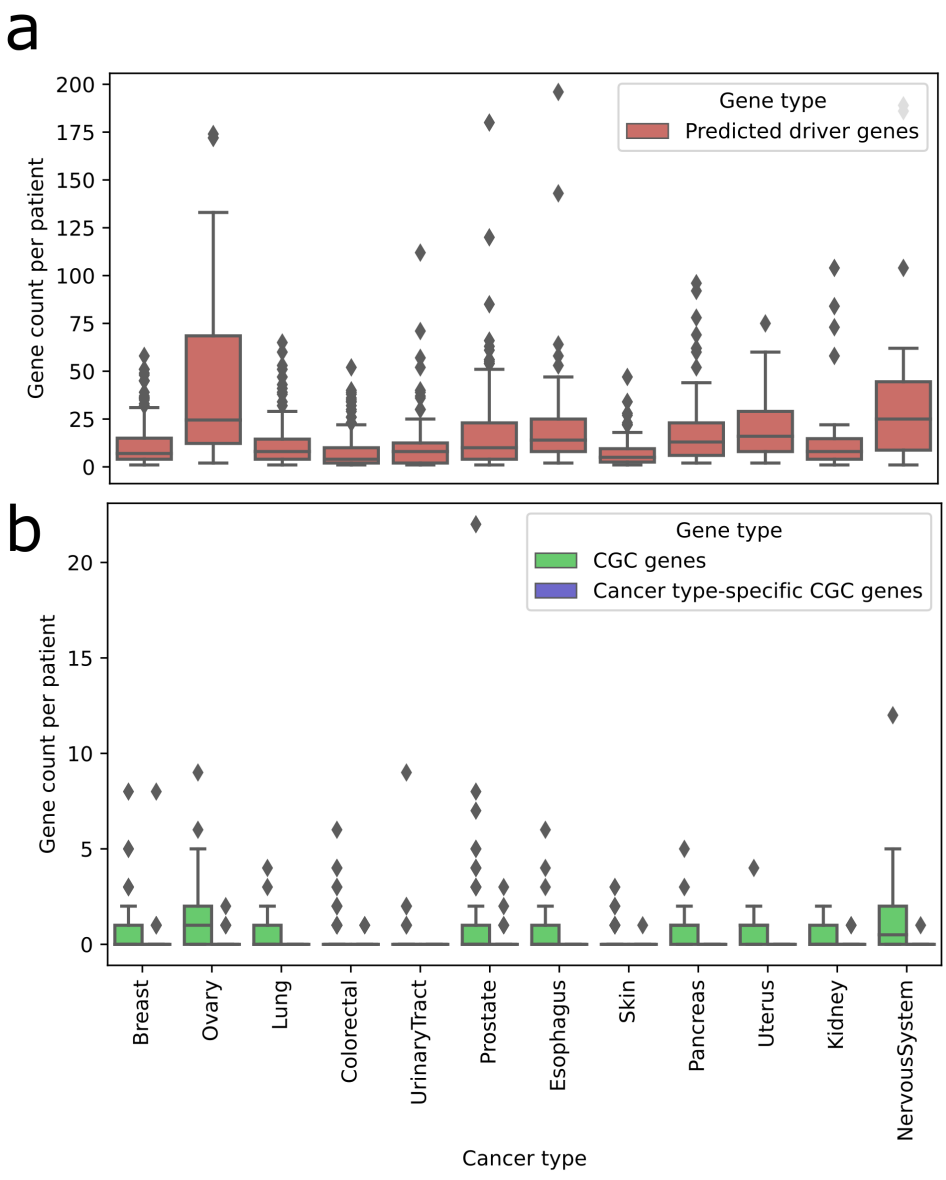


Fig. S5. Distribution of the number of predicted (a) driver genes and (b) (cancer type-specific) CGC genes disrupted by non-coding SVs across patients.

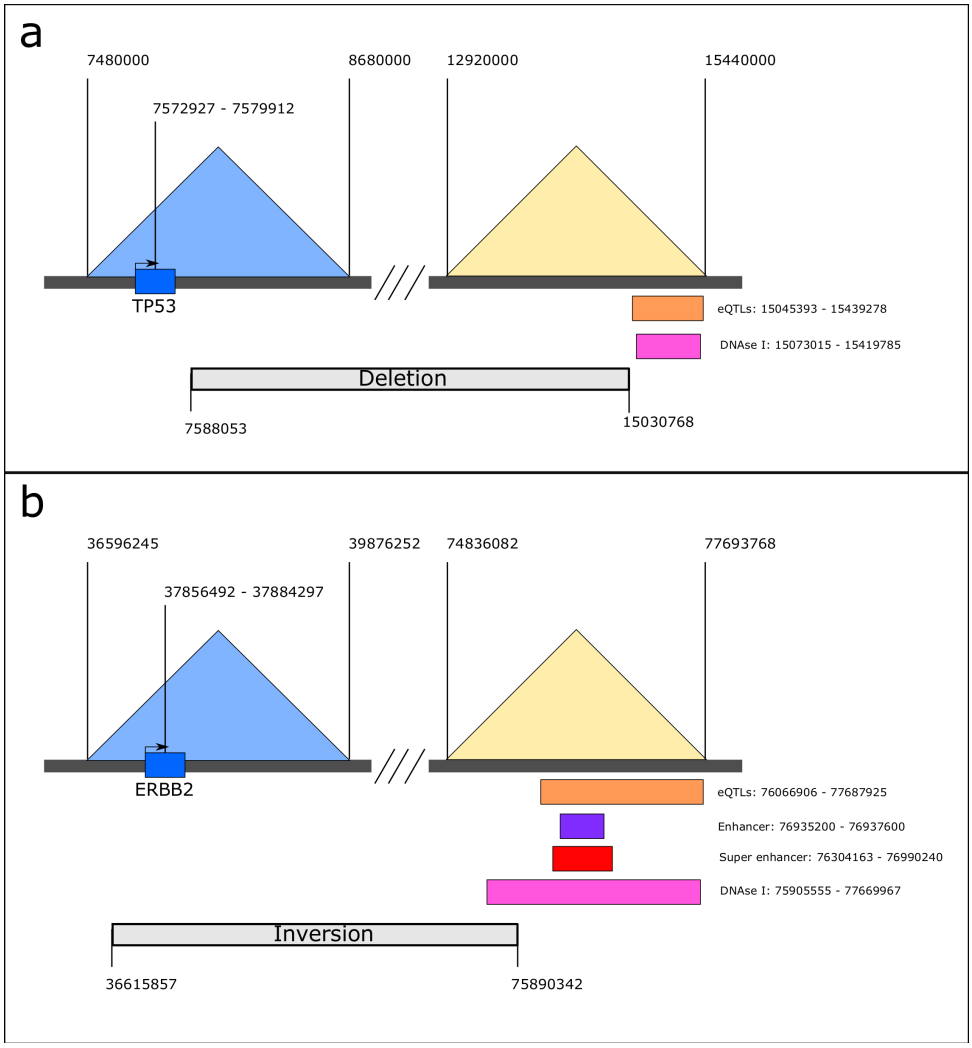


Fig. S6. Schematic illustration of non-coding SVs exerting possible pathogenic effects on (a) TP53 in a prostate cancer patient and (b) ERBB2 in an ovarian cancer patient. For ERBB2, the inversion brings the gene into a new TAD where potential new interactions can be formed with a cluster of eQTLs, an enhancer and a super enhancer that are located in a region with high DNAse I (open chromatin). For TP53, the deletion removes TAD boundaries, bringing the gene close to a cluster of eQTLs with high DNAse I (open chromatin).

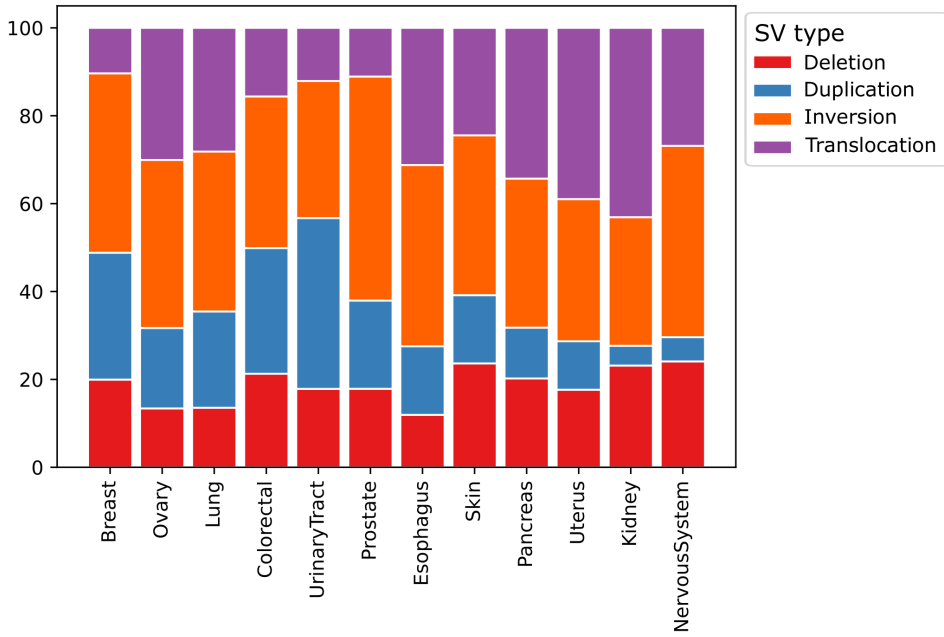


Fig. S7. Relative contribution of each SV type to the predicted pathogenic SVs across all samples in each cancer type.

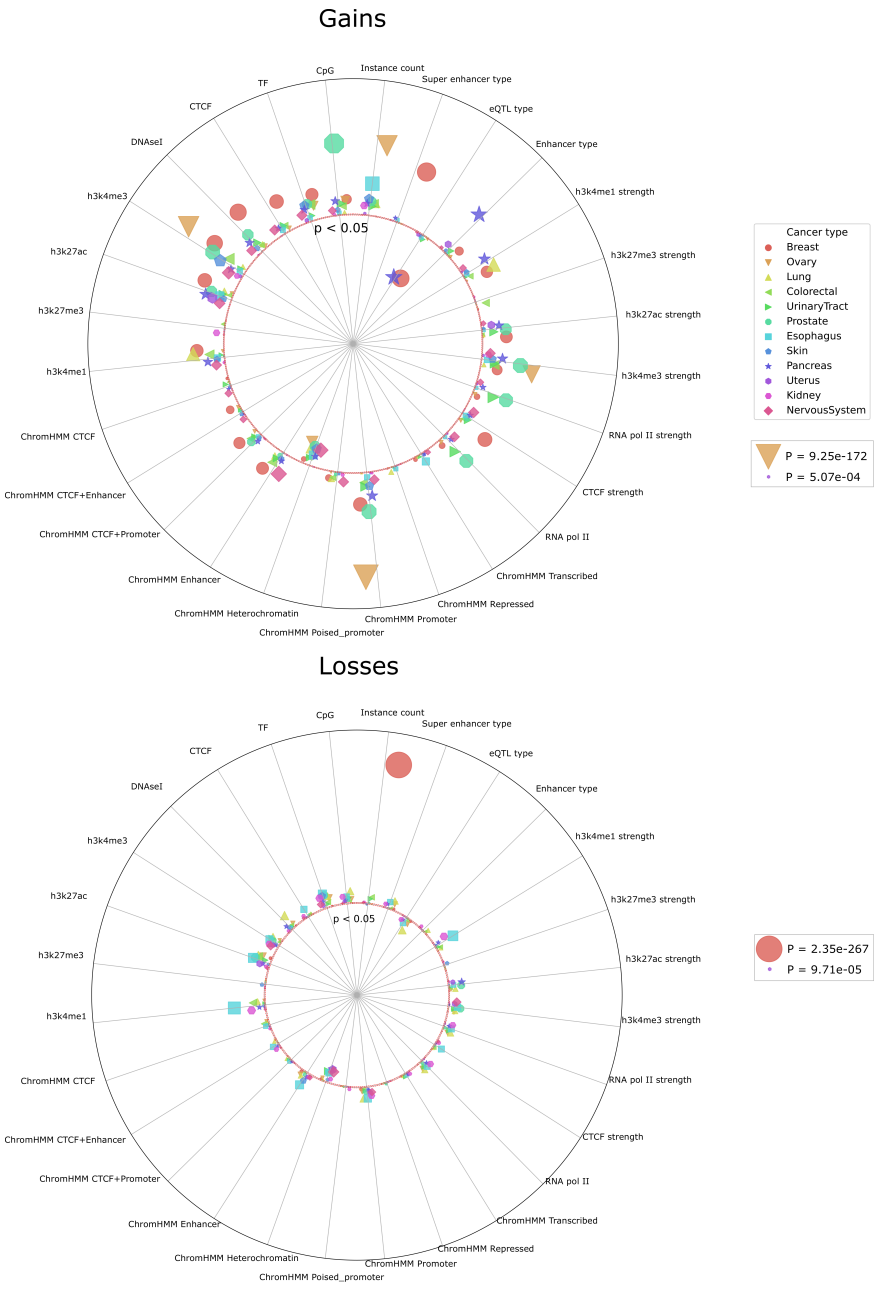


Fig. S8. Regulatory elements affected by non-coding SVs, split into gains and losses. P-values are computed from a z-score based on the frequency of that feature in a gained regulatory element compared to 100 random gained regulatory elements. Points above the red dashed lines indicate P < 0.05 and z > 0, whereas points below the red dashed lines indicate P < 0.05 and z < 0. Note that the significances are slightly different from Fig 3, which is not split into gains and losses. In the majority of cancer types, (super) enhancers are affected rather than eQTLs, which is visible in combination with active enhancer marks (h3k27ac) and open chromatin (lack of ChromHMM heterochromatin). Gains reach higher significance as these are observed more often than by random chance. For losses, the number of lost regulatory elements per patient (instance count) stands out in breast cancer specifically.

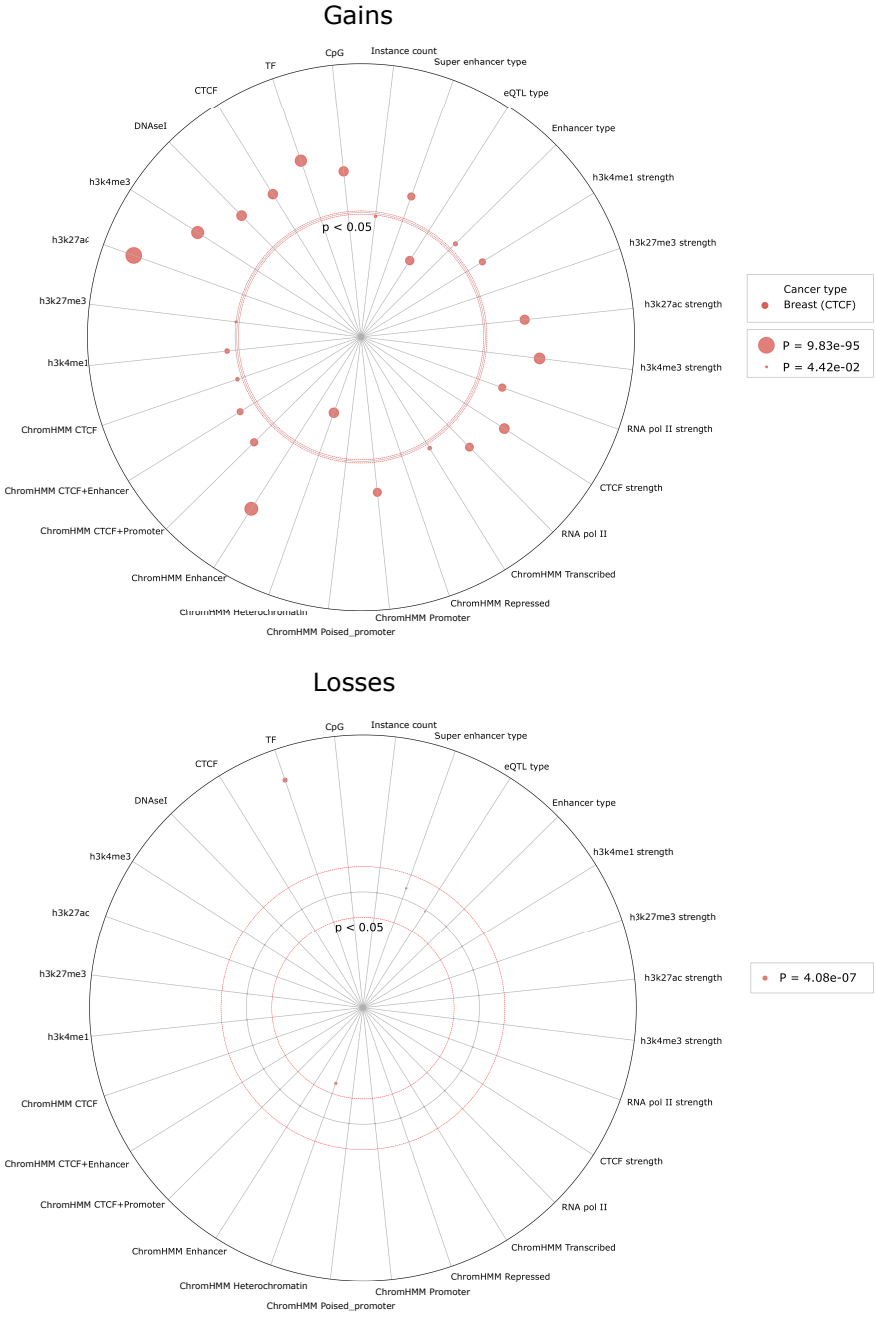


Fig. S9. Regulatory elements affected by non-coding SVs, split into gains and losses, specific for using CTCF loops instead of TAD boundaries in breast cancer. P-values are computed from a z-score based on the frequency of that feature in a gained regulatory element compared to 100 random gained regulatory elements. Points above the red dashed lines indicate $P < 0.05$ and $z > 0$, whereas points below the red dashed lines indicate $P < 0.05$ and $z < 0$. Note that the significances are slightly different from Fig 5c, which is not split into gains and losses. The pattern of gaining (super) enhancers with active (h3k27ac) marks in open chromatin (lack of ChromHMM heterochromatin) is visible here too. Gains reach higher significance as these are observed more often than by random chance.

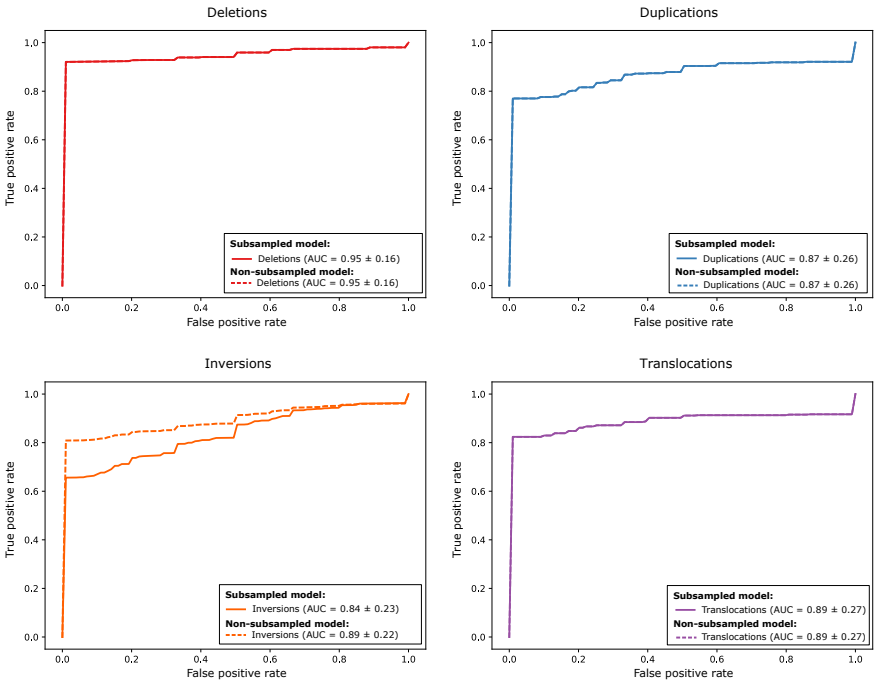


Fig. S10. Subsampling bags does not significantly impact model performance. All performances are unaffected except for inversions, for which performance decreases only slightly.

5

Pan-cancer gene deficiency status prediction in 4,069 whole cancer genomes using machine learning

Marleen M. Nieboer, Jesko Wagner, Luan Nguyen, Jeroen de Ridder

Manuscript in preparation

Abstract

Loss of DNA-repair genes, such as BRCA1, BRCA2 and CDK12, leaves specific signatures of somatic structural variants (SVs) in the genome. These signatures can be used to train machine learning classifiers to detect gene deficiency status (DS). However, obtaining a sample-level classification while representing all genomic features of individual SVs, such as chromatin states, in a feature matrix is a challenging task. Here, we use multiple instance learning (MIL) to describe samples as bags, which may contain any number of SVs represented by an individual feature vector. We show that MIL outperforms the existing non-MIL state-of-the-art. Another problem is that gene deficiency caused by pathways or variants of unknown significance are hard to detect with whole genome sequencing (WGS) data alone, which is often the only available data type for a patient. As a result, the labels of especially the negative set may be noisy. We overcome this problem by combining MIL with positive unlabeled (PU) learning, a classification strategy that deals with label noise by considering the negative set as unlabeled. We demonstrate that PU learning moderately improves the AUCPR. However, as the ground truth labels remain unknown, it is difficult to interpret classifier performance based on the AUCPR. Due to label noise, a reported false positive sample may actually have the deficiency signature, but result in a lower AUCPR. As an alternative strategy to measuring performance, we introduce a swap-one-patient-out cross validation (sopoCV). Each positive sample is artificially swapped to the negative set and the total number of samples correctly identified as positive is reported as classifier performance. We find that AUCPR may result in a biased interpretation of performance, and sopoCV gives a more accurate representation of how well a classifier detects gene DS.

Introduction

Whole-genome sequencing (WGS) of every cancer patient is becoming routine practice in the clinic. These data have enabled personalized treatment of patients with biallelic loss-of-function (LOF) mutations in genes for which effective therapies are readily available, such as immunotherapy for CDK12[1], or PARP-inhibitors for BRCA1 and BRCA2[2, 3]. However, cases of biallelic loss where one or both alleles have been inactivated through epigenetic modifications or pathway downregulation are difficult to detect from WGS data alone. Furthermore, the impact of non-coding mutations or variants of unknown significance (VUS) on gene LOF is difficult to assess. While this problem could partly be solved by the acquisition of more layers of -omics data, in reality these data are too costly for routine diagnostics. However, the ability to detect biallelic loss would greatly benefit more cancer patients.

Recently, a study from the Pan-Cancer Analysis of Whole Genomes (PCAWG) consortium revealed that DNA-repair genes harboring pathogenic mutations leave specific signatures of somatic structural variants (SVs) in the genome[4]. For example, loss of CDK12 was found to lead to an increase in tandem-duplications in late replicating regions[1, 4], whereas loss of BRCA1/2 results in a high frequency of small deletions[4]. Consequently, these mutational signatures may serve as a valuable proxy for the deficiency status (DS) of a gene. More importantly, leveraging machine learning to predict the DS of a gene based on mutational signatures enables detection of indirect or non-genetic

inactivation of genes from WGS data alone.

DS classification has been utilized previously in a method called CHORD for the detection of biallelic loss of BRCA1 and BRCA2 from SV, SNV and indel signatures[3]. To obtain a classification at the sample-level, this method aggregates mutation signatures within a sample. For example, rather than describing each SV by length and type, features represent the number of SVs of a specific type and length identified in that sample. This summarization into a feature matrix enables application of standard machine learning. However, an important limitation of this summarization is that relevant information, such as replication timing, about the individual mutations is lost.

One alternative method to circumvent summarization into a fixed feature matrix is Multiple Instance Learning (MIL). In MIL, so-called 'bags' are defined which contain one or more 'instances'. Each individual instance can be represented with a feature vector that describes its characteristics. In the DS classification setting, each sample would be represented as a bag, and the instances would be represented as the SVs. The SVs can then in turn be described by (genomic) features, such as the replication timing or chromatin state of the region in which they are present. The model needs to learn in this MIL-space what the characteristics are of positive bags that are not shared with the negative bags. A very simple approach (simpleMI) is to convert the MIL-space to a regular feature space by averaging the features of all mutations in a bag (Fig 1A)[5]. In this feature space, a regular classifier can be trained to learn the bag labels. While this approach has been reported to achieve good performance in many classification tasks[6], it does not overcome the problem that information about individual mutations is lost. Furthermore, like other traditional MIL algorithms, it is assumed that the presence of at least one positive instance defines the whole bag as positive. In terms of the underlying problem, the presence of one SV of the expected type and size would define a sample as having biallelic loss, while in reality some SVs may also be generated by other cellular processes. Therefore, the standard MIL assumption may not be as suitable here. Some approaches therefore employ instance selection, which would remove SVs not caused by deficiency of the gene of interest. However, these methods discard a large portion of data, might inadvertently discard the wrong instances, and are often not universally applicable[7]. To overcome this problem, we introduce an alternative MIL-based approach which we call MIL-BreakPoint (MIL-BP) (Fig 1B), that first learns whether each individual SV breakpoint belongs to a deficient sample. Then, the per-breakpoint probabilities are averaged per sample to obtain a final prediction of the biallelic loss status of a gene in that sample. In this way, the approach robustly considers features of all mutations individually.

One of the major challenges in DS classification is defining the classification labels. In the simplest setup, the positive class can be defined as all samples with biallelic loss of a gene as detectable from the WGS data, and a negative class containing all samples without evidence of gene deficiency. While seemingly straightforward, in reality, absence of biallelic LOF mutations does not guarantee true absence of biallelic loss in case one or both alleles are inactivated through a non-mutation pathway. In other words, the negative set may be contaminated with samples that in reality do present with gene deficiency signatures, and which should belong to the positive class.

Label uncertainty is a well-known problem in the machine learning field for which various solutions have been proposed. Here, we apply Positive Unlabeled (PU) learning,

a method combining semi-supervised learning with one-class classification (Fig 1C and 1D). Using this method, classifier is trained on the positive class versus all unlabeled data. The learned boundary is applied to assign labels to the unlabeled examples scoring above a pre-defined probability threshold. This process is repeated iteratively until all objects in the uncertain class are labeled. PU learning is highly suitable for our dataset, as it intrinsically models the presence of uncertain labels in the negative class.

A second, related, major hurdle is to evaluate the performance of the DS classifier. In absence of ground truth labels, it is never known if a model truly separates classes better when the negative class is contaminated. Since our aim is exactly to identify cases of positives that are incorrectly labeled as negatives, a higher false positive (FP) ratio may actually represent better performance. Therefore, commonly used performance metrics, such as precision, that rely on FP rates may misrepresent the actual ability of a classifier to identify biallelic loss.

In this work, we demonstrate the effects of this precision paradox in interpreting classifier performance on 3 genes known to result in clear SV signatures when deficient: CDK12, BRCA1 and BRCA2. We explore the simpleMI and MIL-BP models to solve the problem of representing features on the mutation-level rather than the sample-level and compare these to a version of CHORD modified to work specifically on SVs. To solve the problem of label uncertainty, we build one-class classifiers and PU learning classifiers on top of the 3 methods. Finally, we propose a different strategy to measure the ability of each classifier to correctly identify incorrectly labeled negatives as positive using a swap-one-patient-out CV (sopoCV) approach. Using this method, every positive patient is iteratively incorrectly labeled as negative, and we measure how well the classifier correctly reports these swaps as false positives.

We demonstrate on 4,069 high-depth (> 90X) whole cancer genomes from the Hartwig Medical Foundation (HMF)[8] that CHORD, simpleMI and MIL-BP predict biallelic loss better than by random chance in terms of AUCPR. PU learning is an effective approach to improve the model performance even further. Furthermore, it appears from these results that CHORD is the best model for predicting BRCA1 deficiency, whereas MIL-BP is most suitable for CDK12. SimpleMI is outperformed by both models on all genes but BRCA2. However, the sopoCV approach reveals that the differences between models are much less pronounced, and that CHORD and MIL-BP detect approximately equal numbers of false positives, with only a small benefit for MIL-BP on CDK12. Interestingly, no additional benefit of using PU learning for detecting false positives with sopoCV is found. In summary, our results demonstrate that it is not recommended to select classifiers for predicting biallelic loss on AUCPR in the presence of label noise, as the interpretation of this metric can be misleading.

Methods

Cancer datasets

SNV, indel, CNV and SV calls were obtained for 4,069 samples from 3,651 patients from the HMF. All variants were called using the HMF pipeline (https://github.com/hartwigmedical/pipeline), as described previously[3].

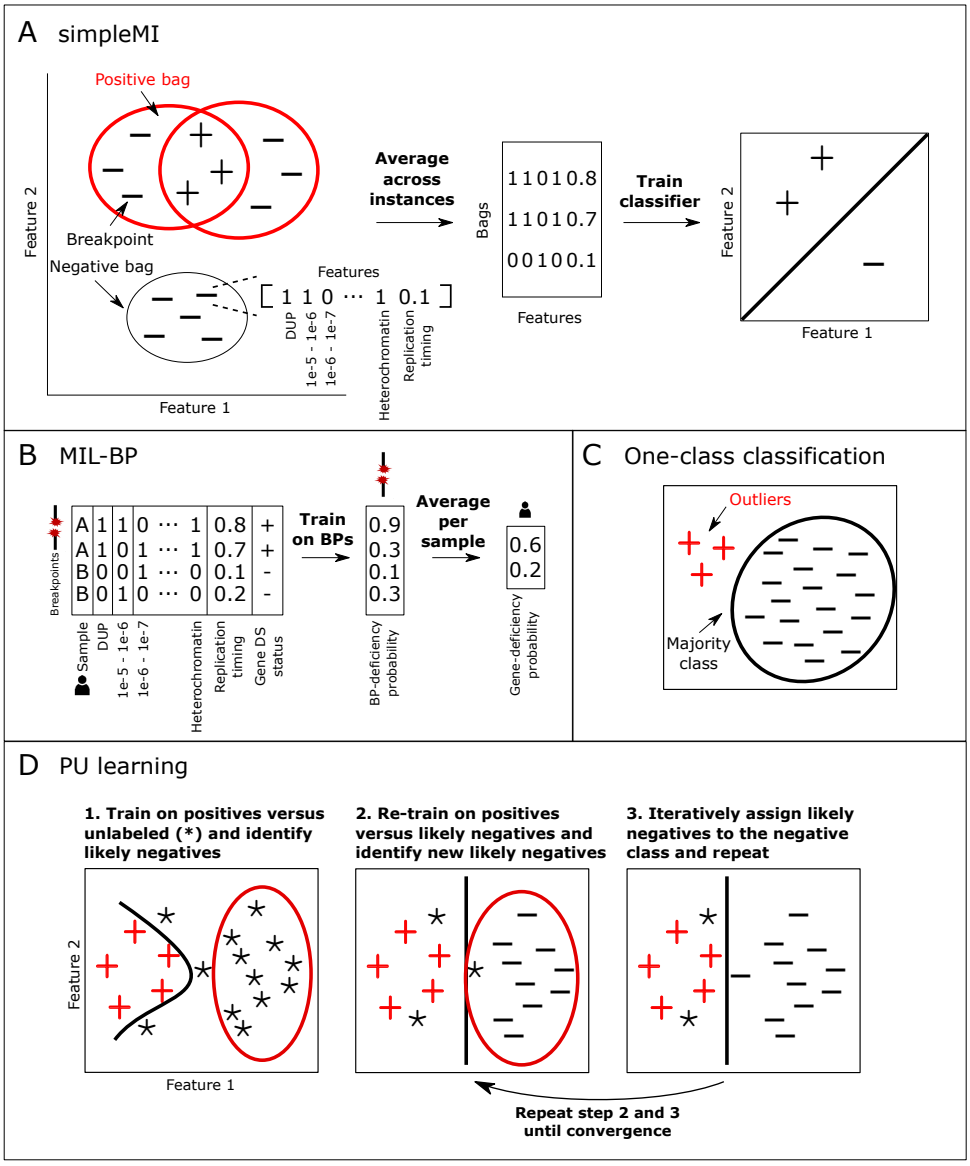


Fig. 1. Method overview. (A) With simpleMI, the MIL-space is converted to a feature space by averaging features across instances. Then, a classifier is trained in the feature space to classify bags (i.e. samples). (B) In MIL-BP, a classifier is first trained on the breakpoints individually. The resulting probabilities are averaged per sample to obtain a probability of the gene DS in that sample. (C) In one-class classification, outliers are detected by learning the distribution of the majority class. (D) PU learning iteratively learns most likely negatives within a set of unlabeled examples by determining a new boundary on the positives versus the identified likely negatives.

Machine learning-based methods to detect biallelic loss

Labels

Biallelic status was assessed using an in-house pipeline that considers copy-number as well as germline and somatic SNV/indel data to interpret biallelic gene status (https://github.com/UMCUGenetics/hmfGeneAnnotation). Pathogenicity was scored following ClinVar's (https://www.ncbi.nlm.nih.gov/clinvar/; GRCh37; database date 2020-02-24) ranking[9]: pathogenic, likely pathogenic, variant of unknown significance, likely benign and benign. A pathogenicity score (P-score) ranked variants from 1 (benign) to 5 (pathogenic). The process of determining gene status encompassed three steps. First, if a gene's copy number was < 0.3, it was considered as deep deletion, in which case both alleles were assigned a P-score of 5. Second, if a gene's copy number was ≥ 0.3, several mutation events were screened for. These included somatic and germline SNVs/indels as well as loss-of-heterozygosity (LOH), which was defined by a minor allele copy number of <0.2. P-scores of these events were then determined using ClinVar for SNVs/indels or assigned as high pathogenicity (P-score = 5) for LOH. SnpEff (http://snpeff.sourceforge.net/; v4.1 h) was used for SNVs/indels with no entry in ClinVar to estimate their pathogenicity. Out-of-frame frameshift were assigned a pathogenic score (P-score = 5), while splice and nonsense variants were considered likely pathogenic (P-score = 4). Missense variants, inframe frameshift and essential splice variants received a P-score of 3. Lastly, other variant types were scored as P-scores of ≤ 2. ClinVar's P-score for a variant was always considered over the P-score generated with SnpEff where applicable. Having computed P-scores per allele of a gene, the scores were summed up to a biallelic pathogenicity score (BP-score) of a maximum value of 10. For genes with multiple events the combination resulting in the highest BP-score was chosen and in cases of ties greedily selected.

Each sample was labeled as belonging to one of the three following classes. (i) If the sample's BP-score was ≥ 9 it was considered deficient. (ii) Else, proficiency was defined as lack of deep deletion and LOH, with all SNVs/indels having a P-score ≤ 3, and all combinations of SNVs/indels having a BP-score ≤ 6. (iii) If a sample did not fulfill either of the above criteria it was defined as having an 'uncertain' deficiency status and excluded from performance analyses.

Samples with fewer than 200 breakpoints in the HMF dataset were removed (n=1144) for all analyses.

Features

To use information about the regions in which SV breakpoints occurred, each breakpoint was annotated with a set of features. First, breakpoints were annotated with their respective SV type (i.e. duplication, deletion, inversion, or translocation), size (distance between the breakpoints of one SV) and local breakpoint homology. For translocations the size was annotated as 'NA'. For the remaining SV types, SV sizes were binned into bins of variable sizes. The smallest bin included all SVs < 1kb, the largest bin represented SVs > 10 Mb, and all other bins captured SVs in steps of powers of 10, e.g. 1 kb – 10 kb. Second, the breakpoints were annotated with information about the region in which they occurred. The region's gene density was computed as the number of protein-coding genes within 500 kb up- or downstream of a breakpoint. Data on chromatin states was

obtained through the Epigenomics Roadmap (https://egg2.wustl.edu/roadmap/web_portal/chr_state_learning.html), representing 15 chromatin states computed from 127 epigenomes. For each 2 kb bin, the most common chromatin state for this region was considered the consensus. Bins in which fewer than 50% of epigenomes showed a consensus of chromatin state were annotated as cell-type specific chromatin regions ('0_NA'). Replication timing as measured by ENCODE using repliSeq[10] for seven cell lines (GM12878, HeLa-S3, HepG2, HUVEC, IMR90, K562, MCF-7) were obtained through the UCSC table browser (http://genome.ucsc.edu/cgi- bin/hgTrackUi?db=hg19&g=wgEncodeUwRepliSeq). Per 1 kb bin, the mean replication timing was computed. Each breakpoint was then annotated with information about its region's gene density, replication timing, and chromatin state. Categorical features were one-hot encoded.

simpleMI

In the simpleMI model (Fig 1A), bags are constructed for each sample. All SV breakpoints are instances, each with a feature vector containing all features as described above. To obtain a final classifier, a feature space was constructed by taking the mean across all instances in a bag, resulting in one feature vector per bag. In this space, a random forest classifier (scikit-learn, v0.22.1, n_estimators: 100) was trained to predict the biallelic loss status of a sample.

MIL-BP

In the MIL-BP model, every breakpoint is first viewed individually, and a prediction is made whether it belongs to a deficient sample using a random forest classifier (scikit-learn, v0.22.1, n_estimators: 100). In a second step, the resulting probabilities are aggregated per sample by averaging them. The result is a prediction of a sample's biallelic loss status.

To lower computational complexity introduced by the large number of breakpoints, negative samples were iteratively added to the training data until a 1:10 ratio of breakpoints belonged to the positive and negative class, respectively. Subsampling was exclusive to MIL-BP, and is not applied to the other models.

CHORD

To compare performance of the MIL-based models to a previously published framework, the recent context-based approach CHORD was used[3]. It differs from the MIL-based approaches in that it aggregates breakpoints per sample into contexts, i.e. all possible combinations of features. This way, the frequency of breakpoints belonging to a context is computed per sample. Training and testing are then performed on the resulting sample-context matrix. However, with the addition of more features many more combinations are possible, resulting in a sparse sample-context matrix. This dimensionality problem therefore renders the context approach suboptimal when using many features. Therefore, contexts were generated using only the features SV type and size. The training procedure was consistent with the MIL methods as described above.

PU learning

To enable PU learning, the BaggingPuClassifier (n_estimators: 15) from the pulearn package (v0.0.7) was used on top of the random forest classifier used by simpleMI, CHORD and MIL-BP.

Measuring classifier performance

To evaluate the performance of the method, a cross-validation (CV) approach was adopted. In it, the data was split into five folds, stratified by the number of negative and positive samples within each fold. Within a fold, a sample can only be in either the training or the testing data. As described above, the training data was then class-balanced and used to train the classifier, followed by prediction on the held-out testing data. The classification cutoff for a sample being of deficient phenotype was defined as the cutoff at which the Matthew's Correlation Coefficient (MCC) was maximal. This cutoff was determined in the 5-fold CV.

Measuring classifier ability to correctly detect false positives: swap-one-patient-out CV

To assess how well each classifier identifies false positives, we apply a swap-one-patient-out CV (sopoCV). Every positive patient is iteratively swapped to the negative class, and the 5-fold CV is performed as described above to determine if the classifier correctly labels the swap as a false positive. For MIL-BP, all breakpoints of a sample are swapped. As running MIL-BP for BRCA2 in the sopoCV setting is not computationally feasible, the number of breakpoints was for this scenario randomly reduced to 25%. The optimal classifier cutoff was determined based on the maximum MCC in the 5-fold CV in a non-swapped run.

Results

The precision paradox complicates comparing classifiers by AUCPR

To demonstrate the ability of simpleMI and MIL-BP to predict biallelic loss status of CDK12, BRCA1 and BRCA2, we applied both models to 4069 samples from the HMF dataset and visualized the precision-recall curves alongside the performance of our modified version of CHORD (Fig 2). In general, each model reaches AUCPR higher than random chance. However, the most suitable model varies per gene. For CDK12, best performance is achieved by MIL-BP, whereas CHORD outperforms both simpleMI and MIL-BP on BRCA1. Part of this variation may be explained by the intrinsic (dis)advantages of each model. SimpleMI is robust to noise in individual breakpoints by averaging these, but is therefore more sensitive to outliers, resulting in an overall lower performance across the tested genes. CHORD similarly reduces noise by aggregating to a representation at sample-level, which benefits prediction for BRCA1. In contrast, the approach of MIL-BP to initially classify on the level of breakpoints is beneficial for genes that generate a lot of breakpoints as part of their loss phenotype, such as CDK12.

However, although the precision-recall curves are useful to estimate model performance under the assumption that the labels are correct, the presence of label noise makes it difficult to obtain a fair comparison between the models based on the AUCPR

alone. For example, simpleMI and MIL-BP report lower precision than CHORD for BRCA1, but this result may actually be good if the classifiers identify false positives that truly show a biallelic loss phenotype. Thus, a different strategy is needed to properly assess the ability of the models to detect biallelic loss in (independent) datasets.

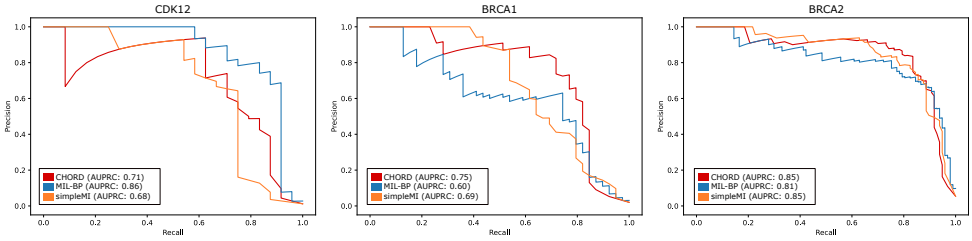


Fig. 2. Precision-recall curves of the 3 tested classifiers on CDK12, BRCA1 and BRCA2.

PU learning is an effective method to improve model performance based on AUCPR

To assess the performance of existing solutions for the label uncertainty problem, we swapped out the base random forest classifier of CHORD, simpleMI and MIL-BP for a bagging PU-learning classifier. From the precision-recall curves (Fig 3), we notice that the PU learning classifier, which naturally handles label impurity in the negative class, improves the AUCPR for all models except for simpleMI. However, in the absence of 'noiseless' ground truth labels, it remains difficult to determine if higher reported AUCPR indeed indicates a better ability to identify biallelic loss. Instead, the lower precision of the original models may reflect a higher detection of false positives that are truly deficient. Therefore, it is required to use a different performance metric than AUCPR that allows for a fair comparison between the models to be made.

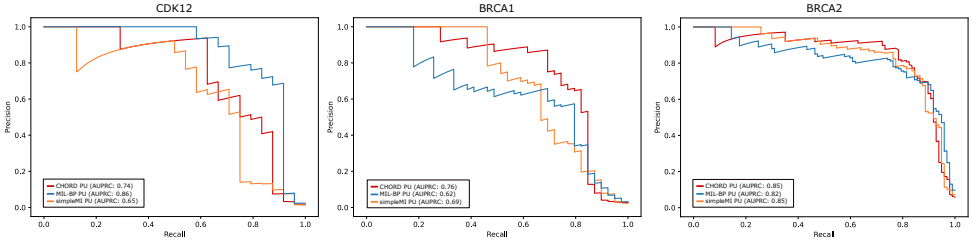


Fig. 3. Precision-recall curves of CHORD, simpleMI and MIL-BP on predicting biallelic CDK12, BRCA1 and BRCA2 loss when the default random forest classifier is combined with a bagging PU classifier.

Swap-one-patient-out CV: measuring the ability of classifiers to identify gene DS

Although PU learning appears to be an adequate solution to deal with uncertain labels in the negative set based on AUCPR (Fig 3), it remains difficult to validate if the model accurately detects non-genetic biallelic loss in the absence of ground truth labels. Therefore, we aimed to solve this problem by introducing a CV-based strategy to measure the ability of a classifier to detect false positives. Within this swap-one-patient-out CV (sopoCV)

setting, every positive sample was sequentially assigned to the negative class, and we counted how many samples each model correctly reported as false positive within the 5-fold CV as described previously. We combined sopoCV with simpleMI, MIL-BP, CHORD, and their PU learning-based implementations.

While it appeared from Fig 2 that each model performed best on a different gene, these differences are less pronounced for the ability of the classifiers to detect gene DS in the sopoCV (Fig 4). Although simpleMI performs worst overall, the performance of CHORD and MIL-BP are highly similar, with a small benefit for MIL-BP on CDK12. Furthermore, the difference in performance of MIL-BP on BRCA1 only differs minimally from CHORD, in contrast to the decrease measured with AUCPR. Notaby, PU learning improves performance in the sopoCV setting for almost all classifiers. Overall, these results show that MIL-BP is a highly efficient approach to detect gene DS.

As the number of breakpoints was reduced to 25% for MIL-BP and BRCA2 in the sopoCV setting due to computational limitations, it is uncertain if higher performance could be achieved if all breakpoints are present. However, as the ratio of correctly identified patients is on par with CHORD and simpleMI, it reveals that the information in the breakpoints is highly redundant for BRCA2.

In conclusion, we showed that models reaching higher AUCPR are not necessarily better at detecting the biallelic loss phenotype in the sopoCV. Thus, when selecting the best classifier to apply to independent data, sopoCV may be a more suitable metric than AUCPR.

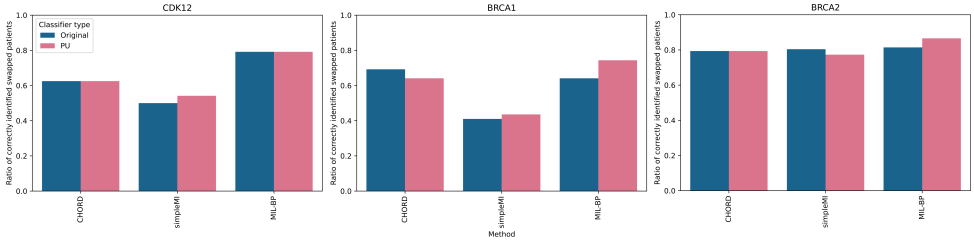


Fig. 4. Ratio of positive samples correctly identified as false positives by each model after swapping their labels to negative in sopoCV.

Discussion

With the increased availability of WGS cancer datasets, many options are opening up for machine learning models aiming to learn from the genetics of cancer genomes. Detecting gene DS is one such example, which could benefit the selection of suitable treatment for many cancer patients. We explored the advantage of using MIL-based strategies for gene DS classification and showed that MIL-BP performs better than the state-of-the-art on CDK12. However, gene deficiency acquired through non-genetic pathways cannot be detected from WGS data alone, resulting in potential label noise in the negative set. Therefore, a higher false positive rate may thus represent better ability to detect gene DS, and thus metrics such as AUCPR may be unreliable. To this end, we utilized PU learning to further improve our tested classifiers. We demonstrated that PU learning slightly

improves method performance in terms of AUCPR. However, since the noiseless ground truth labels are not known, it remains a challenge to interpret if the models are now more correct than their original counterparts.

To overcome this problem, we demonstrated through a sopoCV approach how well each classification model identifies positives as false positives if the label is artificially set to negative. Using this alternative way of measuring classifier performance, we note that the differences in model performance are less pronounced than was initially measured through AUCPR. Furthermore, we find that MIL-BP performs well in the sopoCV approach on all genes, and is thus a highly effective method for gene DS classification.

In conclusion, AUCPR needs to be interpreted with caution in datasets with noisy labels. Large efforts to gather patient-specific data, such as methylation, are therefore a great promise to obtain clean datasets for machine learning. However, we demonstrated that for now, sopoCV is an effective alternative method to measure classifier performance.

References

[1] Y.-M. Wu, M. Cieślik, R. J. Lonigro, P. Vats, M. A. Reimers, X. Cao, Y. Ning, L. Wang, L. P. Kunju, N. de Sarkar, E. I. Heath, J. Chou, F. Y. Feng, P. S. Nelson, J. S. de Bono, W. Zou, B. Montgomery, A. Alva, D. R. Robinson, and A. M. Chinnaiyan, *Inactivation of CDK12 Delineates a Distinct Immunogenic Class of Advanced Prostate Cancer,* Cell **173**, 1770 (2018).

[2] H. Davies, D. Glodzik, S. Morganella, L. R. Yates, J. Staaf, X. Zou, M. Ramakrishna, S. Martin, S. Boyault, A. M. Sieuwerts, P. T. Simpson, T. A. King, K. Raine, J. E. Eyfjord, G. Kong, Å. Borg, E. Birney, H. G. Stunnenberg, M. J. van de Vijver, A.-L. Børresen-Dale, J. W. M. Martens, P. N. Span, S. R. Lakhani, A. Vincent-Salomon, C. Sotiriou, A. Tutt, A. M. Thompson, S. Van Laere, A. L. Richardson, A. Viari, P. J. Campbell, M. R. Stratton, and S. Nik-Zainal, *HRDetect is a predictor of BRCA1 and BRCA2 deficiency based on mutational signatures,* Nature Medicine **23**, 517 (2017).

[3] L. Nguyen, J. W. M. Martens, A. Van Hoeck, and E. Cuppen, *Pan-cancer landscape of homologous recombination deficiency,* Nature Communications **11**, 5584 (2020).

[4] Li *et al.*, *Patterns of somatic structural variation in human cancer genomes,* Nature **578**, 112 (2020).

[5] A. Zafra and S. Ventura, *G3P-MI: A genetic programming algorithm for multiple instance learning,* Information Sciences **180**, 4496 (2010).

[6] J. Foulds and E. Frank, *A review of multi-instance learning assumptions,* The Knowledge Engineering Review **25**, 1 (2010).

[7] J. A. Olvera-López, J. A. Carrasco-Ochoa, J. F. Martínez-Trinidad, and J. Kittler, *A review of instance selection methods,* Artificial Intelligence Review **34**, 133 (2010).

[8] P. Priestley, J. Baber, M. P. Lolkema, N. Steeghs, E. de Bruijn, C. Shale, K. Duyvesteyn, S. Haidari, A. van Hoeck, W. Onstenk, P. Roepman, M. Voda, H. J. Bloemendal,

V. C. G. Tjan-Heijnen, C. M. L. van Herpen, M. Labots, P. O. Witteveen, E. F. Smit, S. Sleijfer, E. E. Voest, and E. Cuppen, *Pan-cancer whole-genome analyses of metastatic solid tumours,* Nature **575**, 210 (2019).

[9] M. J. Landrum, S. Chitipiralla, G. R. Brown, C. Chen, B. Gu, J. Hart, D. Hoffman, W. Jang, K. Kaur, C. Liu, V. Lyoshin, Z. Maddipatla, R. Maiti, J. Mitchell, N. O'Leary, G. R. Riley, W. Shi, G. Zhou, V. Schneider, D. Maglott, J. B. Holmes, and B. L. Kattman, *ClinVar: improvements to accessing data,* Nucleic Acids Research **48**, D835 (2020).

[10] C. Marchal, T. Sasaki, D. Vera, K. Wilson, J. Sima, J. C. Rivera-Mulia, C. Trevilla-García, C. Nogues, E. Nafie, and D. M. Gilbert, *Genome-wide analysis of replication timing by next-generation sequencing with E/L Repli-seq,* Nature Protocols **13**, 819 (2018).

6

Discussion

Summary

In this thesis, our aim was to create a better understanding of cancer development. To fulfill this goal, we focused on 3 main areas. First, **chapter 2** explored how mutations accumulate in cancer over time by reconstructing phylogenetic trees directly from microdissected samples with reduced heterogeneity. Our method contrasts the typical approach where heterogeneous samples need to be deconvolved first using complicated models, and is thus useful to reduce an additional layer of noise. Second, the multiple instance learning (MIL)-based models introduced in **chapter 3** and **chapter 4** were used to demonstrate that non-coding structural variants (SVs) have the capability to drive cancer by disrupting 3D genome structures, which was until now relatively poorly understood and thus has important implications for cancer diagnostics. Third, **chapter 5** introduced machine learning techniques to identify gene deficiency status in presence of label noise. We introduced a metric that measures how well the classifiers perform without relying on AUCPR, which may be biased if measured in datasets with uncertain labels. Not only does a better classification of patients benefit selection of optimal treatment, but it also aids in further research into commonalities between groups of cancer patients.

Although this work has provided new insights into cancer development, many questions are still unanswered, and open challenges remain.

What do we need to do to complete our understanding of cancer evolution?

Knowing which subclones are present in a tumor is essential to determine personalized and effective treatments. The field of studying cancer evolution has grown rapidly over the past decade, but to date a lot of details of the phylogenetic relations between subclones are still missed by the state-of-the-art deconvolution approaches[1]. To circumvent the need for deconvolution, our method TargetClone (as discussed in **chapter 2**) reduced heterogeneity by sampling from multiple sites in the tumor. However, our approach relied on obtaining relatively homogeneous samples, which is in practice difficult to achieve[2]. In the past few years, deconvolution techniques applied on heterogeneous multi-region sequencing datasets have become increasingly popular[3]. However, despite the continuous increase in performance, a number of challenges still remain. For example, if subclones do not overlap between samples from different regions, it remains difficult to detect rare subclones due to low read depth and technical errors[4]. Alternative approaches, such as REVOLVER[5], HINTRA[6] or RECAP[7], that aim to construct consensus trees across multiple patients in a cohort also face the same challenges. If subclones are not detected, treatments may fail, or the cancer may re-grow from subclones that were not eradicated by the therapy[8].

The only currently existing approach to fully overcome the need for deconvolution is single-cell sequencing. Although the technique sounds like an ideal solution, a lot of hurdles need yet to be overcome. Most single-cell-based phylogeny reconstruction methods were designed to handle common issues such as allelic dropout and coverage biases due to low input material[9–11]. However, the most problematic issue may be the tradeoff between sequencing depth and the number of cells sequenced[1]. In a tumor

containing billions of cells, the chance of missing important subclones becomes larger with a lower number of sequenced cells. However, increasing the number of cells will result in a lower read depth, making it harder to characterize all subclones due to increased noise and technical errors[1]. If these challenges can be overcome, single-cell sequencing techniques may very well be the future standard in reconstructing clonal evolution trees in cancer.

Finally, although single-cell sequencing can help us understand cancer evolution on a genetic level, considering only the DNA may not be sufficient to select proper anti-cancer therapies for patients. Despite having the same genetic code, subclones may vary from each other on other -omics levels to confer evolutionary advantage, which could potentially lead to unexpected treatment response if not accounted for[12]. For example, recent studies have successfully integrated single-cell RNA sequencing data with spatial information to obtain a more accurate overview of the tumor microenvironment[13, 14]. It would be ideal if more -omics data could be included in future models, but also present a large effort to obtain, which this chapter will further elaborate on in the next 2 sections.

Completing the cancer driver catalogue

New machine learning approaches to predict mutation pathogenicity

Over the recent years, a lot of prediction tools have been developed to study the pathogenicity of mutations across many patients and cancer types[15–22]. However, a lot of these methods remain limited to investigating coding mutations, while the importance of non-coding mutations is becoming increasingly clear[23]. Although interpreting the effects of non-coding mutations is complicated by the large number of regulatory functions that these can disrupt, detailed research into every single one of them is needed to gain a complete understanding of how we should treat cancer. Methods like DeepSEA[21] and ExPecto[22] have solved this problem for single-nucleotide variants (SNVs) by training deep learning models to recognize which genomic features, such as histone modifications, are characteristic of pathogenic SNVs. In **chapter 3** and **chapter 4**, we used a similar approach by training a MIL model to learn genomic features characterizing pathogenic non-coding SVs that disrupt boundaries of topologically associated domains (TADs) and chromatin loops. An important benefit to this bag-based approach over traditional machine learning and deep learning is that each bag can hold information about any disruptions to the genome across a very large range rather than just the direct surroundings of an SV. However, as our model focuses on 3D structures, which comprise only a small part of genome regulation, more studies into the role of non-coding SVs in disrupting other regulatory functions are required. Within our MIL model, this could potentially be achieved by updating the feature vectors to be less specific for TADs and chromatin loops, and adding features that specify direct disruption of regulatory elements. A potential alternative approach to solve this problem could be a deep learning model that learns the characteristic genomic features of regions around SV breakpoints, similar to how SVs were annotated in **chapter 5**. The predictions of this deep learning model and the MIL model could eventually be integrated based on model probabilities to clear up any uncertainty about whether a non-coding SV drives cancer through disrupting the 3D structure or a regulatory element it directly overlaps with, for example.

However, a few important challenges remain to be addressed before we can fully utilize the power of machine learning approaches in this context.

Increasing sample size can further improve pathogenicity prediction
Our research described in **chapter 4** indicated that non-coding SVs target known cancer drivers in many cancer types. Furthermore, although the mechanisms by which non-coding SVs disrupt gene regulation appear to be similar across cancer types, the overall contribution of driver non-coding SVs compared to driver SNVs varies greatly between cancer types. However, our study was limited to a small sample size for most cancer types and may thus not perfectly recapitulate the true impact of non-coding SVs. Large-scale efforts such as the Hartwig Medical Foundation (HMF)[24] and Pan-Cancer Analysis of Whole genomes (PCAWG)[23] are already generating many high-quality cancer datasets. As the number of samples in these consortia continues to grow, our ability to study the role of non-coding SVs in cancer will improve further.

Establishing a reference pathogenicity database for non-coding SVs to use as labels
The increase in knowledge about mutation pathogenicity fueled the construction of large, publicly available databases such as the Cancer Gene Census (CGC)[25] and ClinVar[26] that allow central access of the impact of somatic mutations. These datasets have allowed the application of machine learning methods to learn the characteristics of pathogenic mutations in comparison to non-pathogenic mutations. Such methods are extremely useful to elucidate recurring patterns in large amounts of cancer data. A particular strength of machine learning is to learn combinations of features that characterize pathogenic mutations specifically, which are often hard to determine otherwise. As the role of non-coding SVs in cancer is becoming increasingly clear, the need for a good reference database listing the pathogenicity of these is growing. Although databases such as ClinVar contain clinical interpretation of some SVs, the number of characterized SVs is small by machine learning standards, and the focus remains on SVs in coding regions. In **chapter 3** and **chapter 4**, we overcame the lack of a good reference database by using expression data to label pairs of SVs and affected genes to use in machine learning. Large consortia such as the HMF and PCAWG are already measuring expression data for the majority of patients in their datasets. However, as these data are costly, paired WGS and gene expression data may not be available for all cancer samples from other sources and may not always be generated retrospectively. Therefore, the construction of a solid reference database can be a good intermediate step to further elucidate the role of non-coding SVs in cancer.

Switching from labels based on pathogenic mutations to pathogenic genes
Although mutations are a clear signal for identifying genes that can drive cancer, these are not necessarily the only information we can use in the hunt for undiscovered drivers. While our ability to correctly determine the presence of mutations has increased significantly over the years, mutation calls are not yet perfect. Biases such as lack of material, low read depth, sequencing errors, errors made by mutation callers, and even human errors, can result in missing important mutations[27]. This is especially relevant for SVs, where short-read sequencing is still commonly applied[28]. Due to the short length of the reads, it is often impossible to detect SVs in all regions of the genome, such as repeat

regions[29]. While long-read sequencing, such as nanopore sequencing, is gaining a lot of traction, a lot of existing data will not be re-generated using new techniques. Furthermore, even for the mutations that we do detect correctly, it sometimes remains unclear if the mutation has any functional effect. As we addressed in **chapter 5**, driver genes themselves do not always have pathogenic mutations, but may instead be deregulated through upstream pathway effects. Therefore, it is important that we do not focus all of our efforts into predicting the pathogenicity of mutations. Instead, there is still a lot to gain in combining multi-omics data to identify the driver genes, rather than just the driver mutations.

Integrating multi-omics data to clear up label noise

Machine learning models rely on correct labels to make accurate predictions. In reality, as was demonstrated in **chapter 3** and was a main focus of **chapter 5**, labels are often extremely noisy and can lead to decreased performance. In **chapter 3**, the lack of patient-specific data made it difficult to determine with full confidence if genes are affected by non-coding SVs, or by unmeasured effects, such as methylation. In addition, despite filtering out genes affected by coding mutations, we could not account for effects where gene expression could be altered due to mutations in upstream pathway partners. Although recent novel methods have shown that integrating multi-omics layers with gene and protein interaction can improve driver gene prediction, such methods have not yet taken non-coding SVs into account[19, 30, 31]. While combining non-coding SVs into existing frameworks provides a promising direction for further research, some hurdles would need to be overcome to enable this integration. For example, the recurrence of non-coding SVs appears limited compared to SNVs (see **chapter 4**) and CNVs[23]. Thus, the (indirect) effect of non-coding SVs on genes may be crowded out by other mutations if low frequency is not properly accounted for in the statistical models. In **chapter 5**, we addressed the problems in measuring classifier performance using typical methods including precision, which may misrepresent actual performance if false positives are expected due to high levels of label uncertainty in the negative set. To overcome this problem, model-based strategies such as one-class classification and semi-supervised learning have been previously introduced[32, 33]. We demonstrated that semi-supervised learning approaches, in particular PU learning, are good options to measure performance in presence of label noise. Additionally, we introduced a swap-one-patient-out CV approach to measure performance without relying on precision. Yet, it remains difficult to validate the performance of any solution in absence of noiseless ground truth labels, which is often the case in cancer datasets. Therefore, reducing label noise remains a very important topic. In many cases, a logical step forward would be to clear uncertainty in the labels by incorporating data from more –omics layers measured within each patient, and the potential problems with this approach which will be discussed further in the next section.

Validating predicted pathogenicity in model systems to improve label certainty

After prediction tools have been used to obtain a ranked list of the most likely pathogenic mutations, it is essential to validate their ability to cause cancer in model systems. This is especially true for non-coding SVs, for which validation studies are sparse. The pathogenicity of germline TAD-breaking SVs has previously been validated in mice, but this study

focused on a single locus in the genome[34]. As for the somatic case, the reality is that up to hundreds of potentially interesting loci exist, rendering testing each of these individually an extremely laborious and time-consuming task. Even though our machine learning approach described in **chapter 3** and **chapter 4** has been able to pinpoint a smaller set of candidate driver SVs disrupting TADs and CTCF loops, the number remains in the thousands across cancer types. Furthermore, as (non-coding) SVs may disrupt more regulatory elements than just the 3D structure, the list of potential driver SVs may be even larger. This problem underscores the need for multiple independent studies generating pathogenicity scores for non-coding SVs. If different strategies often report the same driver mutations, it creates a stronger ground for lab-based validations. Furthermore, improving driver prediction methods by incorporating more patient-specific data, for example to be used as features, could help remove more false positives.

Improving features to improve pathogenicity prediction

Currently, it is often difficult to obtain regulatory data specific for every tissue. Filling in these missing pieces may greatly benefit model design and performance, but also highlights potential issues. As was demonstrated in **chapter 4**, training cancer-specific classifiers using as features regulatory information from a different tissue type than the tissue of origin does often not significantly reduce performance. While it is well-known that overlap exists between regulatory marks across cell types[35] and thus intrinsically increases model robustness, there also exist tissue types that unexpectedly increase performance. These findings may partially be explained by our dataset consisting of metastasis samples, which may no longer necessarily represent the tissue of origin well. In addition, as regulatory data is often measured in healthy tissue, these may also not be truly representative of primary cancers. As such, it is hard to determine the optimal regulatory dataset to use to train models without performing an exhaustive search across available tissue types. Therefore, rather than only generating data for reference tissues, there may be a benefit to measuring patient-specific regulatory data. With the availability of such datasets, it will become a lot easier to construct reliable models. However, generating and processing such huge amounts of patient data would be a gigantic effort, which may not be computationally feasible in the way our systems are currently set up. So what do we need to do before we can even start thinking about such enrichments?

Computational challenges in processing huge amounts of cancer data

Unification and standardization of data from many different sources

With the collection of more (patient-specific) cancer data, the next challenge becomes the integration of all these data from different sources, which are often provided in different data formats. Even though several file formats are recognized as standards in the field, such as FASTA, BAM or GFF files, not every format is equally well-defined. Therefore, although big consortia such as HMF and PCAWG are putting huge efforts into processing all of their data uniformly and providing everything in the same formats, minor differences can often be found in files between sources. Although file versions account for a large part of these differences, these can also occur due to choices made by the data

providers themselves. A representative format for this issue is the VCF format, which is commonly used to store mutation calls. Although the first 8 columns of a VCF file are standardized and well-defined, the INFO field is designed to hold any possible data. Therefore, integrating data from different consortia may be difficult if both sources made a choice to store the relevant data under a slightly different identifier. The same problem occurs with gene identifiers, for which a choice can be made from ENSEMBL, Gene Symbols or Entrez IDs, among others. Despite the existence of tools to map these different identifiers together, problems may occur if data was generated with different reference builds, and if an identifier does not exist in one of the versions.

Many of these integration tasks are now performed by individual researchers by hand, which is a very time-consuming task prone to mistakes. The existence of standards would definitely help in saving time and improving on the reproducibility issue in science. The largest attempt to date to standardize has been the development of ontologies, which aim to form a non-redundant, rich description of all possible entities in (biological) data[36]. Such uniform data descriptions have allowed the usage of existing linked data techniques to query these data altogether without having to perform any integration (e.g. Bio2RDF)[37]. However, despite the promising applications linked data opens up, the majority of newly generated data is not yet suited for these purposes. Furthermore, a number of important questions remain to be answered before ontologies can be widely applied. Who is responsible for defining ontologies, especially when existing vocabularies are not suited for your specific type of data? Who will ensure that newly-introduced ontologies are non-redundant? Furthermore, how will we go about retrospectively converting pre-existing data that is often non-standardized and sometimes very specific to the needs of a certain project? The difficulty of introducing a standard that fits every researcher's needs may be one of the main reasons why these were not developed as soon as the first large-scale data generation projects were started.

Developing smart tools to make large amounts of data accessible to researchers

After successful data integration remains the step to make the data, or findings, accessible to other researchers. A lot of results of individual studies often remain scattered across different sources and publications. However, linked data and ontologies, as discussed previously, would allow researchers to easily query many of these findings at once. Such databases, in a way similar to Google, would make it a lot easier to prioritize findings that a researcher may be looking for, without having to search through (and manually combine) the supplementary material of many different publications. Although platforms such as canSAR[38] have already made great efforts towards facilitating research and drug discovery by combining various data sources, it remains a challenge to integrate private research data with the majority of such systems. Therefore, they are often not used by researchers who need to generate results to their specific needs. With the availability and adherence to standards such as ontologies, it would become a lot easier to link these data and findings together automatically and provide platforms that allow straightforward data mining and knowledge generation across the world.

Conclusions and brief outlook

As more and more cancer data are being generated, options are opening up to learn more about the characteristics of cancer. The availability of larger datasets will not only enable many existing driver prediction models to pick up new mutations that did previously not meet frequency thresholds, but also allow the development of novel tools that are specifically designed to handle large amounts of (multi-omics) data. As our understanding of the non-coding genome continues to grow, like we showed in this thesis, a lot of potential is opened up for models that are specifically designed to prioritize non-coding mutations. However, every new possibility also comes with many new challenges. Although large steps are already being made in processing and integrating large amounts of multi-omics data, we are still far away from having solid standards that can be applied by scientists in the field. Along the way, a lot of new findings will be generated, which stresses the importance of also not overlooking solid validation of candidate mutations and genes in the lab. Only that way will we be able to achieve personalized medicine, and design optimal treatment programs for every patient.

References

[1] M. J. Williams, A. Sottoriva, and T. A. Graham, *Measuring Clonal Evolution in Cancer with Genomics,* Annual Review of Genomics and Human Genetics **20**, 309 (2019).

[2] C. Bevilacqua and B. Ducos, *Laser microdissection: A powerful tool for genomics at cell level,* Molecular Aspects of Medicine **59**, 5 (2018).

[3] D. Ramazzotti, A. Graudenzi, L. De Sano, M. Antoniotti, and G. Caravagna, *Learning mutational graphs of individual tumour evolution from single-cell and multi-region sequencing data,* BMC Bioinformatics **20**, 210 (2019).

[4] C. Wang, J. Yang, H. Luo, K. Wang, Y. Wang, Z.-X. Xiao, X. Tao, H. Jiang, and H. Cai, *CancerTracer: a curated database for intrapatient tumor heterogeneity,* Nucleic Acids Research (2019), 10.1093/nar/gkz1061.

[5] G. Caravagna, Y. Giarratano, D. Ramazzotti, I. Tomlinson, T. A. Graham, G. Sanguinetti, and A. Sottoriva, *Detecting repeated cancer evolution from multi-region tumor sequencing data,* Nature Methods **15**, 707 (2018).

[6] S. Khakabimamaghani, S. Malikic, J. Tang, D. Ding, R. Morin, L. Chindelevitch, and M. Ester, *Collaborative intra-tumor heterogeneity detection,* Bioinformatics **35**, i379 (2019).

[7] S. Christensen, J. Kim, N. Chia, O. Koyejo, and M. El-Kebir, *Detecting evolutionary patterns of cancers using consensus trees,* Bioinformatics **36**, i684 (2020).

[8] F. Janku, *Tumor heterogeneity in the clinic: is it a real problem?* Therapeutic Advances in Medical Oncology **6**, 43 (2014).

[9] J. Singer, J. Kuipers, K. Jahn, and N. Beerenwinkel, *Single-cell mutation identification via phylogenetic inference,* Nature Communications **9**, 5144 (2018).

[10] E. M. Ross and F. Markowetz, *OncoNEM: inferring tumor evolution from single-cell sequencing data,* Genome Biology **17**, 69 (2016).

[11] D. J. McCarthy, R. Rostom, Y. Huang, D. J. Kunz, P. Danecek, M. J. Bonder, T. Hagai, R. Lyu, W. Wang, D. J. Gaffney, B. D. Simons, O. Stegle, and S. A. Teichmann, *Cardelino: computational integration of somatic clonal substructure and single-cell transcriptomes,* Nature Methods **17**, 414 (2020).

[12] N. E. Navin, *Cancer genomics: one cell at a time,* Genome Biology **15**, 452 (2014).

[13] A. Saviano, N. C. Henderson, and T. F. Baumert, *Single-cell genomics and spatial transcriptomics: Discovery of novel cell states and cellular interactions in liver physiology and disease biology,* Journal of Hepatology **73**, 1219 (2020).

[14] M. L. Suvà and I. Tirosh, *Single-Cell RNA Sequencing in Cancer: Lessons Learned and Emerging Challenges,* Molecular Cell **75**, 7 (2019).

[15] M. S. Lawrence, P. Stojanov, P. Polak, G. V. Kryukov, K. Cibulskis, A. Sivachenko, S. L. Carter, C. Stewart, C. H. Mermel, S. a. Roberts, A. Kiezun, P. S. Hammerman, A. McKenna, Y. Drier, L. Zou, A. H. Ramos, T. J. Pugh, N. Stransky, E. Helman, J. Kim, C. Sougnez, L. Ambrogio, E. Nickerson, E. Shefler, M. L. Cortés, D. Auclair, G. Saksena, D. Voet, M. Noble, D. DiCara, P. Lin, L. Lichtenstein, D. I. Heiman, T. Fennell, M. Imielinski, B. Hernandez, E. Hodis, S. Baca, A. M. Dulak, J. Lohr, D.-A. Landau, C. J. Wu, J. Melendez-Zajgla, A. Hidalgo-Miranda, A. Koren, S. a. McCarroll, J. Mora, R. S. Lee, B. Crompton, R. Onofrio, M. Parkin, W. Winckler, K. Ardlie, S. B. Gabriel, C. W. M. Roberts, J. a. Biegel, K. Stegmaier, A. J. Bass, L. a. Garraway, M. Meyerson, T. R. Golub, D. a. Gordenin, S. Sunyaev, E. S. Lander, and G. Getz, *Mutational heterogeneity in cancer and the search for new cancer-associated genes.* Nature **499**, 214 (2013).

[16] D. Tamborero, A. Gonzalez-Perez, and N. Lopez-Bigas, *OncodriveCLUST: exploiting the positional clustering of somatic mutations to identify cancer genes,* Bioinformatics **29**, 2238 (2013).

[17] P. Rentzsch, D. Witten, G. M. Cooper, J. Shendure, and M. Kircher, *CADD: predicting the deleteriousness of variants throughout the human genome,* Nucleic Acids Research **47**, D886 (2019).

[18] M. F. Rogers, H. A. Shihab, M. Mort, D. N. Cooper, T. R. Gaunt, and C. Campbell, *FATHMM-XF: accurate prediction of pathogenic point mutations via extended features,* Bioinformatics **34**, 511 (2018).

[19] A. Bashashati, G. Haffari, J. Ding, G. Ha, K. Lui, J. Rosner, D. G. Huntsman, C. Caldas, S. A. Aparicio, and S. P. Shah, *DriverNet: uncovering the impact of somatic driver mutations on transcriptional networks in cancer,* Genome Biology **13**, R124 (2012).

[20] M. A. Reyna, M. D. M. Leiserson, and B. J. Raphael, *Hierarchical HotNet: identifying hierarchies of altered subnetworks,* Bioinformatics **34**, i972 (2018).

[21] J. Zhou and O. G. Troyanskaya, *Predicting effects of noncoding variants with deep learning–based sequence model,* Nature Methods **12**, 931 (2015).

[22] J. Zhou, C. L. Theesfeld, K. Yao, K. M. Chen, A. K. Wong, and O. G. Troyanskaya, *Deep learning sequence-based ab initio prediction of variant effects on expression and disease risk,* Nature Genetics **50**, 1171 (2018).

[23] E. Rheinbay, M. M. Nielsen, F. Abascal, J. A. Wala, O. Shapira, G. Tiao, H. Hornshøj, J. M. Hess, R. I. Juul, Z. Lin, L. Feuerbach, R. Sabarinathan, T. Madsen, J. Kim, L. Mularoni, S. Shuai, A. Lanzós, C. Herrmann, Y. E. Maruvka, C. Shen, S. B. Amin, P. Bandopadhayay, J. Bertl, K. A. Boroevich, J. Busanovich, J. Carlevaro-Fita, D. Chakravarty, C. W. Y. Chan, D. Craft, P. Dhingra, K. Diamanti, N. A. Fonseca, A. Gonzalez-Perez, Q. Guo, M. P. Hamilton, N. J. Haradhvala, C. Hong, K. Isaev, T. A. Johnson, M. Juul, A. Kahles, A. Kahraman, Y. Kim, J. Komorowski, K. Kumar, S. Kumar, D. Lee, K.-V. Lehmann, Y. Li, E. M. Liu, L. Lochovsky, K. Park, O. Pich, N. D. Roberts, G. Saksena, S. E. Schumacher, N. Sidiropoulos, L. Sieverling, N. Sinnott-Armstrong, C. Stewart, D. Tamborero, J. M. C. Tubio, H. M. Umer, L. Uuskü-la-Reimand, C. Wadelius, L. Wadi, X. Yao, C.-Z. Zhang, J. Zhang, J. E. Haber, A. Hobolth, M. Imielinski, M. Kellis, M. S. Lawrence, C. von Mering, H. Nakagawa, B. J. Raphael, M. A. Rubin, C. Sander, L. D. Stein, J. M. Stuart, T. Tsunoda, D. A. Wheeler, R. Johnson, J. Reimand, M. Gerstein, E. Khurana, P. J. Campbell, N. López-Bigas, J. Weischenfeldt, R. Beroukhim, I. Martincorena, J. S. Pedersen, and G. Getz, *Analyses of non-coding somatic drivers in 2,658 cancer whole genomes,* Nature **578**, 102 (2020).

[24] P. Priestley, J. Baber, M. P. Lolkema, N. Steeghs, E. de Bruijn, C. Shale, K. Duyvesteyn, S. Haidari, A. van Hoeck, W. Onstenk, P. Roepman, M. Voda, H. J. Bloemendal, V. C. G. Tjan-Heijnen, C. M. L. van Herpen, M. Labots, P. O. Witteveen, E. F. Smit, S. Sleijfer, E. E. Voest, and E. Cuppen, *Pan-cancer whole-genome analyses of metastatic solid tumours,* Nature **575**, 210 (2019).

[25] Z. Sondka, S. Bamford, C. G. Cole, S. A. Ward, I. Dunham, and S. A. Forbes, *The COSMIC Cancer Gene Census: describing genetic dysfunction across all human cancers,* Nature Reviews Cancer **18**, 696 (2018).

[26] M. J. Landrum, J. M. Lee, M. Benson, G. Brown, C. Chao, S. Chitipiralla, B. Gu, J. Hart, D. Hoffman, J. Hoover, W. Jang, K. Katz, M. Ovetsky, G. Riley, A. Sethi, R. Tully, R. Villamarin-Salomon, W. Rubinstein, and D. R. Maglott, *ClinVar: public archive of interpretations of clinically relevant variants,* Nucleic Acids Research **44**, D862 (2016).

[27] C. Xu, *A review of somatic single nucleotide variant calling algorithms for next-generation sequencing data,* Computational and Structural Biotechnology Journal **16**, 15 (2018).

[28] D. L. Cameron, L. Di Stefano, and A. T. Papenfuss, *Comprehensive evaluation and characterisation of short read general-purpose structural variant calling software,* Nature Communications **10**, 3240 (2019).

[29] M. Mahmoud, N. Gobet, D. I. Cruz-Dávalos, N. Mounier, C. Dessimoz, and F. J. Sedlazeck, *Structural variant calling: the long and the short of it,* Genome Biology **20**, 246 (2019).

[30] J. P. Hou and J. Ma, *DawnRank: discovering personalized driver genes in cancer,* Genome Medicine **6**, 56 (2014).

[31] D. Bertrand, K. R. Chng, F. G. Sherbaf, A. Kiesel, B. K. H. Chia, Y. Y. Sia, S. K. Huang, D. S. Hoon, E. T. Liu, A. Hillmer, and N. Nagarajan, *Patient-specific driver gene prediction and risk assessment through integrated network analysis of cancer omics profiles,* Nucleic Acids Research **43**, e44 (2015).

[32] J. Kaufmann, K. Asalone, R. Corizzo, C. Saldanha, J. Bracht, and N. Japkowicz, *One-Class Ensembles for Rare Genomic Sequences Identification,* (2020) pp. 340–354.

[33] C. Yin and Z. Chen, *Developing Sustainable Classification of Diseases via Deep Learning and Semi-Supervised Learning,* Healthcare **8**, 291 (2020).

[34] D. G. Lupiáñez, K. Kraft, V. Heinrich, P. Krawitz, F. Brancati, E. Klopocki, D. Horn, H. Kayserili, J. M. Opitz, R. Laxova, F. Santos-Simarro, B. Gilbert-Dussardier, L. Wittler, M. Borschiwer, S. A. Haas, M. Osterwalder, M. Franke, B. Timmermann, J. Hecht, M. Spielmann, A. Visel, and S. Mundlos, *Disruptions of Topological Chromatin Domains Cause Pathogenic Rewiring of Gene-Enhancer Interactions,* Cell **161**, 1012 (2015).

[35] J. Schreiber, T. Durham, J. Bilmes, and W. S. Noble, *Avocado: a multi-scale deep tensor factorization method learns a latent representation of the human epigenome,* Genome Biology **21**, 81 (2020).

[36] M. Salvadores, P. R. Alexander, M. A. Musen, and N. F. Noy, *BioPortal as a dataset of linked biomedical ontologies and terminologies in RDF,* Semantic Web **4**, 277 (2013).

[37] F. Belleau, M.-A. Nolin, N. Tourigny, P. Rigault, and J. Morissette, *Bio2RDF: Towards a mashup to build bioinformatics knowledge systems,* Journal of Biomedical Informatics **41**, 706 (2008).

[38] E. A. Coker, C. Mitsopoulos, J. E. Tym, A. Komianou, C. Kannas, P. Di Micco, E. Villasclaras Fernandez, B. Ozer, A. A. Antolin, P. Workman, and B. Al-Lazikani, *canSAR: update to the cancer translational research and drug discovery knowledgebase,* Nucleic Acids Research **47**, D917 (2019).

Summary

Despite years of research, our understanding of cancer remains incomplete to optimally treat the disease. With the increasing availability of whole-genome sequencing (WGS) cancer datasets, our knowledge on the role of mutations in cancer development has grown. From these data, it became clear that each patient contains on average 4-5 mutations that drive cancer growth, in contrast to up to thousands of passenger mutations. The increase of the number of computational tools that aim to identify the cancer drivers in patients has resulted in an ever-growing list of clinically actionable mutations, leading to improved treatments for cancer patients. Yet, patients remain for whom no driver mutations can be identified, or who do not respond well to existing treatments. In this thesis, we aimed to contribute to finding solutions to 3 challenges that currently need to be addressed to gain a better understanding of cancer.

In **chapter 2**, we explore the challenge of tumor heterogeneity. Rather than being a mass containing 1 type of cell, tumors usually present as a mixture of different cells, called subclones. However, it is currently not typical to sample all cells in the tumor individually, and instead samples consist of a heterogeneous bulk of different cells. As a result, any mutations called from sequencing data will often be an average across subclones in the sample. Subclones that are underrepresented in samples may contain mutations that are measured in such low frequencies that these cannot be distinguished from sequencing errors and noise. This is problematic for selecting the right treatment for a patient, since exactly those mutations that are missed may confer the tumor resistance to the treatment. Therefore, reconstructing a concise overview of all subclones in a tumor is essential for treating the disease. While conventional methods typically use statistical models to deconvolve the average mutation profile into a subclonal profile, one important problem is that many infrequent subclones are often still missed. In **chapter 2**, we instead aim to reduce heterogeneity by microdissecting the tissue as a means of obtaining more homogeneous samples. We introduce TargetClone, a method that uses a combination of Single Nucleotide Polymorphisms (SNPs) and Single Nucleotide Variants (SNVs) to infer which subclones are present in microdissected samples. We use our method to reconstruct phylogenetic trees representing the subclones and the most likely order in which these developed in 4 type II Testicular Germ Cell Cancer (TGCC) patients.

In **chapter 3** and **chapter 4**, we focus on the challenge of predicting the effect of non-coding structural variants (SVs). From recent studies, it has become clear that mutations in the non-coding genome play an important role in cancer. While (non-coding) SNVs have been under active study, computational methods to study the effect of SVs in especially non-coding regions are lacking, leaving them often ignored in cancer diagnostics. One important way in which non-coding SVs may be pathogenic is by disrupting the boundaries of Topologically Associated Domains (TADs), which are regions in the genome wherein DNA interacts more frequently with each other than with regions outside of the TAD. These structures maintain proper regulatory interactions in the genome.

Therefore, disruptions of TAD boundaries can enable the formation of novel interactions between genes and regulatory elements, such as enhancers. However, it is not known when such re-wiring events are pathogenic. While machine learning is an ideal solution for these problems, 2 important challenges exist. First, it is difficult to represent the large number of possibly affected interactions in a typical feature matrix. Second, no ground truth SV pathogenicity datasets exist that can be used as labels.

In **chapter 3**, we introduce svMIL, a method that learns which TAD boundary-disrupting non-coding SVs are pathogenic. Our method is based on Multiple Instance Learning (MIL), which circumvents the need for a feature matrix by representing non-coding SVs as bags that may contain any number of disrupted interactions. Non-coding SV pathogenicity is determined from patient matched gene expression data. We show that our method can predict pathogenic SVs in breast and ovarian cancer.

In **chapter 4**, we further improve on the svMIL method introduced in **chapter 3**. We apply svMIL2 to 1646 whole cancer genomes and demonstrate that our method is generally applicable across 12 cancer types and identifies non-coding SVs affecting well-known driver genes. We find that non-coding SVs exert their pathogenic effects similarly across cancer types by disrupting active (super) enhancers in open chromatin regions. Furthermore, non-coding SVs appear to contribute to the development of cancer more than driver SNVs in especially ovarian and pancreatic cancer. Finally, we also apply svMIL2 to intra-TAD CTCF loops rather than TAD boundaries. While SVs clearly can affect TADs, the role of disrupting the smaller CTCF loops inside TADs is not yet understood. We identify non-coding SVs affecting known driver genes through intra-TAD CTCF loop disruption in breast cancer, albeit with a smaller effect size than for TAD boundaries. However, due to lack of data on CTCF loops, the role of disrupting these loops in other cancer types remains an open question.

Finally, in **chapter 5**, we address the challenge of gene deficiency status (DS) classification. Pathogenic mutations in DNA-repair genes, such as BRCA1, BRCA2 or CDK12, leave signatures of somatic SVs in the genome. These signatures are a valuable proxy for machine learning classifiers to learn how to detect gene DS. However, obtaining a patient-level label while representing all genomic features of each SV, such as replication timing, in a single feature matrix is not trivial. We solve this problem using MIL, defining patients as bags which contain all SVs of a patient described with individual feature vectors. While machine learning-based methods often are often great solutions to open problems such as gene DS classification, there are potential pitfalls to using these models that should not be overlooked. Often, only WGS data is available for each patient. While it is straightforward to label genes based on whether these are deficient or not in the WGS data, in reality, genes may also be inactivated through non-genetic pathways. These gene inactivations may be caused by processes such as methylation or non-coding mutations, which cannot be determined from the WGS data alone. In this manner, patients in the negative set may actually contain deficient genes, leading to the use of noisy labels in the training process. We demonstrate that Positive Unlabeled (PU) learning, in which noise is overcome by viewing the negative set as unlabeled, is a suitable approach to improve classifier performance. However, achieving a higher false positive ratio in the presence of label noise may actually represent higher performance, as this patient may have been incorrectly labeled as negative while showing the deficiency phenotype.

Thus, metrics that are based on the false positive rate, such as AUCPR, may not actually be most suitable to compare model performance. In this chapter, we introduce a swap-one-patient-out CV (sopoCV) approach which iteratively labels each positive sample as negative, and measures how well a classifier can identify swapped patients as false positives in datasets with noisy labels. We show for 3 different machine learning models that sopoCV may be a more accurate way of measuring classifier performance than AUCPR.

In summary, this thesis has introduced new methods to help study cancer datasets. Using the method described in **chapter 3** and **chapter 4**, we demonstrated the importance of including non-coding SVs in cancer diagnostics. Furthermore, in **chapter 5** we showed that it is essential to properly measure the predictions of machine learning classifiers to properly interpret their performance on cancer data. Altogether, these tools provide researchers with new ways of learning to better understand the development of cancer.

Samenvatting

Ondanks vele jaren aan onderzoek, begrijpen we kanker nog steeds niet genoeg om de ziekte optimaal te behandelen. Sinds sequentiedatasets van gehele kankergenomen steeds meer beschikbaar zijn geworden, is onze kennis over de rol van mutaties in kanker gegroeid. Uit deze data is duidelijk geworden dat iedere kankerpatiënt gemiddeld 4-5 mutaties heeft die de groei van de kanker bevorderen, tegenover de tot in de duizenden mutaties met weinig effect. De toename van het aantal computationele methoden om deze kanker bevorderende mutaties in patiënten op te sporen heeft geleid tot een samenstelling van een lijst van mutaties met klinisch belang, wat heeft geleid tot een verbetering van de behandelingen tegen kanker. Toch blijft er een groep patiënten bestaan voor wie geen kanker bevorderende mutaties kunnen worden gevonden, of wiens ziektebeeld niet verbetert met de bestaande behandelingen. Het doel van dit proefschrift was om bij te dragen aan het vinden van oplossingen voor 3 uitdagingen die moeten worden opgelost om kanker beter te begrijpen.

In **hoofdstuk 2** onderzoeken we de uitdaging van tumor heterogeniteit. Tumoren bestaan vaak uit meerdere verschillende typen cellen, in plaats van 1. Deze cellen heten subklonen. Het is echter niet gebruikelijk om een apart monster te nemen van iedere individuele cel in de tumor, en in plaats daarvan bestaan monsters uit een heterogene cel populatie. Als gevolg hiervan zijn alle mutaties die worden bepaald uit sequentiedata vaak een gemiddelde van alle subklonen in een monster. Subklonen die een klein deel van de populatie opmaken kunnen mutaties bevatten die in zulke lage frequenties worden gemeten dat deze niet van ruis en sequentiefouten kunnen worden onderscheiden. Dit is problematisch voor de selectie van de beste behandeling voor een patiënt, aangezien het precies deze gemiste mutaties kunnen zijn die een tumor resistent maken tegen de behandeling. Hierdoor is het van groot belang om een gedetailleerd overzicht te genereren van alle subklonen in een tumor, zodat de meest geschikte behandeling kan worden bepaald. Conventionele methoden gebruiken vaak statistische modellen om het gemiddelde mutatieprofiel te deconvolueren naar een profiel op het niveau van de subklonen. Toch blijft het lastig om op deze manier alle laagfrequente subklonen te detecteren. In **hoofdstuk 2** is ons doel om in plaats daarvan de heterogeniteit te reduceren door weefsels te microdissecteren om monsters te verkrijgen met hogere homogeniteit. We introduceren TargetClone, een methode waarin een combinatie van enkel-nucleotide polymorfismes (Single-Nucleotide Polymorphisms, SNPs) en enkel-nucleotide varianten (Single-Nucleotide Variants, SNVs) worden gebruikt om de subklonen in microdissecties te bepalen. We gebruiken onze methode om fylogenetische bomen te genereren waarin de subklonen en de volgorde waarin deze meest waarschijnlijk zijn ontstaan staan gerepresenteerd voor 4 patiënten met testiculaire type II kiemceltumoren.

In **hoofdstuk 3** en **hoofdstuk 4** ligt de focus op de uitdaging om het effect van niet-gecodeerde structurele varianten (SVs) te bepalen. Uit recent onderzoek is gebleken dat mutaties in het niet-gecodeerde genoom een belangrijke rol spelen in kanker. Terwijl

(niet-gecodeerde) SNVs actief worden bestudeerd, zijn er weinig computationele methoden beschikbaar om het effect van SVs, voornamelijk niet-gecodeerd, te bestuderen. Hierdoor worden deze vaak genegeerd in de kankerdiagnostiek. Eén belangrijke manier waarop niet-gecodeerde SVs pathogeen kunnen zijn is door de grenzen tussen Topologisch Geassocieerde Domeinen (Topologically Associated Domains, TADs) te verstoren. Deze TADs zijn gebieden in het genoom waarin DNA vaker met elkaar interacties vormt dan met DNA buiten de TAD. De TAD structuren houden de juiste interacties tussen genen en regulatoire elementen in stand. Verstoringen van de grenzen tussen TADs kan dus leiden tot het ontstaan van nieuwe interacties tussen genen en regulatoire elementen, zoals enhancers. Het is echter niet duidelijk wanneer deze verstoringen pathogeen zijn. Machinaal leren leent zich goed voor dit soort problemen, maar hierbij bestaan 2 belangrijke problemen. Het is ten eerste lastig om de grote hoeveelheid verstoorde interacties te beschrijven in een typische feature matrix. Ten tweede bestaan er geen datasets met ware SV pathogeniteit die als labels kunnen worden gebruikt.

In **hoofdstuk 3** introduceren we svMIL, een methode die leert welke TAD grensverstorende niet-gecodeerde SVs pathogeen zijn. Onze methode is gebaseerd op Multiple Instance Learning (MIL), waarin het gebruik van een feature matrix omzeild wordt door niet-gecodeerde SVs als zakken weer te geven, die kunnen bestaan uit elke mogelijke hoeveelheid verstoorde interacties. Niet-gecodeerde SV pathogeniteit wordt afgeleid uit patiënt-geassocieerde genexpressiedata. We laten zien dat onze methode pathogene SVs kan voorspellen in borst- en ovariumkanker.

In **hoofdstuk 4** verbeteren we de in **hoofdstuk 3** geïntroduceerde svMIL methode. We passen svMIL2 toe op 1646 gehele kankergenomen en tonen aan dat onze methode generiek toepasbaar is op 12 kankertypes, en dat deze niet-gecodeerde SVs identificeert die bekende kanker-bevorderende genen verstoren. We vinden dat pathogene niet-gecodeerde SVs actieve (super) enhancers in euchromatine gebieden aantasten, en dat dit patroon vergelijkbaar is in alle kankertypes. Verder blijken niet-gecodeerde SVs meer bij te dragen aan de ontwikkeling van kanker dan kanker-bevorderende SNVs in voornamelijk ovarium- en alvleesklierkanker. Als laatste passen we svMIL2 ook toe op intra-TAD CTCF lussen in plaats van TAD-grenzen. Ondanks dat het duidelijk is dat SVs TADs kunnen verstoren, wordt de rol van het verstoren van de kleinere CTCF-lussen binnen TADs nog niet goed begrepen. We identificeren niet-gecodeerde SVs die bekende kanker-bevorderende genen beïnvloeden door de verstoring van CTCF-lussen in borstkanker, zij het met een kleinere effectgrootte dan voor TAD-grenzen. Echter blijft de rol van het verstoren van deze lussen in andere kankertypes een open vraag door het ontbreken van voldoende data over CTCF-lussen.

Tot slot behandelen we in **hoofdstuk 5** de uitdaging van gen-deficiëntie status (DS) classificatie. Pathogene mutaties in DNA-reparatiegenen, zoals BRCA1, BRCA2 en CDK12, laten signaturen van somatische SVs achter in het genoom. Deze signaturen zijn een waardevolle bron om met machinaal leren gen DS te detecteren. Het is echter niet triviaal om een label op patiënt-niveau te verkrijgen waarbij alle genomische features van SVs, zoals replicatietiming, in één feature matrix zijn weergegeven. We lossen dit probleem op met MIL, waarin alle zakken patiënten vertegenwoordigen, die alle SVs in de patiënt beschrijven met een eigen feature vector. Machinaal leren is vaak een zeer geschikte oplossing voor onopgeloste problemen, zoals gen DS-classificatie. Toch bestaan

er valkuilen wanneer deze methoden worden gebruikt die niet over het hoofd gezien moeten worden. Er is vaak per patiënt alleen sequentiedata van het gehele genoom beschikbaar. Een eenvoudig optie is om genen een label te geven op basis van of deze deficiënt zijn in de genomische sequentiedata. Het is echter ook mogelijk dat deze genen zijn geïnactiveerd via niet-genetische paden. Deze inactivaties kunnen zijn veroorzaakt door processen als methylatie, of niet-gecodeerde mutaties, die niet direct kunnen worden gedetecteerd uit alleen de genomische sequentiedata. Hierdoor kunnen patiënten in de negatieve set deficiënte genen hebben, waardoor het trainingsproces labels met ruis gebruikt. We tonen aan dat Positive Unlabeled (PU) learning, waarin ruis wordt aangepakt door de negatieve set als ongelabeld te behandelen, een geschikte manier is om de classificatieprestatie te verhogen. Een hogere ratio van fout-positieven in een scenario met labelruis kan echter betere prestatie betekenen, omdat deze patiënt een foute (negatieve) label heeft gekregen terwijl deze wel het deficiëntie fenotype heeft. Metriek op basis van de ratio fout-positieven, zoals AUCPR, zijn hierdoor soms niet het meest geschikt om modellen te vergelijken. In dit hoofdstuk introduceren we swap-one-patient-out cross-validatie (sopoCV), waarin we iteratief iedere positieve patiënt als negatief labelen, en meten hoe goed een classificatiealgoritme deze nu negatieve patiënten als fout positief kan identificeren in datasets met labelruis. We laten zien voor 3 verschillende modellen op basis van machinaal leren dat sopoCV een geschiktere manier kan zijn om classificatieprestatie te meten dan de AUCPR.

In dit proefschrift hebben we nieuwe methoden geïntroduceerd om kankerdata te bestuderen. Met de methode beschreven in **hoofdstuk 3** en **hoofdstuk 4** hebben we het belang laten zien om niet-gecodeerde SVs mee te nemen in kankerdiagnostiek. In **hoofdstuk 5** hebben we aangetoond dat het essentieel is om de voorspellingen van classificatiealgoritmes juist te meten om een passende interpretatie te maken van de prestatie op kankerdata. Alles samengenomen geven deze tools wetenschappers nieuwe mogelijkheden om de ontwikkeling van kanker beter leren te begrijpen.

Acknowledgements

While my PhD may have had many ups and downs, I finally made it to the end! Of course, this was not without the help of many people along the way, and I want to express my gratitude here.

First of all, thanks to my promoter **Edwin Cuppen** for giving me a chance at doing a PhD. Also, I want to thank **Berend Snel, Dick de Ridder, Aniek Janssen, Leendert Looijenga and Rene Eickemans** for critically reading my thesis and for giving me a green light to defend! I would also like to thank **Berend, Aniek and Wigard Kloosterman** for being part of my supervisory committee and giving helpful feedback throughout the years.

Jeroen, thank you so much for all your guidance from my Master's thesis all the way until the end of my PhD! I am grateful for all the opportunity I had to learn and grow into the person I am today. While the process of getting to a final version of the papers wasn't always easy, your enthusiasm during our meetings always inspired me work on the projects from a different angle and get to a final product in the end.

Lambert, being supervised by an experienced researcher like you taught me a lot early on in my career. Although the final results may have been more underwhelming than we would have liked and there were a lot of setbacks in getting the work published, I'm still happy that we managed to get through and contribute something useful to science in the end. Thank you! **Leendert**, thank you for your supervision on my very first project. I'm very happy that I got the opportunity to work in your lab.

Of course, I want to thank all former and current members of the **de Ridder group** for creating a supportive and inspiring environment to do my PhD in. I have been here since the very beginning of the lab, and I have seen many people come and go. I have definitely witnessed the group go through many positive changes over time, and I'm sure the current members can keep this up and create an even brighter future! **Joske**, I remember you started your PhD around the time I started working on my Master's thesis and you have been the one person who has always been around. I always looked to you for figuring out how to actually do a PhD, and here we are at the end, so your advice was definitely helpful! **Amin**, although you weren't around for my whole PhD, your great interest in scientific projects that aren't even your own and being able to still give great advice taught me a lot. I'm sure these skills will be invaluable for the rest of your career!

To all current and former members of the **east side office**, thank you so much for always providing a fun and supportive work environment to return to every day. Although I did not meet all of you in-person anymore during the past year due to covid-19, the daily chats (and not to forget, rants) and coffee breaks remain fresh in my memory. **Joanna**

W, we moved to the east side office together when we decided that we wanted a more comfortable workspace, but I also remember we worried that we would chat too much. Finding someone with interests similar to my own to talk about was definitely one of the best things that could have happened during my PhD. You have made amazing progress as a scientist and as a person, and I'm sure you will finish with a beautiful thesis. I've seen you build and design great things during your PhD, and I'm sure that these will be invaluable to you wherever you decide to go next! **Jesko**, while you were only here for a short period of time, you may well be one of the people I chatted with the most during covid times. You did great work on your project and I'm very happy to see the end results. Good luck in the next steps of your career, and I hope you really won't have to survive off cup noodles! **Joanna von B**, I know of no one who is better than you at just getting the work done. You are great at setting boundaries (and I don't mean just the whiteboard) and creating a healthy space for yourself, and I think we can all learn a lot from that. Good luck with the remainder of your PhD! **Roy**, I still remember when you taught me the basics of how to use the HPC, and now you are the one person everyone looks to for advice on how to get the most out of their computer-related tasks. I'm sure you will do great in your new job position! **Myrthe**, you have been essential for making the office 'gezellig'. I'm sure you will fill that role again once everyone can go back to the office! **Liting**, seeing your adorable cats in the background on Zoom always put a smile on my face. I'm sure I'm not the only one who felt that way! **Emmy**, now that Joske has left, I think you are exactly what the office needs to keep feeling lively! Your dedication to social activities and organizational tasks is amazing, keep it up! **Sara**, you were someone I always looked up to. The first time I met you was while you were in a heated discussion with Mircea, and I didn't know what to think. But you turned out to be a very kind and supportive person, and I am extremely grateful for all you did to help me. Thank you so much! **Joep**, I definitely learned so many things from you – whether it was related to science, mental health, or life in general. When you left to New Zealand it was a great loss for the lab, but today everyone still remembers your wisdom. Thank you for everything!

To everyone from the other offices too, it was great getting to know you all. **Alexandra**, whenever we talk, we seem to end up just ranting away about, well, everything. I'm glad there has always been someone who I could share my feelings with about the whole process, and you've been great support for that. I'm sure you will have some great results soon, even if it may not feel that way right now! **Marc**, your knowledge of both wet-lab and dry-lab work is pretty uncommon in our lab, and I think it will definitely be useful to everyone! **Carlo**, thank you for the many helpful discussions we had about the 3D genome! **Jasmin**, although data management was at times too complex for me to understand, I was able to learn a lot from you! Also, Mozart would like to thank you for taking very good care of him while I was away. **Luca**, thank you for doing all the work for acquiring the HMF data that was an invaluable source for me during my PhD. Without your help, the project would have been significantly less interesting. **Tilman**, you were always a great help whenever I needed to run a complex tool or pipeline. Thank you for setting all of this up! **Adrien**, your dedication to teaching is clearly paying off, and I hope that we can fill the world with more amazing bioinformaticians thanks to your help. **Flip**, where would everyone be without your help with computer-related issues? Thank you

for setting up everything I needed to make the process go smoothly.

I would also like to thank everyone in the **CMM** for all work discussions, inspiration and collaborations. **Luan**, thank you for being a valuable co-author on my final chapter and working together with me and Jesko to create a great final result. **Sjors** and **Judith**, thanks for the great inspiration and information that I needed to get my SV-related project off the ground.

This work would not have been possible without the permission from so many cancer patients to include their data in scientific research. Thank you for your help in creating a better world for everyone.

List of Publications

Part of this thesis

Nieboer, M. M., Nguyen, L. & de Ridder, J. (2021). Predicting pathogenic non-coding SVs disrupting the 3D genome in 1,646 whole cancer genomes using Multiple Instance Learning. Nature Scientific Reports. *In press*.

Nieboer, M. M., & de Ridder, J. (2020). svMIL: predicting the pathogenic effect of TAD boundary-disrupting somatic structural variants through multiple instance learning. Bioinformatics, 36(Supplement_2), i692-i699.

Nieboer, M. M., Dorssers, L. C., Straver, R., Looijenga, L. H., & de Ridder, J. (2018). TargetClone: A multi-sample approach for reconstructing subclonal evolution of tumors. PloS one, 13(11), e0208002

Other publications

Dorssers, L. C., Gillis, A. J., Stoop, H., van Marion, R., **Nieboer, M. M.**, van Riet, J., ... & Looijenga, L. H. (2019). Molecular heterogeneity and early metastatic clone selection in testicular germ cell cancer development. British journal of cancer, 120(4), 444-452.

Stancu, M. C., Van Roosmalen, M. J., Renkens, I., **Nieboer, M. M.**, Middelkamp, S., De Ligt, J., ... & Kloosterman, W. P. (2017). Mapping and phasing of structural variation in patient genomes using nanopore sequencing. Nature communications, 8(1), 1-13.

Curriculum Vitæ

Marleen Nieboer was born on 27 November 1993 in Winschoten, the Netherlands. After completing secondary school at Dollard College Winschoten in 2010, she started a Bachelor's degree in Bioinformatics at the Hanze University in Groningen. After discovering that she wanted to learn more about the computational side of the field, she continued her education at Leiden University and TU Delft and obtained her Master's degree in Computer Science with a specialization in Bioinformatics in 2016. With a desire to apply her knowledge to the medical field, she started her PhD in October 2016 at the UMC Utrecht under the supervision of dr. Jeroen de Ridder. During her PhD, she worked on developing various bioinformatics algorithms with the ultimate aim of contributing to a better understanding of the development and treatment of cancer.