



See related Commentary on page 784



Clinical Validation of Whole Genome Sequencing for Cancer Diagnostics

Paul Roepman, Ph.D.,^{*} Ewart de Bruijn,^{*} Stef van Lieshout,^{*} Lieke Schoenmaker,^{*} Mirjam C. Boelens,[†] Hendrikus J. Dubbink,[‡] Willemine R.R. Geurts-Giele,[‡] Floris H. Groenendijk,[‡] Manon M.H. Huibers,[§] Mariëtte E.G. Kranendonk,[§] Margaretha G.M. Roemer,[¶] Kris G. Samsom,[†] Marloes Steehouwer,^{||} Wendy W.J. de Leng,[§] Alexander Hoischen,^{||**} Bauke Ylstra,[¶] Kim Monkhorst,[†] Jacobus J.M. van der Hoeven,^{*} and Edwin Cuppen^{*††}

From the Hartwig Medical Foundation,^{*} Amsterdam; the Department of Pathology,[†] Netherlands Cancer Institute, Amsterdam; the Department of Pathology,[‡] Erasmus MC Cancer Institute, Rotterdam; the Department of Pathology[§] and the Center for Molecular Medicine and Oncode Institute,^{††} University Medical Center Utrecht, Utrecht; the Department of Pathology,[¶] Amsterdam University Medical Center, Amsterdam; and the Departments of Human Genetics^{||} and Internal Medicine,^{**} Radboud Institute for Molecular Life Sciences, Radboud University Medical Center, Nijmegen, the Netherlands

Accepted for publication
April 12, 2021.

Address correspondence to Paul Roepman, Ph.D., Hartwig Medical Foundation, Science Park 408, Amsterdam 1098 XH, the Netherlands. E-mail: p.roepman@hartwigmedicalfoundation.nl.

Whole genome sequencing (WGS) using fresh-frozen tissue and matched blood samples from cancer patients may become the most complete genetic tumor test. With the increasing availability of small biopsies and the need to screen more number of biomarkers, the use of a single all-inclusive test is preferable over multiple consecutive assays. To meet high-quality diagnostics standards, we optimized and clinically validated WGS sample and data processing procedures, resulting in a technical success rate of 95.6% for fresh-frozen samples with sufficient ($\geq 20\%$) tumor content. Independent validation of identified biomarkers against commonly used diagnostic assays showed a high sensitivity (recall; 98.5%) and precision (positive predictive value; 97.8%) for detection of somatic single-nucleotide variants and insertions and deletions (across 22 genes), and high concordance for detection of gene amplification (97.0%; *EGFR* and *MET*) as well as somatic complete loss (100%; *CDKN2A/p16*). Gene fusion analysis showed a concordance of 91.3% between DNA-based WGS and an orthogonal RNA-based gene fusion assay. Microsatellite (in)stability assessment showed a sensitivity of 100% with a precision of 94%, and virus detection (human papillomavirus), an accuracy of 100% compared with standard testing. In conclusion, whole genome sequencing has a $>95\%$ sensitivity and precision compared with routinely used DNA techniques in diagnostics, and all relevant mutation types can be detected reliably in a single assay. (*J Mol Diagn* 2021, 23: 816–833; <https://doi.org/10.1016/j.jmoldx.2021.04.011>)

Needs and complexity in molecular cancer diagnostics are rapidly increasing, driven by a growing number of targeted drugs and developments toward more personalized treatments.^{1,2} Simultaneously, advances in next-generation DNA sequencing technology have greatly enhanced the capability of cancer genome analyses, thereby rapidly progressing diagnostic approaches from small targeted panels to large panels and exome sequencing. Currently, whole genome sequencing (WGS) using tissue and matched blood samples from patients with (metastatic) cancer³ is getting in reach as the most complete genetic tumor diagnostics test. In the context of the Dutch national Center for Personalized Cancer Treatment (CPCT) clinical study (NCT01855477; <https://clinicaltrials.gov/ct2/show/NCT01855477>, last accessed May 21, 2021), Hartwig Medical Foundation has established a national WGS facility, including robust

sampling procedure and logistics in >45 (of the 87) hospitals located across the Netherlands for the centralized analysis of tumor biopsies by WGS. Since the start in 2016, >5000 tumors and matched control samples have been analyzed by WGS, of which the first cohort of 2500 patients has been extensively characterized and described.⁴ Originally, this clinical study aimed to analyze data for biomarker discovery, but with growing clinical demands for more extensive and broader DNA analysis for patient stratification toward targeted treatments,⁵ the scope of WGS is now entering

Disclosures: None declared.

Current address of M.M.H.H., Department of Genetics, University Medical Center Utrecht, Heidelberglaan, Utrecht, the Netherlands; M.E.G.K., Princess Máxima Center for Pediatric Oncology, Heidelberglaan, Utrecht, the Netherlands.

routine diagnostic usage. As part of this development, the required amount of tumor tissue for as well as the turnaround time of the WGS procedure were decreased, together with implementation of more extensive quality control metrics and independent validation required for accreditation. Currently, there is an ongoing trend toward the availability of only small biopsies, especially for advanced stage cancer where metastatic lesions are sampled using core needle biopsies, with at the same time a growing need to screen for an increasing number of biomarkers. For future proof and efficient molecular diagnostics, the use of a single all-inclusive test is preferred over multiple consecutive assays that, together, often take more time, require more tissue, and provide a far less complete profile of the molecular characteristics.

To meet the high-quality diagnostics standards, we have optimized and clinically validated the performance of the WGS workflow on fresh-frozen tumor samples, both technically as well as bioinformatically, as these are highly interconnected in determining the precision (positive predictive value) and sensitivity (recall) of the test. The validation efforts include current standard-of-care biomarkers (oncogenic hotspots, inactivating mutations in tumor suppressor genes), but also broader analyses of gene fusions and other genomic rearrangements as well as emerging genome-wide or complex biomarkers, like tumor mutational burden estimation, microsatellite instability (MSI),⁶ and homologous repair deficiency (HRD) signatures.^{7,8} More important, an open-source and data-driven filtering and reporting strategy has been put into place to reduce the wealth of information into a diagnostically manageable size and to provide an overview of all clinically relevant DNA aberrations.

Herein, we show that WGS has an overall >95% sensitivity (recall) and precision (positive predictive value) compared with other routinely used tests and that all relevant mutation types can be readily and reliably detected in a single assay. Although WGS required minimal quantity of input material and can be applied pan-cancer, the tumor purity was a limiting factor (requiring >20% tumor cells) as well as the availability of fresh-frozen tumor material, which were prerequisites for high-quality results, as described herein. Together, WGS has now matured from a research technology into an International Organization for Standardization (ISO) accredited test that is ready to be used for clinical decision making.

Materials and Methods

Tumor Samples

For this study, samples were used from patients that were included as part of the Center for Personalized Cancer Treatment (CPCT) (NCT01855477), the Drug Rediscovery Protocol (DRUP) (NCT02925234; <https://clinicaltrials.gov/ct2/show/NCT02925234>, last accessed May 21, 2021), and the Whole genome sequencing Implementation in standard Diagnostics for Every cancer patient (WIDE) (NL68609.031.18; https://www.toetsingonline.nl/to/ccmo_search.nsf/

[fABRpop?readform&unids7E4B29996099EEEEAC12585A5001611EB](https://clinicaltrials.gov/ct2/show/NCT02925234), last accessed May 21, 2021) clinical studies, which were approved by the medical ethical committees of the University Medical Center Utrecht and the Netherlands Cancer Institute. All patients have consented to the reuse of their pseudonymized data for research aimed at improving cancer care.

Whole Genome Sequencing

Whole genome sequencing (WGS) was performed under ISO-17025 accreditation at the Hartwig Medical Foundation laboratory (Amsterdam, the Netherlands). The WGS test used DNA extracted from fresh-frozen or frozen archived tumor tissue (primary or metastatic) and from matching blood samples (reference). DNA extraction is performed on the QIASymphony (Qiagen, Hilden, Germany) following standard reagents and protocols: 1 mL of blood was used for DNA isolation using the QIASymphony DSP DNA Midi kit (Qiagen). The QIASymphony DSP DNA Mini kit was used for tissue DNA isolation. Next, 50 to 200 ng DNA was fragmented by sonication on the Covaris LE220 Focused ultrasonicator (Covaris, Brighton, UK; median fragment size, 450 bp) for TruSeq Nano DNA Library (Illumina, San Diego, CA) preparation, including PCR amplification (eight cycles). All procedures were automated on the Beckman Coulter Biomek 4000 and Biomek i7 liquid handling robots (Beckman Coulter, Brea, CA). The Illumina HiSeqX and NovaSeq6000 platforms were used for sequencing tumor (approximately 90×) and blood (approximately 30×) genomes. No minimal threshold was applied regarding the mean coverage but instead the Gbase sequencing output for the tumor and blood samples had to be >300 and >100 Gb, respectively, to be eligible for downstream diagnostic analysis. Additional data quality criteria were as follows: read mapping percentage, >95%; reference genome-wide coverage 10×, >90%; reference genome-wide coverage 20×, >70%; tumor genome-wide 30× coverage, >80%; and tumor genome-wide 60× coverage, >65%.

Tumor Purity

WGS analysis required tumor samples with sufficient tumor cell percentage (≥20%). Prescreening of eligible tumor samples was performed by manual pathologic tumor cell percentage (pTCP) scoring of hematoxylin and eosin–stained sections, cut from the same frozen biopsy [following standard formalin-fixed, paraffin-embedded (FFPE) protocol] that was used for DNA isolation (to minimize the potential effect of tumor heterogeneity). In addition, molecular-based tumor cell purity (mTCP) was determined based on the WGS data (see *Bioinformatics*) for optimal analysis and interpretation of the DNA results. The mTCP was also determined after shallow whole-genome sequencing (8× to 15× coverage depth) to be able to identify tumors with a potential discrepancy in pTCP and

mTCP before continuing with deep sequencing (approximately 90× to 110×). This also allowed prescreening tumors for which no (reliable) pathologic assessment was available. Only cases with an mTCP of $\geq 20\%$ were considered eligible for diagnostics analysis.

Bioinformatics

Sequencing data were analyzed with an in-house developed open-source software-based pipeline. Reliable variant calling by sequencing techniques (especially WGS) depends on a complex, often Bayesian, approach, including read quality, variant allele frequency, sequence depth, and tumor purity and ploidy. A schematic overview of all of the used tools is provided in [Supplemental Figure S1](#). Sequencing read alignment of matching tumor and blood reference samples was performed using the Burrows-Wheeler Aligner version 0.7.17. Somatic variant calling [single-nucleotide variants (SNVs), multinucleotide variants, and insertions and deletions (indels)] between the tumor reference pair was performed using STRELKA version 1.0.14 (<https://github.com/Illumina/strelka>, last accessed May 21, 2021), with which indels up to 50 bp could reliably be identified.⁹ Larger insertions and deletions (≥ 50 bp) were detected using the tool GRIDSS version 2.8.3 (<https://github.com/PapenfussLab/gridss>, last accessed May 21, 2021) as being structural variants. GRIDSS is a structural variant detection tool including a genome-wide break-end assembler and a somatic structural variation caller, and is able to detect genomic break junctions.¹⁰

Variant and gene ploidy aspects were assessed using the AMBER tool version 3.3 (<https://github.com/hartwigmedical/hmftools/tree/master/amber>, last accessed May 21, 2021), which determined allele copy numbers of heterozygous germline variants in the tumor samples. In combination with COBALT version 1.7 (<https://github.com/hartwigmedical/hmftools/tree/master/count-bam-lines>, last accessed May 21, 2021), which determined read depth ratios and copy numbers of the supplied tumor and reference data, information was gathered concerning the local copy number and ploidy for bins of approximately 1 kb across the tumor genome. In addition, a sex check was performed using the COBALT output based on the observed sex chromosome pattern.

Output from the AMBER (bi-allele frequencies), COBALT (read depth ratios), STRELKA (somatic variants), and GRIDSS (structural variants) was combined in the tool PURPLE version 2.43 ([Supplemental Figure S1](#)) that was designed specifically for WGS data. PURPLE was able to estimate the purity (mTCP) and copy number profile of a tumor sample by searching for the best genome-wide purity/ploidy fit with the input data. The tool provided tumor purity corrected variant allele frequencies (VAFs) and allele-specific copy numbers that could be used for detection of loss of heterozygosity.¹¹ More important, tumor purity correction allowed for reliable identification of somatic

complete loss of a gene [eg, loss of heterozygosity of *BRCA1* and deep (bi-allelic) deletions of *CDKN2A*].

Downstream interpretation of structural variants and the calling and annotation of gene fusions were performed using LINX version 1.7 (<https://github.com/hartwigmedical/hmftools/tree/master/sv-linx>, last accessed May 21, 2021). This tool was able to group together the individual structural variant calls into distinct events, predicted the local structure of the derivative chromosome, and properly classified and annotated events for their functional impact.¹¹

Genome-wide mutational characteristics were determined, including the tumor's mutational load (ML; defined as the total number of somatic missense variants across the whole genome of the tumor) and the tumor mutational burden (TMB; defined as the number of all somatic variants per genome Mb). MSI was assessed using the method described by the MSISeq tool.⁶ In brief, the number of indels was calculated per million bases and occurring in homopolymers of five or more bases or dinucleotide, trinucleotide, and tetranucleotide sequences of repeat count four or more. Samples with a score >4 were classified as MSI.

Homologous recombination DNA repair deficiency was assessed using the previously described CHORD tool version 60.02_1.03.⁸ The CHORD tool is random forest classifier of HRD and was able to distinguish between *BRCA1* and *BRCA2*-type HRD phenotypes. The main discriminants for HRD were the numbers of deletes with microhomology and the number of large duplications with length between 1 and 100 kb. CHORD achieved a maximum F1 score (approximately 0.88) for predicting HRD with a cutoff of 0.5, and samples above this cutoff were classified as HRD.⁸

Furthermore, the presence of viral DNA was detected using VIRUSBreakend (GRIDSS subtool) that identified viral integrations anywhere in the host genome using a single breakend-based strategy followed by taxonomic classification of the detected viral DNA.¹²

All code and scripts used for analysis of the WGS data are open source and available at GitHub (<https://github.com/hartwigmedical>, last accessed February 15, 2021). The raw and analyzed WGS data used in this article are available for validation and cancer research purposes through a standardized controlled data access procedure (<https://www.hartwigmedicalfoundation.nl/applying-for-data>, last accessed February 15, 2021).

Orthogonal Validation Experiments

Independent validation was performed for all to-be-reported types of clinically relevant DNA aberrations, including mutations (SNVs, multinucleotide variants, and indels) with specific focus on *BRAF*, gene amplification (*ERBB2* and *MET* as examples) and complete loss of genes (*CDKN2A*, *BRCA1*, and *BRCA2*), microsatellite (in)stability, gene fusions, and viral infection [human papillomavirus (HPV) as example]. WGS results were retrospectively compared

against (as far as possible) routine diagnostic assays performed independently in ISO15189-accredited pathology laboratories. If a clinical assay was not available for the validation purpose, a custom research-use-only test was performed. The following independently performed validation experiments were performed. An overview of the used tumor samples and tumor types for each validation experiment is available as [Supplemental Table S1](#).

Validation of SNV, Multinucleotide Variant, and Indel Detection

A custom-designed (research-use-only) single-molecule molecular inversion probe (smMIP) sequencing panel was designed for independent confirmation of variants detected by WGS. The smMIP panel sequencing was designed and processed similar to previous reports (Radboudumc, Nijmegen, the Netherlands).^{13,14} In total, 415 smMIPs (covering 1.4 kbp) were designed to test 192 randomly selected variants (165 SNVs and 27 indels) that were detected by WGS across 29 tumor samples. smMIP validation was performed using the same isolated DNA as was used for WGS, and analyzed by JSI SeqPilot version 5.1.0 (JSI Medical Systems, Ettenheim, Germany).

Orthogonal clinical validation of variant detection was performed using 48 samples and compared against a custom-made OncoPrint next-generation sequencing (NGS) gene panel (Thermo Scientific, Waltham, MA), processed independently (double blind) in a routine pathology laboratory under ISO15189 accreditation (Erasmus MC, Rotterdam, the Netherlands).¹⁵ The custom OncoPrint assay covered 25.2 kb exonic regions across 40 genes [design (version 5.1) available in supplementary data of the article by Pruis et al.¹⁵] and was performed using the same isolated DNA as was used for WGS, thereby ruling out potential tumor heterogeneity. JSI SeqPilot version 5.2.0 was used for analysis, and a formal clinical report was generated. In addition, for 10 samples, a comparison was made between the WGS-based ML assessment and the OncoPrint Tumor Mutational Load assay (Thermo Scientific).

Validation of Copy Number Assessment

WGS-based copy number assessment was validated against fluorescent *in situ* hybridization (FISH) using COLO829 and a cohort of diagnostic tumor samples. For COLO829, a comparison was made for the ploidy of chromosomes 9, 13, 16, 18, and 9p24 (*CD274/PDCD1LG2*), and 2q23 (*ALK*) (Amsterdam UMC, Amsterdam, the Netherlands). Chromosome enumeration probes (CEPs) for the centromeric region of chromosome 9, 13, 16, and CEP9, CEP13, CEP16, and CEP18 were used, as well as locus-specific break-apart probes for 2p23 (*ALK*) fusion (Vysis, Abbott, IL) and 9p24 (*CD274/PDCD1LG2*) fusion (Leica Biosystems, Wetzlar, Germany). Slides were visualized on a Leica DM5500 fluorescence microscope (Leica

Biosystems), and for each marker, 100 cells per slide were scored for the percentages of cells with respective numbers of chromosomes (signals) counted.

Diagnostic *ERBB2* copy number readout was validated using 16 tumor samples and using *Her2/neu* FISH analysis at an independent routine pathology laboratory (University Medical Center Utrecht). Fresh-frozen sections for FISH analysis were from the same biopsy used for WGS or from a matching second biopsy obtained at the same moment. FISH scoring was performed according to guidelines.¹⁶ For fresh-frozen samples, new sections were fixed using overnight incubation with formalin. Subsequently, routine FFPE FISH protocol was used, excluding the xylene deparaffinization step. Slides were used for probe hybridization (LPS001; Cytocell, Cambridge, UK), scanned using the Leica DM6000 scanner, and analyzed with Cytovision software (Leica Biosystems). A formal clinical report was generated that was compared with the WGS results, for which the absolute copy numbers detected by WGS were compared with the absolute copy numbers detected by FISH.

In addition to *ERBB2*, WGS-based *MET* copy number readouts were investigated for samples classified as positive for *MET* amplification based on routine chromogenic dual *in situ* hybridization on matching FFPE biopsies. Routine *MET* amplification status was assessed using the *MET* DNP and Chromosome 7 DIG probes (Ventana, Tuscan, AZ) on sections (5 µm thick; SuperFrost slide; Thermo Scientific), according to the manufacturer's instructions. Samples were classified as positive for *MET*/CEP7 ratio >2.2.

Detection of complete loss of genes by WGS was validated using *CDKN2A* in which the WGS data were compared against p16 protein expression. *CDKN2A*/p16 was assessed by immunohistochemistry (IHC) on sections (3 µm thick) of matching FFPE tumor samples, using the monoclonal primary antibody E6H4 (Ventana).

Validation of Fusion Gene Detection

Validation of gene fusion detection by WGS was performed against RNA-based Anchored Multiplex PCR NGS assay (Archer FusionPlex Solid Tumor; ArcherDx, Boulder, CO). Twenty-four samples were selected on the basis of the WGS results to include multiple fusion genes. Matching RNA (200 ng), isolated from the same tissue as the DNA that was used for WGS, was analyzed according to routine pathologic procedures (ISO15189 certified; Erasmus MC). A formal clinical report was generated and compared with the WGS results.

Validation of Microsatellite (In)stability Readout

For a set of 50 tumor samples, the microsatellite status was validated using the MSI analysis system (Promega, Madison, WI) and performed at a routine pathology laboratory (Erasmus MC)¹⁷ and using the same isolated DNA that was

Table 1 Performance Characteristics for Clinical-Grade WGS Using GIAB and Tumor Biopsy Samples

Quality metric	Sample type (n)	Median value	Range value
Total read count	GIAB (22)	914*10 ⁶	644*10 ⁶ –1,429*10 ⁶
% Mapped	GIAB (22)	0.970	0.958–0.988
Precision SNVs	GIAB (22)	0.998	0.994–0.998
Sensitivity (recall) SNVs	GIAB (22)	0.989	0.973–0.990
F-score SNVs	GIAB (22)	0.993	0.985–0.994
Coverage whole genome	Tumor (25)	106×	84×–130×
Coverage protein-coding region (cancer genes)	Tumor (25)	105×	78×–134×
% of Protein-coding bases ≥10× (cancer genes)	Tumor (25)	99.68	99.45–99.81
% of Protein-coding bases ≥30× (cancer genes)	Tumor (25)	99.29	98.51–99.63

The GIAB samples (*n* = 21) have been analyzed in duplicate runs using multiple sequencers and across a period of 8 months in 2018. Data from 25 randomly selected tumor samples (from 2018) were used for coverage performance. The protein-coding region included 460 cancer-associated genes (2.33 Mbp in total). The F-score is a measure of a test's accuracy and is calculated from the precision (true positive/true positive + false positive) and sensitivity (recall; true positive/true positive + false negative) of the test.

GIAB, genome in a bottle; SNV, single-nucleotide variant; WGS, whole genome sequencing.

used for WGS. These fluorescent multiplex PCR assays analyzed five nearly monomorphic mononucleotide micro-satellite loci (BAT-25, BAT-26, NR-21, NR-24, and MONO-27). Matching tumor and blood samples were analyzed for accurate detection. Both the number of positive loci as well as binary classification of MSI and microsatellite stable were reported. Additional MSI-positive cases (*n* = 10) were included in the validation based on routine mis-match repair (MMR) IHC status (mlh1, pms2, msh2, and msh6) and/or *MLH1* methylation status (MS-MLPA kit; MRC-Holland, Amsterdam, the Netherlands).

Validation of Tumor-Associated Virus Detection

WGS-based detection of presence of high-risk HPV and/or Epstein-Barr virus (EBV) DNA was compared against routine pathologic testing (Netherlands Cancer Institute) using the QIAscreen HPV PCR Test (Qiagen) for HPV and EBV-encoded RNA (EBER) IHC for detection of presence of EBV in the tumor (both according to standard protocols). If available, results of routine testing for HPV and/or EBV were used for comparison with WGS. If not available, HPV status was determined retrospectively using an aliquot of the DNA (20 ng) that was used for WGS.

Results

Analytical Performance

In addition to the orthogonal clinical validation experiments that are described in the next paragraphs, the analytical performance of WGS was continuously monitored using a genome-in-a-bottle mix-in sample (tumor, 30% NA12878; normal, 100% NA24385) for which all DNA aberrations were known. The accuracy of genome-in-a-bottle genome-wide variant detection (SNVs and short indels) by WGS was high and stable across different runs and using multiple

sequencers (in a time period of 8 months) with a precision of 0.998 (range, 0.994 to 0.998) and a sensitivity (recall) of 0.989 (range, 0.973 to 0.990) (Table 1). F-scores (combining the precision and recall of the test) for variant detection exceeded the preset 0.98 lower limit for high-quality sequencing data (median, 0.993; range, 0.985 to 0.994). Direct comparison of all genome-wide somatic base calls (COLO829) between HiSeq and NovaSeq runs indicated a concordant result for 99.99953% of the bases. All discordant bases (1445 of approximately 3.1 billion) were located outside protein coding regions of cancer-associated genes (460 genes; 2.33 Mbp), resulting in identical reported results based on both platforms (SNV and indel analysis only). WGS coverage analysis across a set of 25 randomly selected tumor samples indicated stable and high coverage across the entire genome (median coverage after mapping, 106×; range, 84× to 130×). The protein coding regions of 460 cancer-associated genes showed a median coverage of 105× (range, 78× to 134×) with 99.68% and 99.29% of all bases covered at least 10× and 30×, respectively (Table 1).

To discern the (minimal) required alternative (ALT) variant read counts) and VAFs (nonpurity corrected) of the Bayesian calling pipeline, an analysis was performed to the VAF/ALT of reliable detected nonsynonymous variants for 118 cancer-associated genes across a set of 2520 tumor samples.⁴ Of >10,000 called variants, only 4 were based on an ALT count of ≤4, indicating that for the WGS setup used (combination of wet laboratory and bioinformatics), at least five ALT reads are required for reliable variant calling, representing a minimal sample VAF of 5% (with a coverage of approximately 100×) (Supplemental Figure S2).

The minimally required tumor cell percentage (purity) for sensitive variant detection was assessed using an *in silico* sensitivity model with a preset minimal coverage of 100× and ALT read count of 5. On the basis of this model, the minimal tumor purity with a sensitivity >95% to reliably detect a single-nucleotide variant was determined as 0.19

(Supplemental Table S2). This minimally required tumor purity was experimentally confirmed using a dilution experiment (COLO829; performed in duplicate) in which the tumor content was lowered incrementally (mTCP of 100%, 34%, 20%, and 13%). All four oncogenic driver mutations that are known to be present in COLO829 (*BRAF* p.Val600Glu, *CDKN2A* p.Gly124fs, *SF3B1* p.Pro718Leu, and *TP63* p.Met499Ile) were still all reported for the 20% tumor purity sample, whereas the 13% sample only showed three of the four mutations (missing *CDKN2A* p.Gly124fs).

The reproducibility of the complete workflow was confirmed on two diagnostic cases (non—small-cell lung cancer and an undifferentiated pleomorphic sarcoma) in which the replicated tests were started with new library preparations from the isolated blood/biopsy DNA samples and resulted in highly similar molecular profiles with

identical diagnostic reports (Figure 1 and Supplemental Table S3).

Sample Quality, Tumor Purity, and Success Rate

Samples used for WGS analysis were composed predominantly of freshly frozen fine-needle biopsies taken from a metastatic lesion. WGS required at least 50 ng of input DNA and that amount could successfully be isolated from >99% of all eligible biopsies. To determine whether WGS quality is dependent on the (primary) tumor type, a large-scale analysis was performed on the CPCT-02 sample cohort, for which samples were collected in 44 different hospitals. Of all of the samples sequenced by WGS ($n = 2921$), 86% passed all quality criteria ($n = 2520$), with a

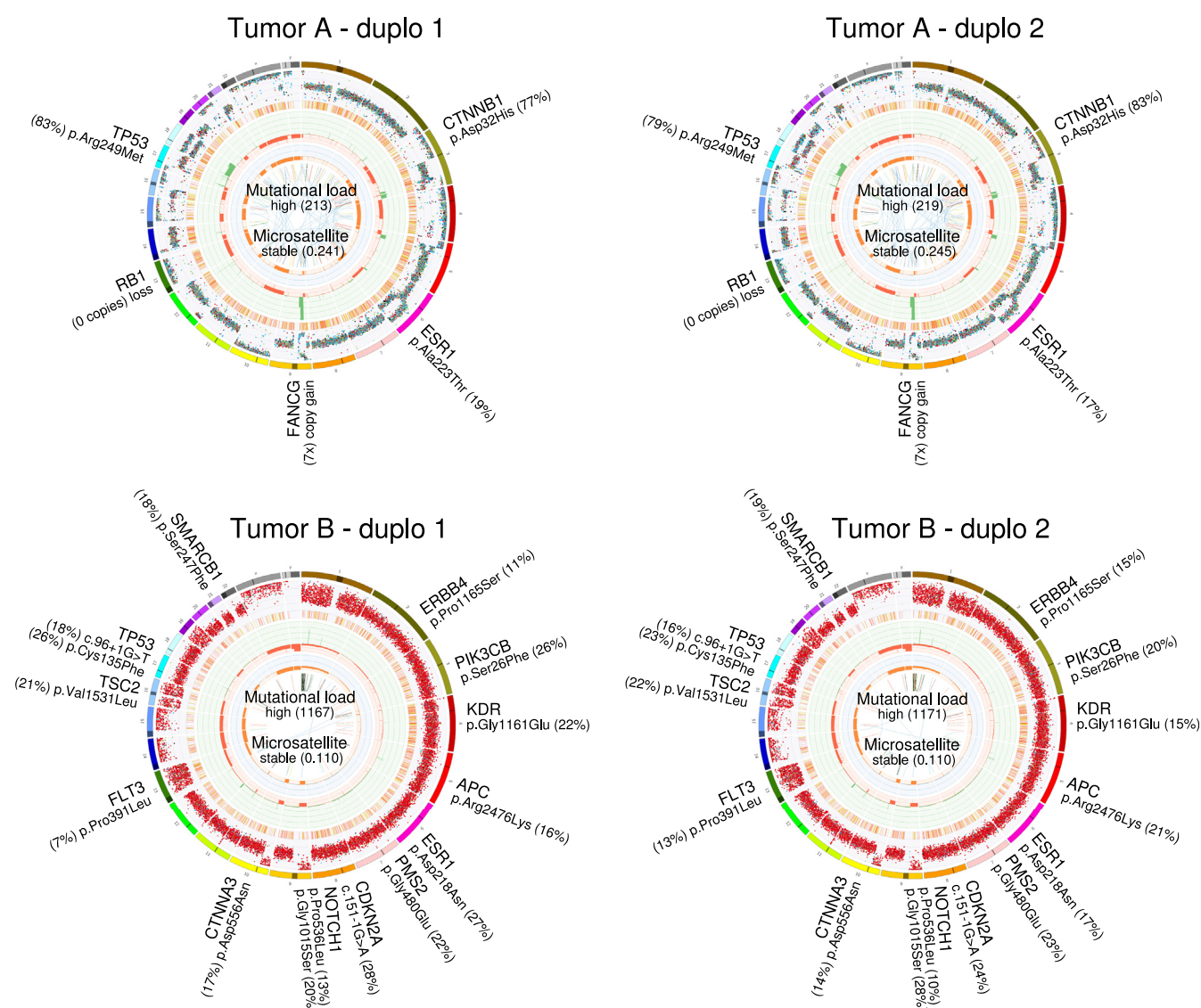


Figure 1 Representation of all tumor-specific DNA aberrations as detected using whole genome sequencing (WGS). For each case, the complete CIRCOS is shown as well as the reported genomics events, including the mutational burden and microsatellite readout. WGS is performed in duplicate (duplo 1 and duplo 2) for two tumor samples (tumor A, non—small-cell lung cancer; and tumor B, undifferentiated pleomorphic sarcoma).

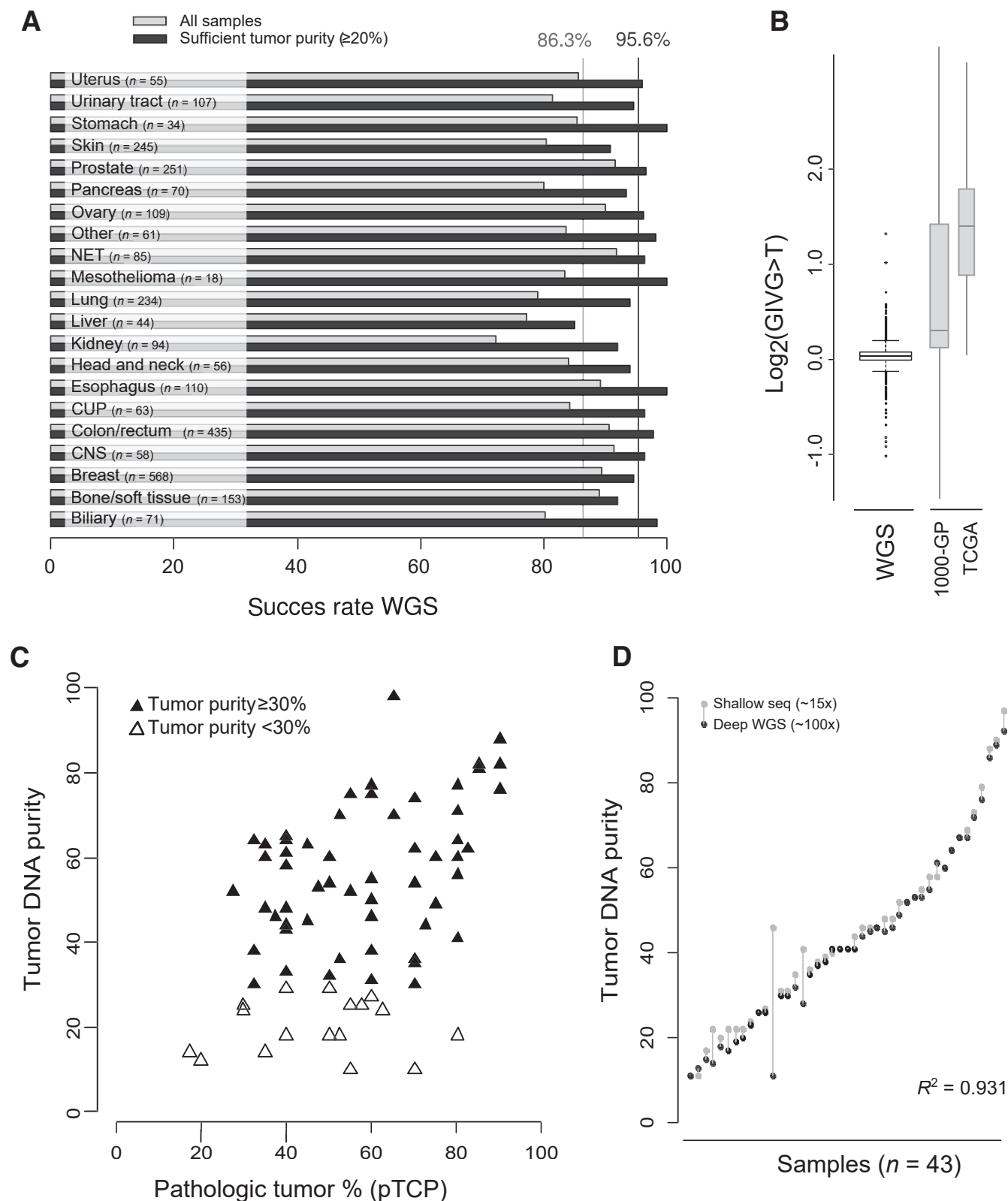


Figure 2 **A:** Whole genome sequencing (WGS) success rates for different primary tumor types. Success rates are shown for all samples and for samples that have sufficient tumor content. The average overall success rate across all tumor types is indicated by the vertical lines. **B:** Global Imbalance Value G to T scores ($GIV_{G>T}$). As a reference, the $GIV_{G>T}$ score range is depicted for the 1000 Genomes Project (1000-GP) and The Cancer Genome Atlas (TCGA) subset that are described previously by Do and Debrovic.¹⁸ **C:** Comparison of pathologic tumor percentage scoring (pTCP) with sequencing (seq)-based tumor DNA purity. **D:** Comparison of tumor purity assessment using shallow sequencing (gray; approximately 15x) and based on deep whole genome sequencing (black; approximately 100x). $n = 2520$ (**B**); $n = 43$ (**D**). CNS, central nervous system; CUP, cancers with unknown primary; NET, neuro-endocrine tumor.

lower success rate for kidney (72.3%), liver (77.3%), and lung (79.1%) cancer patients (Figure 2A).

Damaged DNA can cause lower-quality sequencing data, as previously described for DNA isolated from FFPE material.¹⁸ Although damage was expected to be much lower for fresh-frozen samples, the previously described Global Imbalance Value (GIV) score¹⁹ was used to directly assess this. The GIV scores are indicative of DNA damage (typically due to oxidation of deoxyguanosine to 7,8-dihydro-8-oxoguanine) with completely undamaged samples having a $GIV_{G>T}$ score of 1 and severely damaged samples with $GIV_{G>T}$ scores >1.5 , resulting in a large excess of false-positive $G>T$ variants due to technical artifacts.¹⁹ The analyzed set of 2520 samples showed low $GIV_{G>T}$ scores, with a median of only 1.02 (range, 0.495 to 2.495) and only three samples (0.11%) with a GIV score >1.5 (Figure 2B). In comparison, 41% of the 1000 Genomes Project samples showed a $GIV_{G>T}$ score of at least 1.5, whereas 73% of The Cancer Genome Atlas samples (also including FFPE-based samples) showed a $GIV_{G>T}$ score >2 .¹⁹

Correct assessment of a sample's tumor purity is essential for accurate determination of tumor-specific allele frequencies and copy number values. Both manual pathologic (pTCP) as well as molecular/DNA-based (mTCP) assessment were performed and using the same fresh-frozen biopsy to minimize potential heterogeneity. The pTCP and mTCP scoring showed a modest but significant correlation for samples with higher tumor content ($r = 0.40$; $P = 0.002$), but this association was absent for samples with lower ($<30\%$) tumor purity ($r = 0.08$; $P = 0.76$) (Figure 2C). Additional investigation of sections that were collected before and after cutting the sections (20 tumor samples) revealed intratumor heterogeneity but could not explain all differences (Supplemental Table S4). Possibly, the amount of tumor-infiltrating lymphocytes (each harboring a genome but difficult to quantify in histologic slides) plays a role in the observed differences between mTCP and pTCP, especially for tumors with fewer tumor cells.

An insufficient amount of tumor cells was the most prevalent failure rate despite prior pathologic prescreening (pTCP $> 20\%$ to 30%): 6.4% of samples showed an mTCP between 5% and 20% and 2.9% showed a seemingly absence of tumor DNA (mTCP $< 5\%$). In case reliable pTCP assessment was not available, mTCP calculations based on shallow sequencing data (approximately $8\times$ to $15\times$ average coverage) could be used for prescreening of biopsies eligible for deep sequencing. Comparison of mTCPs by shallow and deep WGS (approximately $90\times$ to $110\times$) showed a good correlation (R^2 of 0.931; $n = 43$) (Figure 2D), with an average deviation between both purities of only 3.2% (range, 0% to 35%, caused by an outlying non-small-cell lung cancer case). This result showed that shallow sequencing data were sufficiently reliable for mTCP-based estimations and could be used as an alternative for histopathologic assessment. When

samples were selected with sufficient mTCP ($\geq 20\%$) and sufficient DNA yield (>50 ng), the technical success rate for generating high-quality WGS data and reportable outcomes was 95.6% (Figure 2A).

SNVs, Indels, and Mutational Burden

Confirmation of variants detected by WGS was initially assessed by tailored smMIP panel sequencing.^{13,14} Across 29 samples, 192 randomly selected variants (165 SNVs and 27 indels, including passenger and driver variants) were sequenced and analyzed by a custom-designed smMIP panel (no reliable panel design was possible for 17.6% of the initial selected WGS variants, mainly because of the vicinity of intergenic repeat regions). Nearly all (98.4%) of the variants were confirmed by smMIP sequencing, indicating a high sensitivity (recall) of the smMIP assay and a high true positive rate of WGS. The observed variant allele frequencies showed a high correlation ($R^2 = 0.733$) between both assays (Figure 3A). Three variants could not be confirmed by smMIP: one intergenic SNV (chr3:75887550G $>$ C) due to a double mutation at that position for which the smMIP panel called the other variant (chr3:75887550G $>$ T) and two intergenic indels (chr8:106533360_106533361insAC and chr12:125662751_125662752insA).

Orthogonal clinical validation of mutations in a specific oncogene, *BRAF*, was performed using 48 selected samples and compared against the custom-made Oncomine gene-panel NGS assay (Thermo Scientific). Twenty-five samples showed a *BRAF* exon 15 or exon 11 mutation by WGS that were confirmed by panel NGS (Figure 3C). Vice versa, 26 *BRAF* mutations that were detected using panel-based sequencing were also identified using WGS. A single *BRAF* p.Gly469Ala mutation identified by panel NGS was not confirmed using the WGS analysis because of low mutation frequency (approximately 2%). WGS identified two less common *BRAF* variants (p.Ala762Val and p.Pro403fs) that were not covered by the used panel design. Both variants were unlikely to result in *BRAF* activation and were likely passenger variants, especially because both tumors were MSI with a high TMB. All other 20 *BRAF* wild-type samples by WGS were confirmed by panel sequencing.

Next, all somatic nonsynonymous mutations across the NGS panel design were evaluated (25.2 kb covering hotspot exons of 40 genes). Combined with the *BRAF* results, in total 138 mutations (121 SNVs and 17 indels) were detected by at least one of the tests, of which 136 were reported by WGS and 133 using panel sequencing (Figure 3C and Supplemental Table S5), resulting in an overall 98.5% sensitivity (recall) and 95.6% precision (positive predictive value) for WGS compared with panel based. A *PTEN* p.Lys327Arg mutation that was identified using the panel was not reported by the WGS test. Re-analysis of the WGS read data confirmed the presence of this variant with a lower

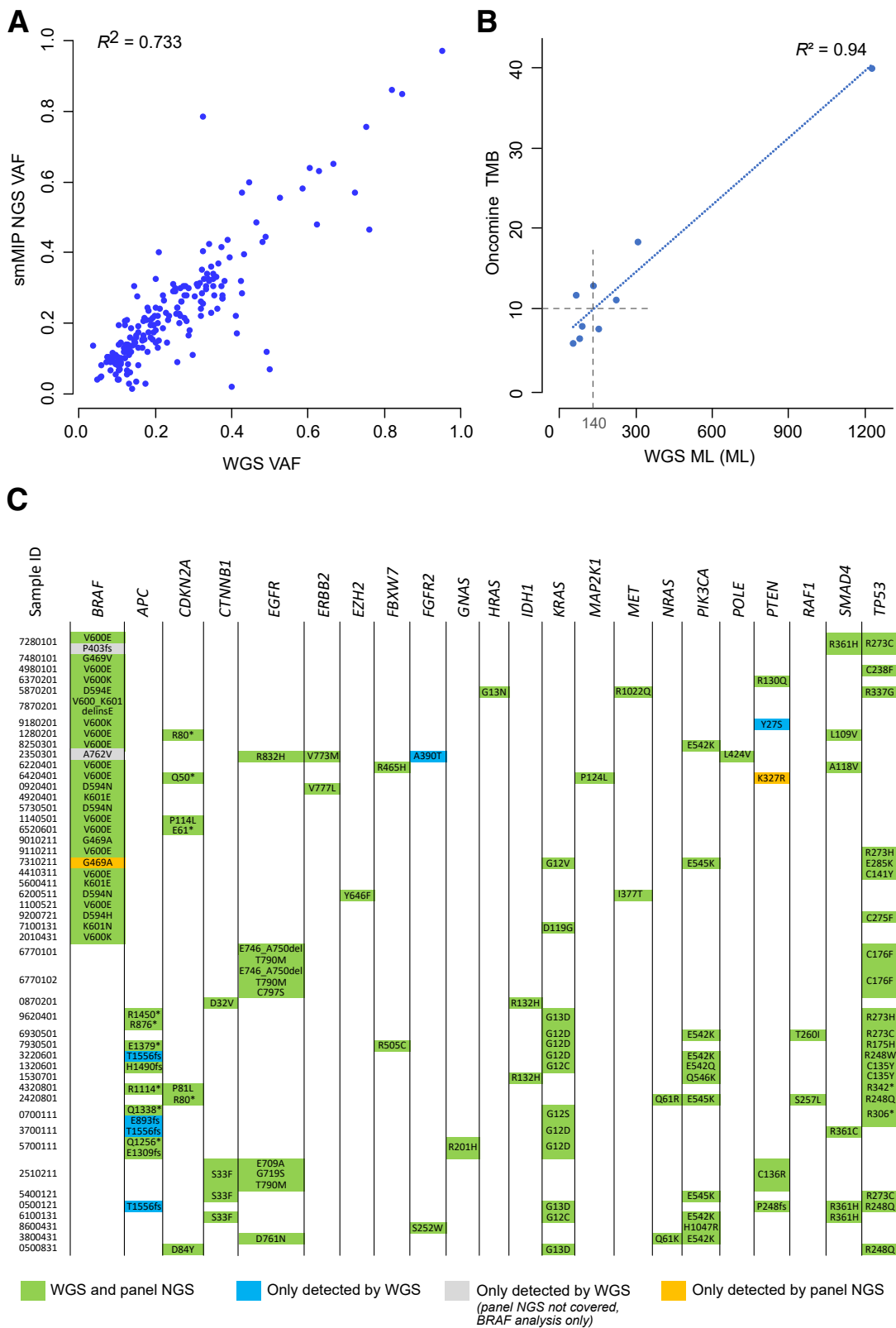


Figure 3 **A:** Variant allele frequencies (VAFs) for single-nucleotide variants, multinucleotide variants, and short insertion and deletion variants that are detected using whole genome sequencing (WGS) and confirmed by single-molecule molecular inversion probe (smMIP) next-generation sequencing (NGS) panel sequencing. **B:** Comparison of WGS-based mutational load (ML) readout with NGS panel-based tumor mutational burden (TMB). **C:** Overview of all protein-changing mutations that are detected by WGS and/or the custom-made OncoPrint NGS assay. Mutations reported by both assays are marked in green, variants only reported by WGS are marked in blue, and those reported only using the panel NGS assay are marked in orange. For *BRAF*, mutations detected by WGS but which are not included in the panel assay design are shown (in gray). For all other genes, only mutations included in the panel design are considered. ID, identifier.

VAF in the tumor (7% with a coverage of 8 of 116 reads) but also with reduced coverage in the blood reference. This combination affected the Bayesian somatic variant calling algorithm (which depends on information from both the tumor and normal reference samples) and as a consequence no somatic variant could be reliably called. On the contrary, the panel assay did not report a pathogenic *PTEN* variant (p.Tyr27Ser), which was identified by WGS (VAF, 12%) using the same input DNA. The variant was present in the

NGS panel data (VAF, 6%) but was not reported because of incorrect manual curation. The panel missed identifying the *APC* p.Thr1556fs inactivating mutation in three samples. This *APC* codon lies within a homopolymeric DNA region, and the IonTorrent sequencing technology used for the panel sequencing is known to face more difficulties in repetitive DNA regions. Considering the *APC* p.Thr1556fs as true positive results, the WGS precision (positive predictive value) was recalculated as 97.8%

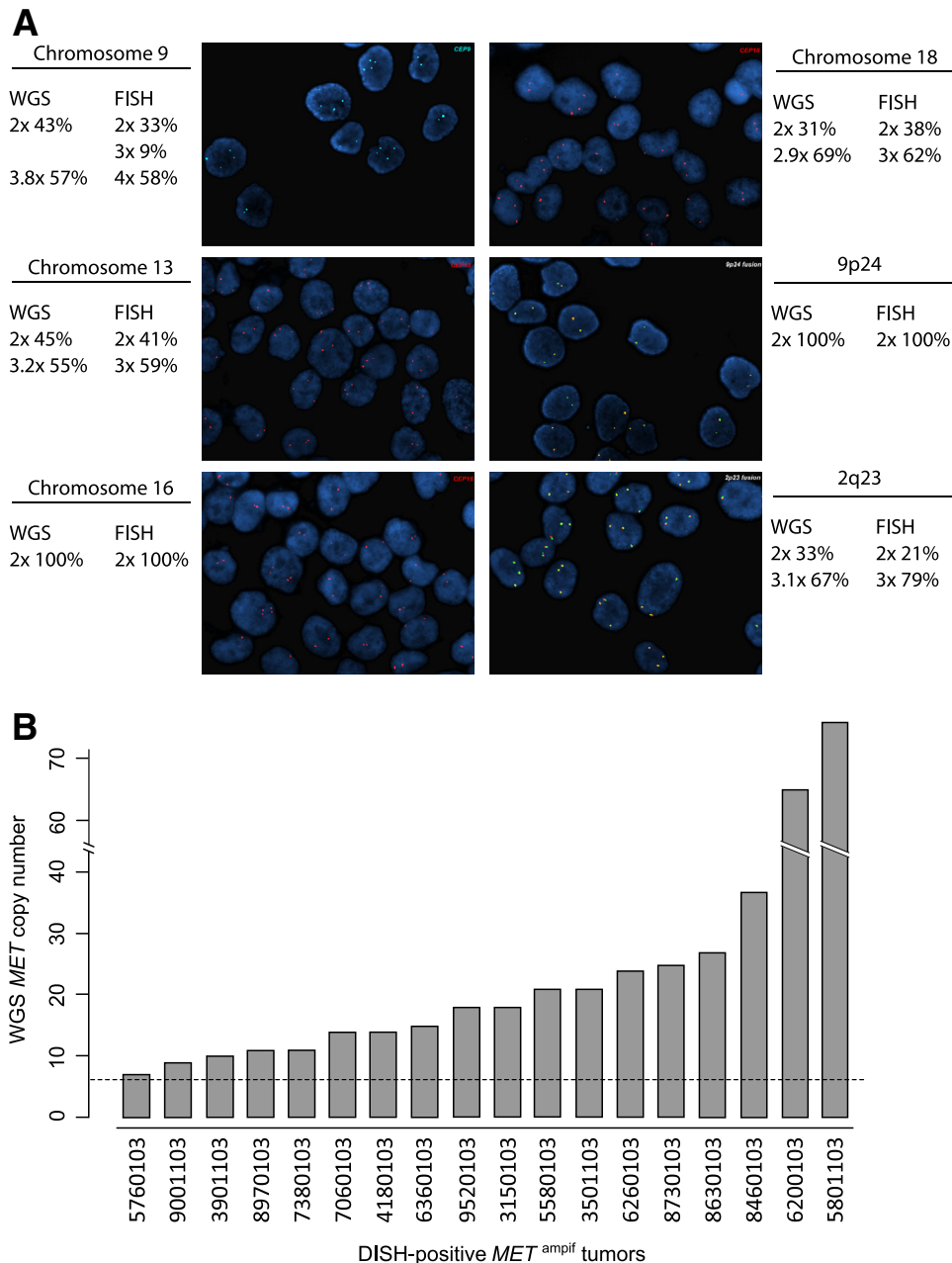


Figure 4 **A:** Comparison of COL0829 copy number analysis based on whole genome sequencing (WGS) and using fluorescent *in situ* hybridization (FISH) probes for copy number assessment of chromosomes 9, 13, 16, and 18, and for 9p24 (*CD274/PDCD1LG2*) and 2q23 (*ALK*). For both tests, the copy number as well as the percentage of tumor cells are determined. **B:** WGS-based copy number readouts of *MET* of 18 tumor samples that were considered positive for *MET* amplification by routine dual *in situ* hybridization (DISH) analysis. The dashed horizontal line represents the 6× copy threshold.

Although the performance of tumor ML estimations is directly following the performance of accurate non-synonymous variant calling (analytically, ML is only a simple summation of the observed variants), mutational burden readout was compared on 10 additional samples between WGS and Oncomine Tumor Mutational Load assay (Thermo Scientific). Both readouts showed a high correlation ($R^2 = 0.94$), but this was mainly caused by a single high ML sample ($ML > 1200$) (Figure 3B). Binary classification based on both tests (WGS-based ML cutoff of 140 mutations versus tumor mutational load-based TMB cutoff of 10 mutations/Mb) indicated a concordance for seven of nine samples (one sample was not evaluable by Oncomine Tumor Mutational Load assay), but also indicated a lower correlation in the cutoff region ($R^2 = 0.16$ when excluding two highest ML/TMB samples). This result illustrated the challenge of accurate mutational burden readout using a more limited gene panel compared with exome or genome-wide measurements, as discussed elsewhere.^{20,21}

Summarized, orthogonal panel NGS validation indicated a high overall sensitivity (recall; 98.5%) and a high precision (positive predictive value; 97.8%) for detection of variants by WGS [SNV ($n = 121$): 98.3% and 98.3%; indels ($n = 17$): 100% and 94.1%, respectively] using fresh-frozen biopsies with $\geq 20\%$ tumor purity, which was similar compared with commonly used panel-based approaches on FFPE material.²²

Copy Number Alterations

WGS chromosomal ploidy and copy number were initially benchmarked against FISH analysis on six genomic

locations of COLO829 [centromeric region of chromosomes 9, 13, 16, and 18, and 2q23 (*ALK*) and 9p24 (*CD274/PDCD1LG2*)]. WGS and FISH analysis showed highly similar purity and ploidy calculations, with chromosome 9 showing $4\times$ in approximately 55% of cells, chromosome 13 showing $3\times$ in approximately 55%, chromosome 18 showing $3\times$ in approximately 60%, 2q23 locus showing $3\times$ in 70% to 80%, and complete diploid chromosome 16 and 9q24 locus for all cells (Figure 4A).

Orthogonal validation of *ERBB2* (*Her2/neu*) amplification detection was performed using 16 samples from various tumor types (Supplemental Table S1) and including samples with weak and strong amplification levels. WGS *ERBB2* copy gains $>6\times$ were considered as actionable amplifications based on previous experience in the CPCT-02 study⁴ and because this cutoff value is used as eligibility criteria in the Dutch DRUP trial (NCT02925234) for various genes in a pan-tumor setting (eg, *EGFR*, *ERBB2*, *MET*, and *FGFR1*).²³ Matching fresh-frozen sections were analyzed by *ERBB2* FISH at an independent routine pathology laboratory (Table 2). For one sample (number 8700401), FISH analysis failed because of insufficient tumor cells (confirmed by immunohistochemistry); the other FISH results were considered representative. All samples with a WGS copy number $>6\times$ were confirmed by FISH to harbor substantial *ERBB2* amplified signals. For copy numbers between 2 and 6, at best an *ERBB2* gain was observed by FISH but considered insufficient for amplification (classified as *ERBB2* gain or equivocal). A borderline discordant *ERBB2* status was observed for a single case (sample number 5550101; FISH, $2\times$ to $4\times$ in 82% of the cells compared with WGS, $6\times$). No technical explanation could be identified, but this might be caused because of tumor

Table 2 *ERBB2* Copy Number Analysis by WGS and FISH

Sample ID	WGS <i>ERBB2</i>	FISH <i>HER2</i> normal, %	FISH <i>HER2</i> 2–4, %	FISH <i>HER2</i> 4–6, %	FISH <i>HER2</i> >6, %	FISH classification
3000421	9×	11	53	7	30	Amplification
1810501	9×	8	34	10	48	Amplification
1300611	8×	12	31	7	51	Amplification
8720501	8×	2	32	15	50	Amplification
1210601	71×	2	34	10	54	Amplification
9200111	45×	5	10	8	76	Amplification
3200111	43×	1	18	4	77	Amplification
3300211	25×	6	41	8	45	Amplification
8740201	8×	3	22	10	65	Amplification
5550101	6×	10	82	7	1	No amplification (equivocal)
3660101	5×	4	59	37	0	No amplification (gain)
7000711	4×	10	20	68	2	No amplification (equivocal)
7100331	2×	11	63	25	0	No amplification (gain)
0200111	4×	28	44	24	4	No amplification (gain)
5200221	6×	52	46	3	0	No amplification

ERBB2 (*HER2*) FISH results were scored solely on tumor cells and categorized as normal signals, 2 to 4 signals, 4 to 6 signals, and >6 *ERBB2* signals (according to guidelines¹⁶). For WGS as well as FISH, only absolute copy numbers/counts are used.

FISH, fluorescent *in situ* hybridization; ID, identifier; WGS, whole genome sequencing.

Table 3 HRD Using the CHORD Signature for 16 Tumors That Showed Complete Loss of the *BRCA1* ($n = 1$) or *BRCA2* ($n = 15$) Gene by WGS

Sample ID	Tumor type	Gene deletion (0 copies)	CHORD score	CHORD status
5000122	Breast	<i>BRCA2</i>	1.000	HRD
2410601	Ovary	<i>BRCA1</i>	0.992	HRD
4610801	Urothelial tract	<i>BRCA2</i>	0.920	HRD
3001101	Prostate	<i>BRCA2</i>	0.998	HRD
1110601	Pancreas	<i>BRCA2</i>	0.978	HRD
7300331	Breast	<i>BRCA2</i>	0.902	HRD
9930101	Prostate	<i>BRCA2</i>	0.940	HRD
3000712	Bile duct	<i>BRCA2</i>	0.940	HRD
7900102	Prostate	<i>BRCA2</i>	1.000	HRD
2040701	Prostate	<i>BRCA2</i>	0.986	HRD
6620401	Breast	<i>BRCA2</i>	0.956	HRD
6000921	Prostate	<i>BRCA2</i>	0.912	HRD
7330701	Adrenal gland	<i>BRCA2</i>	0.880	HRD
0100711	Bile duct	<i>BRCA2</i>	0.750	HRD
1120701	Prostate	<i>BRCA2</i>	1.000	HRD
3820401	Prostate	<i>BRCA2</i>	0.988	HRD

A CHORD score >0.50 is indicative for HRD.⁸

HRD, homologous repair deficiency; ID, identifier; WGS, whole genome sequencing.

heterogeneity between the sections used for WGS and FISH. Of note, this specific case involved a colorectal tumor for which the FISH assay is less common in routine practice.

Additional evidence for accurate WGS copy number detection was obtained for *MET*, using 18 samples (Supplemental Table S1) that had been independently scored as positive for *MET* amplification by dual *in situ* hybridization analysis during routine diagnostics. All 18 cases showed WGS-based *MET* copy numbers >6 with a large range from 7 to 76 copies and a median 23 copies (Figure 4B). Combined, the *ERBB2* and *MET* data showed a high concordance between WGS and ISH analysis (97.0%; 32 of the 33 cases), indicating that WGS reliably detected sufficiently high gene amplifications. For lower gains, the concordance showed more variability, but the question remains whether such low gains are biologically and/or clinically relevant.²⁴

To validate the detection of complete bi-allelic loss of genes by WGS, results regarding the presence of complete loss of *CDKN2A* were compared with routine p16 IHC data of 39 samples. Twenty-two samples with no (zero) intact copies of *CDKN2A* in the tumor cells, as detected by WGS (corrected for tumor purity), were all confirmed negative for p16 expression by IHC (100%; 22 of 22) (Supplemental Table S6). The 17 samples with presence of wild-type *CDKN2A* according to WGS (at least 1 intact wild-type allele) were all found to be positive for p16 IHC. Furthermore, for samples in the 2520 tumor cohort that showed complete loss of all *BRCA1* or *BRCA2* alleles according to WGS, a characteristic HRD profile⁸ was present in all cases (16 of 16) (Table 3), thereby confirming complete bi-allelic *BRCA* inactivation. Of note, this type of bi-allelic *BRCA* inactivation due to complete deletion is challenging to detect reliably by panel NGS, as the PCR amplicon libraries

Table 4 Fusion Genes Detected by WGS and the Archer FusionPlex on Matching DNA and RNA

Gene fusion details	WGS	Archer NGS	N	Sample ID
No fusion	None	None	7	7650101, 1380101, 0590101, 2310011, 9010211, 2510211, 4900321
<i>EIF2AK2</i> ex12– <i>ALK</i> ex3	Yes	Yes	1	1460201
<i>EML4</i> ex13– <i>ALK</i> ex20	Yes	Yes	5	4980101, 4980102, 7190101, 5120401, 5120402
<i>EML4</i> ex2– <i>ALK</i> ex18	Yes	Yes	1	9320501
<i>EML4</i> ex6– <i>ALK</i> ex20	Yes	Yes	2	6690101, 3430201
<i>SPAG17</i> ex20– <i>ALK</i> ex9	Yes	None	1	7330501
<i>EZR</i> ex10– <i>ROS1</i> ex34	Yes	Yes	1	4500401
<i>GOPC</i> ex8– <i>ROS1</i> ex35	Yes	Yes	2	2080101, 1410801
<i>MEF2D</i> ex1– <i>NTRK1</i> ex2	None	Yes	1	3190101
<i>PTPRF</i> ex11– <i>NRG1</i> ex6	Yes	Yes	1	1100631
<i>TRPS1</i> ex1– <i>NRG1</i> ex2	Yes	Yes	1	4100511
<i>TMPRSS2</i> ex2– <i>ERG</i> ex3	Yes	N/A*	1	0530701
Total			24	

*TMPRSS2-ERG fusions are not included in the used Archer FusionPlex assay.

Ex, exon; ID, identifier; N/A, not applicable; NGS, next-generation sequencing; WGS, whole genome sequencing.

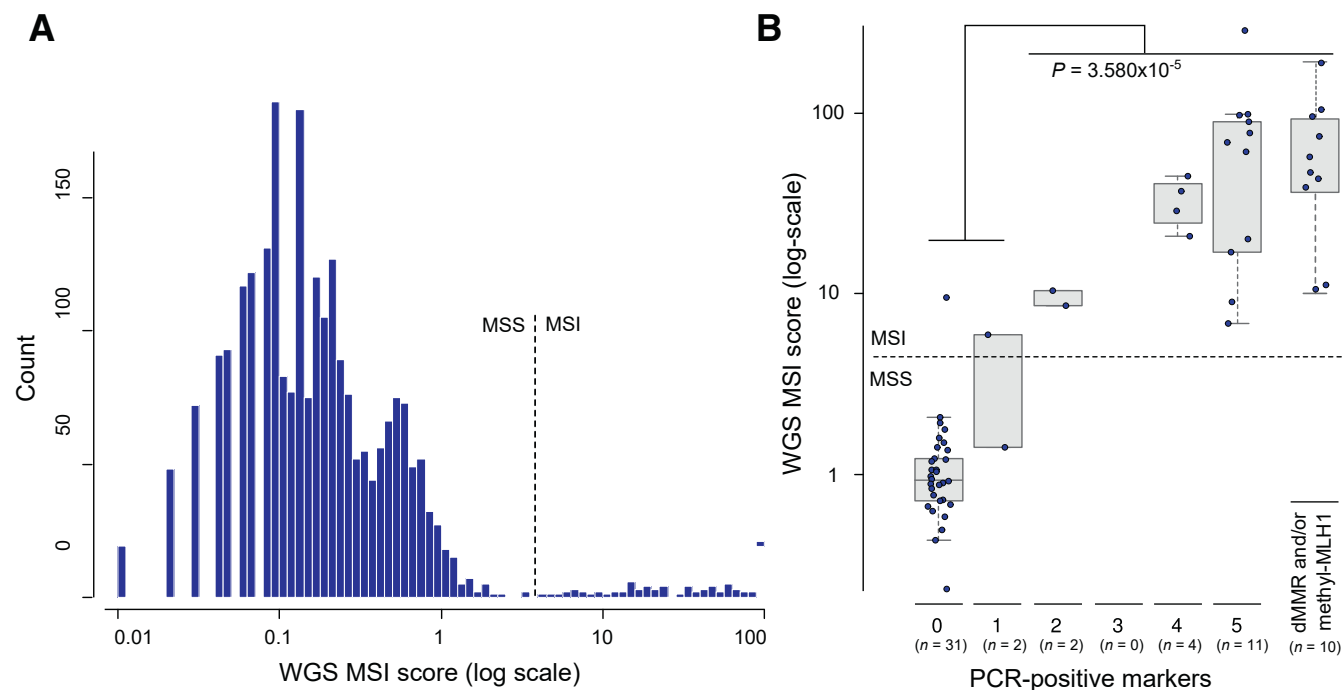


Figure 5 **A:** Whole genome sequencing (WGS)—based microsatellite instability (MSI) quantification across a cohort of 2520 metastatic cancer samples. **B:** WGS MSI readout compared with the five-marker PCR-based test using an independent set of 60 validation samples. dMMR, deficient mis-match repair; methyl-MLH1, methylated *MLH1* gene; MSS, microsatellite stable.

are in such cases based on wild-type *BRCA* alleles from the normal cells in the tissue samples.

Fusion Genes

Detection of gene fusions by WGS was compared with results obtained with an RNA-based Anchored Multiplex PCR NGS assay (ArcherDx) and was performed independently on 24 samples using matching DNA and RNA from the same biopsy. Samples were selected on the basis of the WGS results to include one or more clinically relevant fusion genes. The Archer NGS assay confirmed the WGS findings for 21 of the 23 samples (91.3%), including fusion of *ALK*, *NRG1*, and *ROS1* (Table 4). For one sample, no comparison could be made, as the *TMPRSS2-ERG* fusion was not covered by the used Archer FusionPlex assay.

An *NTRK1* fusion detected by Archer NGS (*MEF2D-NTRK1*: 22 reads; 60% VAF) could not be identified using WGS, possibly due to a complex structural variation pattern involving multiple break junctions in the intronic regions and thus more difficult to call using WGS data compared with analysis of RNA. Vice versa, one fusion (*SPAG17-ALK*) detected by WGS showed no evidence in the tumor RNA. Although based on fusion at DNA level, a viable in-frame fusion protein was predicted, it can well be that the corresponding RNA was expressed at low levels (eg, due to temporal or spatial expression variation) that are insufficient for reliable detection by the Archer assay.

Microsatellite Instability

WGS microsatellite (in)stability classification was validated independently using 60 samples, including multiple tumor types (Supplemental Table S1), and compared with the routinely used five-marker PCR MSI panel^{17,25} (50 samples) or compared with MMR/MLH1-methylation analysis (10 samples). Assessment of MSI by WGS was defined as the number of small indels per million bases occurring in ≥ 5 -mer homopolymers and in dinucleotide, trinucleotide, and tetranucleotide repeats.⁶ The cohort of 2520 tumor samples showed an average MSI score of 1.11, with the vast majority of samples having a low score and a long tail toward higher MSI scores (range, 0.004 to 93) (Figure 5A). A total of 2.7% of the samples were classified as MSI using a cutoff of 4 (cutoff was based on the apparent binominal distribution of the MSI scores). On the validation set ($n = 60$), the sensitivity of WGS MSI classification was 100% (95% CI, 88.8%–100%) with a precision of 94% (95% CI, 84.8%–93.9%) and a Cohen κ score of 0.933 (95% CI, 0.732–0.933). In addition to the binary MSI/microsatellite stable concordance, the MSI score correlated with the number of positive PCR markers, in which samples with only one or two positive PCR markers showed a marginal MSI score (Figure 5B).

One of the two discordant cases was a lymphoma sample (number 2300211) with a complex pathology, showing one of the five positive PCR markers (classified

Table 5 Detection and Typing of HPV in Tumor Biopsies Using WGS and PCR Analysis

PCR HPV result	WGS HPV result	N	Sample ID
No high-risk HPV	No HPV detected	13	4600103, 6600103, 2810103, 0120103, 3120103, 5220103, 2720103, 1920103, 5530103, 6930103, 1570103, 1190103, 7411103
HPV high-risk type 16	HPV high-risk type 16	19	2350103, 1360103, 7201103, 9501103, 6601103, 5311103, 3101103, 6960101, 8980101, 7000103, 7100103, 2110103, 7410103, 0710103, 5720103, 4640103, 9950103
HPV high-risk type 18	HPV high-risk type 18	4	3990101, 2900103, 0750103, 0701103
HPV high-risk other	HPV high-risk type 16	2	9360103, 3790103
HPV high-risk other	HPV high-risk type 18	1	6920103
Total		37	

HPV, human papillomavirus; ID, identifier; WGS, whole genome sequencing.

as microsatellite stable) but a WGS score of 5.9 (classified as MSI). IHC analysis showed no substantial loss of MMR proteins, although WGS analysis indicated a somatic *PMS2* p.Ile193Met variant in combination with a likely inactivating *PMS2* structural variant. The p.Ile193Met mutation is classified with a high prior in the Leiden Open Variant Database (http://hci-lovd.hci.utah.edu/home.php?select_db=PMS2_priors, last accessed April 16, 2021) and thus likely represents a pathogenic variant. Both the MSI PCR test as well as the MMR IHC had not been validated for use in lymphoma cases, so a definitive conclusion remained difficult. The second discordant case (number 0740103), a colorectal cancer sample with a WGS MSI score of 9.7 but without a positive PCR marker (0.5 markers) showed a hypermutation phenotype (ML, 8050; TMB, 601) and harbored two *POLE* mutations (p.Phe1435Val and p.Ser459Phe). Although technically microsatellite stable by the routine PCR assay (and thus considered a discordant validation result), the sample was likely a hypermutator with a DNA repair deficiency.

Tumor-Genome Viruses

Recently, it has been shown that the presence of viruses can be detected with great accuracy using WGS.²⁶ Validation of viral detection focused on HPV because of the prevalence and clinical importance, and the availability of routine testing (eg, QIAscreen HPV PCR assay; Qiagen). Thirty-seven tumor samples were used for independent validation between WGS and PCR assay, including 24 HPV positive and 13 negative (Supplemental Table S1). WGS HPV status was in concordance with standard pathology assessment for all 37 cases (100% accuracy) with a Cohen κ score of 1.00 (95% CI, 0.70–1.00). HPV high-risk types were concordant between both tests for 21 of the 24 positive cases (Table 5). For three samples that were classified as high-risk other using the PCR assay, WGS analysis indicated either HPV type 16 (number 9360103, 3790103) or type 18 (number 6920103). This appeared more a classification than a detection error, but for which the cause remained elusive.

In addition to the orthogonal HPV validation, six samples, with presence of EBV viral DNA based on WGS analysis (human gammaherpesvirus 4; <https://www.ncbi.nlm.nih.gov/nuccore>; accession number NC_007605.1), were assessed by EBER IHC. Interestingly, the three cases with seemingly integrated EBV DNA by WGS scored positive for EBER, whereas the three nonintegrated EBV cases were scored as negative for EBER IHC.

Discussion

During the past few years, WGS and the associated data analysis and interpretation have matured from a research-use-only tool to a diagnostic-level technology.²⁷ Together with the clinical need to screen for an increasing number of biomarkers in an increased number of tumor types (or even pan-cancer)^{1,23} and the often-limited available biopsy tissue, the use of a single all-inclusive DNA test is a more than welcome development for efficient molecular diagnostics. Herein, we report on (retrospective) orthogonal validation efforts of WGS on fresh-frozen biopsies and show, to our knowledge for the first time, that the performance of WGS using biopsy with at least 20% tumor cells is equal to the range of routinely used diagnostic tests with technical concordances of >95%. Specifically, we show that a single WGS-based tumor-normal test can provide information regarding the following: i) actionable small variants (SNVs and indels; routinely detected by targeted panel tests); ii) gene amplifications (FISH); iii) fusion genes (FISH or RNA panels); iv) microsatellite instability (amplicon fragment analysis); v) viral infections (PCR); and vi) tumor mutational load determination (larger NGS panels). The turn-around time has been reduced toward a clinically acceptable maximum of 10 working days, comprising a minimal net processing time of 1 day sample registration and DNA isolation, 1 day library preparation, 2 days sequencing, and 2 days data analysis and report generation.

Currently, WGS still requires a tumor content that is higher than focused panel-based approaches (minimal, 20% for WGS versus 5% to 10% for panel NGS). This limitation

is due to a lower sequencing depth by WGS and its associated costs, but with ongoing developments, it is anticipated that WGS with approximately 250× coverage will become feasible for such samples in the next coming years.

The biggest challenge to start using WGS in routine practice is the need of appropriate samples as this will, for most hospitals, require an adaptation in the pathology laboratories that are currently mostly FFPE oriented. The feasibility of implementing WGS in routine practice is currently being evaluated in a prospective clinical validation study.²⁸

The high performance of WGS is primarily the result of two important aspects that are fundamentally different from most current diagnostics procedures for cancer. First, the use of fresh-frozen tumor material yields consistent high-quality DNA and sequencing results. Second, parallel processing of the patient's fresh blood sample to serve as a control/baseline for the matching tumor sample. This way, all germline variants can be automatically subtracted, and tumor-specific changes can be precisely pinpointed. Even across a focused set of approximately 500 cancer-associated genes,⁴ the bulk of all missense variants observed in the tumor are in fact inherited germline polymorphisms without clinical significance, making comprehensive tumor-only interpretation and filtering a daunting task. This challenge is not unique for WGS but in principle also applies for (large) NGS panels.^{29,30} Filtering of germline variants using population database information remains challenging because of various reasons (eg, biases toward White population and rare or subpopulation specific variants), and the impact on TMB measurements is likely large when germline and somatic variants cannot be discriminated accurately.

Bioinformatics and high-end reporting tools are essential for data-rich assay. Following WGS, the complex whole genome data and results should become manageable and understandable for the end users (eg, pathologists, medical oncologists, and treating physicians), requiring a delicate balance between what can be detected and what should be reported. To facilitate downstream interpretation, the WGS setup described herein ranks all observed nonsynonymous variants based on calculated oncogenic driver likelihoods.⁴ This strategy allows for a focus on the oncogenic high-driver events, while providing all information on likely passenger variants (median/low drivers). For reporting of gene amplification, the tumor's average ploidy is used as a filter to avoid reporting too many increased copy number events due to whole genome/chromosome duplications. Clinical annotation of the observed DNA aberrations (mutations in high-driver genes, fusions, and copy number variants) was performed by automatic integration of open-source knowledgebases (CIViC,³¹ OncoKB,³² and CGI³³) for which only evidence items with convincing clinical relevance (level A + B) were included. Information regarding potential active (and recruiting) clinical trials is integrated in the reporting using a curated national (Dutch)

clinical study registry (<https://iclusion.com>, last accessed February 1, 2021).

DNA sequencing tests are often performed as laboratory-developed tests (LDTs) and the technical parameters, validation requirements, and quality assurance are typically governed by national regulation and legislation that can differ. Various expert groups have drafted guidelines and recommendations for the standardization of multigene panel testing^{2,34–36}; and for our validation efforts, we have followed the guidelines for setup and validation of (new) sequencing tests in ISO-accredited pathologic laboratories in the Netherlands. However, with the ongoing approval of NGS panel assays by the Food and Drug Administration (<https://www.fda.gov/medical-devices/vitro-diagnostics/list-cleared-or-approved-companion-diagnostic-devices-vitro-and-imaging-tools>, last accessed February 1, 2021) and the upcoming new European regulations for *in vitro* diagnostic medical devices (*in vitro* diagnostics regulations 2017/746) in 2022,³⁷ it is anticipated that genome sequencing tests will become regulated following international guidelines, standardization, and quality schemes. Clinical validation, as described herein, by comparison with common standards will be a key component of such regulations.

With the increase in sequencing capabilities, the bioinformatics part has become essential for a good analysis and interpretation of the sequencing data of WGS but also for the emerging larger comprehensive panels. Traditionally, laboratories have focused most on the wet laboratory performance and automation but it has become clear that the downstream bioinformatics, and the data infrastructure to manage all data, are equally important. All of the analysis tools and reporting software should also meet the requirements of *in vitro* diagnostics regulations and applicable ISO regulations. For diagnostic purpose, local frozen snapshots of the software tool and/or knowledgebase are part of a diagnostic validation workflow for which updates can be periodically implemented. Production functionality is thus not impacted by changes in or continuity of external resources, but to keep up with developments in the field, such resources need to be ingested and maintained by a dedicated team to ensure diagnostic continuity.

With the rapid development of more targeted drugs and their associated biomarkers, in addition to standardization of the test results, it is important to be able to efficiently and quickly add new biomarkers/genes to the clinical reports (eg, *NRG1* and *NTRK* fusions and *PIK3CA* activating mutations). WGS will allow such a rapid and efficient codevelopment of (all) future diagnostic DNA markers, because it only requires an update of the bioinformatics and reporting aspects, without the need of laborious and costly new test developments or adaptations of panel designs, including the required laboratory analytical validation experiments. In addition, the data from previously tested patients can, in principle and on request from the treating physician, be reanalyzed for the presence of the new biomarkers and recontacting of the patient can be considered.³⁸

In addition to tumor-specific mutations, the WGS workflow, with the use of the patient's blood sample as a reference, allows the reporting of germline mutations. Depending on the institutional and/or local guidelines on germline reporting, the bioinformatics pipeline is sufficiently flexible to provide various levels of germline reporting. From blacklisting of all variants that are present in the patient's germline, to only reporting of pathogenic mutations that show a second hit or loss of heterozygosity in the tumor, or reporting of all pathogenic germline findings, including heterozygous carrier mutations. Furthermore, the set of genes to report on can be adjusted on the basis of the tumor type or the clinical context and consent of the patient.

In certain cases, comprehensive analysis of all genomic aberrations at DNA level might still be insufficient to provide a full picture of molecular tumor characteristics. For example, a gene fusion that is considered in frame and viable for a fusion protein might still not be expressed and could be considered a biological false-positive finding. Also, the presence of integrated viral DNA (measured as viral integration sites) does not always result in active viral gene expression.³⁹ In such situations, analysis of the transcriptome using whole transcriptome sequencing in addition to WGS can provide a better molecular characterization of a tumor.

Comprehensive DNA and/or RNA screening can likely also assist in a (more) detailed classification and diagnosis of tumor types. Currently, tumor classification still relies on histopathologic investigation, but progress has been made to also start using genomic classifications, especially in the context of rare cancer and cancers with unknown primary.^{40,41} We can envision a future in which WGS does not only provide information on possible treatment options but also provides another piece of the puzzle to resolve a complex diagnosis.

Setting aside the direct impact WGS can have for clinical use and comprehensive screening for clinical study eligibility, a whole-genome view of the tumor will yield a wealth of valuable research data and provide the opportunity to increase our insights in oncogenic processes and to better explain or predict the response to targeted or immunotherapy. A real-world evidence-based learning health care system will greatly enhance our understanding of cancer and facilitate the discovery of newly identified biomarkers, targeted therapies, and improved treatment decision making for patients.

Acknowledgments

We thank Peggy Atmodimedjo, Isabelle Meijssen, Ronald van Marion, and Hanna Schoep for assistance with data collection; Sandra van den Broek, Nina Jacobs, and David Koetsier for analyzing data; and Immy Riethorst for sample logistics. The data from Hartwig Medical Foundation and

the Center of Personalised Cancer Treatment supported research and publication of parts of this manuscript.

Author Contributions

P.R., E.B., and E.C. designed the study; P.R., E.B., S.L., and L.S. analyzed the whole genome sequencing data; P.R., M.C.B., H.J.D., W.R.R.G., F.H.G., A.H., M.M.H.H., M.E.G.K., W.W.J.L., K.M., M.G.M.R., K.G.S., M.S., and B.Y. analyzed and validated experiments; P.R., E.B., J.J.M.H., and E.C. wrote the manuscript; all authors approved submission and publication of the manuscript.

Supplemental Data

Supplemental material for this article can be found at <http://doi.org/10.1016/j.jmoldx.2021.04.011>.

References

- Hyman DM, Taylor BS, Baselga J: Implementing genome-driven oncology. *Cell* 2017, 168:584–599
- Mosele F, Remon J, Mateo J, Westphalen CB, Barlesi F, Lolkema MP, Normanno N, Scarpa A, Robson M, Meric-Bernstam F, Wagle N, Stenzinger A, Bonastre J, Bayle A, Michiels S, Bièche I, Rouleau E, Jezdic S, Douillard J-Y, Reis-Filho J, Dienstmann R, André F: Recommendations for the use of next-generation sequencing (NGS) for patients with metastatic cancers: a report from the ESMO Precision Medicine Working Group. *Ann Oncol* 2020, 31:1491–1505
- Weinstein JN; The Cancer Genome Atlas Research Network, Collisson EA, Mills GB, Mills Shaw KR, Ozenberger BA, Ellrott K, Shmulevich I, Sander C, Stuart JM: The Cancer Genome Atlas Pan-Cancer analysis project. *Nat Genet* 2013, 45: 1113–1120
- Priestley P, Baber J, Lolkema MP, Steeghs N, de Bruijn E, Shale C, Duyvesteyn K, Haidari S, van Hoeck A, Onstenk W, Roepman P, Voda M, Bloemendal HJ, Tjan-Heijnen VCG, van Herpen CML, Labots M, Witteveen PO, Smit EF, Sleijfer S, Voest EE, Cuppen E: Pan-cancer whole-genome analyses of metastatic solid tumours. *Nature* 2019, 575:210–216
- Manolio TA, Rowley R, Williams MS, Roden D, Ginsburg GS, Bult C, Chisholm RL, Deverka PA, McLeod HL, Mensah GA, Relling MV, Rodriguez LL, Tamburro C, Green ED: Opportunities, resources, and techniques for implementing genomics in clinical care. *Lancet* 2019, 394:511–520
- Huang MN, McPherson JR, Cutcutache I, Teh BT, Tan P, Rozen SG: MSIsq: software for assessing microsatellite instability from catalogs of somatic mutations. *Sci Rep* 2015, 5:13321
- Davies H, Glodzik D, Morganella S, Yates LR, Staaf J, Zou X, Ramakrishna M, Martin S, Boyault S, Sieuwerts AM, Simpson PT, King TA, Raine K, Eyfjord JE, Kong G, Borg Å, Birney E, Stunnenberg HG, van de Vijver MJ, Børresen-Dale A-L, Martens JWM, Span PN, Lakhani SR, Vincent-Salomon A, Sotiriou C, Tutt A, Thompson AM, Van Laere S, Richardson AL, Viari A, Campbell PJ, Stratton MR, Nik-Zainal S: HRDetect is a predictor of BRCA1 and BRCA2 deficiency based on mutational signatures. *Nat Med* 2017, 23:517–525
- Nguyen L, Martens JWM, Van Hoeck A, Cuppen E: Pan-cancer landscape of homologous recombination deficiency. *Nat Commun* 2020, 11:5584

9. Kim S, Scheffler K, Halpern AL, Bekritsky MA, Noh E, Källberg M, Chen X, Kim Y, Beyter D, Krusche P, Saunders CT: Strelka2: fast and accurate calling of germline and somatic variants. *Nat Methods* 2018, 15:591–594
10. Cameron DL, Schröder J, Penington JS, Do H, Molania R, Dobrovic A, Speed TP, Papenfuss AT: GRIDSS: sensitive and specific genomic rearrangement detection using positional de Bruijn graph assembly. *Genome Res* 2017, 27:2050–2060
11. Cameron DL, Baber J, Shale C, Papenfuss AT, Valle-Inclan JE, Besselink N, Cuppen E, Priestley P: GRIDSS, PURPLE, LINX: Unscrambling the tumor genome via integrated analysis of structural variation and copy number. *bioRxiv* 2020:781013
12. Cameron DL, Jacobs N, Roepman P, Priestley P, Cuppen E, Papenfuss AT: VIRUSBreakend: Viral Integration Recognition Using Single Breakends. *Bioinformatics* 2021:btab343
13. Eijkelenboom A, Kamping EJ, Kastner-van Raaij AW, Hendriks-Cornelissen SJ, Neveling K, Kuiper RP, Hoischen A, Nelen MR, Ligtenberg MJL, Tops BBJ: Reliable next-generation sequencing of formalin-fixed, paraffin-embedded tissue using single molecule tags. *J Mol Diagn* 2016, 18:851–863
14. Acuna-Hidalgo R, Sengul H, Steehouwer M, van de Vorst M, Vermeulen SH, Kiemeny LALM, Veltman JA, Gilissen C, Hoischen A: Ultra-sensitive sequencing identifies high prevalence of clonal hematopoiesis-associated mutations throughout adult life. *Am J Hum Genet* 2017, 101:50–64
15. Pruis MA, Geurts-Giele WRR, Von der TJH, Meijssen IC, Dinjens WNM, Aerts JGJV, Dingemans AMC, Lolkema MP, Paats MS, Dubbink HJ: Highly accurate DNA-based detection and treatment results of MET exon 14 skipping mutations in lung cancer. *Lung Cancer* 2020, 140:46–54
16. Wolff AC, Elizabeth Hale Hammond M, Allison KH, Harvey BE, Mangu PB, Bartlett JMS, Bilous M, Ellis IO, Fitzgibbons P, Hanna W, Jenkins RB, Press MF, Spears PA, Vance GH, Viale G, McShane LM, Dowsett M: Human epidermal growth factor receptor 2 testing in breast cancer: American Society of Clinical Oncology/College of American Pathologists Clinical Practice Guideline focused update. *J Clin Oncol* 2018, 36:2105–2122
17. van Lier MGF, Wagner A, van Leerdam ME, Biermann K, Kuipers EJ, Steyerberg EW, Dubbink HJ, Dinjens WNM: A review on the molecular diagnostics of Lynch syndrome: a central role for the pathology laboratory. *J Cell Mol Med* 2010, 14:181–197
18. Do H, Dobrovic A: Sequence artifacts in DNA from formalin-fixed tissues: causes and strategies for minimization. *Clin Chem* 2015, 61:64–71
19. Chen L, Liu P, Evans TC Jr, Ettwiller LM: DNA damage is a pervasive cause of sequencing errors, directly confounding variant identification. *Science* 2017, 355:752–756
20. Mankor JM, Paats MS, Groenendijk FH, Roepman P, Dinjens WNM, Dubbink HJ, Sleijfer S, CPCT Consortium, Cuppen E, Lolkema MPJK: Impact of panel design and cut-off on tumour mutational burden assessment in metastatic solid tumour samples. *Br J Cancer* 2020, 122:953–956
21. Budczies J, Allgäuer M, Litchfield K, Rempel E, Christopoulos P, Kazdal D, Endris V, Thomas M, Fröhling S, Peters S, Swanton C, Schirmacher P, Stenzinger A: Optimizing panel-based tumor mutational burden (TMB) measurement. *Ann Oncol* 2019, 30:1496–1506
22. Williams HL, Walsh K, Diamond A, Oniscu A, Deans ZC: Validation of the OncoPrint™ focus panel for next-generation sequencing of clinical tumour samples. *Virchows Arch* 2018, 473:489–503
23. van der Velden DL, Hoes LR, van der Wijngaart H, van Berge Henegouwen JM, van Werkhoven E, Roepman P, Schilsky RL, de Leng WWJ, Huitema ADR, Nuijen B, Nederlof PM, van Herpen CML, de Groot DJA, Devriese LA, Hoebe A, de Jonge MJA, Chalabi M, Smit EF, de Langen AJ, Mehra N, Labots M, Kapiteijn E, Sleijfer S, Cuppen E, Verheul HMW, Gelderblom H, Voest EE: The drug rediscovery protocol facilitates the expanded use of existing anticancer drugs. *Nature* 2019, 574:127–131
24. Eijkelenboom A, Tops BBJ, van den Berg A, van den Brule AJC, Dinjens WNM, Dubbink HJ, Ter Elst A, Geurts-Giele WRR, Groenen PJTA, Groenendijk FH, Heideman DAM, Huibers MMH, Huijsmans CJJ, Jeuken JWM, van Kempen LC, Korpershoek E, Kroeze LI, de Leng WWJ, van Noesel CJM, Speel E-JM, Vogel MJ, van Wezel T, Nederlof PM, Schuurin E, Ligtenberg MJL: Recommendations for the clinical interpretation and reporting of copy number gains using gene panel NGS analysis in routine diagnostics. *Virchows Arch* 2019, 474:673–680
25. Patil DT, Bronner MP, Portier BP, Fraser CR, Plesec TP, Liu X: A five-marker panel in a multiplex PCR accurately detects microsatellite instability-high colorectal tumors without control DNA. *Diagn Mol Pathol* 2012, 21:127–133
26. Zapatka M, Pathogens P, Borozan I, Brewer DS, Iskar M, Grundhoff A, Alawi M, Desai N, Sültmann H, Moch H, Cooper CS, Eils R, Ferretti V, Lichter P; PCAWG Consortium: The landscape of viral associations in human cancers. *Nat Genet* 2020, 52:320–330
27. Wrzeszczynski KO, Felice V, Abhyankar A, Kozon L, Geiger H, Manaa D, London F, Robinson D, Fang X, Lin D, Lamendola-Essel MF, Khaira D, Dikoglu E, Emde A-K, Robine N, Shah M, Arora K, Basturk O, Bhanot U, Kentsis A, Mansukhani MM, Bhagat G, Jobanputra V: Analytical validation of clinical whole-genome and transcriptome sequencing of patient-derived tumors for reporting targetable variants in cancer. *J Mol Diagn* 2018, 20:822–835
28. Samsom KG, Bosch LJW, Schipper LJ, Roepman P, de Bruijn E, Hoes LR, Riethorst I, Schoenmaker L, van der Kolk LE, Retèl VP, Frederix GWJ, Buffart TE, van der Hoeven JJM, Voest EE, Cuppen E, Monkhorst K, Meijer GA: Study protocol: whole genome sequencing implementation in standard diagnostics for every cancer patient (WIDE). *BMC Med Genomics* 2020, 13:169
29. Jones S, Anagnostou V, Lytle K, Parpart-Li S, Nesselbush M, Riley DR, Shukla M, Chesnick B, Kadan M, Papp E, Galens KG, Murphy D, Zhang T, Kann L, Sausen M, Angiuoli SV, Diaz LA Jr, Velculescu VE: Personalized genomic analyses for cancer mutation discovery and interpretation. *Sci Transl Med* 2015, 7:283ra53
30. Cheng DT, Prasad M, Chekaluk Y, Benayed R, Sadowska J, Zehir A, Syed A, Wang YE, Somar J, Li Y, Yelskaya Z, Wong D, Robson ME, Offit K, Berger MF, Nafa K, Ladanyi M, Zhang L: Comprehensive detection of germline variants by MSK-IMPACT, a clinical diagnostic platform for solid tumor molecular oncology and concurrent cancer predisposition testing. *BMC Med Genomics* 2017, 10:33
31. Barnell EK, Waalkes A, Mosior MC, Penewit K, Cotto KC, Danos AM, Sheta LM, Campbell KM, Krysiak K, Rieke D, Spies NC, Skidmore ZL, Pritchard CC, Fehniger TA, Uppaluri R, Govindan R, Griffith M, Salipante SJ, Griffith OL: Open-sourced CIViC annotation pipeline to identify and annotate clinically relevant variants using single-molecule molecular inversion probes. *JCO Clin Cancer Inform* 2019, 3:1–12
32. Chakravarty D, Gao J, Phillips SM, Kundra R, Zhang H, Wang J, et al: OncoKB: a precision oncology knowledge base. *JCO Precis Oncol* 2017, 1:1–16
33. Tamborero D, Rubio-Perez C, Deu-Pons J, Schroeder MP, Vivanco A, Rovira A, Tusquets I, Albanell J, Rodon J, Tabernero J, de Torres C, Dienstmann R, Gonzalez-Perez A, Lopez-Bigas N: Cancer genome interpreter annotates the biological and clinical relevance of tumor alterations. *Genome Med* 2018, 10:25
34. Deans ZC; On behalf of IQN Path ASBL, Costa JL, Cree I, Dequeker E, Edsjö A, Henderson S, Hummel M, Ligtenberg MJL, Loddio M, Machado JC, Marchetti A, Marquis K, Mason J, Normanno N, Rouleau E, Schuurin E, Snelson K-M, Thunnissen E, Tops B, Williams G, van Krieken H, Hall JA: Integration of

- next-generation sequencing in clinical diagnostic molecular pathology laboratories for analysis of solid tumours; an expert opinion on behalf of IQN Path ASBL. *Virchows Arch* 2017, 470:5–20
35. Jennings LJ, Arcila ME, Corless C, Kamel-Reid S, Lubin IM, Pfeifer J, Temple-Smolkin RL, Voelkerding KV, Nikiforova MN: Guidelines for validation of next-generation sequencing-based oncology panels: a joint consensus recommendation of the Association for Molecular Pathology and College of American Pathologists. *J Mol Diagn* 2017, 19:341–365
 36. Roy S, Coldren C, Karunamurthy A, Kip NS, Klee EW, Lincoln SE, Leon A, Pullambhatla M, Temple-Smolkin RL, Voelkerding KV, Wang C, Carter AB: Standards and guidelines for validating next-generation sequencing bioinformatics pipelines: a joint recommendation of the Association for Molecular Pathology and the College of American Pathologists. *J Mol Diagn* 2018, 20:4–27
 37. European Parliament and the Council of the European Union: Regulation (EU) 2017/746 of the European Parliament and of the Council of 5 April 2017 on in vitro diagnostic medical devices and repealing Directive 98/79/EC and Commission Decision 2010/227/EU. *Off J Eur Union* 2017, 117:176–332
 38. Sirchia F, Carrieri D, Dheensa S, Benjamin C, Kayserili H, Cordier C, van El CG, Turnpenny PD, Melegh B, Mendes Á, Halbersma-Konings TF, van Langen IM, Lucassen AM, Clarke AJ, Forzano F, Kelly SE: Recontacting or not recontacting? a survey of current practices in clinical genetics centres in Europe. *Eur J Hum Genet* 2018, 26:946–954
 39. Groves IJ, Coleman N: Human papillomavirus genome integration in squamous carcinogenesis: what have next-generation sequencing studies taught us? *J Pathol* 2018, 245:9–18
 40. Abraham J, Heimberger AB, Marshall J, Heath E, Drabick J, Helmstetter A, Xiu J, Magee D, Stafford P, Nabhan C, Antani S, Johnston C, Oberley M, Korn WM, Spetzler D: Machine learning analysis using 77,044 genomic and transcriptomic profiles to accurately predict tumor type. *Transl Oncol* 2021, 14:101016
 41. Jiao W, Atwal G, Polak P, Karlic R, Cuppen E; PCAWG Tumor Subtypes and Clinical Translation Working Group, Danyi A, de Ridder J, van Herpen C, Lolkema MP, Steeghs N, Getz G, Morris Q, Stein LD; PCAWG Consortium: A deep learning system accurately classifies primary and metastatic cancers using passenger mutation patterns. *Nat Commun* 2020, 11:728