Pan-cancer whole genome analyses of metastatic solid tumors

Peter Priestley[1,2,*,#], Jonathan Baber[1,2,*], Martijn P. Lolkema[3,4], Neeltje Steeghs[3,5], Ewart de Bruijn[1], Korneel Duyvesteyn[1], Susan Haidari[1,3], Arne van Hoeck[6], Wendy Onstenk[1,3,4], Paul Roepman[1], Charles Shale[2], Mircea Voda[1], Haiko J. Bloemendal[7], Vivianne C.G. Tjan-Heijnen[8], Carla M.L. van Herpen[9], Mariette Labots[10], Petronella O. Witteveen[11], Egbert F. Smit[3,5], Stefan Sleijfer[3,4], Emile E. Voest[3,5], Edwin Cuppen[1,3,6,#]

[1] Hartwig Medical Foundation, Science Park 408, Amsterdam, The Netherlands
[2] Hartwig Medical Foundation Australia, Sydney, Australia
[3] Center for Personalized Cancer Treatment, The Netherlands
[4] Erasmus MC Cancer Institute, Doctor Molewaterplein 40, Rotterdam, The Netherlands
[5] Netherlands Cancer Institute/Antoni van Leeuwenhoekhuis, Plesmanlaan 121, Amsterdam, The Netherlands
[6] Center for Molecular Medicine and Oncode Institute, University Medical Center Utrecht, Heidelberglaan 100, Utrecht, The Netherlands
[7] Meander Medisch Centrum, Maatweg 3, Amersfoort, The Netherlands
[8] Maastricht University Medical Center, P. Debyelaan 25, Maastricht, The Netherlands
[9] Radboud University Medical Center, Geert Grooteplein Zuid 10, Nijmegen, The Netherlands
[10] VU Medical Center, De Boelelaan 1117, Amsterdam, The Netherlands
[11] Cancer Center, University Medical Center Utrecht, Heidelberglaan 100, Utrecht, The Netherlands

[*] shared first author
[#] corresponding authors: p.priestley@hartwigmedicalfoundation.nl, e.cuppen@hartwigmedicalfoundation.nl

Abstract

Metastatic cancer is one of the major causes of death and is associated with poor treatment efficiency. A better understanding of the characteristics of late stage cancer is required to help tailor personalised treatment, reduce overtreatment and improve outcomes. Here we describe the largest pan-cancer study of metastatic solid tumor genomes, including 2,520 whole genome-sequenced tumor-normal pairs, analyzed at a median depth of 106x and 38x respectively, and surveying over 70 million somatic variants. Metastatic lesions were found to be very diverse, with mutation characteristics reflecting those of the primary tumor types, although with high rates of whole genome duplication events (56%). Metastatic lesions are relatively homogeneous with the vast majority (96%) of driver mutations being clonal and up to 80% of tumor suppressor genes bi-allelically inactivated through different mutational mechanisms. For 62% of all patients, genetic variants that may be associated with outcome of approved or experimental therapies were detected. These actionable events were distributed over the various mutation types (single and multiple nucleotide variants, insertions and deletions, copy number alterations and structural variants) underlining the importance of comprehensive genomic tumor profiling for cancer precision medicine for advanced cancer treatment.

Introduction

Metastatic cancer is one of the leading causes of death globally and is a major burden for society despite the availability of an increasing number of (targeted) drugs. Health care costs associated with treatment of metastatic disease are increasing rapidly due to the high cost of novel targeted treatments and immunotherapy, while many patients do not benefit from these approaches with inevitable adverse effects for most patients. Metastatic cancer therefore poses a major challenge for society to balance between individual and societal treatment responsibilities. Since cancer genomes evolve over time, both in the highly heterogeneous primary tumor mass and as disseminated metastatic cells[1,2], a better understanding of metastatic cancer genomes is crucial to further improve on tailoring treatment for late stage cancers.

In recent years, several large-scale whole genome sequencing (WGS) analysis efforts such as TCGA and ICGC have yielded valuable insights in the diversity of the molecular processes driving different types of adult[3,4] and pediatric[5,6] cancer and have fueled the promises of genome-driven oncology care[7]. However, these analyses were primarily done on primary tumor material whereas metastatic cancers, which cause the bulk of the disease burden and 90% of all cancer deaths, have been less comprehensively studied at the whole genome level, with previous efforts focusing on tumor-specific cohorts[8–10] or at a targeted gene panel[11] or exome level[12].

Here we describe the first large-scale pan-cancer whole-genome landscape of metastatic cancers based on the Hartwig Medical Foundation (HMF) cohort of 2,520 paired tumor and normal genomes from 2,405 patients. The samples have been collected prospectively as fresh frozen biopsies taken from a broad range of metastases (Extended Data Fig. 1) and blood controls from patients with advanced cancer in a clinical study setup coordinated by the Center for Personalized Cancer Treatment (CPCT) in 41 hospitals in the Netherlands (Supplementary Table 1). All samples were paired with standardized clinical information (Supplementary Table 2). The sample distribution over age and primary tumor types broadly reflects solid cancer incidence in the Western world, including rare cancers (Fig. 1a-b).

The cohort has been analyzed with uniform and high depth paired-end (2 x 150 bp) whole genome sequencing with a median depth of 106x for tumor samples and 38x for the blood control (Extended Data Fig. 1). Sequencing data were analyzed for all types of somatic variants using an optimized bioinformatic pipeline based on open source tools (Methods). We identified a total of 59,472,629 single nucleotide variants (SNVs), 839,126 multiple nucleotide variants (MNVs), 9,598,205 insertions and deletions (INDELs) and 653,452 structural variants (SVs) (Supplementary Table 2). We found that the relative high sequencing depth is important for variant calling sensitivity as downsampling of the tumor sample coverage to ~53x resulted in an average decrease in sensitivity of 10% for SNV, 2% for INDEL, 15% for MNV, and 19% for SV (Extended Data Fig. 2).

Here we present a first characterization of this unique and comprehensive resource for a better genomic understanding of advanced cancer.

Mutational landscape of metastatic cancer

We analysed the tumor mutational burden (TMB) of each class of variants per cancer type based on tissue of origin (Fig. 1c-h, Supplementary Table 2). In line with previous studies on primary cancers[13], we found extensive variation in mutational load of up to 3 orders of magnitude both within and across cancer types.

The median SNV counts per sample were highest in skin, predominantly consisting of melanoma (44k) and lung (36k) tumors with ten-fold higher SNV counts than sarcomas (4.1k), neuroendocrine tumors (NET) (3.5k) and mesotheliomas (3.4k). The variation for MNVs was even greater with lung (median=815) and skin (median=764) tumors having five times the median MNV counts of any other tumor type. This can be explained by the well-known mutational impact of UV radiation (CC->TT MNV) and smoking (CC->AA MNV) mutational signatures, respectively (Fig. 1f). Although only di-nucleotide substitutions are typically reported as MNVs, 10.7% of the MNVs involve three nucleotides and 0.6% had four or more nucleotides affected.

INDEL counts were typically ten-fold lower than SNVs, with a lower relative rate for skin and lung cancers (Fig. 1d, Extended Data Fig. 3). Genome-wide analysis of INDELs at microsatellite loci identified 60 samples with microsatellite instability (MSI) (Supplementary Table 2), representing 2.4% of all tumors. The highest rates of MSI were observed in central nervous system (CNS) (9.4%), uterus (9.0%) and prostate (6.1%) tumors. For metastatic colorectal cancer lesions we found an MSI frequency of only 4.0%, which is lower than reported for primary colorectal cancer, and in line with better prognosis for patients with localized MSI colorectal cancer, which less often metastasizes[14].

Remarkably, 67% of all INDELs in the entire cohort were found in the 60 MSI samples, and 85% of all INDELs in the cohort were found in microsatellites or short tandem repeats. Only 0.33% of INDELs (32k, ~1% of non-microsatellite INDELs) were found in coding sequences, of which the majority (88%) had a predicted high impact by affecting the open reading frame of the gene.

The median rate of SVs across the cohort was 193 per tumor, with the highest median counts observed in ovary (415) and esophageal (379) tumors, and the lowest in kidney tumors (71) and NET (56) (Fig. 1h, Supplementary Table 2). Simple deletions were the most commonly observed SV subtype (33% of all SVs) and were the most prevalent in every cancer type except stomach and esophageal tumors which were highly enriched in translocations.

Copy number alteration landscape of metastatic cancer

Copy number alterations (CNAs) are important hallmarks of tumorigenesis[15]. Pan-cancer, the most highly amplified regions in our metastatic cancer cohort contain the established oncogenes such as EGFR, CCNE1, CCND1 and MDM2 (Fig. 2). Chromosomal arms 1q, 5p, 8q and 20q are also highly enriched in moderate amplification across the cohort each affecting >20% of all samples. For the amplifications of 5p and 8q this is likely related to the common amplification targets of TERT and MYC, respectively. However, the targets of the amplifications on 1q, predominantly found in breast cancers (>50% of samples) and amplifications on 20q, predominantly found in colorectal cancers (>65% of samples), are less clear.

We identified some intriguing patterns of recurrent loss of heterozygosity (LOH) caused by CNAs. Overall an average of 23% of the autosomal DNA per tumor has LOH. Unsurprisingly, TP53 has the highest LOH recurrence at 67% of samples. Many of the other LOH peaks are also explained by well-known tumor suppressor genes (TSG). However, several clear LOH peaks are observed which cannot easily be explained by known TSG selection, such as one on 8p (57% of samples). 8p LOH has previously been linked to lipid metabolism and drug response[16], although involvement of individual genes has not been established. Alternatively, 8p LOH could potentially be the result of the mechanism by which the amplification of 8q, the most commonly amplified part of the genome, is established.

There are remarkable differences in LOH between cancer types (Fig. 2, Extended Data Fig. 4). For instance, we observed LOH events on the 3p arm in 90% of kidney samples[17] and LOH of the complete chromosome 10 in 72% of CNS tumors (predominantly glioblastoma multiforme[18]). Even in the case of the TP53 region on chromosome 17, different tumor types display clearly different patterns of LOH. Ovarian cancers exhibit LOH of the full chromosome 17 in 75% of samples whereas in prostate cancer, which also has 70% LOH for TP53, this is nearly always caused by highly focal deletions.

Unlike LOH events, homozygous deletions are nearly always restricted to small chromosomal regions. Not a single example was found in which a complete autosomal arm was homozygously deleted. Homozygous deletions of genes are surprisingly rare as well: we found only 4,915 autosomal events (mean = 2.0 per tumor) where one or multiple consecutive genes are fully or partially homozygously deleted. In 46% of these events a putative TSG was deleted. The scarcity of passenger homozygous deletions underlines the fact that despite widespread copy number alterations in metastatic tumors, the vast majority of genes or gross chromosomal organization likely remain essential for tumor cell survival. Chromosome Y loss, which has been described anecdotally for various tumor types[19,20], is a special case and is deleted in 36% of all male tumor genomes but varies strongly between tumor types from 5% to 68% for CNS and biliary tumors respectively (Extended Data Fig. 5).

An extreme form of copy number change can be caused by whole genome duplication (WGD). We found WGD events in 56% of all samples ranging between 17% in CNS to 80% in esophageal tumors (Fig. 2d,e). This is much higher than reported previously for primary tumors (25%-37%)[21,22] and also higher than estimated from panel-based sequencing analyses of advanced tumors (30%)[23]. Ploidy levels, in combination with accurate tumor purity information, are essential for correct interpretation of the measured raw SNV and INDEL frequencies, e.g. to discriminate bi-allelic inactivation of TSG from heterozygous events which are more likely to be passengers or to determine (sub)clonality. Hence determining the WGD status of a tumor is highly relevant for diagnostic applications. Furthermore, WGD has previously been found to correlate with a greater incidence of cancer recurrence for ovarian cancer[22] and has been associated with poor prognosis across cancer types, independently of established clinical prognostic factors[23].

Significantly mutated genes

To identify significantly mutated genes (SMGs) potentially specific for metastatic cancer, we used the dNdScv approach[24] with strict cutoffs ($q<0.01$) for the pan-cancer and tumor-type specific datasets. In addition to reproducing previous results on cancer drivers, a few novel genes were identified (Extended Data Fig. 6, Supplementary Table 3). In the pan-cancer analyses we found only a single novel SMG, which was not either present in the curated COSMIC Cancer Gene Census or found by Martincorena et al[24]. This gene, MLK4 (q = 2e-4), is a mixed lineage kinase that regulates the JNK,P38 and ERK signaling pathways and has been reported to inhibit tumorigenesis in colorectal cancer[25]. In addition, in our tumor type-specific analyses, which for several tumor types is limited by the number of samples, we identified a novel metastatic breast cancer-specific SMG - ZFPM1 (also known as Friend of GATA1 (FOG1), q = 8e-5), a zinc-finger transcription factor protein without clear links with cancer. Nonetheless, we found six unique frameshift variants (all in a context of biallelic inactivation) and three nonsense variants, which suggests a driver role for this gene in metastatic breast cancer.

Our cohort also lends support to some prior SMG findings. In particular, eight significantly mutated putative TSG in the HMF cohort were also found by Martincorena et al[24] - GPS2 (pan-cancer, q=1e-5 & breast, q=2e-3), SOX9 (colorectal & pan-cancer, q=0), TGIF1 (pan-cancer, q=3e-3 & colorectal q=6e-3), ZFP36L1 (urinary tract q=3e-4, pan-cancer q=9e-4) and ZFP36L2 (colorectal & pan-cancer, q=0), HLA-B (lymphoid, q=5e-5), MGA (pan-cancer, q=4e-03), KMT2B (skin, q=3e-3) and RARG (urinary tract 8e-4). None of these genes are currently included in the COSMIC Cancer Gene Census[26]. ZFP36L1 and ZFP36L2 are of particular interest as these genes are zinc-finger proteins that normally play a repressive regulatory role in cell proliferation, presumably through a cyclin D dependent and p53 independent pathway[27]. ZFP36L2 is also independently found as a significantly deleted gene in our cohort, most prominently in colorectal and prostate cancers.

We also searched for genes that were significantly amplified or deleted (Supplementary Table 4). CDKN2A and PTEN were the most significantly deleted genes overall, but many of the top genes involved common fragile sites (CFS) particularly FHIT and DMD, deleted in 5% and 4% of samples, respectively. The role of CFS in tumorigenesis is unclear and aberrations affecting these genes are frequently treated as passenger mutations reflecting localized genomic instability[28]. However, the uneven distribution of the deletions across cancer types may indicate that some of these could be genuine tumor-type specific cancer drivers. For example, we find deletions in DMD to be highly enriched in esophageal tumors (deleted in 38% of samples, whilst SV burden in this tumortype is only about 2-fold higher than average), GIST (Gastro-Intestinal Stromal Tumors; 24%) and pancreatic neuroendocrine tumors (panNET; 41%), which is consistent with a recent study that indicated DMD as a TSG in cancers with myogenic programs[29]. However, tissue type-specific gene expression and differences in origins of replication may also contribute to the observed patterns[28]. We also identified several significantly deleted genes not reported previously, including MLLT4 (13 samples) and PARD3 (9 samples).

Unlike homozygous deletions, amplification peaks tend to be broad and often encompass large number of genes, making identification of the amplification target challenging. However, SRY-related high-mobility group box 4 gene (SOX4, 6p22.3) stands out as a significantly amplified single gene peak (26 amplifications) and is highly enriched in urinary tract cancers (19% of samples highly amplified) (Extended Data Fig. 4). SOX4 is known to be over-expressed in prostate, hepatocellular, lung, bladder and medulloblastoma cancers with poor prognostic features and advanced disease status and is a modulator of the PI3K/Akt signaling[30].

Also notable was a broad amplification peak of 10 genes around ZMIZ1 at 10q22.3 (32 samples) which has not previously been reported. ZMIZ1 is a member of the Protein Inhibitor of Activated STAT (PIAS)-like family of coregulators and is a direct and selective cofactor of Notch1 in T-cell development and leukemia[31]. CDX2, previously identified as an amplified lineage-survival oncogene in colorectal cancer[32], is also highly amplified in our cohort with 20 out of 22 amplified samples found in colorectal cancer, representing 5.4% of all colorectal samples.

Driver mutation catalog

We created a comprehensive catalog of all cancer driver mutations across all samples in our cohort and all variant classes similar as described previously in primary tumors[33] (N. Lopez, personal communication). To do this, we combined our SMG discovery efforts with those from Martincorena et al.[24] and a panel of well known cancer genes (Cosmic Curated Genes)[34], and added gene fusions, TERT promoter mutations and germline predisposition variants found in our cohort. Accounting for the proportion of SNV and INDELs estimated by dNdScv to be passengers, we found 13,423 somatic driver events among the 20,125 identified mutations in the combined gene panel (Supplementary table 5) together with 189 germline predisposition variants (Supplementary table 6). The somatic drivers include 7,423 coding mutation, 615 non-coding point mutation drivers, 2,715 homozygous deletions (25% of which are in common fragile sites), 2,393 focal amplifications and 277 fusion events.

For the cohort as a whole, 55% of point mutations in the gene panel driver catalog were predicted to be genuine driver events. To facilitate analysis of variants of unknown significance (VUS) at a per patient level, we calculated a sample-specific likelihood for each point mutation to be a driver taking into account the TMB of the sample as well as the biallelic inactivation status of the gene for TSG and hotspot positions in oncogenes. Predictions of pathogenic variant overlap with known biology, e.g. clustering of benign missense variants in the 3' half of the APC gene (Extended Data Fig. 7b) fits with the absence of FAP-causing germline variants in this part of the gene[35].

Overall, the catalog matches previous inventories of cancer drivers. TP53 (52% of samples), CDKN2A (21%), APC (16%), PIK3CA (16%), KRAS (14%), PTEN (13%) and TERT (12%) were the most common driver genes together making up 25% of all the driver mutations in the catalog (Fig. 3). However, all of the ten most prevalent driver genes in our cohort were reported at a higher rate than for primary cancers[36], which may reflect the more advanced disease state. AR and ESR1 in particular are more prevalent, with driver mutations in 44% of prostate and 18% of breast cancers, respectively. Both genes are linked to resistance to hormonal therapy, a common treatment for these tumor types, and have been previously reported as enriched in advanced metastatic cancer[11] but are identified at higher rates in this study.

Looking at a per patient level, the mean number of total driver events per patient was 5.7, with the highest rate in urinary tract tumors (mean rate = 8.0) and the lowest in NET (mean rate = 2.8) (Fig. 4). Esophageal and stomach tumors also had elevated driver counts, largely due to a much higher rate of deletions in CFS genes (mean rate = 1.6 for stomach, 1.7 for esophageal) compared to other cancer types (pan-cancer mean rate = 0.3). Fragile sites aside, the differential rates of drivers between cancer types in each variant class do correlate with the relative mutational load, with the exception of skin cancers, which have a lower than expected number of SNV drivers (Extended Data Fig. 3f).

In 98.6% of all samples at least one somatic driver mutation or germline predisposition variant was found. Of the 34 samples with no identified driver, 18 were NET of the small intestine (representing 49% of all patients of this subtype). This likely indicates that small intestine NETs have a distinct set of drivers that are not captured yet in any of the cancer gene resources used and are also not prevalent enough in our relatively small NET cohort to be detected as significant. Alternatively, NET tumors could be mainly driven by epigenetic mechanisms not detected by WGS[37].

The number of amplified driver genes varied significantly between cancer types with highly elevated rates per sample in breast cancer (mean = 2.1), esophageal, urinary tract and stomach (all mean = 1.7) cancers and nearly no amplification drivers in kidney cancer (mean = 0.1) and none in the mesothelioma cohort (Extended Data Fig. 8a). In tumor types with high rates of amplifications, these amplifications are generally found across a broad spectrum of oncogenes (Extended Data Fig. 8b), suggesting there are mutagenic processes active in these tissues that favor amplifications, rather than tissue-specific selection of individual driver genes. AR and EGFR are notable exceptions, with highly selective amplifications in prostate, and in CNS and lung cancers, respectively, in line with previous reports[18,38,39]. Intriguingly, we also found two-fold more amplification drivers in samples with WGD (Extended Data Fig. 8c) despite amplifications being defined as relative to the average sample ploidy.

The 189 germline variants identified in 29 cancer predisposition genes (present in 7.9% of the cohort) consisted of 8 deletions and 181 point mutations (Fig. 3c, Supplementary Table 6). The top 5 affected genes were the well-known germline drivers CHEK2, BRCA2, MUTYH, BRCA1 and ATM, and together contain nearly 80% of the observed predisposition variants (Fig 3c). The corresponding wild type alleles were found to be lost in the tumor sample in more than half of the cases, either by LOH or somatic point mutation, indicating a high penetrance for these variants, particularly in BRCA1 (89% of cases), APC (83%) and BRCA2 (79%).

The 277 fusions consisted of 168 in-frame coding fusions, 91 cis-activating fusions involving repositioning of regulatory elements in 5' genic regions, and 18 in-frame intragenic deletions where one or more exons were deleted (Supplementary table 7). ERG (89 samples), BRAF(17 samples), ERBB4 (16 samples), ALK(12 samples), NRG1(9 samples) and ETV4 (7 samples) were the most commonly observed 3' partners together making up more than half of the fusions. 77 of the 89 ERG fusions were TMPRSS2-ERG affecting 38% of all prostate cancer samples in the cohort. 146 fusion pairs were not previously recorded in CGI, OncoKb, COSMIC or CIViC[34,40–42]. A novel recurrent KMT2A-BCOR fusion was observed in 2 samples (sarcoma and stomach cancer) and there were also 3 recurrent novel localized fusions resulting from adjacent gene pairs: YWHAE-CRK (2 samples), FGFR2-ATE1 (2 samples) and BCR-GNAZ (2 samples).

Only promoter mutations in TERT were included in the study due to the current lack of robust evidence for other recurrent oncogenic non-coding mutations[43]. A total of 257 variants were found at 5 known recurrent variant hotspots[11] and included in the driver catalog.

Oncogene hotspots and novel activating variants

We found that the 70% of somatic driver mutations in oncogenes occur at or within 5 nucleotides of already known pathogenic mutational hotspots (Extended Data Fig. 7a). In the six most prevalent oncogenes (KRAS, PIK3CA, BRAF, NRAS, CTNNB1 & ESR1) the rate was 96% (Fig. 5). Furthermore, in many of the key oncogenes, we document several likely activating but non-canonical variants near known mutational hotspots (Fig. 5). For example, we found activating MNVs in the well known BRAF V600 hotspot (22 cases), but also novel non-hotspot MNVs in KRAS (8 cases) and NRAS (4 cases) (Extended Data Fig 7b).

In-frame indels were even more striking, since despite being exceptionally rare overall (mean = 1.7 per sample), we found an excess in known oncogenes including PIK3CA (19 cases), ERBB2 (10 cases) and BRAF(8 cases) frequently occurring at or near known hotspots[44]. Notably, many of the in-frame indels are enriched in specific tumor types. For instance, all 18 KIT in-frame indels were found in sarcomas, 6 out of 8 MUC6 in-frame indels in esophageal tumors, and 6 of 10 ERBB2 in-frame indels in lung tumors. Finally, we identified 10 in-frame indels in FOXA1, which are highly enriched in prostate cancer (7 of 10 cases) and clustered in two locations that were not previously associated with pathogenic mutations[45].

In CTNNB1 we identified an interesting novel recurrent in-frame deletion of the complete exon 3 in 12 samples, 9 of which are colorectal cancers. Surprisingly, these deletions were homozygous but suspected to be activating as CTNNB1 normally acts as an oncogene in the WNT/beta-catenin pathway and none of these nine colorectal samples had any APC driver mutations.

Biallelic tumor suppressor gene inactivation

Our results strongly support the Knudson two-hit hypothesis[46] for tumor suppressor genes with 80% of all TSG drivers explained by biallelic inactivation by genetic alterations (i.e. either by homozygous deletion (32%), multiple somatic point mutations (7%), or a point mutation in combination with LOH (41%)). This rate is the highest observed in any large-scale cancer WGS study. For many key tumor suppressor genes the biallelic inactivation rate is almost 100% (more specifically: TP53 (93%), CDKN2A (97%), RB1 (94%), PTEN (92%) and SMAD4 (96%); Fig. 3b), suggesting that biallelic inactivation of these genes is a strict requirement for metastatic cancer.

Other prominent TSGs, however, have lower biallelic rates, including ARID1A (55%), KMT2C (49%) and ATM (49%). It is unclear whether we systematically missed the second hit in these cases, as this could potentially be mediated through non-mutational epigenetic mechanisms[47], or if these genes impact on tumorigenesis via a haploinsufficiency mechanism[48].

Clonal and subclonal variants

To obtain insight into ongoing tumor evolution dynamics, we examined the clonality of all variants. Surprisingly, only 6.5% of all SNV, MNV & INDELs across the cohort and just 3.7% of the driver point mutations were found to be subclonal (Fig. 6). The low proportion of samples with subclonal variants could be partially due to the detection limits of the sequencing approach (sequencing depth, bioinformatic analysis settings), particularly for low purity samples. However, even for samples with purities higher than 80% the total proportion of subclonal variants only reaches 10.2% (Fig. 6b). Furthermore, sensitized detection of variants at hotspot positions in cancer genes showed that our analysis pipeline detected over 96% of variants with allele frequencies of > 3%. Although the cohort contains some samples with (very) high fractions of subclonal variants, overall the metastatic tumor samples are relatively homogeneous without the presence of multiple diverged subclones. Low intratumor heterogeneity may be in part attributed to the fact that nearly all biopsies were obtained by a core needle biopsy, which results in highly localized sampling, but is nevertheless much lower compared to previous observations in primary cancers[2].

In the 111 patients with independently collected repeat biopsies from the same patient (Supplementary Table 8) we found 11% of all SNVs to be subclonal. Whilst 76% of clonal variants were shared between biopsies, less than 30% of the subclonal variants were shared.

While we can not exclude the presence of larger amounts of lower frequency subclonal variants, the low rate of high-frequency subclonal variants taken together with the observation that a very high proportion of subclonal variants are private to a local metastasis, suggest a model where individual metastatic lesions are dominated by a single clone at any one point in time and that more limited tumor evolution and subclonal selection takes places after distant metastatic seeding. This contrasts with observations in primary tumors, where larger degrees of subclonality and multiple major subclones are more frequently observed[2,49], but supports other recent studies which demonstrate minimal driver gene heterogeneity in metastases[8,50].

Co-occurrence of Drivers

We examined the pairwise co-occurrence of driver gene mutations per cancer type and found ten combinations of genes that were significantly mutually exclusively mutated, and ten combinations of genes which were significantly co-occurrently mutated (excluding pairs of genes on the same chromosome which are frequently co-amplified or co-deleted) (Fig. 7). The 20 significant findings include previously reported co-occurrence of mutated DAX|MEN1 in pancreatic NET (q=0.0007), and CDH1|SPOP in prostate tumors (q=0.0005), as well as negative associations of mutated genes within the same signal transduction pathway such as KRAS|BRAF (q=4e-4) and KRAS|NRAS (q=0.009) in colorectal cancer, BRAF|NRAS in skin cancer (q=6e-12), CDKN2A|RB1 in lung cancer (q=8e-5) and APC|CTNNB1 in colorectal cancer (q=8e-6). APC is also strongly negatively correlated with both BRAF (q=1e-4) and RNF43 (q=2e-5) which together are characteristic of the serrated molecular subtype of colorectal cancers[51]. We also found that SMAD2|SMAD3 are highly positively correlated in colorectal cancer (q=0.02), mirroring a result reported previously in a large cohort of colorectal cancers[52].

In breast cancer, we found a number of significant novel relationships, including a positive relationship for GATA3|VMP1(q=1e-4) and FOXA1|PIK3CA (q=2e-3), and negative relationships for ESR1|TP53 (q=9e-4) and GATA3|TP53 (q=2e-3), which will need further validation and experimental follow-up to understand underlying biology.

Actionability

We analyzed opportunities for biomarker-based treatment for all patients by mapping driver events to three clinical annotation databases: CGI[42], CIViC[40] and OncoKB[41]. In 1,485 patients (62%) at least one 'actionable' event was identified (Supplementary Table 9). Whilst these numbers are in line with results from primary tumors[33], longitudinal studies will be required to conclude if genomic analyses for therapeutic guidance should be repeated when a patient experiences progressive disease. Half of the patients with an actionable event (31% of total) contained a biomarker with a predicted sensitivity to a drug at level A (approved anti-cancer drugs) and lacked any known resistance biomarkers for the same drug (Fig. 8a). In 13% of patients the suggested therapy was a registered indication, while in 18% of cases it was outside the labeled indication. In a further 31% of patients a level B (experimental therapy) biomarker was identified. The actionable biomarkers spanned all variant classes including 1,815 SNVs, 48 MNVs, 195 indels, 745 CNAs, 68 fusion genes and 60 patients with microsatellite instability (Fig. 8b).

Tumor mutation burden is an important emerging biomarker for response to immune checkpoint inhibitor therapy[53] as it is a proxy for the amount of neo-antigens in the tumor cells. For NSCLC it has been shown in at least 2 subgroup analyses of large phase III trials that both PFS and OS are significantly improved with first line immunotherapy as compared to chemotherapy for patients whose tumors have a TMB >10 mutations per Mb[54,55]. Although various clinical studies based on this parameter are currently emerging, TMB was not yet included in the above actionability analysis. However, when applying the same cut-off to all samples in our cohort, an overall 18% of patients would qualify, varying from 0% for liver, mesothelioma and ovarian cancer patients to more than 50% of lung and skin cancer patients (Extended Data Fig. 3b).

Discussion

Genomic testing of tumors faces numerous challenges in meeting clinical needs, including i) the interpretation of variants of unknown significance (VUS), ii) the steadily expanding universe of actionable genes, often with an increasingly small fraction of patients affected (e.g. NRG1[56] and NTRK fusions[57] in less than 2% of all patients), and iii) the development of advanced genome-derived biomarkers such as tumor mutational load, DNA repair status and mutational signatures. Our results demonstrate in several ways that WGS analyses of metastatic cancer can provide novel and relevant insights and be instrumental in addressing some of these key challenges in cancer precision medicine.

First, our systematic and large-scale pan-cancer analyses on metastatic cancer tissue allowed for the identification of several novel (cancer type-specific) cancer drivers and mutation hotspots. Second, the driver catalog analyses can be used to mitigate the problem of VUS interpretation[33] both by leveraging previously identified pathogenic mutations (accounting for more than 2/3rds of oncogenic point-mutation drivers) and by careful analysis of the biallelic inactivation of putative TSG which accounts for over 80% of TSG drivers in metastatic cancer. Third, we demonstrate the importance of accounting for all types of variants, including large scale genomic rearrangements (via fusions and copy number alteration events), which account for more than half of all drivers, but also activating MNV and INDELs which we have shown are commonly found in many key oncogenes. Fourth, we have shown that using WGS, even with very strict variant calling criteria, we could find driver variants in more than 98% of all metastatic tumors, including putatively actionable events in a clinical and experimental setting for up to 62% of patients.

Although we did not find metastatic tumor genomes to be fundamentally different to primary tumors in terms of the mutational landscape or genes driving advanced tumorigenesis, we described characteristics that could contribute to therapy responsiveness or resistance in individual patients. In particular we showed that WGD is a more pervasive element of tumorigenesis than previously understood affecting over half of all metastatic cancers. We also found metastatic lesions to be less heterogeneous than reported in primary tumors, although the limited depth sequencing does not allow for drawing conclusions regarding low-frequency subclonal variants.

It should be noted that differences between WGS cohorts should be interpreted with some caution as inevitable differences between experimental and computational approaches may impact on observations and can only be addressed in longitudinal studies including the different stages of disease. Furthermore, the HMF cohort includes a mix of treatment-naive metastatic patients and patients who have undergone (extensive) prior systemic treatments. While this may impact on specific tumor characteristics, it also provides opportunities for studying treatment response and resistance as this data is recorded in the studies.

Finally, we believe the resource described here is a valuable complementary resource to comparable whole genome sequencing-based resources of primary tumors in advancing fundamental and translational cancer research. Therefore, all non-privacy sensitive data is publicly available through a local interface developed by ICGC[58] (work in progress) and all other data is made freely available for scientific research by a controlled access mechanism (see www.hartwigmedicalfoundation.nl/en for details).

Acknowledgements

We thank the Hartwig Foundation and Barcode for Life for financial support of clinical studies and WGS analyses. Development of the data portal was supported by a grant from KWF Kankerbestrijding (HMF2017-8225, GENONCO). We are particularly grateful to all patients, nurses and medical specialists for their essential contributions making this study possible. We would like to specifically thank Hans van Snellenberg for operational management of the Hartwig Medical Foundation. We would like to thank Stefan Willems, Wendy de Leng, Alexander Hoischen and Winand Dinjens for support with pathology assessments and mutation validations and Jeroen de Ridder, Wigard Kloosterman and Harmen van de Werken for critically reading the manuscript.

Figure Legends

Figure 1: Mutational load of metastatic cancer per tumor type
a) The number of samples of each tumor type cohort. Tumor types are ranked from lowest to highest overall mutation burden (TMB)
b) Violin plot showing age distribution of each tumor type with 25th, 50th and 75th percentiles marked.
c)-d) cumulative distribution function plot (individual samples were ranked independently for each panel) of mutational load for each tumor type for SNV and MNV (c) and INDEL and SV (d). The median for each cohort is indicated with a vertical line.
e)-h) Mutational context or variant subtype per individual sample for each of (e) Single Nucleotide Variant (SNV), (f) Multi Nucleotide Variant (MNV), (g) INsertion/DELetion (INDEL), (h) Structural Variant (SV). Each column chart is ranked within tumor type by mutational load in that variant class. MNVs are classified by the dinucleotide substitution with NN referring to any dinucleotide combination. SVs are classified by type: INV = inversion, DEL = deletion, DUP = tandem duplication, TRL = translocation, INS = insertion.

Figure 2: Copy number landscape of metastatic cancer
Proportion of samples with amplification and deletion events by genomic position per cohort - pan-cancer (a), central nervous system (CNS) (b) and kidney (c). The inner ring shows the % of tumors with homozygous deletion (orange), LOH and significant loss (copy number < 0.6x sample ploidy - dark blue) and near copy neutral LOH (light blue). Outer ring shows % of tumors with high level amplification (>3x sample ploidy - orange), moderate amplification (>2x sample ploidy - dark green) and low level amplification (>1.4x amplification - light green). The scale on both rings is 0-100% and inverted for the inner ring. The most frequently observed high-level gene amplifications (black text) and homozygous deletions (red text) are shown.
d) Proportion of tumors with a whole genome duplication event (dark blue) grouped by tumor type.
e) Average sample ploidy distribution over the complete cohort. Samples with a WGD event (true) are shown in darker blue.

Figure 3: Most prevalent driver genes in metastatic cancer
Most prevalent somatically mutated TSG (a) and oncogenes (b), and germline predisposition variants (c) . From left to right, the heatmap shows the % of samples in each cancer type which are found to have each gene mutated; absolute bar chart shows the pan-cancer % of samples with the given gene mutated; relative bar chart shows the breakdown by type of alteration. For TSG, the % of samples with a driver in which the gene is found biallelically inactivated, and for germline predisposition variants the % of samples with loss of wild type in the tumor are also shown.

Figure 4: Drivers per sample by tumor type
a) Violin plot showing the distribution of the number of drivers per sample grouped by tumor type. Black dots indicate the mean values for each tumor type.
b) Relative bar chart showing the breakdown per cancer type of the type of alteration.

Figure 5: Oncogenic Hotspots
Count of driver point mutations by variant type. Known pathogenic mutations curated from external databases are categorized as hotspot mutations. Mutations within 5 bases of a known pathogenic mutation are shown as near hotspot and all other mutations are shown as non-hotspot.

Figure 6: Subclonality
a) Count of samples per tumor purity bucket. b) Violin plot showing the percentage of point mutations which are subclonal in each purity bucket. Black dots indicate the mean for each bucket. c) Percentage of driver point mutations that are subclonal in each purity bucket.

Figure 7: Driver co-occurrence
a) Mutated driver gene pairs which are significantly positively (on the right) or negatively (on the left) correlated in individual tumor types sorted by q-value. The color indicates the tumor type as depicted below the chart.

Figure 8: Actionability
a) Percentage of samples in each cancer type with an actionable mutation based on data in CGI, CIViC and OncoKB knowledgebases. Level 'A' represents presence of biomarkers with either an approved therapy or guidelines and level B represents biomarkers having strong biological evidence or clinical trials indicating they are actionable. On label indicates treatment registered by federal authorities for that tumor type, while off-label indicates a registration for other tumor types.
b) Break down of the actionable variants by mutation type.

Extended Data Figures and Tables

Extended Data Figure 1: Hartwig sample workflow, biopsy locations and sequence coverage
a) Sample workflow from patient to high-quality WGS data. A total of 4,018 patients were enrolled in the study between April 2016 and April 2018. For 9% of patients no blood and/or biopsy material was obtained, mostly because conditions of patients prohibited further study participation. Up to 4 fresh-frozen biopsies per patient were received, which were sequentially analyzed to identify a biopsy with more than 30% tumor cellularity as determined by routine histology assessment. For 859 patients no suitable biopsy was obtained and 2,796 patients were further processed for WGS. 44 and 29 samples failed in either DNA isolation or library preparation and raw WGS data quality QC, respectively. For an additional 385 samples the WGS data was of good quality, but the tumor purity determination based on WGS data (PURPLE) was less than 20% making reliable and comprehensive somatic variant calling and were therefore excluded. Eventually, 2,338 tumor-normal sample pairs with high-quality WGS data were obtained, which were supplemented with 182 pairs from pre-April 2016, adding up to 2,520 tumor normal pairs that were included in this study.
b) Breakdown of cohort by biopsy location. Tumor biopsies were taken from a broad range of locations. Primary tumor type is shown on the left and the biopsy location on the right.
c) Distribution of sample sequencing depth for tumor and blood reference.

Extended Data Figure 2: Impact of downsampling on variant calling
Comparison of variant calling of 10 randomly selected samples at normal depth and 50% downsampled for purity (a), SNV counts (b), SV counts (c), Ploidy (d), MNV counts (e) and INDEL counts (f). For the panels B, C, E and F, the black dots represent the % reduction per sample of counts (right axis) and the dotted line represents the average % reduction across all tested samples.

Extended Data Figure 3: Mutational load, genome wide analyses and drivers
a) Proportion of samples by cancer type classified as microsatellite instable (MSISeq score > 4)
b) Proportion of samples with a high mutational burden (TMB > 10 SNV / Mb)
c)-e) Scatter plot of mutational load per sample for INDEL vs SNV (c), INDEL vs SV (d), and SV vs SNV (e). MSI (MSISeq score > 4) and 'high TMB' (>10 SNV/ MB) thresholds are indicated.
f)-h) Mean mutational load vs driver rate for SNV (f), INDEL (g) and SV (h) grouped by cancer type. MSI samples were excluded.

Extended Data Figure 4: Copy Number profile per cancer types
Circos plots showing the proportion of samples with amplification and deletion events by genomic position per cancer type. The inner ring shows the % of tumors with homozygous deletion (red), LOH

and significant loss (copy number < 0.6x sample ploidy - dark blue) and near copy neutral LOH (light blue). The outer ring shows the % of tumors with high level amplification (>3x sample ploidy - orange), moderate amplification (>2x sample ploidy - dark green) and low level amplification (>1.4x amplification - light green). Scales on both rings are 0-100% and inverted for the inner ring. The most frequently observed high level gene amplifications (black text) and homozygous deletions (red text) are labelled.

Extended Data Figure 5: Somatic Y chromosome Loss
Proportion of Male tumors with somatic loss of >50% of Y chromosome (dark blue) grouped by tumor type.

Extended Data Figure 6: Significantly mutated genes
Tile chart showing genes found to be significantly mutated per cancer type cohort and pan-cancer using dNdScv. Gene names marked in orange are also significant in Martincorena et al[24], but not found in COSMIC curated or census. Gene names marked in red are novel in this study.

Extended Data Figure 7: Coding mutation profiles by driver gene
Location and driver classification of all coding mutations (SNVs and indels) in oncogenes (a) and tumor suppressor genes (TSG) (b) in the driver catalog. The lollipops on the chart show the location (coding sequence coordinates) and count of mutations for all candidate drivers. The height of lollipop represents the total count of each individual variant in the cohort (log scale). The height of the solid line represents the sum of driver likelihoods for that variant, ie. the proportion that are expected to be drivers. (Partially) dotted lines hence indicate variants for which driver role is uncertain. For TSG only, variants are unshaded if all instances of that variant are monoallelic single hits with no LOH. The right column chart shows the estimated number of drivers (calculated as the sum of driver likelihoods) and passenger variants in each gene by cancer type.

Extended Data Figure 8: Amplifications
a) Mean rate of amplification drivers per cancer type. b) Breakdown of the number of amplification drivers per gene by cancer type. c) Mean rate of drivers per variant type for samples with and without WGD.

Supplementary Table 1: Overview of contributing organizations and local principal investigators.

Supplementary Table 2: Overview of cohort and sample characteristics

Supplementary Table 3: Pan-cancer and cancer type-specific dNdScv results

Supplementary Table 4: Recurring amplifications (a) and deletions (b) and associated target genes

Supplementary Table 5: Somatic driver catalog

Supplementary Table 6: Germline driver catalog

Supplementary Table 7: Gene Fusions

Supplementary Table 8: Overview of patients with multiple biopsies

Supplementary Table 9: Actionable mutations

References

1. Klein, C. A. Selection and adaptation during metastatic cancer progression. *Nature* **501**, 365–372 (2013).
2. McGranahan, N. & Swanton, C. Clonal Heterogeneity and Tumor Evolution: Past, Present, and the Future. *Cell* **168**, 613–628 (2017).
3. Cancer Genome Atlas Research, Network *et al.* The Cancer Genome Atlas Pan-Cancer analysis project. *Nat. Genet.* **45**, 1113–1120 (2013).
4. International Cancer Genome, Consortium *et al.* International network of cancer genome projects. *Nature* **464**, 993–998 (2010).
5. Grobner, S. N. *et al.* The landscape of genomic alterations across childhood cancers. *Nature* **555**, 321-327 (2018).
6. Ma, X. *et al.* Pan-cancer genome and transcriptome analyses of 1,699 paediatric leukaemias and solid tumours. *Nature* **555**, 371-376 (2018).
7. Hyman, D. M., Taylor, B. S. & Baselga, J. Implementing Genome-Driven Oncology. *Cell* **168**, 584–599 (2017).
8. Yates, L. R. *et al.* Genomic Evolution of Breast Cancer Metastasis and Relapse. *Cancer Cell* **32**, 169–184 e7 (2017).
9. Naxerova, K. *et al.* Origins of lymphatic and distant metastases in human colorectal cancer. *Science* **357**, 55–60 (2017).
10. Gundem, G. *et al.* The evolutionary history of lethal metastatic prostate cancer. *Nature* **520**, 353–357 (2015).
11. Zehir, A. *et al.* Mutational landscape of metastatic cancer revealed from prospective clinical sequencing of 10,000 patients. *Nat. Med.* **23**, 703–713 (2017).
12. Robinson, D. R. *et al.* Integrative clinical genomics of metastatic cancer. *Nature* **548**, 297–303 (2017).
13. Alexandrov, L. B. *et al.* Signatures of mutational processes in human cancer. *Nature* **500**, 415–421 (2013).
14. Gryfe, R. *et al.* Tumor microsatellite instability and clinical outcome in young patients with colorectal cancer. *N. Engl. J. Med.* **342**, 69–77 (2000).
15. Taylor, A. M. *et al.* Genomic and Functional Approaches to Understanding Cancer Aneuploidy. *Cancer Cell* **33**, 676–689.e3 (2018).
16. Cai, Y. *et al.* Loss of Chromosome 8p Governs Tumor Progression and Drug Response by Altering Lipid Metabolism. *Cancer Cell* **29**, 751–766 (2016).
17. Sato, Y. *et al.* Integrated molecular analysis of clear-cell renal cell carcinoma. *Nat. Genet.* **45**, 860–867 (2013).
18. Brennan, C. W. *et al.* The somatic genomic landscape of glioblastoma. *Cell* **155**, 462–477 (2013).
19. Hunter, S., Gramlich, T., Abbott, K. & Varma, V. Y chromosome loss in esophageal carcinoma: an in situ hybridization study. *Genes Chromosomes Cancer* **8**, 172–177 (1993).
20. Sauter, G. *et al.* Y chromosome loss detected by FISH in bladder cancer. *Cancer Genet. Cytogenet.* **82**, 163–169 (1995).
21. Zack, T. I. *et al.* Pan-cancer patterns of somatic copy number alteration. *Nat. Genet.* **45**, 1134–1140 (2013).
22. Carter, S. L. *et al.* Absolute quantification of somatic DNA alterations in human cancer. *Nat. Biotechnol.* **30**, 413–421 (2012).
23. Bielski, C. M. *et al.* Genome doubling shapes the evolution and prognosis of advanced cancers. *Nat. Genet.* **50**, 1189-1195 (2018).
24. Martincorena, I. *et al.* Universal Patterns of Selection in Cancer and Somatic Tissues. *Cell* **171**, 1029–1041 e21 (2017).
25. Marusiak, A. A. *et al.* Recurrent MLK4 Loss-of-Function Mutations Suppress JNK Signaling to Promote Colon Tumorigenesis. *Cancer Res.* **76**, 724–735 (2016).
26. Forbes, S. A. *et al.* COSMIC: somatic cancer genetics at high-resolution. *Nucleic Acids Res.* **45**, D777–D783 (2017).

27. Suk, F.-M. *et al.* ZFP36L1 and ZFP36L2 inhibit cell proliferation in a cyclin D-dependent and p53-independent manner. *Sci. Rep.* **8**, 2742 (2018).
28. Glover, T. W., Wilson, T. E. & Arlt, M. F. Fragile sites in cancer: more than meets the eye. *Nat. Rev. Cancer* **17**, 489–501 (2017).
29. Wang, Y. *et al.* Dystrophin is a tumor suppressor in human cancers with myogenic programs. *Nat. Genet.* **46**, 601–606 (2014).
30. Mehta, G. A. *et al.* Amplification of SOX4 promotes PI3K/Akt signaling in human breast cancer. *Breast Cancer Res. Treat.* **162**, 439–450 (2017).
31. Pinnell, N. *et al.* The PIAS-like Coactivator Zmiz1 Is a Direct and Selective Cofactor of Notch1 in T Cell Development and Leukemia. *Immunity* **43**, 870–883 (2015).
32. Salari, K. *et al.* CDX2 is an amplified lineage-survival oncogene in colorectal cancer. *Proc. Natl. Acad. Sci. U. S. A.* **109**, E3196–205 (2012).
33. Sabarinathan, R. *et al.* The whole-genome panorama of cancer drivers. *BioArchive* (2017). doi:10.1101/190330
34. Futreal, P. A. *et al.* A census of human cancer genes. *Nat. Rev. Cancer* **4**, 177–183 (2004).
35. Friedl, W. *et al.* Can APC mutation analysis contribute to therapeutic decisions in familial adenomatous polyposis? Experience from 680 FAP families. *Gut* **48**, 515–521 (2001).
36. Bailey, M. H. *et al.* Comprehensive Characterization of Cancer Driver Genes and Mutations. *Cell* **173**, 371–385.e18 (2018).
37. Cives, M., Simone, V., Rizzo, F. M. & Silvestris, F. NETs: organ-related epigenetic derangements and potential clinical applications. *Oncotarget* **7**, 57414–57429 (2016).
38. Viswanathan, S. R. *et al.* Structural Alterations Driving Castration-Resistant Prostate Cancer Revealed by Linked-Read Genome Sequencing. *Cell* **174**, 433-447 (2018).
39. Cancer Genome Atlas Research Network. Comprehensive genomic characterization of squamous cell lung cancers. *Nature* **489**, 519–525 (2012).
40. Griffith, M. *et al.* CIViC is a community knowledgebase for expert crowdsourcing the clinical interpretation of variants in cancer. *Nat. Genet.* **49**, 170–174 (2017).
41. Chakravarty, D. *et al.* OncoKB: A Precision Oncology Knowledge Base. *JCO Precis Oncol* **2017**, (2017).
42. Tamborero, D. *et al.* Cancer Genome Interpreter annotates the biological and clinical relevance of tumor alterations. *Genome Med.* **10**, 25 (2018).
43. Cuykendall, T. N., Rubin, M. A. & Khurana, E. Non-coding genetic variation in cancer. *Current Opinion in Systems Biology* **1**, 9–15 (2017).
44. Chang, M. T. *et al.* Accelerating Discovery of Functional Mutant Alleles in Cancer. *Cancer Discov.* **8**, 174–183 (2018).
45. Yang, Y. A. & Yu, J. Current perspectives on FOXA1 regulation of androgen receptor signaling and prostate cancer. *Genes Dis* **2**, 144–151 (2015).
46. Knudson, A. G., Jr. Mutation and cancer: statistical study of retinoblastoma. *Proc. Natl. Acad. Sci. U. S. A.* **68**, 820–823 (1971).
47. Schlicker, A., Michaut, M., Rahman, R. & Wessels, L. F. A. OncoScape: Exploring the cancer aberration landscape by genomic data fusion. *Sci. Rep.* **6**, 28103 (2016).
48. Davoli, T. *et al.* Cumulative haploinsufficiency and triplosensitivity drive aneuploidy patterns and shape the cancer genome. *Cell* **155**, 948–962 (2013).
49. Andor, N. *et al.* Pan-cancer analysis of the extent and consequences of intratumor heterogeneity. *Nat. Med.* **22**, 105–113 (2016).
50. Reiter, J. G. *et al.* Minimal functional driver gene heterogeneity among untreated metastases. *Science* **361**, 1033–1037 (2018).
51. Bond, C. E. *et al.* RNF43 and ZNRF3 are commonly altered in serrated pathway colorectal tumorigenesis. *Oncotarget* **7**, 70589–70600 (2016).
52. Fleming, N. I. *et al.* SMAD2, SMAD3 and SMAD4 mutations in colorectal cancer. *Cancer Res.* **73**, 725–735 (2013).
53. Goodman, A. M. *et al.* Tumor Mutational Burden as an Independent Predictor of Response to

Immunotherapy in Diverse Cancers. *Mol. Cancer Ther.* **16**, 2598–2608 (2017).

54. Hellmann, M. D. *et al.* Nivolumab plus Ipilimumab in Lung Cancer with a High Tumor Mutational Burden. *N. Engl. J. Med.* **378**, 2093–2104 (2018).
55. Carbone, D. P. *et al.* First-Line Nivolumab in Stage IV or Recurrent Non-Small-Cell Lung Cancer. *N. Engl. J. Med.* **376**, 2415–2426 (2017).
56. Fernandez-Cuesta, L. & Thomas, R. K. Molecular Pathways: Targeting NRG1 Fusions in Lung Cancer. *Clin. Cancer Res.* **21**, 1989–1994 (2015).
57. Laetsch, T. W. *et al.* Larotrectinib for paediatric solid tumours harbouring NTRK gene fusions: phase 1 results from a multicentre, open-label, phase 1/2 study. *Lancet Oncol.* **19**, 705–714 (2018).
58. Zhang, J. *et al.* International Cancer Genome Consortium Data Portal--a one-stop shop for cancer genomics data. *Database* **2011**, bar026 (2011).

Figure 1

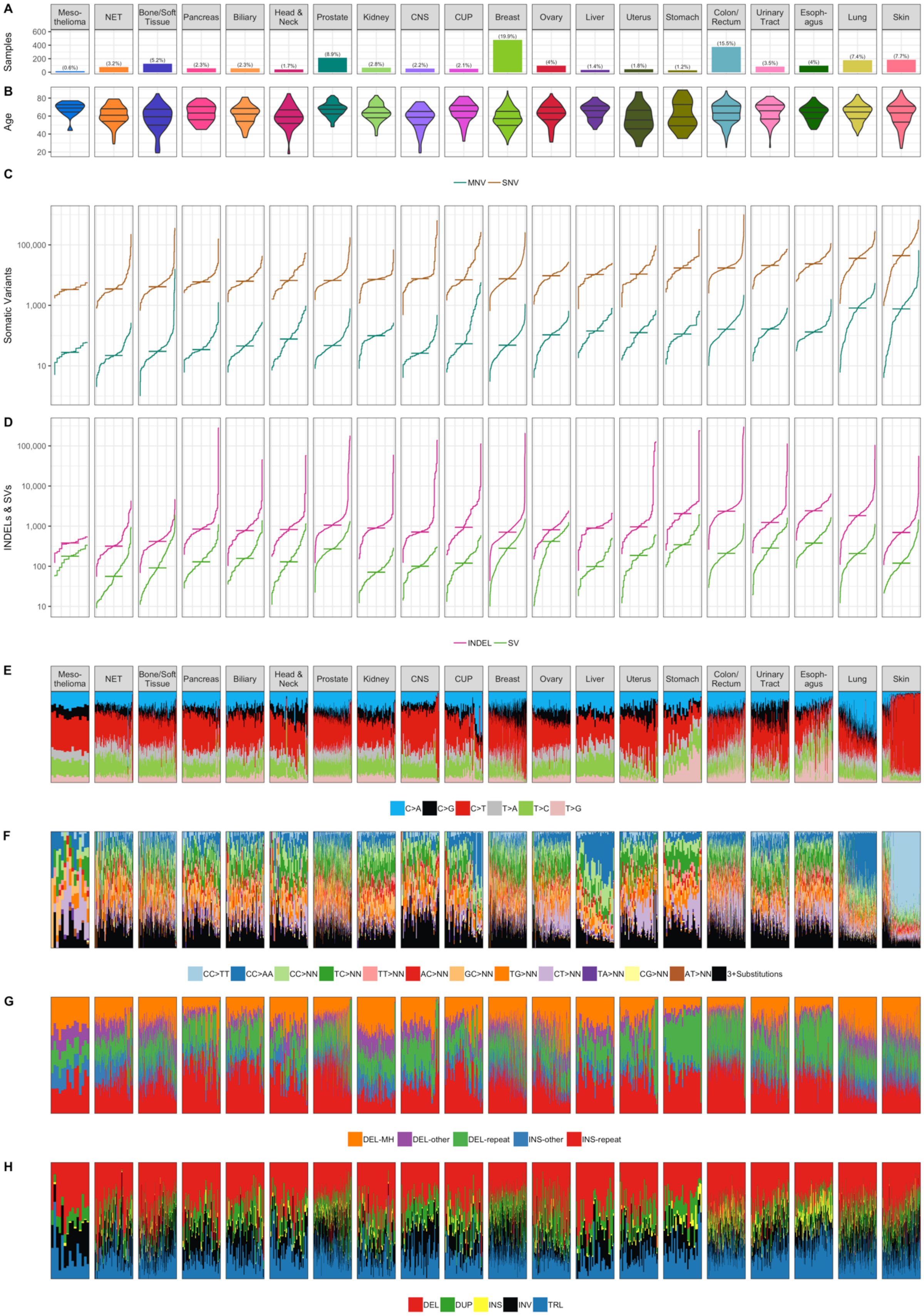

A
B
C
D
E
F
G
H
Meso-thelioma
NET
Bone/Soft Tissue
Pancreas
Biliary
Head & Neck
Prostate
Kidney
CNS
CUP
Breast
Ovary
Liver
Uterus
Stomach
Colon/ Rectum
Urinary Tract
Esoph-agus
Lung
Skin
Samples
600
400
200
0
(0.6%)
(3.2%)
(5.2%)
(2.3%)
(2.3%)
(1.7%)
(8.9%)
(2.8%)
(2.2%)
(2.1%)
(19.9%)
(4%)
(1.4%)
(1.8%)
(1.2%)
(15.5%)
(3.5%)
(4%)
(7.4%)
(7.7%)
Age
80
60
40
20
MNV
SNV
Somatic Variants
100,000
1,000
10
INDELs & SVs
100,000
10,000
1,000
100
10
INDEL
SV
C>A
C>G
C>T
T>A
T>C
T>G
CC>TT
CC>AA
CC>NN
TC>NN
TT>NN
AC>NN
GC>NN
TG>NN
CT>NN
TA>NN
CG>NN
AT>NN
3+Substitutions
DEL-MH
DEL-other
DEL-repeat
INS-other
INS-repeat
DEL
DUP
INS
INV
TRL

Figure 2

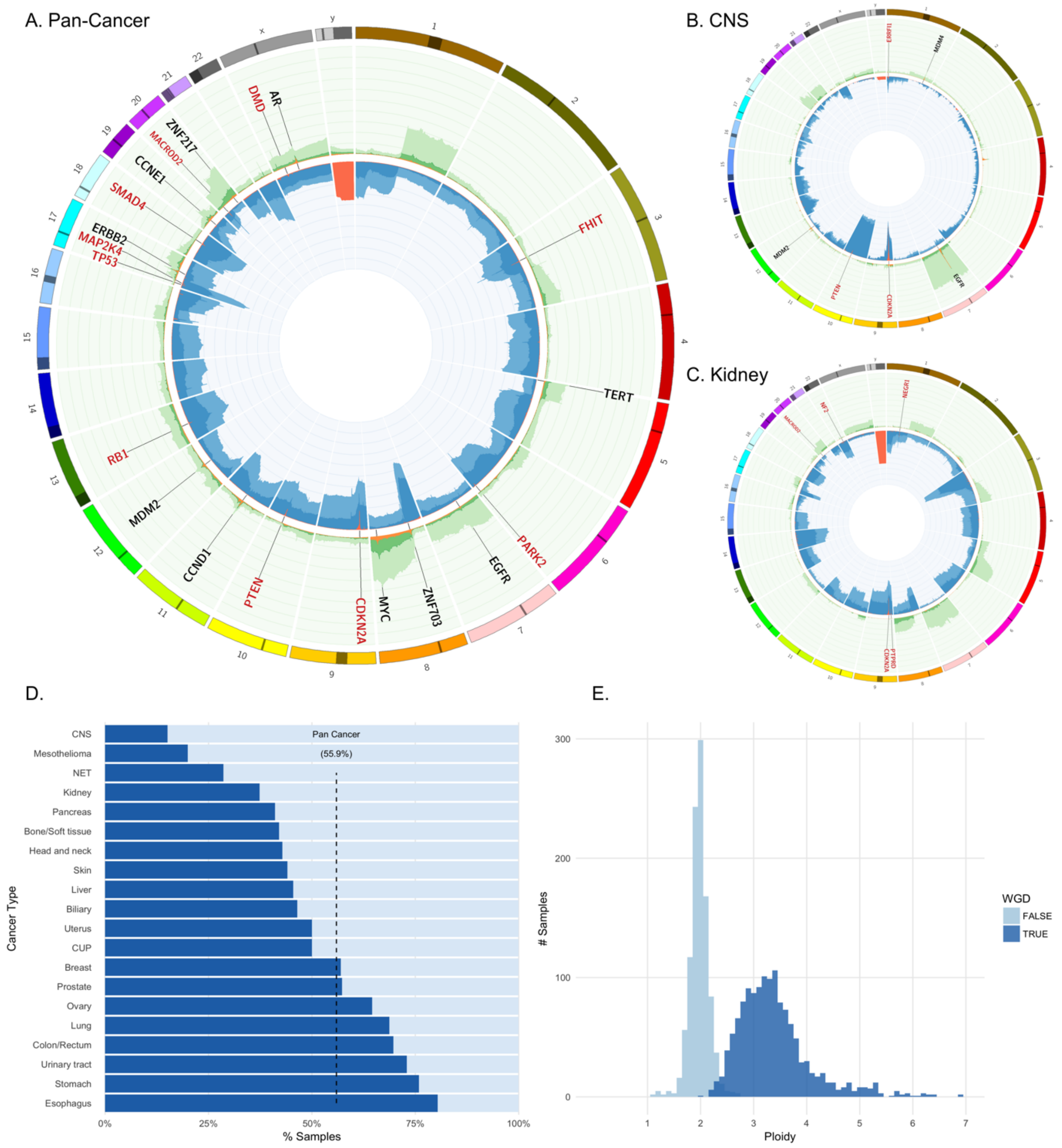
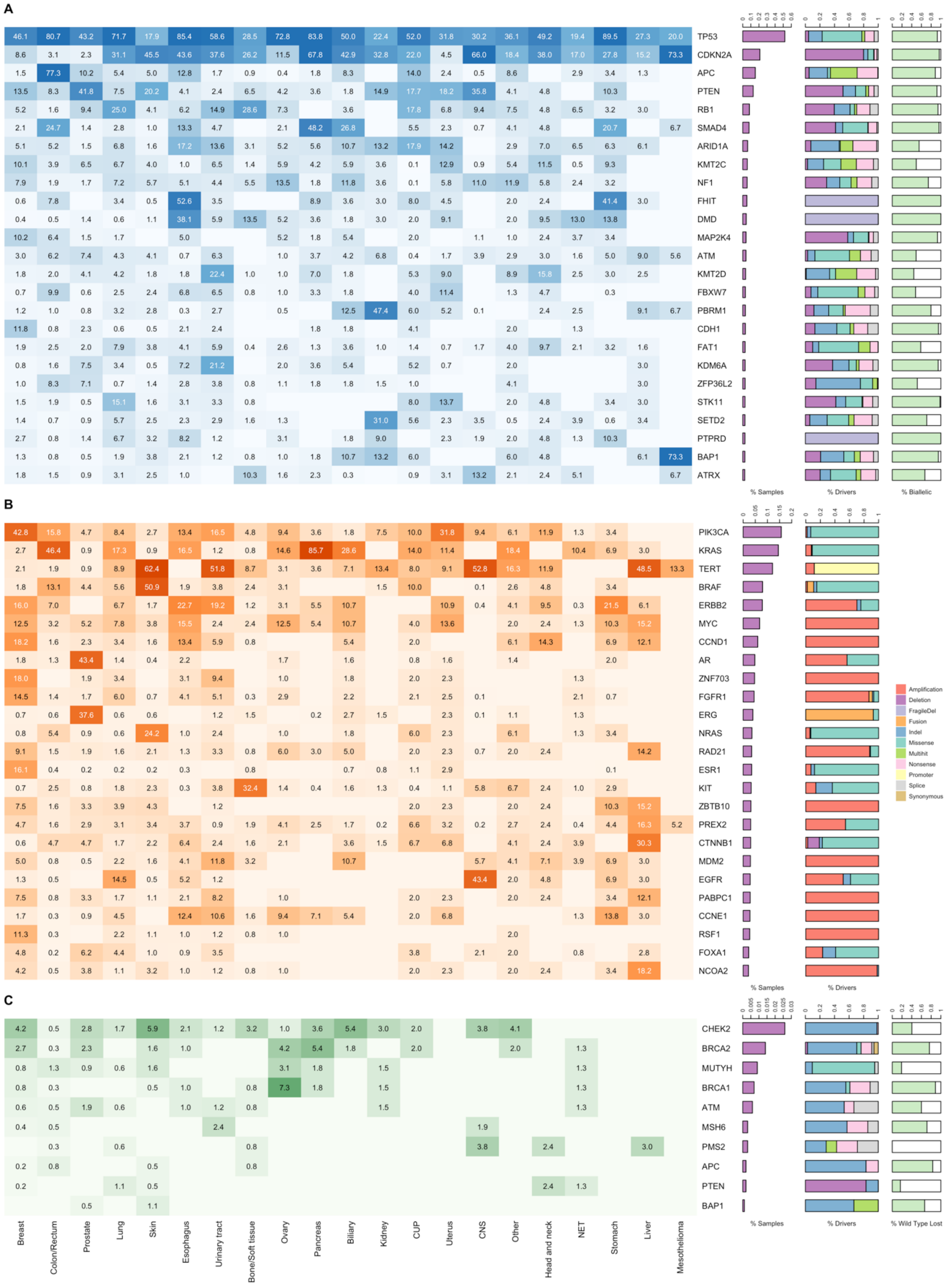

A. Pan-Cancer
B. CNS
C. Kidney
D.
E.
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
x
y
DMD
AR
ZNF217
MACROD2
CCNE1
SMAD4
ERBB2
MAP2K4
TP53
RB1
MDM2
CCND1
PTEN
CDKN2A
MYC
ZNF703
EGFR
PARK2
TERT
FHIT
ERRFI1
MDM4
NEGR1
NF2
PTPRD
Cancer Type
CNS
Mesothelioma
NET
Kidney
Pancreas
Bone/Soft tissue
Head and neck
Skin
Liver
Biliary
Uterus
CUP
Breast
Prostate
Ovary
Lung
Colon/Rectum
Urinary tract
Stomach
Esophagus
Pan Cancer
(55.9%)
0%
25%
50%
75%
100%
% Samples
# Samples
300
200
100
0
1
2
3
4
5
6
7
Ploidy
WGD
FALSE
TRUE

Figure 3



A

	Breast	Colon/Rectum	Prostate	Lung	Skin	Esophagus	Urinary tract	Bone/Soft tissue	Ovary	Pancreas	Biliary	Kidney	CUP	Uterus	CNS	Other	Head and neck	NET	Stomach	Liver	Mesothelioma
TP53	46.1	80.7	43.2	71.7	17.9	85.4	58.6	28.5	72.8	83.8	50.0	22.4	52.0	31.8	30.2	36.1	49.2	19.4	89.5	27.3	20.0
CDKN2A	8.6	3.1	2.3	31.1	45.5	43.6	37.6	26.2	11.5	67.8	42.9	32.8	22.0	4.5	66.0	18.4	38.0	17.0	27.8	15.2	73.3
APC	1.5	77.3	10.2	5.4	5.0	12.8	1.7	0.9	0.4	1.8	8.3		14.0	2.4	0.5	8.6		2.9	3.4	1.3	
PTEN	13.5	8.3	41.8	7.5	20.2	4.1	2.4	6.5	4.2	3.6	1.8	14.9	17.7	18.2	35.8	4.1	4.8		10.3		
RB1	5.2	1.6	9.4	25.0	4.1	6.2	14.9	28.6	7.3		3.6		17.8	6.8	9.4	7.5	4.8	6.5	3.2	3.0	
SMAD4	2.1	24.7	1.4	2.8	1.0	13.3	4.7		2.1	48.2	26.8		5.5	2.3	0.7	4.1	4.8		20.7		6.7
ARID1A	5.1	5.2	1.5	6.8	1.6	17.2	13.6	3.1	5.2	5.6	10.7	13.2	17.9	14.2		2.9	7.0	6.5	6.3	6.1	
KMT2C	10.1	3.9	6.5	6.7	4.0	1.0	6.5	1.4	5.9	4.2	5.9	3.6	0.1	12.9	0.9	5.4	11.5	0.5	9.3		
NF1	7.9	1.9	1.7	7.2	5.7	5.1	4.4	5.5	13.5	1.8	11.8	3.6	0.1	5.8	11.0	11.9	5.8	2.4	3.2		
FHIT	0.6	7.8		3.4	0.5	52.6	3.5			8.9	3.6	3.0	8.0	4.5		2.0	2.4		41.4	3.0	
DMD	0.4	0.5	1.4	0.6	1.1	38.1	5.9	13.5	5.2	3.6	1.8	3.0	2.0	9.1		2.0	9.5	13.0	13.8		
MAP2K4	10.2	6.4	1.5	1.7		5.0			5.2	1.8	5.4		2.0		1.1	1.0	2.4	3.7	3.4		
ATM	3.0	6.2	7.4	4.3	4.1	0.7	6.3		1.0	3.7	4.2	6.8	0.4	1.7	3.9	2.9	3.0	1.6	5.0	9.0	5.6
KMT2D	1.8	2.0	4.1	4.2	1.8	1.8	22.4	1.0	1.0	7.0	1.8		5.3	9.0		8.9	15.8	2.5	3.0	2.5	
FBXW7	0.7	9.9	0.6	2.5	2.4	6.8	6.5	0.8	1.0	3.3	1.8		4.0	11.4		1.3	4.7		0.3		
PBRM1	1.2	1.0	0.8	3.2	2.8	0.3	2.7		0.5		12.5	47.4	6.0	5.2	0.1		2.4	2.5		9.1	6.7
CDH1	11.8	0.8	2.3	0.6	0.5	2.1	2.4			1.8	1.8		4.1			2.0		1.3			
FAT1	1.9	2.5	2.0	7.9	3.8	4.1	5.9	0.4	2.6	1.3	3.6	1.0	1.4	0.7	1.7	4.0	9.7	2.1	3.2	1.6	
KDM6A	0.8	1.6	7.5	3.4	0.5	7.2	21.2		2.0	3.6	5.4		5.2	0.7		2.0				3.0	
ZFP36L2	1.0	8.3	7.1	0.7	1.4	2.8	3.8	0.8	1.1	1.8	1.8	1.5	1.0			4.1				3.0	
STK11	1.5	1.9	0.5	15.1	1.6	3.1	3.3	0.8					8.0	13.7		2.0	4.8		3.4	3.0	
SETD2	1.4	0.7	0.9	5.7	2.5	2.3	2.9	1.6	1.3			31.0	5.6	2.3	3.5	0.5	2.4	3.9	0.6	3.4	
PTPRD	2.7	0.8	1.4	6.7	3.2	8.2	1.2		3.1		1.8	9.0		2.3	1.9	2.0	4.8	1.3	10.3		
BAP1	1.3	0.8	0.5	1.9	3.8	2.1	1.2	0.8	1.0	1.8	10.7	13.2	6.0			6.0	4.8			6.1	73.3
ATRX	1.8	1.5	0.9	3.1	2.5	1.0		10.3	1.6	2.3	0.3		0.9	3.1	13.2	2.1	2.4	5.1			6.7

B

	Breast	Colon/Rectum	Prostate	Lung	Skin	Esophagus	Urinary tract	Bone/Soft tissue	Ovary	Pancreas	Biliary	Kidney	CUP	Uterus	CNS	Other	Head and neck	NET	Stomach	Liver	Mesothelioma
PIK3CA	42.8	15.8	4.7	8.4	2.7	13.4	16.5	4.8	9.4	3.6	1.8	7.5	10.0	31.8	9.4	6.1	11.9	1.3	3.4		
KRAS	2.7	46.4	0.9	17.3	0.9	16.5	1.2	0.8	14.6	85.7	28.6		14.0	11.4		18.4		10.4	6.9	3.0	
TERT	2.1	1.9	0.9	8.9	62.4		51.8	8.7	3.1	3.6	7.1	13.4	8.0	9.1	52.8	16.3	11.9			48.5	13.3
BRAF	1.8	13.1	4.4	5.6	50.9	1.9	3.8	2.4	3.1			1.5	10.0		0.1	2.6	4.8		3.4		
ERBB2	16.0	7.0		6.7	1.7	22.7	19.2	1.2	3.1	5.5	10.7			10.9	0.4	4.1	9.5	0.3	21.5	6.1	
MYC	12.5	3.2	5.2	7.8	3.8	15.5	2.4	2.4	12.5	5.4	10.7		4.0	13.6		2.0	2.4	1.3	10.3	15.2	
CCND1	18.2	1.6	2.3	3.4	1.6	13.4	5.9	0.8			5.4		2.0			6.1	14.3		6.9	12.1	
AR	1.8	1.3	43.4	1.4	0.4	2.2			1.7		1.6		0.8	1.6		1.4			2.0		
ZNF703	18.0		1.9	3.4		3.1	9.4		1.0		1.8		2.0	2.3				1.3			
FGFR1	14.5	1.4	1.7	6.0	0.7	4.1	5.1	0.3	2.9		2.2		2.1	2.5	0.1			2.1	0.7		
ERG	0.7	0.6	37.6	0.6	0.6		1.2	1.5		0.2	2.7	1.5		2.3	0.1	1.1		1.3			
NRAS	0.8	5.4	0.9	0.6	24.2	1.0	2.4		1.0		1.8		6.0	2.3		6.1		1.3	3.4		
RAD21	9.1	1.5	1.9	1.6	2.1	1.3	3.3	0.8	6.0	3.0	5.0		2.0	2.3	0.7	2.0	2.4			14.2	
ESR1	16.1	0.4	0.2	0.2	0.2	0.3		0.8			0.7	0.8	1.1	2.9					0.1		
KIT	0.7	2.5	0.8	1.8	2.3	0.3	3.8	32.4	1.4	0.4	1.6	1.3	0.4	1.1	5.8	6.7	2.4	1.0	2.9		
ZBTB10	7.5	1.6	3.3	3.9	4.3		1.2						2.0	2.3		2.0	2.4		10.3	15.2	
PREX2	4.7	1.6	2.9	3.1	3.4	3.7	0.9	1.9	4.1	2.5	1.7	0.2	6.6	3.2	0.2	2.7	2.4	0.4	4.4	16.3	5.2
CTNNB1	0.6	4.7	4.7	1.7	2.2	6.4	2.4	1.6	2.1		3.6	1.5	6.7	6.8		4.1	2.4	3.9		30.3	
MDM2	5.0	0.8	0.5	2.2	1.6	4.1	11.8	3.2			10.7				5.7	4.1	7.1	3.9	6.9	3.0	
EGFR	1.3	0.5		14.5	0.5	5.2	1.2								43.4	2.0	4.8		6.9	3.0	
PABPC1	7.5	0.8	3.3	1.7	1.1	2.1	8.2		1.0				2.0	2.3		2.0	2.4		3.4	12.1	
CCNE1	1.7	0.3	0.9	4.5		12.4	10.6	1.6	9.4	7.1	5.4		2.0	6.8				1.3	13.8	3.0	
RSF1	11.3	0.3		2.2	1.1	1.0	1.2	0.8	1.0							2.0					
FOXA1	4.8	0.2	6.2	4.4	1.0	0.9	3.5						3.8		2.1	2.0		0.8		2.8	
NCOA2	4.2	0.5	3.8	1.1	3.2	1.0	1.2	0.8	1.0				2.0	2.3		2.0	2.4		3.4	18.2	

C

	Breast	Colon/Rectum	Prostate	Lung	Skin	Esophagus	Urinary tract	Bone/Soft tissue	Ovary	Pancreas	Biliary	Kidney	CUP	Uterus	CNS	Other	Head and neck	NET	Stomach	Liver	Mesothelioma
CHEK2	4.2	0.5	2.8	1.7	5.9	2.1	1.2	3.2	1.0	3.6	5.4	3.0	2.0		3.8	4.1					
BRCA2	2.7	0.3	2.3		1.6	1.0			4.2	5.4	1.8		2.0			2.0		1.3			
MUTYH	0.8	1.3	0.9	0.6	1.6				3.1	1.8		1.5						1.3			
BRCA1	0.8	0.3			0.5	1.0		0.8	7.3	1.8		1.5						1.3			
ATM	0.6	0.5	1.9	0.6		1.0	1.2	0.8				1.5						1.3			
MSH6	0.4	0.5					2.4								1.9						
PMS2		0.3		0.6				0.8							3.8		2.4			3.0	
APC	0.2	0.8			0.5			0.8													
PTEN	0.2			1.1	0.5												2.4	1.3			
BAP1			0.5		1.1																

Figure 4

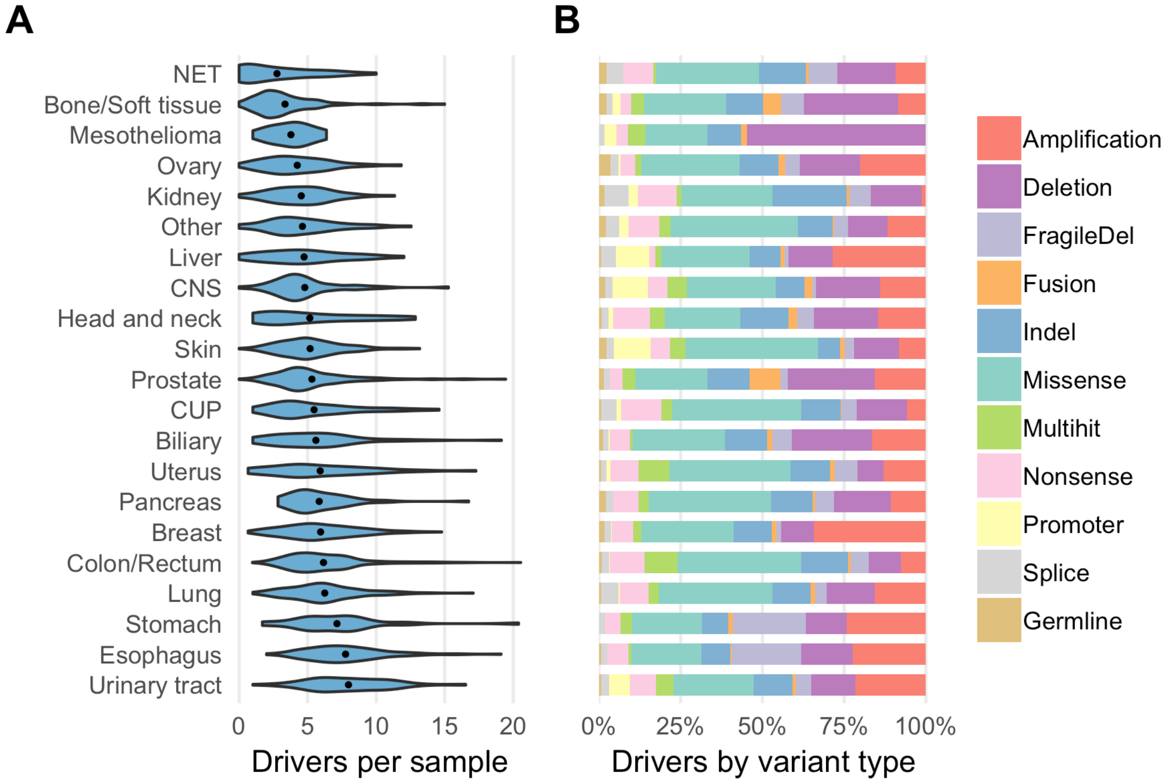

A
B
NET
Bone/Soft tissue
Mesothelioma
Ovary
Kidney
Other
Liver
CNS
Head and neck
Skin
Prostate
CUP
Biliary
Uterus
Pancreas
Breast
Colon/Rectum
Lung
Stomach
Esophagus
Urinary tract
0
5
10
15
20
Drivers per sample
0%
25%
50%
75%
100%
Drivers by variant type
Amplification
Deletion
FragileDel
Fusion
Indel
Missense
Multihit
Nonsense
Promoter
Splice
Germline

Figure 5

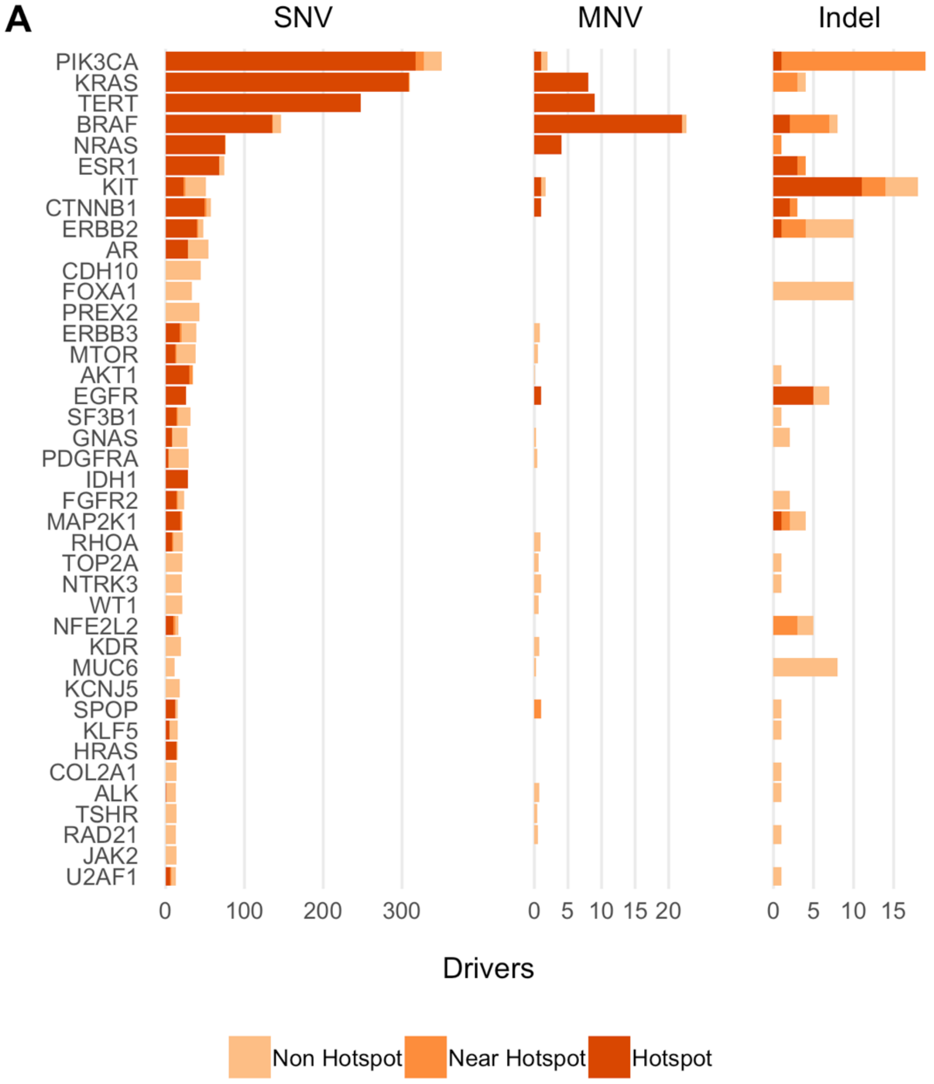

A
SNV
MNV
Indel
PIK3CA
KRAS
TERT
BRAF
NRAS
ESR1
KIT
CTNNB1
ERBB2
AR
CDH10
FOXA1
PREX2
ERBB3
MTOR
AKT1
EGFR
SF3B1
GNAS
PDGFRA
IDH1
FGFR2
MAP2K1
RHOA
TOP2A
NTRK3
WT1
NFE2L2
KDR
MUC6
KCNJ5
SPOP
KLF5
HRAS
COL2A1
ALK
TSHR
RAD21
JAK2
U2AF1
0
100
200
300
0
5
10
15
20
0
5
10
15
Drivers
Non Hotspot
Near Hotspot
Hotspot

Figure 6

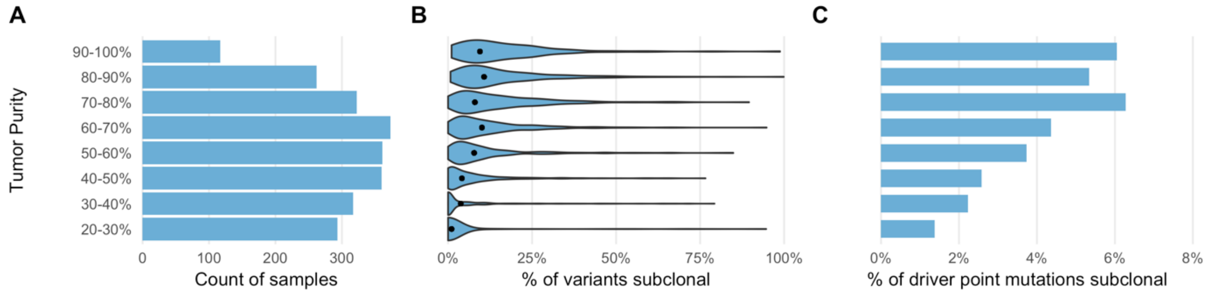

A
B
C
Tumor Purity
90-100%
80-90%
70-80%
60-70%
50-60%
40-50%
30-40%
20-30%
0
100
200
300
Count of samples
0%
25%
50%
75%
100%
% of variants subclonal
0%
2%
4%
6%
8%
% of driver point mutations subclonal

Figure 7

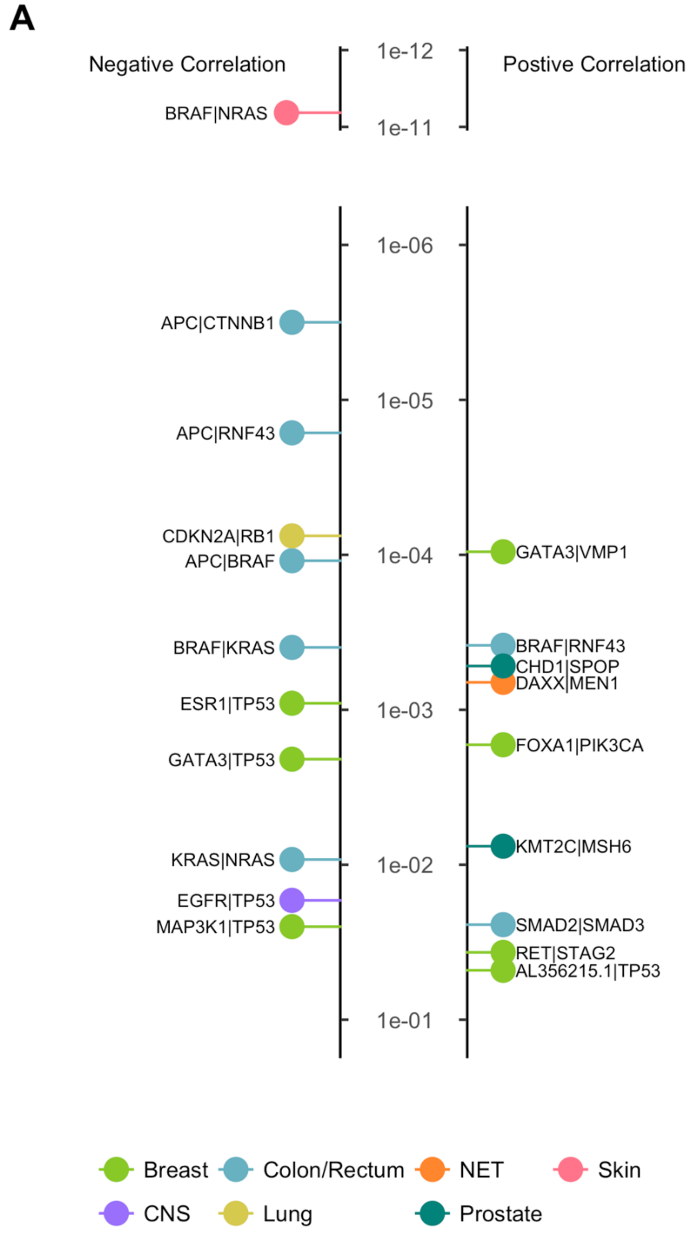
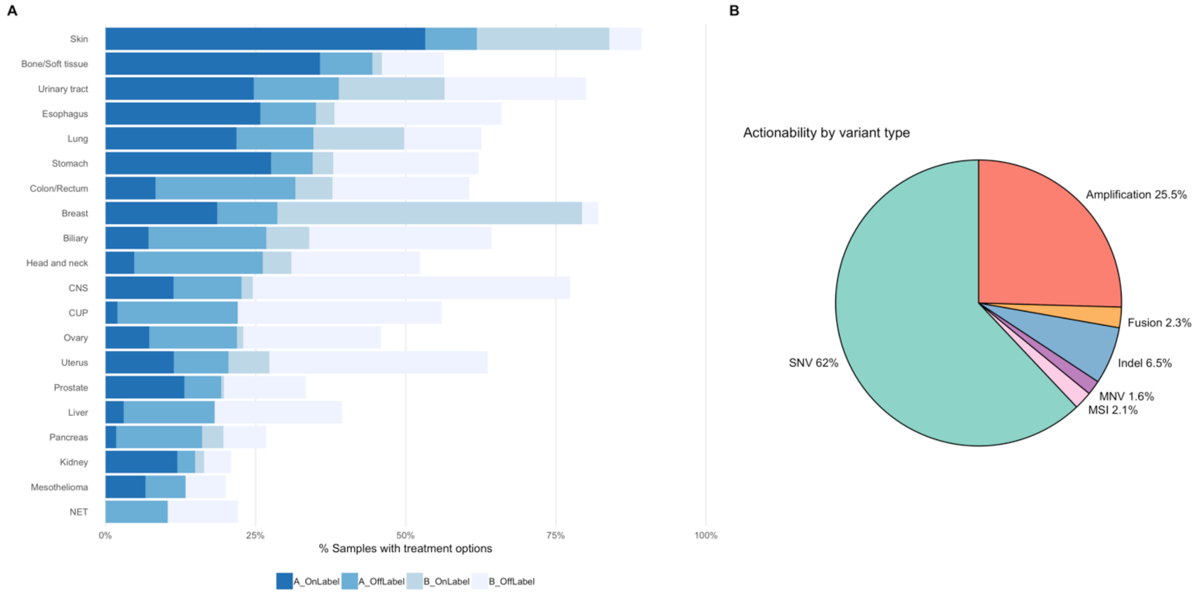

A
Negative Correlation
Postive Correlation
1e-12
1e-11
1e-06
1e-05
1e-04
1e-03
1e-02
1e-01
BRAF|NRAS
APC|CTNNB1
APC|RNF43
CDKN2A|RB1
APC|BRAF
BRAF|KRAS
ESR1|TP53
GATA3|TP53
KRAS|NRAS
EGFR|TP53
MAP3K1|TP53
GATA3|VMP1
BRAF|RNF43
CHD1|SPOP
DAXX|MEN1
FOXA1|PIK3CA
KMT2C|MSH6
SMAD2|SMAD3
RET|STAG2
AL356215.1|TP53
Breast
Colon/Rectum
NET
Skin
CNS
Lung
Prostate

Figure 8


A
Skin
Bone/Soft tissue
Urinary tract
Esophagus
Lung
Stomach
Colon/Rectum
Breast
Biliary
Head and neck
CNS
CUP
Ovary
Uterus
Prostate
Liver
Pancreas
Kidney
Mesothelioma
NET
0%
25%
50%
75%
100%
% Samples with treatment options
A_OnLabel
A_OffLabel
B_OnLabel
B_OffLabel
B
Actionability by variant type
Amplification 25.5%
Fusion 2.3%
Indel 6.5%
MNV 1.6%
MSI 2.1%
SNV 62%

Detailed methods for

Pan-cancer whole genome analyses of metastatic solid tumors

Peter Priestley, Jonathan Baber, Martijn P. Lolkema, Neeltje Steeghs, Ewart de Bruijn, Korneel Duyvesteyn, Susan Haidari, Arne van Hoeck, Wendy Onstenk, Paul Roepman, Charles Shale, Mircea Voda, Haiko J. Bloemendal, Vivianne C.G. Tjan-Heijnen, Carla M.L. van Herpen, Mariette Labots, Petronella O. Witteveen, Egbert F. Smit, Stefan Sleijfer, Emile E. Voest, Edwin Cuppen

Content

1. Sample collection

Patients with advanced cancer not curable by local treatment options and being candidates for any type of systemic treatment and any line of treatment were included as part of the CPCT-02 (NCT01855477) and DRUP (NCT02925234) clinical studies, which were approved by the medical ethical committees (METC) of the University Medical Center Utrecht and the Netherlands Cancer Institute, respectively. A total of 41 academic, teaching and general hospitals across the Netherlands participated in these studies and collected material and clinical data by standardized protocols[1]. Patients have given explicit consent for whole genome sequencing and data sharing for cancer research purposes. Clinical data, including primary tumor type, biopsy location, gender and birth year were collected in electronic case record forms and stored in a central database.

Core needle biopsies were sampled from the metastatic lesion, or when considered not feasible or not safe, from the primary tumor site when still in situ. One to four biopsies were collected (average of 2.1 per patient) and frozen in liquid nitrogen directly after sampling and further processed at a central pathology tissue facility. Frozen biopsies were mounted on a microtome in water droplets for optimal preservation of all types of biomolecules (DNA, RNA and proteins) for subsequent and future omics-based analyses. A single 6 micron section was collected for hematoxylin-eosin (HE) staining and estimation of tumor cellularity by an experienced pathologist. Subsequently, 25 sections of 20 micron, containing an estimated 25,000 to 500,000 cells, were collected in a tube for DNA isolation. In parallel, a tube of blood was collected in CellSave (Menarini-Silicon Biosystems) tubes, which was shipped by room temperature to the central sequencing facility at the Hartwig Medical Foundation. Left-over material (biopsy, DNA) after sample processing was stored in biobanks associated with the studies at the University Medical Center Utrecht and the Netherlands Cancer Institute.

2. Sequencing workflow

DNA was isolated from biopsy and blood on an automated setup (QiaSymphony) according to supplier's protocols (Qiagen) using the DSP DNA Midi kit for blood and QIAsymphony DSP DNA Mini kit for tissue and quantified (Qubit). Before starting DNA isolation from tissue, the biopsy was dissolved in 100 microliter Nuclease-free water by using the Qiagen TissueLyzer and split in two equal fractions for parallel automated DNA and RNA isolation (QiaSymphony). Typically, DNA yield for the tissue biopsy ranged between 50 and 5,000 ng. A total of 50 - 200 ng of DNA was used as input for TruSeq Nano LT library preparation (Illumina), which was performed on an automated liquid handling platform (Beckman Coulter). DNA was sheared using sonication (Covaris) to average fragment lengths of 450 nt. Barcoded libraries were sequenced as pools (blood control 1 lane equivalent, tumor 3 lane equivalents) on HiSeqX (V2.5 reagents) generating 2 x 150 read pairs using standard settings (Illumina).

BCL output from the HiSeqX platform was converted using bcl2fastq tool (Illumina, versions 2.17 to 2.20 have been used) using default parameters. Reads were mapped to the reference genome GRCH37 using BWA-mem v0.7.5a[2], duplicates were marked for filtering and INDELs were realigned using GATK v3.4.46 IndelRealigner[3]. GATK HaplotypeCaller v3.4.46[4] was run to call germline variants in the reference sample. For somatic SNV and INDEL variant calling, GATK BQSR[5] is also applied to recalibrate base qualities.

3. Somatic point mutation calling

We called SNV & INDEL somatic variants using Strelka v1.0.14[6] with the following optimisations:

- **Preservation of known variants:** From the raw Strelka output we marked all known pathogenic variants from external databases such that these would be preserved from all subsequent filtering. The list of pathogenic variants used was the union of:
  - Point mutations in CIViC[7] with level C evidence or higher (download = 01-mar-2018)
  - Somatic variants from CGI[8] (update: 17-jan-2018)
  - Oncogenic or likelyOncogenic variants from OncoKb[9] (download = 01-mar-2018); http://oncokb.org/api/v1/utils/allAnnotatedVariants.txt)
  - TERT promoter variants at genomic coordinates: 5:1295242, 5:1295228, 5:1295250
- **Modified quality score filtering**
  - We split variants into high confidence (HC) and low confidence (LC) regions using the NA12878 GIABv3.2.2 high confidence region definitions[10], based on the observation that we produce far higher rates of false positives variant calls in LC regions
  - Set quality score cutoffs for SNV & INDEL to 10 for HC regions and 20 for LC regions (default = 15 for SNV, 30 for INDEL)
  - Added an additional quality filter to tighten filtering for low allelic frequency variants: quality score * allele frequency > 1.3
- **Improved repeat sensitivity:** Switched off the default Strelka repeat filter to improve indel calling in microsatellites and short repeats.
- **Panel of normals (PON) to remove germline leakage:** Filtered out any variants which were found by GATK haplotypecaller in more than 5 samples in a germline PON consisting of 2000 of our reference blood samples. PON available at (https://resources.hartwigmedicalfoundation.nl/)
- **PON to remove strelka-specific artefacts:** Filtered any variant which was supported by 2 or more reads in strelka in the reference sample in at least 4 patients in our cohort. PON available at (https://resources.hartwigmedicalfoundation.nl/)
- **Removal of INDELS near a PON filtered INDEL -** Regions of complex haplotype alterations are often called as multiple long indels, which can make it more difficult to construct an effective PON, and sometimes we find residual artefacts at these locations. Hence we also filter inserts or deletes which are 3 bases or longer where there is a PON-filtered INDEL of 3 bases or longer within 10 bases in the same sample.
- **MNV Correction** - Variants occurring on consecutive positions, or 1 base apart were considered potential multi nucleotide variants (MNVs). The BAM files were re-examined, and the variants were merged into a single MNV if greater than 80% of the reads with a mapping quality score of at least 10 and which are neither unmapped, duplicated, secondary, nor supplementary containing any of the individual variants also contained the other variants of the potential MNV. The attributes of the resulting MNV variant were determined by picking the minimum values from the individual variants forming the MNV. MNVs were marked as PON filtered only if both individual variants were PON filtered.

The settings and tools for this optimized HMF pipeline are available at https://github.com/hartwigmedical/.

4. Validation of somatic point mutation calling

We performed three separate analyses to validate our somatic variant calling pipeline as follows:

4.1. Validation of somatic precision and sensitivity pipeline on a known benchmark

We tested the default Strelka and HMF optimized settings on a GIAB mix-in sample (ref = NA24385; tumor = 70% NA24385 and 30% NA12878) to test sensitivity at a realistic purity and on a null tumor (ref = NA12878, tumor = NA12878) to test precision. The results of this analysis are as follows:

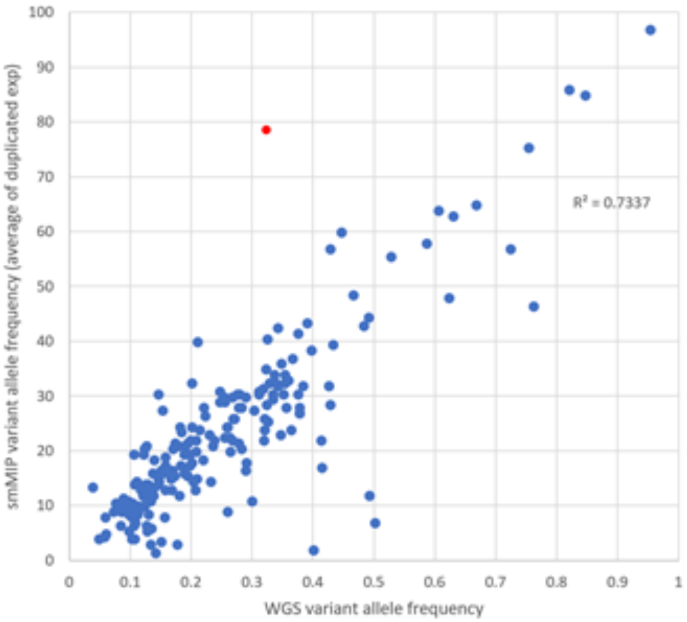
Configuration	SNV sensitivity	SNV false positive / genome	INDEL sensitivity	Indel false positive / genome
Strelka default	93%	3500	24%	41
Optimized HMF pipeline	96%	109	77%	27

4.2. External independent validation of SNV and INDEL calling precision on real samples

We performed external validation of a set of single nucleotide variants (SNV) and short insertion/deletions (indels) that have been detected by Whole Genome Sequencing (WGS) using the single molecule Molecular Inversion Probe (smMIP) technology[11]. SNV and short indels variants were semi-randomly selected from 30 patient samples. The first selection was to include every variant that was reported in a panel of 114 ‘actionable’ cancer genes as used in the routine CPCT-02 study analysis. This way, a total of 82 variants (67 SNVs, 15 indels) were selected in 45 genes. The second selection involved random sampling adding up to a total of 256 coding and non-coding variants from the same 30 patient samples.

A custom smMIP panel was designed to cover the selected variants. For 45 variants (17.6%) no smMIP design was possible, all of which were intergenic variants. For the other 211 variants probes could successfully be designed. Analysis of the smMIP sequencing data indicated that for 17 of the 211 variants (8.1%) the smMIP sequencing data was of insufficient quality (mostly due to repeat stretches), while the WGS data seemed sufficiently reliable for accurate calling (confirmed by visual inspection of the read data), including 3 coding variants (*RB1, ERBB4* and *BRCA2*) and 14 intergenic regions. The retrospective investigation of the WGS data indicated that for another three variants (1.4%) the smMIP as well as the WGS data was of insufficient quality due to large homopolymer stretches.

In total 192 variants could be successfully sequenced and analyzed using the smMIP and could be used for confirmation of the WGS findings. 189 SNVs and indel variants (98.4%) were confirmed by smMIP sequencing, indicating a very high accuracy of WGS-derived variant calling results. All three variants that could not be confirmed by smMIP were from intergenic regions, including 1 variant that showed a mixed double-variant (chr3:75887550_G>T/C) and for which both technologies had difficulties in accurately calling the genotype. For the remaining 2 variants (chr8:106533360_106533361insAC, chr12:125662751_125662752insA), it remains unclear if these could not be detected by smMIP or were falsely called by WGS, as they fall in repetitive genomic stretches.



The 189 successfully confirmed variants showed a good linear correlation in variant allele frequency between WGS and smMIP sequencing (average of duplicates) with an R^2 of 0.733. This result indicated that WGS, with its lower read depth (on average between 100-110x) than smMIP and without a read-barcoding system, is accurate in quantitatively determining the variant frequency at frequencies above 5%. One variant (ch19:55276095C>T, indicated in red in the figure above) showed a large deviation in variant frequency, which was likely due to the much lower than expected coverage of the variant, both in the WGS (37 reads) as well as in the smMIP data (28 and 35 reads).

4.3. Validation of somatic variant calling sensitivity by reanalysis of known hotspots.
To validate somatic calling sensitivity and performance limitations of our pipeline on real samples, we built a customised tool, SAGE (https://github.com/hartwigmedical/hmftools/tree/master/sage) to reanalyse all 10,211 known pathogenic hotspot variants in the coding region of the genome (sourced from CIVIC, OncoKb and CGI as described above). These locations have a much higher prior likelihood of finding a variant in cancer samples.

SAGE searches for each hotspot in the tumor BAM files directly and calls a variant if the sum of read base qualities supporting the ALT > 100, effectively equating to 3 high quality reads of support. Our standard somatic pipeline typically requires 6 or more reads support to call a variant. For the purposes of this validation we excluded from SAGE a small number of variants in high repeat contexts (repeat count >=8) and in regions with very high tumor copy number (tumor read depth > 300) as both these contexts can cause low VAF artefacts which we want to avoid in a sensitivity test.

We evaluated on a randomly selected 1247 samples with the following results

Hotspot variants found in standard somatic pipeline	Additional variants found by SAGE	% variants missed by somatic pipeline
1160	37	3.1%

Of the 37 additional variants found by SAGE but not in our standard somatic pipeline, 27 (2.3%) were found to have been missed by Strelka due to low read count in the tumor (all with only 3 to 6 reads supporting the ALT allele), 8 (0.7%) due to insufficient coverage in the reference sample, and 2 (0.2%) for unknown reason.

Overall this analysis suggests that we capture more than 96% of all variants with 3 or more reads of support in the tumor (equivalent to ~3% VAF).

5. Somatic structural variant calling

Structural Variants were called using Manta(v1.0.3)[12] with default parameters. We then re-examined each breakpoint, calculated variant allele frequencies for each break end and applied seven additional filters to the Manta output to improve precision using an internally built tool called 'Breakpoint-Inspector' (BPI, https://github.com/hartwigmedical/hmftools/tree/master/break-point-inspector) v1.5. Two main types of filters are applied by BPI:

- **Evidence of variant in reference sample** - Variants are filtered out if we can find any evidence of paired read support, split read support or soft clipping concordance (5+ bases at exact breakpoint) in the matching blood sample.
- **Inadequate support for variant in tumor sample** - For all inversions and translocations and for long deletions and tandem duplications (>1000 bases between breakpoints) we require at least 1 read with paired read support. For short deletions and duplications (<1000 bases between breakpoints) we require at least 1 read with split read support. In both cases at least one of those reads must be anchored with at least 30 bases at each breakpoint. We also require the minimum read coverage across each breakpoint in the tumor to be > 10 depth.

Each breakend was annotated with it's position in all transcripts from 'KNOWN' genes in Ensembl v89.37[13]. Each gene was marked as disrupted if there was at least one structural variant that impacted on the canonical transcript.

6. Identification of gene fusions

For each structural variant, every combination of annotated overlapping transcripts from each breakend was tested to see if it could potentially form an intronic inframe fusion. A list of 411 curated known fusion pairs was sourced by taking the union of known fusions from the following external databases:

- Cosmic curated fusions[14] (v83)
- OncoKb[9] (download = 01-mar-2018)
- CGI[8] (update: 17-jan-2018)
- CIViC[7] (download = 01-mar-2018)

We then also created a list of promiscuous fusion partners using the following rules

- **3' promiscuous:** Any gene which appears on the 3' side in more than 3 of the curated fusion pairs OR appears at least once on the 3' side and is marked as promiscuous in either OncoKb, CGI or CIVIC
- **5' promiscuous:** Any gene which appears on the 5' side in more than 3 of the curated fusion pairs OR appears at least once on the 5' side and is marked as promiscuous in either OncoKb, CGI or CIVIC

For each promiscuous partner we also curated a list of essential domains that must be preserved to form a viable fusion partner.

Finally, we report an intronic inframe fusion if the following conditions are met

- Matches an exact fusion from the curated list OR is intergenic and matches 5' promiscuous OR matches 3' promiscuous gene
- Curated domains are preserved
- Does not involve the 3'UTR region of either gene
- For intragenic fusions, must start and end in coding regions of the gene
- 3' partner is a protein coding gene and the transcript does not result in nonsense mediated decay

7. Validation of gene fusions

Whole transcriptome analysis (RNA-seq) of 60 samples with identified fusions was used to validate our gene fusion calling pipeline.

RNA was isolated from the same biopsy material as used for DNA isolation using an automated setup (QiaSymphony) using the QIAsymphony RNA kit (#931636, Qiagen) according to supplier's protocols. RNA was quantified using Qubit RNA HS Assay Kit (Thermo Fisher). Typically, RNA yield for the tissue biopsy ranged between 500 and 5,000 ng. 100 ng of total RNA was used as input for KAPA RNA HyperPrep Kit with RiboErase (HMR) (#KR1351, Roche) and TruSeq DNA CD Indexes 96 Indexes (#PN 20015949, Illumina) performed on an automated liquid handling platform (Beckman Coulter). The standard protocol used involved 240 sec 85 degrees Celcius fragmentation and 15 PCR cycles. Each sample was subsequently sequenced in a multiplexed setup with 2x75 bp reads on a NextSeq 500/550 using the High Output Kit v2 (Illumina, #FC-404-2002), targeting 50M raw reads per sample. BCL output from the NextSeq500 platform was converted using Illumina bcl2fastq tool (versions 2.17 to 2.20 have been used) using default parameters.

STAR-Fusion[15] was used with default settings to call fusion transcripts from the RNA. 38 out of 60 fusions were readily identified independently in the RNA. Manual inspection of the expected chimeric junctions for the remaining 22 fusions revealed RNA support for a further 6 fusions (4 of which were TMPRSS2-ERG), although below the threshold to be called automatically in the RNA with the settings used. Overall, 73% of the tested fusions were thus independently validated by the RNA analysis. The full results are summarised below:

Total fusions tested	**Transcript fusion found by STAR-Fusion**	**Read support in RNA but not called by STAR-Fusion**	**No evidence of fusion transcript in RNA**
60	38 (63%)	6 (10%)	16 (27%)

8. Purity, ploidy and copy number calling

Accurate copy number calling is closely linked with correct sample purity determination. Currently, there is not a clear consensus in the community for a preferred tool for this purpose. We tested several tools (freeC, CANVAS and Sequenza) on the COLO829 benchmark, but none of them provided a correct fit[16]. Therefore we developed PURPLE (PURity & PLoidy Estimator) as an alternative.

PURPLE combines B-allele frequency (BAF), read depth and structural variants to estimate the purity and copy number profile of a tumor sample and follows a similar purity fitting methodology to several other popular tools such as ASCAT, Sequenza and CANVAS, only with a different optimisation function to determine the best fit.

The main advantages of PURPLE (v2.14) for the purposes of this study are:

- extensive attention to removal of artefacts by filtering of inputs (see below sections 7.1, 7.2 and 7.3) and smoothing of output to avoid false positive copy number calling (section 7.5)
- integrated SV and copy number calling allow single base accuracy of copy number calls and accurately call each individual variant as heterozygous or homozygous as well as the detection of partial loss of genes

There are five key steps in the PURPLE pipeline:

1. Calculate BAF in tumor at high confidence heterozygous germline loci
We determine the BAF of the tumor sample by finding heterozygous locations in the reference sample from a panel of 796,447 common germline heterozygous SNP locations. To ensure that we only capture heterozygous points, we filter the panel to only loci with allelic frequencies in the reference sample between 40% and 60% and with depth between 50% and 150% of the reference sample genome wide average. Typically, this yields 140k-200k heterozygous germline variants per patient. We then calculate the allelic frequency of corresponding locations in the tumor.

2. Determine read depth ratios for tumor and reference genomes
The raw read counts per 1,000 base window for both normal and tumor samples, by counting the number of alignment starts in the respective bam files with a mapping quality score of at least 10 that is neither unmapped, duplicated, secondary, nor supplementary. Windows with a GC content less than 0.2 or greater than 0.6 or with an average mappability below 0.85 are excluded from further analysis.

Next we apply a GC normalization to calculate the read ratios. We divide the read count of each window by the median read count of all windows sharing the same GC content then normalise further to the ratio of the median to mean read count of all windows.

Finally, the reference sample ratios have a further ‘diploid’ normalization applied to them to remove megabase scale GC biases. This normalization assumes that the median ratio of each 10Mb window (minimum 1Mb readable) should be diploid for autosomes and haploid for sex chromosomes in males in the germline sample.

3. Segmentation
We segment the genome into regions of uniform copy number by combining segments generated from the read ratios for both tumor and reference sample, from the BAF points with structural variant breakpoints derived from Manta & BPI. Read ratios and BAF points are segmented independently using the Bioconductor copynumber package[17] which uses a piecewise constant fit (PCF) algorithm (with custom settings gamma = 100, k =1). These segment breaks are then combined with the structural variants breaks according to the following rules:

1. Every structural variant break starts a new segment, as does chromosome starts, ends and centromeres. This is regardless of if they are distinguishable from existing segments or not.
2. Ratio and BAF segment breaks are only included if they are distinguishable from an existing segment.

3. To be distinguishable, a break must be at least one complete mappable read depth window away from an existing segment.

Once the segments have been established we map our observations to them. In each segment we take the median BAF of the tumor sample and the median read ratio of the tumor and reference samples. We also record the number of BAF points within the segment as the BAFCount.

A reference sample copy number status is determined at this this stage based on the observed copy number ratio in the reference sample, either ‘DIPLOID’ (0.8<= read depth ratio<=1.2), ‘HETEROZYGOUS_DELETION’ (0.1<=ratio<0.8), ‘HOMOZYGOUS_DELETION’ (ratio<0.1),’AMPLIFICATION’(1.2<ratio<=2.2)or ‘NOISE’ (ratio>2.2). The purity fitting and smoothing steps below use only the DIPLOID germline segments.

4. Purity Fitting

Next we jointly fit tumor purity and sample ploidy (expressed as a normalisation factor) according to the following principles:

1. The absolute copy number of each segment should be close to an integer ploidy
2. The BAF of each segment should be close to a % implied by integer major and minor allele ploidies.
3. Higher ploidies have more degenerate fits but are less biologically plausible and should be penalised
4. Segments are weighted by the count of BAF observations which is treated as a proxy for confidence of BAF and read depth ratio inputs.
5. Segments with lower observed BAFs have more degenerate fits and are weighted less in the fit

For any given tumor purity and sample ploidy we calculate the score by first modelling the major and minor allele ploidy of each segment and minimising the deviation between the observed and modelled values according to the following formulas:

ModelDeviation = abs(ObservedRatio - ModelRatio) + abs(ObservedBaf - ModelBaf)
ModelBaf = (tumorPurity * (segmentMinorPloidy - 1) + 1) / (tumorPurity * (segmentPloidy - 2) + 2)
ModelRatio = sampleNormFactor + (segmentPloidy - 2) * tumorPurity * sampleNormFactor / 2d;

Once modelled, each segment is given a ploidy penalty:

PloidyPenalty = 1 +min(SingleEventDistance, WholeGenomeDoublingDistance);
WholeGenomeDoublingDistance = 1 + abs(segmentMajorAllele - 2) +abs(segmentMinorAllele - 2);
SingleEventDistance = abs(segmentMajorAllele - 1) + abs(segmentMinorAllele - 1);

Summing up over all the segments generates a score for each tumor purity / sample ploidy combination from which we can select the minimum:

$$FitedPurityScore = \frac{1}{TotalBafCount}\sum_{i=1}^{n} PloidyPenalty_i \times ModelDeviation_i \times BafCount_i \times ObservedBaf_i$$

If a sample has a fitted purity solution which is >98.5% diploid and a score within 10% of the best fitted score, the sample is designated as highly diploid and a fit is determined by the highest vaf somatic ploidy peak.

Given a fitted purity and sample ploidy we are then able to determine the purity adjusted copy number and BAF of each segment in the tumor genome from the unadjusted read ratios and BAFs respectively.

5. Smoothing

Since the segmentation algorithm is highly sensitive, and there is a significant amount of noise in the read depth in whole genome sequencing, many adjacent segments created above will have a similar copy number and BAF profile and can be combined and averaged to form a larger, smoothed, region.

We apply a number of rules to merge adjacent regions to create a smooth copy number profile.

1. Never merge a segment break created from a structural variant break end.
2. Use the count of BAF points as a proxy for confidence or weight in the region. Note that some segments may have a BAF count of 0.
3. Merge segments where the difference in BAF and copy number is within tolerances.
4. BAF tolerance is linear between 0.03 and 0.35 dependent on BAF count.
5. Copy number tolerance is linear between 0.3 and 0.7 dependent on BAF count. The tolerance also increases linearly as purity of the tumor sample decreases below 20%.
6. Start from most confident segment and smooth outwards until we reach a segment outside of tolerance. Move on to next highest unsmoothed section.
7. It is possible to merge in (multiple) segments that would otherwise be outside of tolerances if:
    a. The total dubious region is sufficiently small (<30k bases or <50k bases if approaching centromere); and
    b. The dubious region does not end because of a structural variant; and
    c. The dubious region ends at a centromere, telomere or a segment that is within tolerances.
8. When the entire short arm of a chromosome is lacking copy number information (generally on chromosome 13, 14, 15, 21 or 22), the copy number of the long arm is extended to the short arm.
9. Any remaining unknown segments are given the expected copy number of their associated chromosome, i.e. 2 for autosomes and female allosomes, 1 for male allosomes.

Where clusters of SVs exist which are closer together than our read depth ratio window resolution of 1,000 bases, the segments in between will not have any copy number information associated with them. To resolve this, we infer the ploidy from the surrounding copy number regions. The outermost segment of any SV cluster will be associated with a structural variant with a ploidy that can be determined from the adjacent copy number region and the VAF of the SV. We use this ploidy and the orientation of structural variant to calculate the change in copy number across the SV and hence the copy number of the outermost unknown segment. We repeat this process iteratively and infer the copy number of all regions within a cluster.

Once region smoothing is complete, it is possible there will be regions of unknown BAF, if no BAF points were present in a copy number region. We infer this BAF by assuming that they share their minor allele ploidy with their neighbouring region. If there are multiple neighbouring regions with known BAF we use the highest confident region (i.e. highest BAF count) to infer.

At this stage we have determined a copy number and minor allele ploidy for every base in the genome.

9. Validation of purity, ploidy and copy number output

We performed three validations to evaluate the purity and ploidy estimates and copy number profile obtained from PURPLE.

1. Validation of purity estimates through cell line in-silico dilutions

The purity estimates of PURPLE were validated using the tumor cell line COLO829. We created diluted in-silico mixture models of the tumor and blood cell lines from COLO829 with simulated purities of 20%, 30%, 40%, 60%, 80% and 100%, and ran PURPLE on the simulated BAM files against the reference sample.
The PURPLE estimates were found to match the simulation very closely as shown in the table below:

Simulated Purity	PURPLE estimated purity	Difference
20%	20%	0%
30%	30%	0%
40%	40%	0%
50%	50%	0%
60%	60%	0%
80%	81%	1%
100%	100%	0%

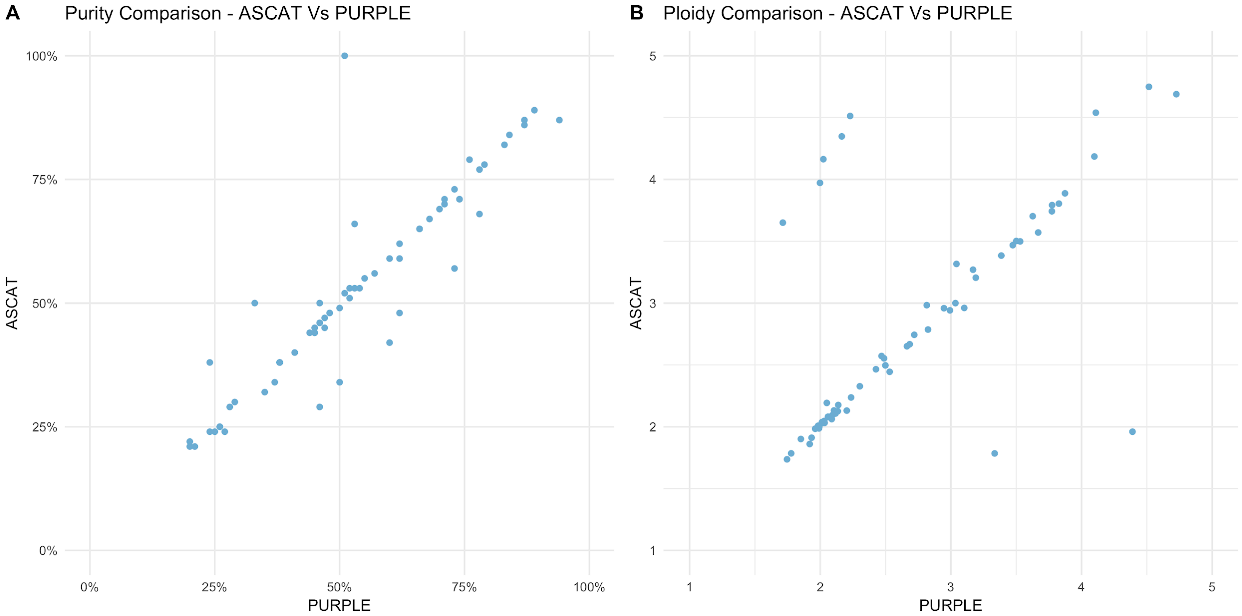
2. Validation of absolute copy number predictions by FISH

We also validated the absolute copy number results for PURPLE by comparing the WGS analysis results of the COLO-829 tumor vs normal cell line pair with DNA Fluorescence In Situ Hybridization (FISH) results for the centromeric region of chromosome 9, 13, 16 and 18 (CEP9, CEP13, CEP16, CEP18) and for the 2p23 ALK locus and the 9p24 JAK2 locus. In total, 100 COLO829 tumor cells were scored for each of the six FISH probes. For both assays the local copy-number as well as the percentage of DNA (PURPLE) or number of cells (FISH) is provided in the table below to indicate the intratumoral heterogeneity. The FISH and sequencing based results showed a very high concordance for the chromosomal copy numbers and the intratumoral heterogeneity (COLO-829 cell line heterogeneity has been described previously[18]).

Genomic region	PURPLE ploidy and purity	FISH copy number
Centromere Chr 9	3.7-4.0 : 53-57%	*2n* : *33%* 3n : 9% 4n : 58%
Centromere Chr 13	3.2 : 55%	*2n* : *41%* 3n : 59%
Centromere Chr 16	*2.0 : 100%*	*2n* : *100%*
Centromere Chr 18	2.8-2.9 : 67-71%	2n : 38% 3n : 62%
ALK (2p23)	3.1 : 67%	2n : 21% 3n : 79%
JAK2 (9p24)	*2.0 : 100%*	*2n* : *100%*

3. Comparison of PURPLE purity and ploidy estimates on patient samples with ASCAT
To validate PURPLE on real patient data, we compared the purity and ploidy outputs from PURPLE to the widely used copy number tool ASCAT[19] for 65 randomly selected samples from our cohort. ASCAT was run on GC corrected data using default parameters except for gamma which was set to 1 which is recommended for massively parallel sequencing data.

The following charts show a comparison of ASCAT vs PURPLE purity and ploidy results with 55 of 65 samples (85%) in agreement to within 10% absolute purity and relative sample ploidy.



There are 2 types of differences observed in the remaining 10 samples:

- Purity differences for highly diploid samples - this is unsurprising as PURPLE has additional functionality which is not dependent on copy number alterations in the tumor for highly diploid samples to fit the somatic ploidies whereas ASCAT does not.
- Whole genome duplication (WGD) vs no whole genome duplication - In 5 of the samples ASCAT calls a WGD event whereas PURPLE does not and in 2 samples the opposite occurs. This reflects the tradeoff in the purity and ploidy determination between penalising higher ploidy solutions which are more degenerate vs lower ploidy solutions with more subclonality. Manual inspection of purity-corrected fitted minor allele ploidy plots reveals in all of the 5 cases where ASCAT calls a WGD that whilst there is subclonality in each of these cases in the PURPLE solution there is no subclonal peak at 0.5 copy number, nor is there a 0.5 somatic ploidy peak, suggesting that the the WGD solution is less likely. Conversely, in the 2 cases where PURPLE only calls a WGD, manual inspection shows that the ASCAT solution would be prefered in one case and the PURPLE solution in the other.

In summary, overall concordance is very high between PURPLE and ASCAT. There appears to be little systematic bias to either calling lower or higher ploidy solutions between methods, and where PURPLE differs from ASCAT it more often than not appears to be the more plausible solution.

10. Sample filtering based on copy number output

Following our copy number calling, samples were QC filtered from the analysis based on 4 criteria:

- **NO_TUMOR** - If PURPLE fails to find any aneuploidy AND the number of somatic SNVs found is less than 1,000 then the sample is marked as NO_TUMOR.
- **MIN_PURITY** - We exclude samples with a fitted purity of <20%
- **FAIL_SEGMENT** - We remove samples with more than 120 copy number segments unsupported at either end by SV breakpoints. This step was added to remove samples with extreme GC bias, with differences in depth of up to or in excess of 10x between high and low GC regions. GC normalisation is unreliable when the corrections are so extreme so we filter.
- **FAIL_DELETED_GENES -** We removed any samples with more than 280 deleted genes. This QC step was added after observing that in a handful of samples with high MB scale positive GC bias we sometimes systematically underestimate the copy number in high GC regions. This can lead us to incorrectly infer homozygous loss of entire chromosomes, particularly on chromosome 19.

Where multiple biopsies exist for a single patient, we always choose the highest purity sample for our analysis of mutational load and drivers.

11. Impact of sequencing depth coverage on somatic variant calling sensitivity

To assess the impact of our sequencing depth on variant calling sensitivity, we selected 10 samples at random, downsampled the BAMs by 50%. We then reran the identical somatic variant calling pipeline.

Comparing the output to the original runs, we found near identical purities and ploidies for the down sampled runs (Extended Data Fig. 2). We observed an average decrease in sensitivity of 10% for SNV, 15% for MNV, 19% for SV, and 2% for INDEL.

The relatively small drop in indel calling sensitivity upon downsampling is caused by hard-coded setting in STRELKA. Strelka has a hard cutoff at 10% VAF for INDELs of less than 5 bases length (which is 99% of INDELs in our dataset) for both 50x and 100x depth whereas for SNVs the cutoff is fixed at ~5 supporting reads independent of read depth. This likely results in underestimation of subclonal INDELs in our dataset but does not affect specificity.

12. Germline predisposition variant calling

We searched for germline variants in a broad list of 152 germline predisposition genes curated by Huang et al[20]. For SNV and INDEL, using the germline variant calling outputs from the GATK HaplotypeCaller[4], we filtered for variants affecting the canonical transcript of these 152 genes which have the following coding or splice effects:

- All SNV Nonsense, INDEL Frameshift or SNV Splice Acceptor/Donor, excluding variants marked in ClinVar[21] as 'Benign/Likely_benign', 'Benign', 'Likely_benign’.
- Missense and synonymous variants, only if marked in clinvar as ‘Pathogenic’ or ‘Likely Pathogenic’, excluding pathogenic disease indications which are clearly unrelated to cancer.

Variants which were found with a median germline VAF across all samples of less than 0.2 or greater than 0.8 were filtered as likely mapping artefacts. We further excluded frameshift variants which are found to be exactly offset by other frameshift variants (thereby creating an in-frame protein product), which actually involved more than 50% of samples in which such events occur.

This yielded 550 potential germline predisposition point mutations across the 2,405 samples in our cohort. For each variant, we determined the genotype in the germline (HET or HOM) and also assessed in the tumor sample whether there is a 2nd somatic hit, and whether the wild type or the variant itself has been lost (see chapter 13: biallelic status evaluation methods). We also searched in the 152 genes for copy number deletions that were heterozygous in the germline with subsequent homozygous loss in the tumor and found an additional 16 of such germline copy number events, giving a total of 566 variants altogether.

We observed that for the variants in many of the 152 predisposition genes that a loss of wild type in the tumor via LOH was lower than the average rate of LOH across the cohort and that fewer than 5% of observed variants had a 2nd somatic hit in the same gene. Moreover, in many of these genes the ALT variant was lost via LOH as frequently as the wild type, suggesting that a significant portion of the 566 variants may be passengers. For our downstream analysis and driver catalog, we therefore restricted our analysis to a more conservative 'High Confidence' list including only the 25 cancer related genes in the ACMG secondary findings reporting guidelines (v2.0)[22], together with 4 curated genes (CDKN2A, CHEK2, BAP1 & ATM), selected because these are the only additional genes from the larger list of 152 genes with a significantly elevated proportion of called germline variants with loss of wild type in the tumor sample.

The following table summarises the statistics for the high confidence and low confidence genes:

Genes	Total germline predisposition SNV & INDEL	% with loss of wild type OR somatic hit in tumor	% with loss of germline ALT variant in tumor
High Confidence: ACMG + 4 curated genes	211	53.1%	10.4%
Low Confidence: Rest of 152 panel	355	16.3%	13.1%

Outside the 29 high confidence genes, the germline variant itself is lost almost as frequently via LOH as the remaining wild type in the tumor, whereas for the high confidence ACMG + curated genes, there is an observed loss of wild type allele in over half of all variants.

For the additional 4 curated genes, the numbers are as follows:

Gene	Count germline predisposition SNV & INDEL	% with loss of wild type in tumor sample	% with loss of germline variant in tumor sample
ATM	17	52.9%	11.8%
BAP1	5	66.7%	0%
CHEK2	72	36.1%	13.9%
CDKN2A	3	66.7%	33.3%

Germline variants with loss of ALT variants in the tumor were also excluded from the final list used in our analyses, leading to a final inclusion of 189 variants from the high confidence panel.

Supplementary Table 6 contains the full catalog of high and low confidence germline variants.

13. Clonality and biallelic status of point mutations

For each point mutation we determined the clonality and biallelic status by comparing the estimated ploidy of the variant to the local copy number at the exact base of the variant. The ploidy of each variant is calculated by adjusting the observed VAF by the purity and then multiplying by the local copy number to work out the absolute number of chromatids that contain the variant.

We mark a mutation as biallelic (i.e. no wild type remaining) if Variant Ploidy > Local Copy Number - 0.5. The 0.5 tolerance is used to allow for the binomial distribution of VAF measurements for each variant. For example, if the local copy number is 2 than any somatic variant with measured ploidy > 1.5 is marked as biallelic.

For each variant we also determine a probability that it is subclonal. This is achieved via a two-step process

1. Fit the somatic ploidies for each sample into a set of clonal and subclonal peaks
We apply an iterative algorithm to find peaks in the ploidy distribution:

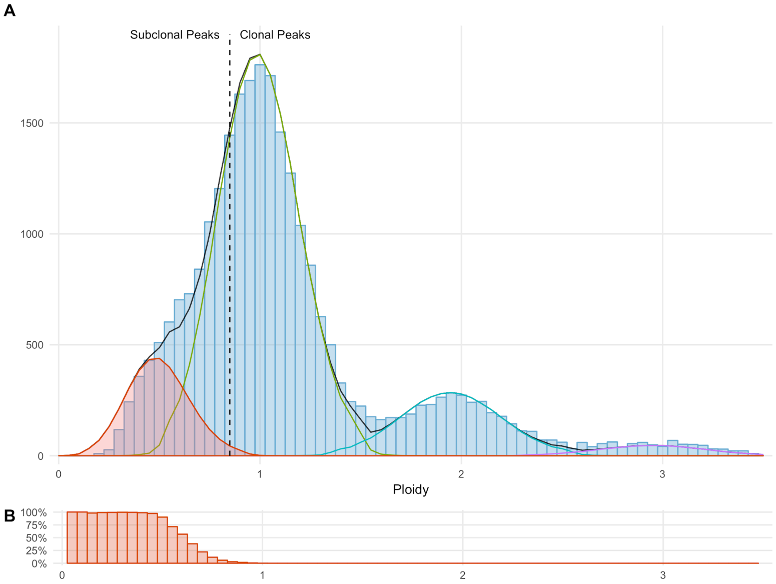
- Determine the peak by finding the highest density of variants within +/- 0.1 of every 0.01 ploidy bucket.
- Sample the variants within a 0.05 ploidy range around the peak.
- For each sampled variant, use a binomial distribution to estimate the likelihood that the variant would appear in all other 0.05 ploidy buckets.
- Sum the expected variants from the peak across all ploidy buckets and subtract from the distribution.
- Repeat the process with the next peak

This process yields a set of ploidy peaks, each with a ploidy and a total density (i.e. count of variants). To avoid overfitting small amounts of noise in the distribution, we filter out any peaks that account for less than 40% of the variants in the ploidy bucket at the peak itself. After this filtering we scale the fitted peaks by a constant so that the sum of fitted peaks = the total variant count of the sample.

Finally we mark a peak as subclonal if the peak ploidy < 0.85.

2. Calculate the probability that each individual variant belongs to each peak
Once we have fitted the somatic ploidy peaks and determined their clonality, we can calculate the subclonal likelihood for any individual variant as the proportion of subclonal variants at that same ploidy.

The following diagram illustrates this process for a typical sample. Figure A shows the histogram of somatic ploidy for all SNV and INDEL in blue. Superimposed are four peaks in different colours fitted from the sample as described above. The red filled peak is below the 0.85 threshold and is thus considered subclonal. The black line shows the overall fitted ploidy distribution. Figure B shows the likelihood of a variant being subclonal at any given ploidy.



Subclonal counts in this paper are calculated as the total density of the subclonal peaks for each sample. Subclonal driver counts are calculated as the sum across the driver catalog of subclonal probability * driver likelihood (driver likelihood is explained in detail in chapter 20).

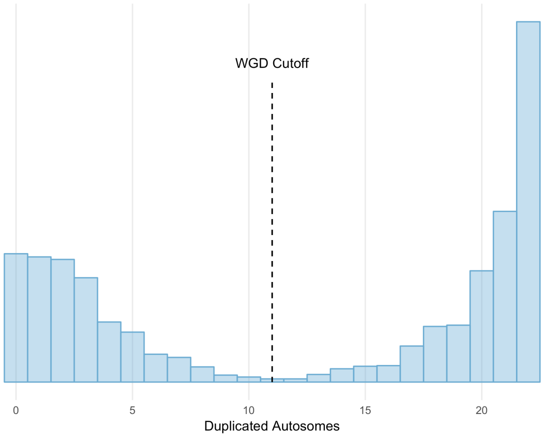
14. WGD status determination

We implement a simple heuristic that determines if Whole Genome Duplication has occurred:

Major allele Ploidy >1.5 on at least 50% of at least 11 autosomes

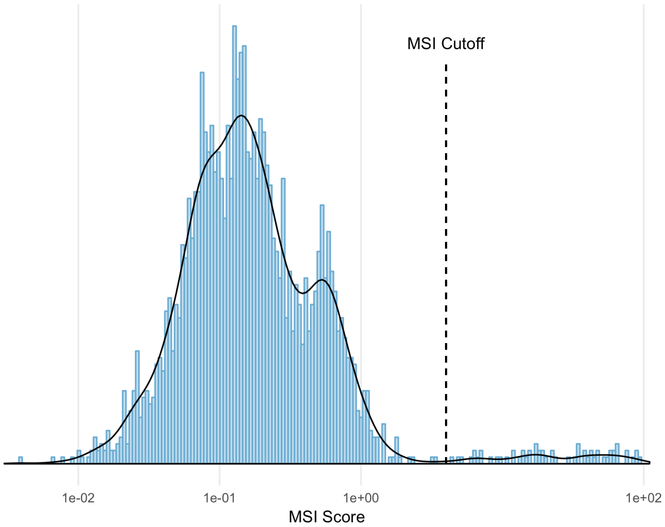
The principle behind this heuristic is that if sufficient independent chromosomes are predominantly duplicated, the most parsimonious explanation is that the duplication occurred in a single genome-wide event.

The number of duplicated autosomes per sample (ie. the number of autosomes which satisfy the above rule) follows a bimodal distribution with 95% of samples have either <= 6 or > =15 autosomes duplicated. Hence, the classification of a genome as WGD is not particularly sensitive to the choice of cut-off as is evident the following chart:



15. MSI status determination

To determine the MSI status of all samples we used the method described by the MSISeq tool[23]. In brief, we count the number of INDELS per million bases occuring in homopolymers of 5 or more bases or dinucleotide, trinucleotide and tetranucleotide sequences of repeat count 4 or more. MSIseq scores ranged from 0.004 up to 98.63, with a long tail towards lower MSI scores as shown in the following chart:



To be able to accurately set and validate the MSIseq cutoff for classification of MSI we compared the WGS results with the standard, routinely used MSI assessment using a 5-marker PCR panel (BAT25, BAT26, NR21, NR24 and MONO27 markers). For a batch of 48 pre-selected samples, the MSI PCR assay was blindly performed by an independent ISO-accredited pathology laboratory. Both the binary MSI and MSS classifications were determined, but also the number of positive markers.

A sample was considered as MSI if two or more of the five markers were score as positive (instable). PCR-based analysis identified 16 MSI samples, all of which were also identified by MSIseq with scores >4. MSIseq identified one sample that was missed by PCR-based analysis, although this sample showed microsatellite instability for one out the five markers. The MSIseq scores thus highly correlate with the number of positive MSI PCR markers and all, except one, samples with an elevated score are classified as MSI by pathology. Based on this data we determined the best cutoff for MSIseq classification to be at a **score of 4**.

Results of the PCR-based and WGS based MSI classification are summarized in the table below. The sensitivity of WGS-based MSI classification on this set was 100% (95%CI 82.6 – 100%) with a specificity of 97% (95%CI 88.2-96.9%). The calculated Cohen's kappa score was 0.954 (95%CI 0-696-0.954), indicative of a very high agreement.

	PCR-MSS	PCR-MSI	Total
MSISeq -MSS	31	0	31
MSIseq - MSI	1	16	17
Total	32	16	48

16. Holistic gene panel for driver discovery

We used Ensembl[13] release 89 as a basis for our gene definitions and have taken the union of Entrez identifiable genes and protein coding genes as our base panel.

Certain genes have multiple definitions. NPIPA7 for example has two definitions which are equally valid, ENSG00000214967 and ENSG00000183889. To solve this we select a single gene definition based on the following steps:

1) Exclude non protein coding genes.
2) Favour genes that are present in both Havana and Ensembl.
3) Select gene with longest transcript.

This returns our final gene panel tally to 25,963 genes of which 20,083 genes are protein coding. For each gene we chose the canonical transcript or the longest if no canonical transcript exists.

For CDKN2A, we included both the p16 and p14arf transcripts in the analysis given the known importance of both transcripts to tumorigenesis[24] and the fact that the two transcripts use alternate reading frames in the same exon.

17. Significantly mutated driver genes discovery

Using all SNV and INDEL variants from the holistic gene panel, we ran dNdScv[25] to find significantly mutated genes (SMGs) and also to estimate the proportion of missense, nonsense, essential splice site and INDEL variants which are drivers in each individual gene in the panel.

Pan cancer and at an individual cancer level we tested the normalised dNdS rates against a null hypothesis that dNdS = 1 for each variant subtype. To identify SMGs in our cohort we used a strict significance cutoff of $q<0.01$.

Two of the newly discovered SMG candidates were subsequently removed via manual curation as they were deemed to be likely artefacts of our methods:

- POM121L12 - found only to be significant due to an extreme covariate value in dNdScv
- TRIM49B - found to have poor mappability on nearly all its variants and a known close paralog

18. Significantly amplified & deleted driver gene discovery

To search for significantly amplified and deleted genes we first calculated the minimum exonic copy number per gene across our holistic gene panel. For amplifications, we searched for all the genes with high level amplifications only (defined as minimum Exonic Copy number > 3 * sample ploidy). For deletions, we searched for all the genes in each sample with either full or partial gene homozygous deletions (defined as minimum exonic copy number < 0.5). The Y chromosome was excluded from the deletion analysis since the Y chromosome is deleted altogether in 35% of all male cancer samples in our cohort and hence is difficult to distinguish at the gene level.

We then searched separately for amplifications and deletions, on a per chromosome basis, for the most significant focal peaks, using an iterative GISTIC-like peel off method[26], specifically:

- Find the highest scoring gene.
    - For deletions the score is simply the count of samples with homozygous deletions in the gene.
    - For amplifications, we need to consider both the count and strength of the amplification so we use:

    - score = sum(log2(copy number/ sample ploidy)).
- Record gene as a peak, and mark all consecutive genes with a score within 15% and 25% of the highest score for deletions and amplifications respectively as part of the candidate peak.
- ‘Peel’ off all samples which contributed to the peak across the entire chromosome.
- Repeat the process.

A filter was applied where we removed deletions from a handful of noisy copy number regions in the genome where we found more than 50% of the observed deletions were not supported on either breakend by a structural variant.

Most of the deletion peaks resolve clearly to a single target gene reflecting the fact that homozygous deletions are highly focal, but for amplifications this is not the case and the majority of our peaks have 10 or more candidates. We therefore annotated the peaks, to choose a single putative target gene using an objective set of automated curation rules in order of precedence:

- If more than 50% of the copy number events in the peeled samples include the telomere or centromere than mark as <CHR>_<ARM>_<TELOMERE/CENTROMERE>
- Else choose highest scoring candidate gene which matches a list of actionable amplifications from OncoKB, CGI and CIViC clinical annotation DBs
- Else choose highest scoring candidate gene found in our panel of significantly mutated genes.
- Else choose highest scoring candidate gene found in cosmic census
- Else choose highest scoring protein coding candidate gene
- Else choose longest non-coding candidate gene

Finally, we filter the peaks to only highly significant deletions and amplifications using the following rules

- Deletions => Keep any peak with > 5 homozygous deletions
- Amplifications => Keep any peak with score > 29

These cut-offs were chosen using a binomial model which assumes the probability of any given gene being observed to be randomly deleted or highly amplified is equal to the average number of genes amplified or deleted in each event divided by the total number of genes considered. The cut-offs were chosen to be the lowest score with a q-value below 0.25. Since amplifications are generally much broader (averaged genes affected per event of 41.6 compared to just 5.4 for deletions) a much higher number of genes is required to reach significance.

The calculation details for the cut-offs are presented in the table below.

	Cohort data						**Statistical Calculations**				
	Count of events	**Sum Scores**	**Count of genes affected**	**Avg genes affected per event**	**Avg score / event**	**Total genes tested**	**Probability event overlaps a given gene**	**Score cutoff**	**P value of cutoff**	**Significant findings**	**Q Value**
Dels	4,915	4,915	26,676	5.4	1.0	25,965	0.00021	5	0.00068	117	0.15
Amps	3,925	6,959	163,393	41.6	1.8	25,965	0.00160	29	0.00030	33	0.23

This model is likely to be highly conservative as it assumes that all the events are passengers, whereas in fact a high proportion contain driver genes.

19. Fragile site annotation

Homozygous deletions were also annotated as common fragile site (CFS) based on their genomic characteristics. This annotation is not definitive, but is useful as CFS are known to be regions of high genomic instability. Hence despite being significantly deleted, their status as a genuine cancer driver remains unclear.

There is no absolute agreement on which regions should be classified as CFS, but two well-known features are a strong enrichment in long genes and a high rate of observed deletions of up to 1 megabase[27]. Hence for this analysis we classified a gene as a fragile site if it met all the following criteria:

- Total length of gene > 500,000 bases
- More than 30% of all SVs with breakpoints that disrupt the gene are deletions with length greater than 20,000 bases and less than 1 megabase.
- The gene is not found to be significantly mutated (by dNdScv) in our cohort or in Martincorena et al.[25].

Using these criteria we annotated the following list of 16 Genes as fragile:

Gene	Chr	Start position	Length (bases)	Total Disruptive SV Count	% of SV that are DELs (>20kb & <1MB)
LRP1B	2	140,988,992	1,900,278	1,272	0.469
FHIT	3	59,735,036	1,502,097	2,128	0.596
LSAMP	3	115,521,235	2,194,860	1,306	0.364
NAALADL2	3	174,156,363	1,367,065	1,198	0.456
CCSER1	4	91,048,686	1,474,378	1,398	0.441
PDE4D	5	58,264,865	1,553,082	1,166	0.458
GMDS	6	1,624,041	621,885	399	0.441
PARK2	6	161,768,452	1,380,351	1,296	0.555
IMMP2L	7	110,303,110	899,463	1,028	0.444
PTPRD	9	8,314,246	2,298,477	1,264	0.309
PRKG1	10	52,750,945	1,307,165	781	0.318
GPHN	14	66,974,125	674,395	291	0.306
WWOX	16	78,133,310	1,113,254	1,319	0.541
MACROD2	20	13,976,015	2,057,827	3,039	0.605
DMD	X	31,115,794	2,241,764	789	0.328
DIAPH2	X	95,939,662	920,334	331	0.381

We also noted that 4 other significantly deleted genes (STS,HDHD1,LRRN3 and LINC00290), though not fulfilling the length criteria above have a particularly high proportion of deletion SVs between 20kb and 1 megabase (over 60%) and hence were also marked as fragile:

Gene	Chr	Start position	Length (bases)	Total Disruptive SV Count	% of SV that are DELs (>20kb & <1MB)
LINC00290	4	181,985,242	95,060	64	0.641
LRRN3	7	110,731,062	34,448	70	0.686
STS	X	7,137,497	135,354	168	0.649
HDHD1	X	6,966,961	99,270	126	0.659

Two of these genes (STS and HDHD1) fall in a previously identified CFS region (FRAXB) and a third, LRNN3, falls in another knowns CFS region (FRAX7). The final one, LINC00290 is a long non-coding RNA with an unknown status as cancer driver.

20. Somatic driver catalog construction

We created a catalog of each and every driver in our cohort across all variant types on a per patient basis. This was done in a similar incremental manner to Sabarinathan et al[28] (N. Lopez, personal communication) whereby we first calculated the number of drivers in a broad panel of known and significantly mutated genes across the full cohort, and then assigned the drivers for each gene to individual patients by ranking and prioritising each of the observed variants. Key points of difference in this study were both the prioritisation mechanism used and our choice to ascribe each mutation a probability of being a driver rather than a binary cutoff based on absolute ranking.

The four detailed steps to create the catalog are described below:

1. Create a panel of driver genes for point mutations using significantly mutated genes and known drivers

We created a gene panel using the union of

- Martincorena significantly mutated genes[25] (filtered to significance of q<0.01)
- HMF significantly mutated genes (q<0.01) at global level or at cancer type level
- Cosmic Curated Genes[14] (v83)

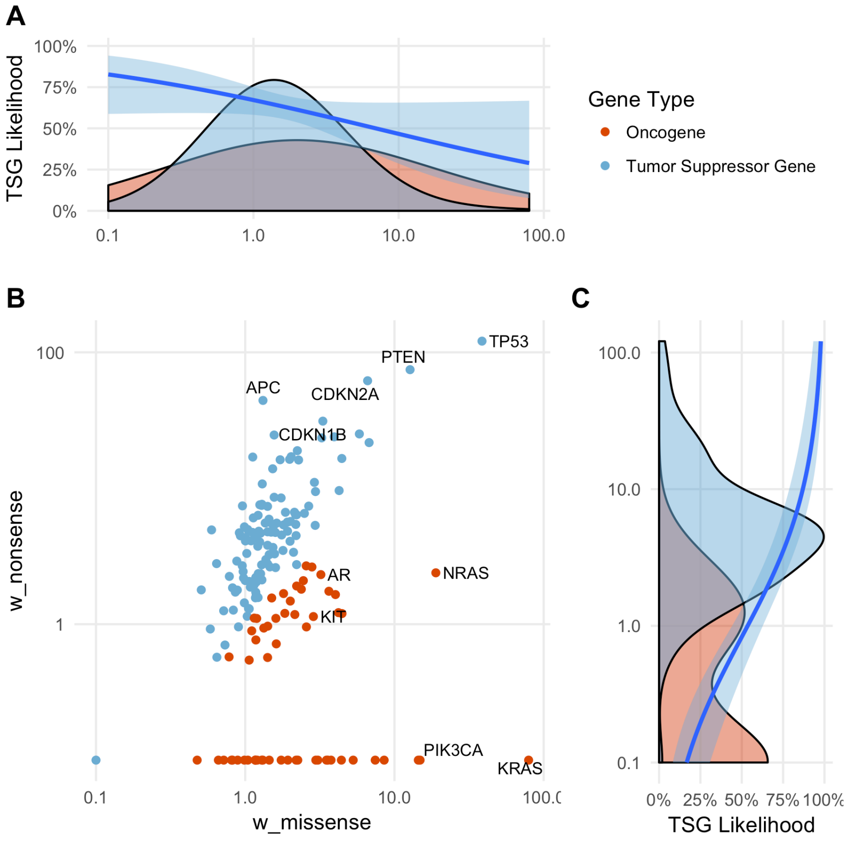
2. Determine TSG or Oncogene status of each significantly mutated gene

We used a logistic regression model to classify the genes in our pane as either tumor suppressor gene (TSG) or oncogene. We trained the model using unambiguous classifications from the Comic curated genes, i.e. a gene was considered either a Oncogene or TSG but not both. We determined that the dNdS missense and nonsense ratios (w_missense and w_nonsense) are both significant predictors of the classification. The coefficients are given in the table below.

	Estimate	Std. Error	z value	Pr(>|z|)
intercept	0.1830	0.3926	0.466	0.64106
w_missense	-0.6869	0.2643	-2.599	0.00936
w_nonsense	0.5237	0.1116	4.691	2.72e-06

We applied the model to all significantly mutated genes in Matincorena and HMF as well as any ambiguous Cosmic curated genes.

The following figure shows all genes that have classified using the logistic regression model. Figures A and C show the likelihood of a gene being classified as a TSG under a single variate logistic model of w_missense and w_nonsense respectively. Figure B shows the classification after the multivariate regression using both predictors.



3. Add drivers from all variant classes to the catalog

Variants were added to the driver catalog which met any of the following criteria

- All missense and inframe indels for panel oncogenes
- All non synonymous and essential splice point mutations for tumor suppressor genes
- All high level amplifications (min exonic copy number > 3 * sample ploidy) for both significantly amplified target genes and panel oncogenes
- All homozygous deletions for significantly deleted target genes and panel TSG (except for the Y chromosome as described before)
- All known or promiscuous inframe gene fusions as described above
- Recurrent TERT promoter mutations

4. Calculate a per sample driver likelihood for each gene in the catalog

A driver likelihood estimate between 0 and 1 was calculated for each variant in the gene panel to ensure that only excess mutations are used for determining the number of drivers in cancer cohort groups or at the individual sample level. High level amplifications, Deletions, Fusions, and TERT promoter mutations are all rare so were assumed to have a likelihood of 1 when found affecting a driver gene, but for coding mutations we need to account for the large number of passenger point mutations that are present throughout the genome and thus also in driver genes.

For coding mutations we also marked coding mutations that are highly likely to be drivers and/or highly unlikely to have occurred as passengers as driver likelihood of 1, specifically:

- Known hotspot variants
- Variants within 5 bases of a known pathogenic hotspot in oncogenes
- Inframe indels in oncogenes with repeat count < 8 repeats. Longer repeat count contexts are excluded as these are often mutated by chance in MSI samples
- Biallelic variants in tumor suppressor genes

For the remaining variants (non-hotspot missense variants in oncogenes and non-biallelic variants in TSG) these were only assigned a > 0 driver likelihood where there was a remaining excess of unallocated drivers based on the calculated dNdS rates in that gene across the cohort after applying the above rules. Any remaining point mutations were assigned a driver likelihood between 0 and 1 using a bayesian statistic to calculate a sample specific likelihood of each gene based on the type of variant observed (missense, nonsense, splice or INDEL) and taking into account the mutational load of the sample. The principle behind the method is that the likelihood of a passenger variant occuring in a particular sample should be approximately proportional to the tumor mutational burden and hence variants in samples with lower mutational burden are more likely to be drivers.

The sample specific likelihood of a residual excess variant being a driver is estimated for each gene using the following formula:

P(Driver|Variant) = P(Driver) / (P(Driver) + P(Variant|Non-Driver) * (1-P(Driver)))

where P(Driver) in a given gene is assumed to be equal across all samples in the cohort, ie:

P(Driver) = (residual unallocated drivers in gene) / # of samples in cohort

And P(Variant|Non-Driver), the probability of observing n or more passenger variants of a particular variant type in a sample in a given gene, is assumed to vary according to tumor mutational burden, and is modelled as a poisson process:

P(Variant|Non-Driver) = 1 - poisson(λ = TMB(Sample) / TMB(Cohort) * (# of passenger variants in cohort),k=n-1)

All counts reported in the paper at a per cancer type or sample level refer to the sum of driver likelihoods for that cancer type or sample.

21. Driver co-occurrence analysis

To examine the co-occurence of drivers, the driver-gene catalog was filtered to exclude fusions and any driver with a driver likelihood of < 0.5. Separately for each cancer type, every pair of driver genes was tested to see whether they co-occur more or less frequently than expected if they were independent using Fisher's Exact Test. The results were adjusted to a FDR using the number of gene-pair comparison being tested in each cancer type cohort. Gene pairs with a positive correlation which were on the same chromosome were excluded from the analysis as they are frequently co-amplified or deleted by chance.

22. Actionability analysis

To determine clinical actionability of the variants observed in each sample, we mapped all variants to 3 external clinical annotation databases

- OncoKB[9] (download = 01-mar-2018)
- CGI[8] (update: 17-jan-2018)
- CIViC[7] (download = 01-mar-2018)

In order to be able to aggregate and compare this data, we have mapped each of the databases to a common data model using the following rules:

1. Level of evidence mapping

The 3 databases we used in this study define different level for evidence items, depending on evidence strength. In order to be able to aggregate and compare this data, we have mapped the CGI and OncoKB evidence levels on the CIViC evidence levels defined at: https://civicdb.org/help/evidence/evidence-levels.

HMF	CIViC	CGI	OncoKB
A	A	FDA guidelines, NCCN guidelines, NCCN/CAP guidelines, CPIC guidelines, European Leukemia Net guideline	1 2 R1
B	B	Clinical trials, Late trials, Late trials,Pre-clinical	3 R2
C	C	Early trials, Case report	
D	D	Pre-clinical	4,R3

In this study we considered only A and B level variants. This classification roughly corresponds to the recently proposed ESMO Scale for Clinical Actionability of molecular Targets (ESCAT)[29] as follows:
HMF A: ESCAT I-A+B (for on label) and I-C (for off-label)
HMF B: ESCAT II-A+B (for on label) and III-A (for off-label)

2. Response type Mapping

We also mapped response type to a common data model. First we filtered out evidence items from the annotation databases that do not lead to clinical actionability (for example prognostic biomarkers). The remaining evidence items were mapped as either responsive or resistant based on the following rules:

HMF	CIViC	CGI	OncoKB
Responsive	Sensitivity	Responsive	1 2 3 4
Resistant	Resistant or Non-Response	Resistant	R1 R2 R3

3. Mutation/Event type mapping

Each evidence item was mapped to HMF data as one of 4 event types according to the following criteria:

HMF Event type	Matching Criteria
Somatic Point Mutation	HGVS / genomic coordinates converted to chromosome, position, ref and alt and mapped to exact variants in our database
Somatic Range Event	Matched to missense / inframe variants in Oncogenes and any non-synonymous variant in TSG contained within a defined range, either exon level, transcript level or specific coordinates. Where a transcript was not specified, the canonical transcript was always used to map coordinates
Somatic CNA	'Deletion' mapped to homozygous deletions and 'Amplification' mapped to high level amplification (>3x sample ploidy)
Fusion	Exact matching to an inframe fusion in our database. For OncoKB 'loss-of-function' fusions were excluded

A small number of items from CIViC level B evidence level were deemed either not specific enough or insufficiently supportive of actionability for this study and were filtered:

- Evidence items supporting TP53, KRAS & PTEN as actionable
- Evidence items supporting actionability with 'chemotherapy' (ie. chemotherapy in general rather than a specific treatment), 'aspirin' or 'steroids'

Finally, a number of suspicious fusions from each of the databases were curated by either changing the 5' and 3' partners or filtered out altogether based on referring to the original evidence sources, specifically:

HMF Curation	CIViC	CGI	OncoKB
Filtered Fusions	BRAF - CUL1	RET - TPCN1	
5' and 3' partners exchanged		ABL1 - BCR PDGFRA - FIP1L1 PDGFB - COL1A1	ROS1 - CD74 EP300 - MLL EP300 - MOZ RET - CCDC6

Some of the more complex event types from the 3 databases have not been fully interpreted and have been excluded from this analysis.

4. Cancer type mapping

Each evidence event mapped was also determined to be either on-label (ie. evidence supports treatment in that specific cancer type) or off-label (evidence exists in another cancer type) for each specific sample. To do this, we have annotated both the patient cancer types and the database cancer types with relevant DOIDs, using the disease ontology database available at: http://disease-ontology.org.

Patient cancer types from the HMF database were annotated according to the following table:

HMF tumor type	DOID
Biliary	4607
Bone/Soft tissue	201;9253

Breast	1612
CNS	3620;3070
Colon/Rectum	9256;219
CUP	-
Esophagus	5041;4944
Head and neck	11934;8618
Kidney	263;8411
Liver	3571
Lung	1324
Mesothelioma	1790
NET	-
Other	-
Ovary	2394
Pancreas	1793
Prostate	10283
Skin	4159
Stomach	10534
Urinary tract	3996
Uterus	363

Database cancer types were mapped to a DOID by automatically querying the ontology on the disease names. Some CIViC evidence items are already annotated with a DOID in the database, this was used if present. We also manually annotated with DOIDs some of the database cancer types that failed the automatic query:

cancerType	DOID	Ontology term
All Tumors	162	cancer
Any cancer type	162	cancer
B cell lymphoma	707	B-cell lymphoma
Billiary tract	4607	biliary tract cancer
Bladder	11054	urinary bladder cancer
Cervix	4362	cervical cancer
CNS Cancer	3620	central nervous system cancer
Dedifferentiated Liposarcoma	3382	liposarcoma
Endometrium	1380	endometrial cancer
Esophagogastric Cancer	5041	esophageal cancer
Gastrointestinal stromal	9253	gastrointestinal stromal tumor
Giant cell astrocytoma	3069	astrocytoma
Hairy-Cell leukemia	285	hairy cell leukemia
Head and neck	11934	head and neck cancer
Head and neck squamous	5520	head and neck squamous cell carcinoma
Hepatic carcinoma	686	liver carcinoma
Hepatocellular Mixed Fibrolamellar Carcinoma	0080182	mixed fibrolamellar hepatocellular carcinoma
Inflammatory myofibroblastic	0050905	inflammatory myofibroblastic tumor
Lung	1324	lung cancer
Lung squamous cell	3907	lung squamous cell carcinoma
Melanoma	8923	Skin melanoma

Mesothelioma	1790	malignant mesothelioma
Neuroendocrine	169	neuroendocrine tumor
Non-small cell lung	3908	non-small cell lung carcinoma
Ovary	2394	ovarian cancer
Pancreas	1793	pancreatic cancer
Renal	263	kidney cancer
Salivary glands	8850	salivary gland cancer
Stomach	10534	stomach cancer
Thymic	3277	thymus cancer
Thyroid	1781	thyroid cancer
Well-Differentiated Liposarcoma	3382	liposarcoma

In case a matching DOID was found for the disease, we annotated the disease with a DOID set consisting of: the disease DOID, all the children DOIDs and all the parent disease DOIDs.

A treatment is defined as on-label if any of the DOIDs of the patient cancer is present in the DOID set of the disease.

5. MSI actionability

Samples classified as MSI in our driver catalog were also mapped as actionable at level A evidence based on clinical annotation in the OncoKb database

6. Aggregation of evidence

For each actionable mutation in each sample, we aggregated all the mapped evidence that was available supporting both on-label and off-label treatments at an A or B evidence level. Treatments that also had evidence supporting resistance based on other biomarkers in the sample at the same or higher level were excluded as non-actionable.

For each sample we reported the highest level of actionability, ranked first by evidence level and then by on-label vs off-label.

23. Data availability

All data described in this study is freely available for academic use from the Hartwig Medical Foundation through standardized procedures and request forms which can be found at https://www.hartwigmedicalfoundation.nl/en. Briefly, a data request can be initiated by filling out the standard form in which intended use of the requested data is motivated. First, an advice on scientific feasibility and validity is obtained from experts in the field which is used as input by an independent Data Access Board who also evaluates if the intended use of the data is compatible with the consent given by the patients and if there would be any applicable legal or ethical constraints. Upon formal approval by the Data Access Board, a standard license agreement which does not have any restrictions regarding Intellectual Property resulting from the data analysis needs to be signed by an official organisation representative before access to the data is granted. Raw data files will be made available through a dedicated download portal with two-factor authentication.

24. References

1. Bins, S. *et al.* Implementation of a Multicenter Biobanking Collaboration for Next-Generation Sequencing-Based Biomarker Discovery Based on Fresh Frozen Pretreatment Tumor Tissue Biopsies. *Oncologist* **22**, 33–40 (2017).
2. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
3. McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).
4. Poplin, R. *et al.* Scaling accurate genetic variant discovery to tens of thousands of samples. *bioRxiv* 201178 (2018). doi:10.1101/201178
5. Van der Auwera, G. A. *et al.* From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Curr. Protoc. Bioinformatics* **43**, 11.10.1–33 (2013).
6. Saunders, C. T. *et al.* Strelka: accurate somatic small-variant calling from sequenced tumor-normal sample pairs. *Bioinformatics* **28**, 1811–1817 (2012).
7. Griffith, M. *et al.* CIViC is a community knowledgebase for expert crowdsourcing the clinical interpretation of variants in cancer. *Nat. Genet.* **49**, 170–174 (2017).
8. Tamborero, D. *et al.* Cancer Genome Interpreter annotates the biological and clinical relevance of tumor alterations. *Genome Med.* **10**, 25 (2018).
9. Chakravarty, D. *et al.* OncoKB: A Precision Oncology Knowledge Base. *JCO Precis Oncol, July 2017*, (2017) doi: 10.1200/PO.17.00011
10. Cleveland, M. H., Zook, J. M., Salit, M. & Vallone, P. M. Determining Performance Metrics for Targeted Next-Generation Sequencing Panels Using Reference Materials. *J. Mol. Diagn.* **20**, 583-590 (2018).
11. Eijkelenboom, A. *et al.* Reliable Next-Generation Sequencing of Formalin-Fixed, Paraffin-Embedded Tissue Using Single Molecule Tags. *J. Mol. Diagn.* **18**, 851–863 (2016).
12. Chen, X. *et al.* Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. *Bioinformatics* **32**, 1220–1222 (2016).
13. Zerbino, D. R. *et al.* Ensembl 2018. *Nucleic Acids Res.* **46**, D754–D761 (2018).
14. Forbes, S. A. *et al.* COSMIC: somatic cancer genetics at high-resolution. *Nucleic Acids Res.* **45**, D777–D783 (2017).
15. Haas, B. *et al.* STAR-Fusion: Fast and Accurate Fusion Transcript Detection from RNA-Seq. *bioRxiv* 120295 (2017). doi:10.1101/120295
16. Craig, D. W. *et al.* A somatic reference standard for cancer genome sequencing. *Sci. Rep.* **6**, 24607 (2016).
17. Nilsen, G. *et al.* Copynumber: Efficient algorithms for single- and multi-track copy number segmentation. *BMC Genomics* **13**, 591 (2012).
18. Velazquez Villarreal, E. I., Kumar, V., Yin, Y., Carpten, J. D. & Craig, D. W. Abstract 437: Leveraging new methods in single-cell copy number analysis and clonotype detection to uncover and characterize hidden subclones within standard cell lines. *Cancer Res.* **78**, 437–437 (2018).
19. Van Loo, P. *et al.* Allele-specific copy number analysis of tumors. *Proc. Natl. Acad. Sci. U. S. A.* **107**, 16910–16915 (2010).
20. Huang, K.-L. *et al.* Pathogenic Germline Variants in 10,389 Adult Cancers. *Cell* **173**, 355–370.e14 (2018).
21. Landrum, M. J. *et al.* ClinVar: public archive of interpretations of clinically relevant variants. *Nucleic Acids Res.* **44**, D862–8 (2016).

22. Kalia, S. S. *et al.* Recommendations for reporting of secondary findings in clinical exome and genome sequencing, 2016 update (ACMG SF v2.0): a policy statement of the American College of Medical Genetics and Genomics. *Genet. Med.* **19**, 249–255 (2017).
23. Huang, M. N. *et al.* MSIseq: Software for Assessing Microsatellite Instability from Catalogs of Somatic Mutations. *Sci. Rep.* **5**, 13321 (2015).
24. Al-Kaabi, A., van Bockel, L. W., Pothen, A. J. & Willems, S. M. p16INK4A and p14ARF gene promoter hypermethylation as prognostic biomarker in oral and oropharyngeal squamous cell carcinoma: a review. *Dis. Markers* **2014**, 260549 (2014).
25. Martincorena, I. *et al.* Universal Patterns of Selection in Cancer and Somatic Tissues. *Cell* **171**, 1029–1041 e21 (2017).
26. Mermel, C. H. *et al.* GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome Biol.* **12**, R41 (2011).
27. Glover, T. W., Wilson, T. E. & Arlt, M. F. Fragile sites in cancer: more than meets the eye. *Nat. Rev. Cancer* **17**, 489–501 (2017).
28. Sabarinathan, R. *et al.* The whole-genome panorama of cancer drivers. *BioArchive* (2017). doi:10.1101/190330
29. Mateo, J. *et al.* A framework to rank genomic alterations as targets for cancer precision medicine: the ESMO Scale for Clinical Actionability of molecular Targets (ESCAT). *Ann. Oncol.* **29**, 1895-1902 (2018).