nature protocols

PROTOCOL

https://doi.org/10.1038/s41596-019-0234-7

Analysis pipelines for cancer genome sequencing in mice

Sebastian Lange[1,2,3,10], Thomas Engleitner[1,3,10], Sebastian Mueller[1,3,10], Roman Maresch[1,3,10], Maximilian Zwiebel[1,3], Laura Gonzalez-Silva[4], Günter Schneider[2], Ruby Banerjee[5], Fengtang Yang[5], George S. Vassiliou[5,6,7], Mathias J. Friedrich[1,2,3], Dieter Saur[2,3,8,9], Ignacio Varela[4] and Roland Rad[1,2,3,8*]

Mouse models of human cancer have transformed our ability to link genetics, molecular mechanisms and phenotypes. Both reverse and forward genetics in mice are currently gaining momentum through advances in next-generation sequencing (NGS). Methodologies to analyze sequencing data were, however, developed for humans and hence do not account for species-specific differences in genome structures and experimental setups. Here, we describe standardized computational pipelines specifically tailored to the analysis of mouse genomic data. We present novel tools and workflows for the detection of different alteration types, including single-nucleotide variants (SNVs), small insertions and deletions (indels), copy-number variations (CNVs), loss of heterozygosity (LOH) and complex rearrangements, such as those of chromothripsis. Workflows have been extensively validated and cross-compared using multiple methodologies. We also give step-by-step guidance on the execution of individual analysis types, provide advice on data interpretation and make the complete code available online. The protocol takes 2–7 d, depending on the desired analyses.

Introduction

The mouse as a model organism has been used in cancer research for almost a century. In the 1920s, the first inbred 'isogenic' mouse lines were generated to establish cancer models that developed different malignancies either spontaneously or after treatment with carcinogens[1]. Transgenesis, embryonic stem cell technology and gene targeting opened the way for the development of genetically engineered mouse models of cancer, revolutionizing our ability to link genes, molecular mechanisms and organismal phenotypes[2]. Mouse models were used to elucidate many of the most fundamental biological principles that have since been discovered[3]. Through CRISPR-based genome engineering, it has now become possible to edit genomes, even somatically in living animals. Fast and scalable in vivo CRISPR applications are substantially changing our ability to perform complex manipulations and functional genomic studies in mice[4]. These and other developments contribute to a growing importance of mouse models in basic and translational cancer research.

In humans, cancer genomics has been revolutionized by NGS. With sequencing costs constantly dropping, NGS has also begun to influence the arena of mouse cancer genomics. As a consequence, the demand for sequencing of mouse cancers is increasing, as is the need for robust analysis pipelines.

A high degree of gene orthology between human and mouse exists. 80% of human protein-coding genes have one-to-one mouse orthologs. The remaining 20% are either (i) in one-to-many, or many-to-many, orthologous relationships; (ii) are members of gene families that have undergone species-specific expansions or reductions; or (iii) contain species-specific open reading frames[5].

Nevertheless, comparative analyses of mouse and human genomes have also revealed some differences between the two species[6,7]. For example, in mice, segmental duplications are typically arranged in clusters, forming contiguous blocks of structural variations, whereas in humans

[1]Institute of Molecular Oncology and Functional Genomics, School of Medicine, Technische Universität München, Munich, Germany. [2]Department of Medicine II, Klinikum rechts der Isar, School of Medicine, Technische Universität München, Munich, Germany. [3]Center for Translational Cancer Research (TranslaTUM), School of Medicine, Technische Universität München, Munich, Germany. [4]Instituto de Biomedicina y Biotecnología de Cantabria, Universidad de Cantabria-CSIC, Santander, Spain. [5]The Wellcome Trust Sanger Institute, Cambridge, UK. [6]Wellcome Trust-MRC Stem Cell Institute, Biomedical Campus, University of Cambridge, Cambridge, UK. [7]Department of Haematology, Cambridge University Hospitals NHS Trust, Cambridge, UK. [8]German Cancer Consortium (DKTK), German Cancer Research Center (DKFZ), Heidelberg, Germany. [9]Institute for Experimental Cancer Therapy, School of Medicine, Technische Universität München, Munich, Germany. [10]These authors contributed equally: Sebastian Lange, Thomas Engleitner, Sebastian Mueller, Roman Maresch. *e-mail: roland.rad@tum.de

duplications are more often interspersed along the genome[8]. These clusters of segmental duplications are hotspots of recombination, leading to diversification both between mouse strains and 'de novo' between individuals of the same strain[9]. Other differences between mouse and human genomes are not well studied, and it is unclear how such differences affect the accuracy of genomic analyses. The development of analytical tools and bioinformatics pipelines was focused on humans and such tools have so far not been systematically validated in the mouse context.

Another limitation in mouse genomic analyses is the lower size/availability of genomic data resources for rodents. For example, single-nucleotide or copy-number databases comprise orders of magnitude more entries in humans than in mice[10,11]. Moreover, large data resources linking mutations to various phenotypes (cancer, Mendelian disorders) exist for human data but are mostly unavailable for the mouse. As an exception to this, a few mutations have been modeled in mice (e.g., *Trp53* point mutants) to dissect functionality at the organismal level.

Finally, the use of inbred strains for mouse cancer studies, which can affect different aspects of data analysis, represents a significant difference from the human situation. For example, the type and extent of inbreeding can have critical impacts on the quality of LOH analysis. Although defined crosses of two different inbred strains can facilitate the analysis of LOH (e.g., in F_1 animals from Sv/129 × C57BL/6), this scenario is rare in mouse cancer studies. Typically, either pure backgrounds are used, in order to control for phenotype stability[12–15], or various mixtures of backgrounds are generated through intercrosses of the different required alleles, which were often engineered in different backgrounds. In both cases, LOH analysis is substantially impaired, either by the low number of variant alleles in the germline or by their uneven distribution.

Development of the protocol

To analyze whole-exome or whole-genome sequencing (WES/WGS) data from mice, we initially tested computational methodologies and settings that were benchmarked for humans. However, validation experiments using array comparative genomic hybridization (aCGH), multicolor fluorescence in situ hybridization (M-FISH), or targeted re-sequencing revealed inaccuracies of results related to different alteration types. There is a scarcity of mouse-specific workflows for the analysis of cancer genomic data. For example, currently no pipelines are available for the inference of LOH and chromothripsis, and workflows for calling of indels and CNVs have not yet been validated in the mouse context. We therefore set out to systematically examine, validate and benchmark tools for the analysis of all cancer-relevant genomic alterations in mice, including SNVs, indels, CNVs, LOH and complex rearrangements.

Our protocol describes computational workflows for each analysis type. It extensively cross-compares, validates and recommends tools for the analysis of SNVs and CNVs, and contributes novel analytical methods and pipelines for the detection of LOH and chromothripsis. We provide all scripts, as well as guidance on their use. The protocol also gives recommendations for a broad spectrum of analytical details, such as parameter settings in various analytical and research contexts. Finally, each section also contains advice on data interpretation.

This work benefited from our extensive collection of various mouse tumor entities, including pancreatic, colon, stomach and hematopoietic cancers. The collection encompasses both tumors derived from genetically engineered mice and cancers triggered by environmental factors such as inflammation. Importantly, we developed primary cancer cell cultures from these mouse tumors, allowing accurate multi-layered analyses and validation approaches. For example, M-FISH using metaphase spreads facilitated the development, refinement and validation of pipelines for the detection of CNVs, LOH or chromothripsis.

We used the workflow (overview in Fig. 1) described in this protocol to analyze mouse cancers from different cancer entities[16,17]. Comparative analysis using matched human cancers revealed important considerations for the use of mouse models. First, the types of genetic alterations occurring in individual cancer types are similar in mouse and human, reflecting the similarities in biology between the species and supporting the role of the mouse as a prime model for human cancer. For example, mutational patterns and complex rearrangement types are similar in the two species, as shown in Fig. 2 for pancreatic cancer. Second, the frequency of mutations is generally lower in mice, particularly in genetically engineered models, which are driven by oncogene and tumor suppressor alterations, often induced using tissue-specific Cre-lines starting at the embryonic or early postnatal stages. Third, the reduced mutational complexity can aid data interpretation and can be exploited to

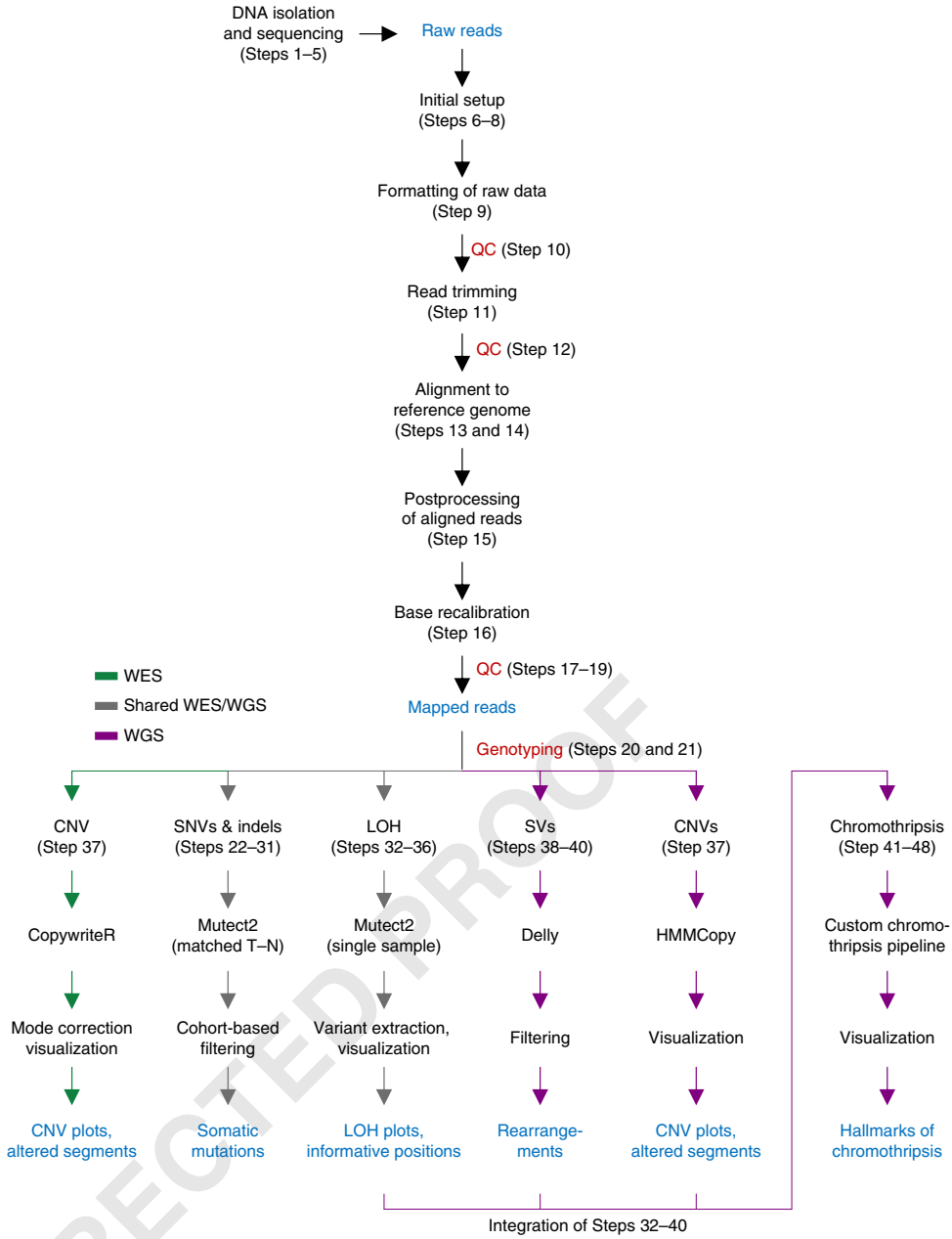


Fig. 1 | Overview of mouse cancer genome analysis workflows. Overview focusing on the bioinformatic section of this protocol, highlighting key procedures and their corresponding steps in the protocol. N, normal; SV, structural variation; T, tumor.

uncover biological principles that are difficult to extract from the more complex human genomes[17] (Anticipated results).

Applications of the method

Genomic analyses in mouse cancer models offer multifaceted opportunities to answer questions in cancer research that are difficult to answer in human studies. One limitation in humans is a lack of tissue resources that are needed in some research areas. For example, although ~1,000 human pancreatic cancers have been deep sequenced, the scarcity of primary/metastasis pairs—particularly of treatment-naive ones—substantially hampers studies into metastasis genetics. Mouse models can overcome this obstacle, allowing systematic sequencing-based surveys for genes that drive metastasis or metastatic organotropism. Also, in sample banks, the phenotypic spectrum of a disease is often misrepresented. For example, ~50% of pancreatic cancer patients present with advanced (stage 4)

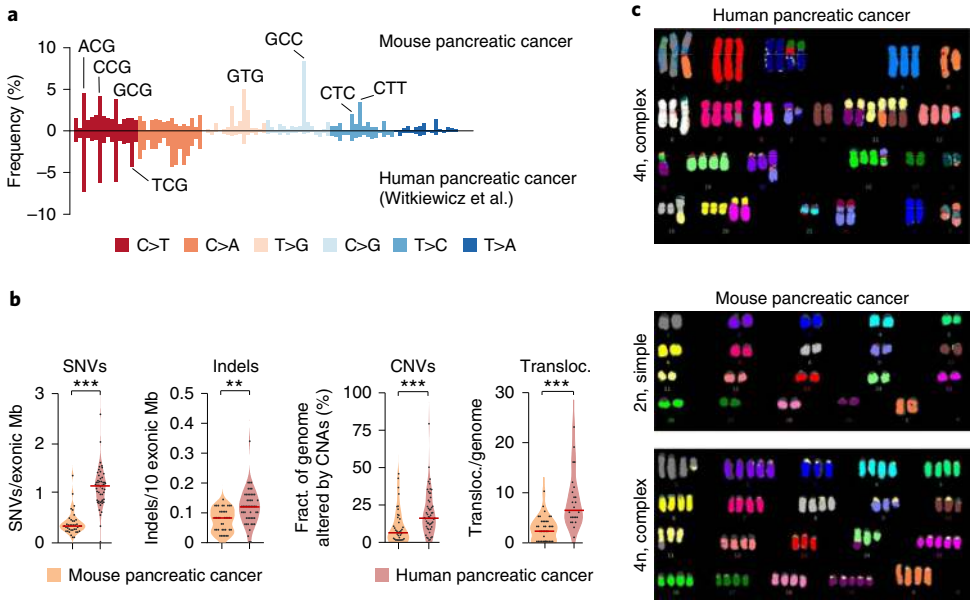


Fig. 2 | Genetic alterations in human and murine tumors. a–c, Similar genetic alterations can be found in tumors from humans and mice (**a**,**c**), although at different frequencies (**b**). **a**, Trinucleotide context-specific somatic SNVs, as detected by WES, for mouse ($n = 38$) and human ($n = 51$ patients from ref. [63]) pancreatic cancer samples. **b**, Frequency of SNVs, indels, CNVs and translocations by WES, aCGH and M-FISH in PK mice ($n = 38$) and human pancreatic ductal adenocarcinomas ($n = 51$ patients for SNVs, indels, CNVs (data from ref. [63]) and $n = 24$ cell lines for translocations). $^{**}P = 0.002$, $^{***}P \leq 0.001$, two-sided Mann-Whitney test; bars, median. **c**, Representative examples of M-FISH karyotypes from pancreatic cancers. Top, highly aneuploid human karyotype (70 chromosomes) with multiple translocations; middle, diploid mouse karyotype (40 chromosomes); bottom, complex mouse karyotype (77 chromosomes, 4 translocations). CNA, copy-number alteration; n, xxxxxxx. **a–c** adapted with permission from ref. [17], Springer Nature Limited.

disease, but because these are not undergoing surgery, such samples account for only <10% of cases in sample banks[18].

Another major advantage of the mouse is the possibility of tailoring experimental conditions to the study of specific context dependencies: (i) a plethora of possibilities for spatiotemporal genetic manipulations in mice[19] allow, for example, analysis of genetic-context dependencies by modeling specific subentities of individual cancer types initiated by different oncogenes; sequencing can be used to identify subentity-specific driver alterations[20,21]. (ii) Cellular-context dependencies can be investigated by activating oncogenes in different cell types of an organ, and genetic analyses of resulting cancers can identify 'cell of origin'–dependent oncogenic processes[22–24]. (iii) Moreover, qualitative manipulation of various environmental contexts (e.g., the immune system, tumor stroma, inflammatory conditions), allows us to study how these factors impinge on genetic tumor evolution and mutational processes.

Monitoring of cancers over time in mice also allows the study of the genetics/epigenetics of dynamic processes and phenotypes, such as epithelial–mesenchymal transition and drug resistance. Combined with the growing experimental toolbox for in vivo cellular barcoding, phylogenetic tracking and other types of evolutionary studies are now feasible at unprecedented scale and depth[25].

Finally, deep sequencing offers opportunities in the arena of forward genetics. Carcinogen-induced rodent models of human cancers have been used for decades. Examples are hepatocellular carcinoma, skin cancer and lung cancers induced by diethylnitrosamine[26], dimethylbenzanthracene[27] and *N*-nitroso-*N*-methylurea[28], respectively. Before the era of NGS, genetic studies in such models were substantially hampered by the low throughput of traditional approaches to cancer genome analysis. Recent studies showed, however, that chemical perturbation of genomes, combined with NGS of cancers in these mice, is a powerful approach for gene discovery and evolutionary studies[29–32].

Comparison with alternative methods

Sequencing-based cancer genome analysis detects—in contrast to other approaches—all classes of genetic alteration in one experiment and is also increasingly outcompeting other methods with respect to costs. For validation purposes, we have extensively used alternative techniques—including

amplicon-based sequencing (SNV validation), aCGH (CNVs) and M-FISH (rearrangements)—to detect individual alteration types.

With respect to the sequencing technology, all data analyzed in this protocol were generated on the widely used Illumina platforms (short-read, paired-end sequencing by synthesis). We prefer Illumina systems for WGS or WES because of their lower costs or lower sequencing error rates as compared with alternative short-read sequencing methods such as semiconductor-based sequencing (Ion Torrent, Thermo Fisher Scientific) or DNA nanoball sequencing (Beijing Genomics Institute), respectively.

The Illumina HiSeq 3000/4000 system has been widely used for WES and WGS sequencing; however, both the HiSeq X Ten (which can be used only for WGS) and the recently released NovaSeq system (WES and WGS sequencing) achieve lower prices per sequenced base. All HiSeq systems use a four-color chemistry, whereas NovaSeq uses the novel two-channel sequencing-by-synthesis chemistry. Although there are minor differences in base-calling quality between these systems[33], our pipelines perform well with raw data produced by any of the Illumina systems.

Ultra-long-read sequencing (>1 kb) through single-molecule real-time sequencing (Pacific Biosciences) and nanopore-based sequencing (Oxford Nanopore) can substantially improve the detection of complex structural variations and also offer applications in epigenetics. Currently, their use in the analysis of cancer genomes is limited, as both error rates and sequencing costs per megabase are higher compared to short-read sequencing[34].

Further information on alternative bioinformatics methods is given in the specific sections below.

Limitations

Limitations of sequencing-based analysis of different genomic alteration types will be discussed specifically in the respective sections below.

Experimental design

Sample collection and DNA isolation

For the analysis of somatic alterations in cancer, a matched normal sample is required. Although any non-cancer tissue can be used as reference sample, circulating tumor cells can be a confounding factor; therefore, blood is not ideal as a control sample. Typically, reference DNA is most easily obtained from tail tips, which are collected during necropsy.

Sequencing of matched control samples—instead of relying only on germline variant filtering using publicly available databases—is important even when using inbred mice. First, germline data are essential for LOH analysis. Second, single-nucleotide germline mutations are acquired at a rate of $\sim 5 \times 10^{-9}$ per generation and base pair (~15 novel SNVs per generation)[35,36]. The same is true for copy-number alterations, with 1×10^{-2} to 1×10^{-6} novel copy-number alterations per generation being reported, depending on the genomic region[9,37]. Novel SNVs acquired through breeding are not represented in databases, which are based on sequencing data from only a few mice per inbred line[10]. Given that inbred lines are often kept over many years or even decades at research institutions, this can profoundly affect SNV calling and determination of mutation rates/patterns. For example, in a primary mouse pancreatic cancer cell culture, 3,573 mutations were identified using databases to filter out potential germline variants. Additional filtering against matched control tissue revealed, however, that 96% of these SNVs are germline variants, leaving only 136 true somatic SNVs.

High-quality input DNA free of contaminants increases the odds of successful and reliable sequencing library preparation. For most sample types, such as tissue, blood or cultured cells, commercial DNA extraction kits using silica-based DNA immobilization are commonly used because they are easy to use and yield consistent results. Archived material is typically formalin-fixed and stored in 70% ethanol or embedded in paraffin (FFPE). Both fixation and embedding can adversely affect the integrity of DNA. We recommend using DNA isolation procedures tailored to specific sample materials in order to ensure amplifiable DNA and adequate sequencing quality[38]. Precise determination of DNA concentrations is necessary to ensure the equimolar representation of each sample in the library. We recommend quantification assays using dsDNA-specific fluorescent dyes.

Library preparation and sequencing

Currently, Illumina short-read, paired-end sequencing is most commonly used for WES and WGS. For WES, on-target coverage of ~80–100× is typically aimed for. To reduce technical bias, we suggest pooling libraries from multiple samples and spreading these across multiple lanes. For WGS, one

Illumina HiSeq X lane, for example, results in ~30× whole-genome coverage using 2 × 150-bp paired-end sequencing. However, depending on the experimental setup and analyzed tissue, it can be necessary to significantly increase sequencing depth. For example, the tumor-cell content in stroma-rich tumors is often <50% of all cells, decreasing the effective sequencing depth of tumor cells. The experimental question can also affect the required sequencing depth. Experiments aimed at studying intratumor heterogeneity and clonal evolution of metastasis, for example, typically require high sequencing depth[39].

Reference files

After initial quality control and trimming, reads are mapped to the reference genome GRCm38 (mm10; http://www.xxxxxxxx), which is based on the C57BL/6J strain. Separate reference genomes have been generated for the most widely used laboratory strains, but these are not routinely used for mapping during analysis of mouse cancer genomes (because they are of inferior quality and are not represented in the standard GRCm38 database). In addition, mice used in cancer studies are often kept on a mixed background.

Information on known mouse germline variations, mostly resulting from strain-specific differences, is, however, needed for base quality recalibration after alignment, as well as for filtering of somatic mutations to reduce false-positive calls. The most widely used database of mouse germline variation is maintained by the Wellcome Sanger Institute (https://www.sanger.ac.uk/sanger/Mouse_SnpViewer).

SNV and small indel calling

Various mutation callers are available for the analysis of WES or WGS data, the most widely used ones being Mutect, Strelka and VarScan. Mutect1(ref. [40]) has been used as an SNV somatic mutation caller in mouse WES studies[17,20,21,29,31,41]. A newer version, Mutect2, was released in 2018 and is already being used in newer publications[42]. Whereas Mutect1 calls only SNVs, Mutect2, which is recommended in this protocol, can detect both SNVs and indels. Both versions use a Bayesian classifier testing a reference model (which assumes that each observed non-reference base is due to sequencing error) against the variant model (which assumes that the specific site contains a true variant). A similar approach is used by Strelka2 (ref. [43]), which can also call both SNVs and indels. Another algorithm, VarScan2, uses a Fisher's exact test to compare the proportion of variant frequencies between tumor and normal samples[44].

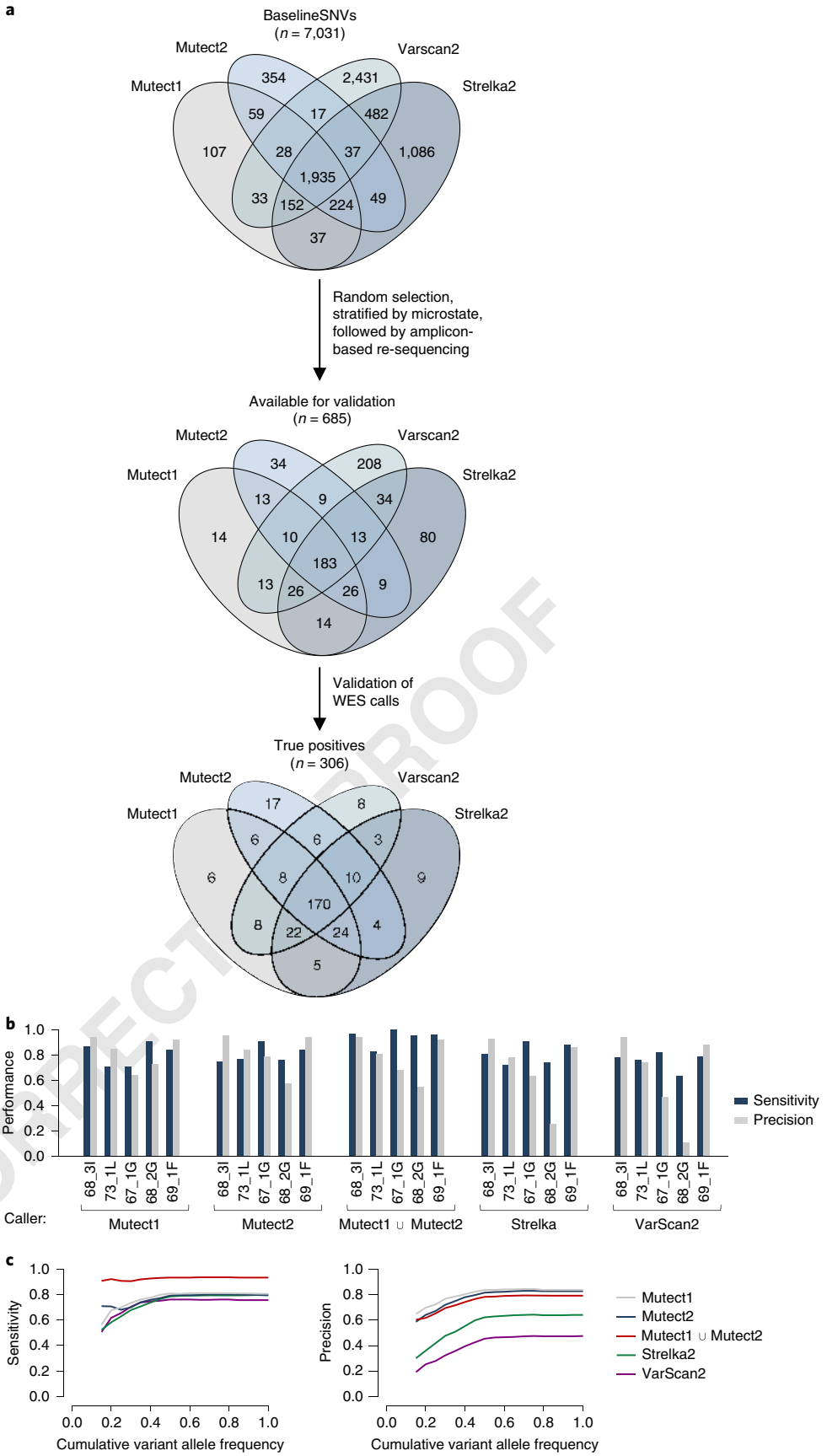
Tools commonly deployed in population genetics, such as GATK HaplotypeCaller, are less well suited to the analysis of cancer genomes[45]. These tools are primarily intended for genotyping germline variants. Their design does not account for cancer-specific aspects, such as varying degrees of healthy stromal cell content, aneuploidy and intratumor heterogeneity with subclonal tumor cell populations. By contrast, dedicated somatic mutation callers typically integrate data from both the control and the tumor sample into a joint statistical model.

Validation and choice of somatic SNV/indel callers

To evaluate the performance of our mutation-calling workflow, we systematically validated SNV calls made by Mutect2 (GATK 4), Mutect1 (GATK 3), Strelka and VarScan2 in mouse primary gastric cancer cell cultures (Fig. 3a), using amplicon-based re-sequencing (685 validated positions; for details, see Supplementary Methods).

Figure 3b shows the performance of the different SNV callers at the individual sample level. Weighted mean values for sensitivity and precision of SNV detection are summarized in Table 1. Mutect1 and Mutect2 outperformed Strelka2 and Varscan2. Although differences in weighted mean sensitivity were not marked (0.81, 0.8, 0.8 and 0.72 for Mutect1, Mutect2, Strelka2 and VarScan2, respectively), we noticed substantial discrepancies in precision, with Mutect1 and Mutect2 consistently reporting fewer false-positive calls than the other algorithms.

These differences between callers were evident over the whole range of variant allele frequencies. Figure 3c shows the cumulative performance in relation to the frequency of analyzed variant alleles. As expected, the confidence of calls was smallest at low mutant allele frequencies. For example, the sensitivity and precision for Mutect2 were 0.7 and 0.61, respectively, at mutant allele frequencies of 0.1–0.2 but increased to 0.89 and 0.95, respectively, at allele frequencies between 0.4 and 0.5. We noted that when combining results from Mutect1 and Mutect2, the increase in sensitivity was more pronounced than the decrease in precision (red curves in Fig. 3c), which could be exploited in projects in which high sensitivity is the key requirement.



◀ **Fig. 3 | Systematic comparison of SNV callers.** Mutect2 outperforms other SNV callers for mouse cancers when validated using targeted re-sequencing. **a**, SNV calling by four different callers identified a total of 7,031 mutations in mouse gastric cancer cell cultures ($n = 5$) on the basis of WES. From the pool of all detected mutations, SNVs were selected for targeted amplicon-based deep re-sequencing. For this, calls were stratified by sample, caller, allele frequency and base change (microstates) and sampled from each stratum randomly. After re-sequencing (median coverage, 13,550; interquartile range, 7,913–20,794), 685 SNVs, from which 306 were true positives, were used for benchmarking the callers. **b**, Sensitivity and precision of SNV callers for individual mouse gastric cancer cell cultures based on the validation of 685 SNV calls by targeted amplicon-based deep re-sequencing. Note that the union (Mutect1 ∪ Mutect2) contains all SNVs detected by either Mutect1 or Mutect2. **c**, Sensitivity (left) and precision (right) of SNV callers in relation to SNV allele frequency. Performance of SNV callers was tested on the basis of 685 validated SNVs. Values for sensitivity or precision were calculated in windows of 0.05, starting at a variant allele frequency of 0.1.

Table 1 | SNV and indel calls detected by Mutect1, Mutect2, VarScan2, Strelka and Pindel in five murine gastric cancer primary cell cultures

Type	Caller	Baseline calls	Validated calls	True positives	False positives	False negatives	Weighted mean sensitivity	Weighted mean precision
SNV	Mutect1	2,703	685	249	48	58	0.81	0.84
	Mutect2	2,575	685	245	51	62	0.80	0.83
	Mutect1 ∪ Mutect2	3,032	685	286	74	20	0.94	0.79
	VarScan2	5,115	685	235	258	72	0.72	0.50
	Strelka2	4,002	685	247	137	60	0.80	0.66
Indel	Mutect2	200	179	133	25	16	0.90	0.84
	Strelka	260	179	59	5	90	0.40	0.90
	Pindel	26	179	5	5	143	0.04	0.39

Sensitivity and precision were calculated as weighted means, accounting for the number of calls in each sample.

For detecting small indels (defined as indels of ≤10 bp), we tested Mutect2 and Strelka2 (Mutect1 does not support indel calling). Moreover, we included Pindel[46], a tool which has been widely used in sequencing studies for indel calling. We validated indels identified by these tools in an approach similar to the one described above for SNVs (179 validated positions; Fig. 4a). We determined sensitivity and precision for each tool, allowing their comparison (although it is worth noting that the 'true' sensitivity and precision cannot be determined in this manner because even the cumulative combination of all three tools might miss some indels). As shown in Fig. 4b, Mutect2 substantially outperformed Strelka2 and Pindel, particularly in regard to sensitivity, which was 0.9, 0.4 and 0.04 for Mutect2, Strelka2 and Pindel, respectively (Table 1).

To evaluate the performance of our pipeline on datasets from other laboratories, we used a validated dataset from a recent study that performed exome sequencing on mouse cancers[20] (Supplementary Methods). The precision and sensitivity of our workflow were 0.95 and 0.89, respectively, confirming the high quality of our SNV-calling pipelines.

In summary, we recommend the use of Mutect2 for SNV and indel calling in mouse cancers. The software is well documented and supported by a large and active user and development community. Mutect1, which was widely used in the past, has two disadvantages: namely, its longer overall runtime and, more importantly, its inability to detect indels.

However, we found that the Mutect2 runtime can vary considerably, depending on the degree of aneuploidy (e.g., 8 h for the analysis of 30× of a purely diploid sample compared to >24 h for a highly aneuploid sample). Strelka2 has substantially lower runtimes (e.g., 35 and 55 min, respectively, for the samples discussed above). Therefore, if computational cost is a constraint, the use of Strelka rather than Mutect2 for somatic mutation calling could be considered (keeping in mind that mutation calling using Strelka has slightly lower precision and sensitivity).

Postprocessing of somatic mutation calls

SNV and indel postprocessing can substantially affect the quality of results and is often tailored to the specific experimental setup. In the analysis of WES from primary cell cultures or cell lines with

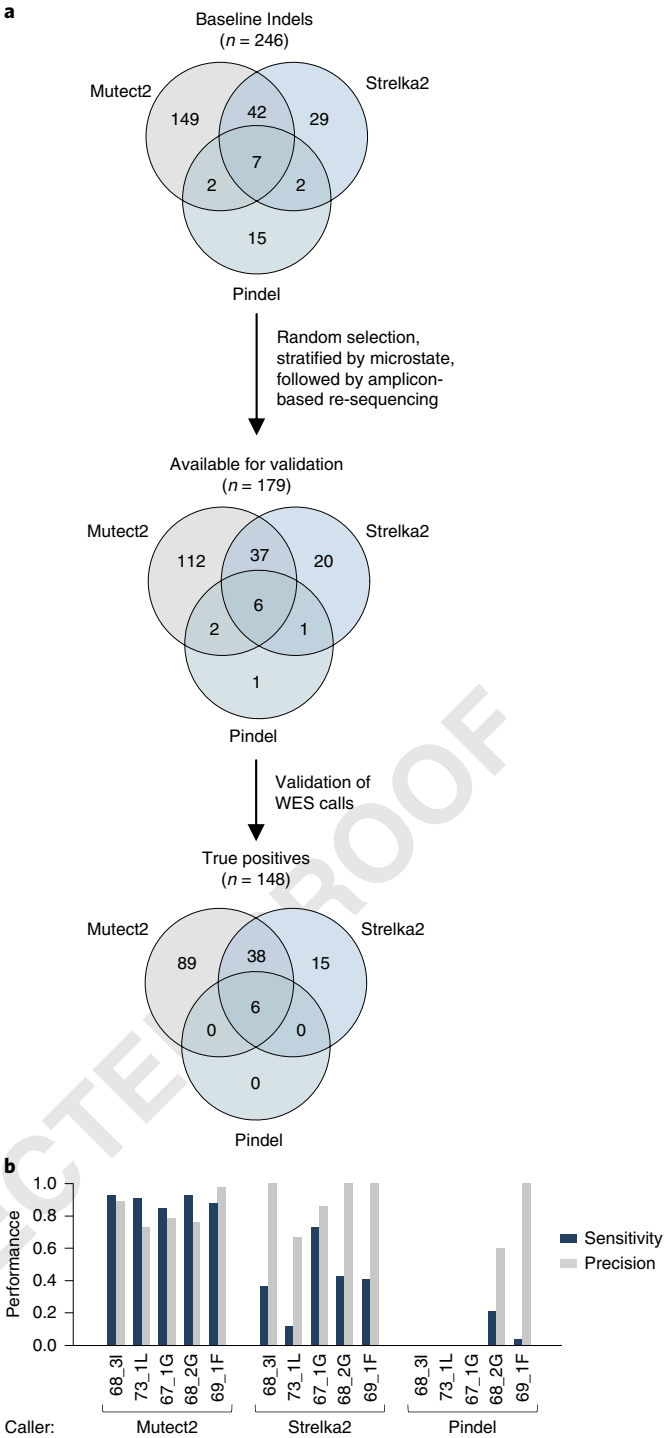


Fig. 4 | Systematic comparison of callers for the detection of small indels. Using targeted re-sequencing, small indel calls were validated. In a comparison of three callers, Mutect2 outperformed both Strelka2 and Pindel. **a**, Indel calling (size, ≤10 bp) by three different callers in mouse gastric cancer cell cultures ($n = 5$) on the basis of WES. 246 indels were identified in total. Indels were stratified by sample, caller, allele frequency and indel size and randomly selected from each stratum. These positions were used for targeted amplicon-based deep re-sequencing (median coverage, 10,625; interquartile range, 6,548–16,757). For benchmarking of the callers, 179 positions, from which 148 were true positives, were used. **b**, Sensitivity (left) and precision (right) of indel callers for individual mouse gastric cancer cell cultures on the basis of the validation of 179 indel calls by targeted amplicon-based deep re-sequencing. The sensitivity and precision for Pindel were 0 for calls from three samples, because Pindel found no true positives for those samples.

Box 1 | Description of SNV output

In the final output file, each line describes the effect of one unique mutation on one unique transcript, of which there are often multiple per gene. An exemplary output is provided in Table 2.

`CHROM` Chromosome name
`POS` Genomic position. For indels, position of the first reference nucleotide.
`REF` The reference base at POS. In case of deletions, the base before the event is included.
`ALT` The alternative/variant base at POS.
`GEN[Tumor].AF` Allele frequency. Frequency of the alternative (mutated) allele.
`GEN[Tumor].AD[0]` Number of reads in the tumor supporting the reference base at POS.
`GEN[Tumor].AD[1]` Number of reads in the tumor supporting the alternative base at POS.
`GEN[Normal].AD[0]` Number of reads in the control supporting the reference base at POS.
`GEN[Normal].AD[1]` Number of reads in the control supporting the alternative base at POS.
`ANN[*].GENE` Gene name (HGNC)
`ANN[*].EFFECT` Effect of this variant using sequence ontology terminology. A detailed explanation of each effect can be found in the SnpEff documentation (http://snpeff.sourceforge.net/SnpEff_manual.html)
`ANN[*].IMPACT` Each effect is categorized into one of our impact categories. Generally, changes in the amino acid sequence of protein-coding genes are categorized as MODERATE or HIGH.
`ANN[*].FEATUREID` Here, this corresponds to the Ensembl Transcript ID.
`ANN[*].HGVS_C` Nucleotide change using HGVS annotation.
`ANN[*].HGVS_P` Amino acid change using HGVS annotation (if the variant affects the coding region)

matched tail tissue (coverage ~100×), we use the following postprocessing filter settings: for SNVs and indels, we recommend using a variant allele frequency ≥10%, a minimum of 10× read coverage in both tumor and normal at this position, a minimum of three reads supporting the variant allele in the tumor sample and no reads supporting the variant allele in the matched normal tissue. The proposed settings are aiming at reducing false-negative calls (increasing precision) while keeping sensitivity high (>90%).

Of note, optimal threshold settings strongly depend on which parameter is more important for a given experiment: sensitivity or precision. For example, in cases in which sensitivity is critical, settings can be relaxed and downstream validation experiments (e.g., using amplicon-based resequencing) can correct for false-negatives. Examples of experimental setups or questions that can warrant qualitative changes in postprocessing are (i) evolutionary studies, which often use low threshold settings (for read support and variant allele frequency) in order to achieve high sensitivity in detecting subclonal events. This is particularly relevant in cancers with a high degree of intratumoral heterogeneity. (ii) Studies using archived material often rely on the recovery of tumor and normal DNA from the same FFPE tissue slide. This carries the risk of tumor cell 'contamination' within the normal tissue. In such cases, detection of 'contaminating' tumor-derived variants in the normal tissue would lead to exclusion of true-positive SNVs in the tumor. In such a scenario, the absolute read count filter for the variant allele must be raised in the normal tissue.

All postprocessing steps are performed using the variant call format (VCF), which describes all called mutations and can be used in virtually all genomic software tools. However, VCF was designed for the interchange between computer programs and is therefore not the best choice as a final output. We therefore export the final results in tabular format. An explanation of all relevant fields can be found in Box 1, and exemplary data are shown in Table 2. Note that genetically engineered mutations in mouse cancer models are present in both the tumor and the germline and are therefore filtered out during standard SNV calling (for example, $Kras^{G12D}$ in the pancreatic cancer model described above).

Sources of error

Historically, one important finding was that during the library preparation, C>A/T>G artifacts were introduced at low frequencies through a combination of heat, induced during DNA shearing, and contaminants in the DNA buffers, resulting in oxidation of guanine[47]. Although this potential source of artifacts was first reported in 2012, implementation of improved protocols for library preparation was often not immediate in sequencing facilities. Although there are tools available for the removal of these artifacts (FilterByOrientiationBias from GATK, DKFZBiasFilter), these often can only attenuate the problem. It is therefore advisable to treat C>A/T>G-calls, originating from raw data generated before or shortly after 2012, with great care. We recommend evaluating all samples for possible sequencing artifacts. Because the read frequency of such artifacts is low, an additional filtering step using high variant frequency thresholds (requiring an allele frequency of >0.2) should be considered, particularly in cases in which high precision is required (e.g., to compare mutational patterns among entities).

Another source of false-positive somatic calls in cancer is the failure to detect 'true' germline variants in the corresponding control tissue (incorrect variant calling in the healthy matched tissue). This source of error is often underestimated. To overcome this problem, we use the following two approaches. (i) First, we filter called somatic mutations using a database of known germline variants (Mouse Genome Project V5), maintained by the Wellcome Sanger Institute[10]. This list encompasses ~18 million SNVs and ~4 million indels. Importantly, these variants were generated using low-coverage WGS of only a few animals per strain. In our experience, the first filtering step using these data resources is not sufficient to remove all 'false-positive' somatic calls. (ii) Second, we routinely generate a cohort-specific list of germline variants, representing all germline SNVs and indels from all available animals (referred to as 'panel-of-normals'). We use this list to perform a second filtering step in order to remove false-positive somatic calls. As an example, the total numbers of somatic calls in one mouse gastric cancer cell line were 1,138 (before filtering), 1,121 (after the first filtering step) and 1,110 (after the second filtering step).

Interpretation of mutation calls

Several methods are available to further explore and interpret the relevance of somatic mutation calls. PROVEAN (Protein Variation Effect Analyzer; http://provean.jcvi.org/) is a software tool that predicts whether an amino acid substitution or indel has an impact on the biological function of a protein[48]. Variant alleles and their coordinates can be uploaded to the PROVEAN web interface. PROVEAN supports the analysis of mouse data, in contrast to comparable tools such as SIFT or PolyPhen-2. Another unique feature of PROVEAN is the possibility of interrogating functional consequences of in-frame indels, which is not supported by other tools.

Statistical approaches to separate commonly abundant passenger mutations from truly significant driver events are based on the assumption that. within a cohort of samples. mutations in driver genes occur more often than expected by chance. Unfortunately, the majority of tools for such approaches are specifically tailored to the analysis of human cancers. An exception to this is MuSiC2, which can be used for mouse data as well[49]. Required inputs for MuSiC2 are mapped reads (BAM files) and mutation calls (MAF or VCF files). In addition to statistics on the gene level, MuSiC2 can also be used to identify significantly enriched pathways within a cohort (increased number of mutated genes in specific pathways as compared to random expectation).

Mutational signatures, estimated from the trinucleotide context of SNVs, can be used to deduce the biological process generating mutations. This type of analysis requires additional reformatting of VCF files and uses the Bioconductor packages Somatic Signatures and Variant Annotation[50]. A major limitation of the identification of mutational signatures in mice, however, is the low number of SNVs per cancer. Tumors arising in genetically engineered transgenic mice often have fewer SNVs than needed for robust mutational signature detection (between 50 and 500 mutations).

Analysis of CNVs from whole-exome data

aCGH and single-nucleotide polymorphism (SNP) arrays have been widely used to call somatic CNV aberrations in mouse and human cancer. The widespread use of NGS to study SNVs in mouse cancers also allows extraction of CNV data from NGS results. To this end, we systematically tested and improved algorithms for the detection of somatic CNVs from WES and WGS data.

Tools that infer CNVs from WES data count reads mapping to genomic regions. These counts are de-noised, corrected for GC content and mappability, which is followed by normalization. Regions are then segmented by circular binary segmentation or a hidden Markov model (HMM). Several tools use reads mapping to exonic regions ('on-target' reads of whole-exome enrichment kits). We found, however, that this approach does not perform well in murine cancers. This is mainly due to sequence-specific variation in pull-down efficiencies during library preparation. Following DNA fragmentation, exonic regions are captured by biotinylated oligonucleotide baits. Thereby, capturing efficacy can be biased by variable factors, e.g., sequence context. Most tools try to adjust for such biases through statistical means.

A recently reported approach uses 'off-target' reads for the analysis of CNVs. These reads originate from genomic regions that are not specifically captured by the enrichment kits but fail to be removed by washing steps during library preparation. Historically, these 'undesired' reads could account for up to 60% of all sequenced reads. Improved library preparation workflows have reduced this number: in our mouse cohorts, the median percentage of reads mapping to off-target regions is ~20%.

CopywriteR[51] was the first tool using off-target reads for the analysis of CNVs. A later algorithm, CNVKit[52], uses both on- and off-target reads in a combined approach. As described in detail below,

Table 2 | Exemplary output of SNV calls

Chrom	Pos	Ref	Alt	Allele freq.	Reads tumor (Ref)	Reads tumor (Alt)	Reads normal (Ref)	Read normal (alt)	Gene	Effect	Impact	Transcript	HGVS_C	HGVS_P
18	34314862	G	C	0.43	89	68	134	0	*Apc*	missense_variant	MODERATE	ENSMUST00000079362.11	c.4810G>C	p.Ala1604Pro
11	69589213	C	T	0.90	5	46	39	0	*Trp53*	stop_gained	HIGH	ENSMUST00000108658.9	c.736C>T	p.Arg246*
10	20963869	C	A	0.17	73	15	95	0	*Ahi1*	synonymous_variant	LOW	ENSMUST00000105525.10	c.678C>A	p.Gly226Gly
1	22630308	TA	T	0.62	6	10	11	0	*Rims1*	intron_variant	MODIFIER	—	—	—
2	26384880	G	C	0.12	53	4	57	0	*Snapc4*	upstream_gene_variant	MODIFIER	ENSMUST00000114115.8	c.-4321C>G	
2	26384880	G	C	0.12	53	4	57	0	*Pmpca*	upstream_gene_variant	MODIFIER	ENSMUST00000076431.12	c.-4485G>C	—

Alt, variant (alternative) base; Chrom, chromosome; HGVS_C, nucleotide change; HGVS_P, amino acid change (for protein-coding genes); Pos, genomic position; Ref, reference base.

we found that, in mice, both tools perform considerably better than tools based on the analysis of 'on-target' reads.

To objectively determine and compare the quality of calls made by these tools, we used 38 murine pancreatic ductal adenocarcinoma primary cell cultures from an earlier study[17]. Tumors were induced in mice with pancreas-specific activation of $Kras^{G12D}$. We performed both aCGH and WES for each cell culture (Supplementary Table 1).

Agilent Genomic Workbench was used to preprocess and segment the aCGH files. Called segments were then manually reviewed and curated, e.g., by also evaluating M-FISH karyotypes, in order to obtain the highest possible quality of calls. Called segments with a log2 ratio between −0.25 and +0.25 were regarded as copy-number neutral. This relatively low cutoff was used to account for intratumoral heterogeneity and the frequent presence of aneuploidy/polyploidy in our cohort.

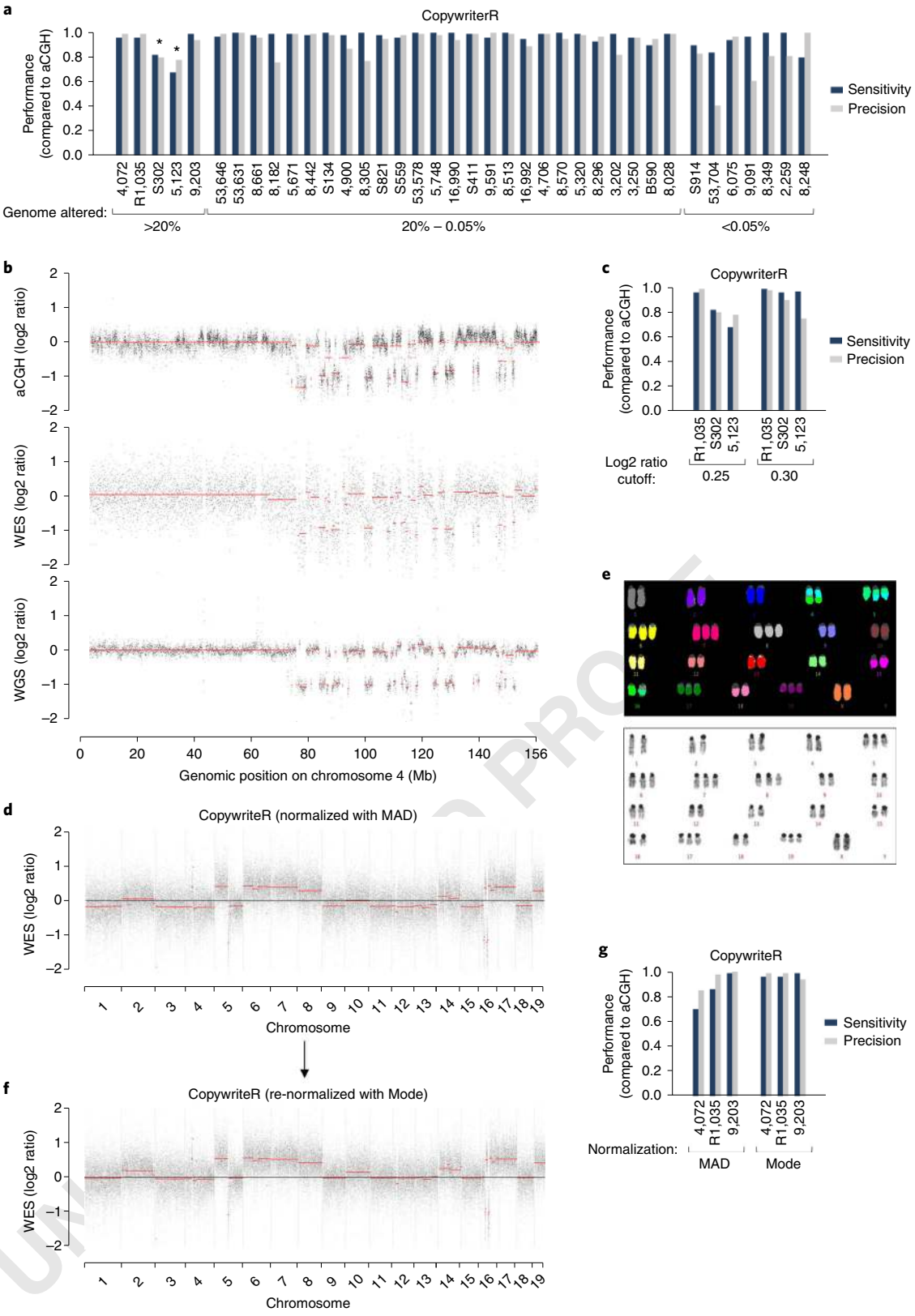
We used CopywriteR and CNVKit to determine copy-number aberrations for each gene from WES data in this cohort. For each tool, sensitivity and precision were determined using aCGH as a reference. For CopywriteR, the weighted mean sensitivity and precision were 94% and 93%, respectively (Fig. 5a), whereas for CNVKit both were 90% (Supplementary Fig. 1).

CNVKit, which uses on- and off-target reads, can be advantageous when looking at small exonic regions. As an example, Supplementary Fig. 2 shows an isolated small intragenic deletion of the *EGFR* gene, which was detected by CNVKit but not by CopywriteR. However, this advantage must be weighed against the slightly higher overall false-positive rates of CNVKit, which becomes particularly apparent in samples with few copy-number changes (note the drop of precision in samples on the right side of the graph in Supplementary Fig. 1). We prefer CopywriteR for calling CNVs and additionally use CNVKit to inspect specific genes of biological interest or cases in which there is evidence of small (exonic) deletions.

Below, we highlight important considerations for the use of CopywriteR. At the individual sample level, the majority of cancers analyzed by CopywriteR have excellent sensitivity and precision scores: 21 out of 38 samples reached ≥98% for both performance indicators (Fig. 5a). Even for chromosomes with highly complex rearrangements, such as in chromothripsis (a frequent phenomenon in our pancreatic cancer cohort), we found very high concordance rates between CNV calls detected by WES (using CopywriteR) and aCGH, or WGS (using HMMCopy), shown in Fig. 5b.

In two samples (S302 and 5123, marked with asterisks in Fig. 5a) CopywriteR performed significantly worse than in the rest of the cohort. M-FISH of these samples revealed extensive aneuploidy/polyploidy and intratumoral heterogeneity (Supplementary Figs. 3–5), resulting in widespread copy-number changes with low log2 ratios (between 0.2 and 0.3, which is very near our cutoff of 0.25 for calling copy number–altered segments). This oscillation around our cutoff value is the cause for the decreased concordance between aCGH and CopywriteR. Importantly, when we raised our segment-calling threshold to ±0.3, concordance increased considerably (Fig. 5c).

The analysis of an extensive series of cancers allowed us to systematically search for limitations inherent to CopywriteR. Notably, we found that, in very aneuploid samples, CopywriteR assigns incorrect log2 ratios to called segments, which is due to incorrect centering to the 'zero baseline' (i.e., see Chr11–13 in Fig. 5d). Figure 5e shows the M-FISH karyotype for such a sample. Because this phenomenon was strongly dependent on the degree of aneuploidy, we suspected that CopywriteR's normalization method, which uses the absolute median deviation as a location parameter, was the cause. To verify this hypothesis, we adopted the normalization strategy used in CNVKit for re-centering called segments from CopywriteR. By contrast, CNVKit uses the mode derived from a



Gaussian kernel estimator as location parameter. This allowed us to correct faulty annotations, resulting in substantially increased concordance with M-FISH and aCGH data, even in highly aneuploid samples (Fig. 5f,g). This re-normalization is implemented in the Procedure.

◀ **Fig. 5 | Performance of CopywriteR for detecting copy-number changes.** Copy-number changes can be inferred from WES data using CopywriteR with precision similar to that of aCGH. **a**, Sensitivity and precision of CopywriteR (median on-target coverage of 75×; from SureSelect^{XT} Mouse All Exon kit ;49.6 Mb) in primary pancreatic cancer cell cultures ($n = 38$). CNV calls were benchmarked with corresponding reference aCGH data (Agilent SurePrint G3 Mouse CGH; 240K) by gene-wise comparison. Called segments with a log2 ratio between −0.25 and +0.25 were regarded as copy-number neutral. Samples were sorted by the fraction of the genome affected by CNVs. Two samples (*) performed significantly worse than the rest of the cohort, owing to a large degree of intratumoral heterogeneity and aneuploidy/polyploidy (see also **c** and Supplementary Figs. 3–5). **b**, Copy-number profiles of Chr4 from one primary pancreatic cancer cell culture sample (S821) detected by aCGH (top), WES using CopywriteR (middle) or WGS using HMMCopy (bottom) show high concordance. **c**, Effect of increasing the log2 cutoff on the performance of CopywriteR, as compared to aCGH, in polyploid cancers with substantial intratumoral heterogeneity (Supplementary Figs. 3–5). **d**, Copy-number profile estimated by CopywriteR for aneuploid sample R1035. For centering, CopywriteR uses the absolute median deviation (MAD), which incorrectly centers copy-number states in highly aneuploidy cancer genomes. Note the shift of the log2 ratio for chromosomes 1, 3, 9, 11 and 12, indicating a subclonal loss, was not confirmed by M-FISH (**e**). **e**, Representative M-FISH karyotype for the same sample. In total, ten separate karyotypes for this sample were analyzed: +2 (2/10 analyzed karyotypes), +5 (10/10), +6 (10/10), +7 (10/10), +8 (7/10), +14 (5/10), +17 (10/10), and +19 (5/10). **f**, Re-centering of the copy-number profile estimated by CopywriteR for sample R1035. Using the mode, estimated by a Gaussian kernel estimator of the called segments, results in expected log2 ratios for all chromosomes. Mode centering results in a shift of the log2 ratio of +0.16. **g**, Performance of CopywriteR using MAD or mode estimator for centering. After correction using the mode estimator, the performance of CopywriteR improves for the samples with the highest CNV load.

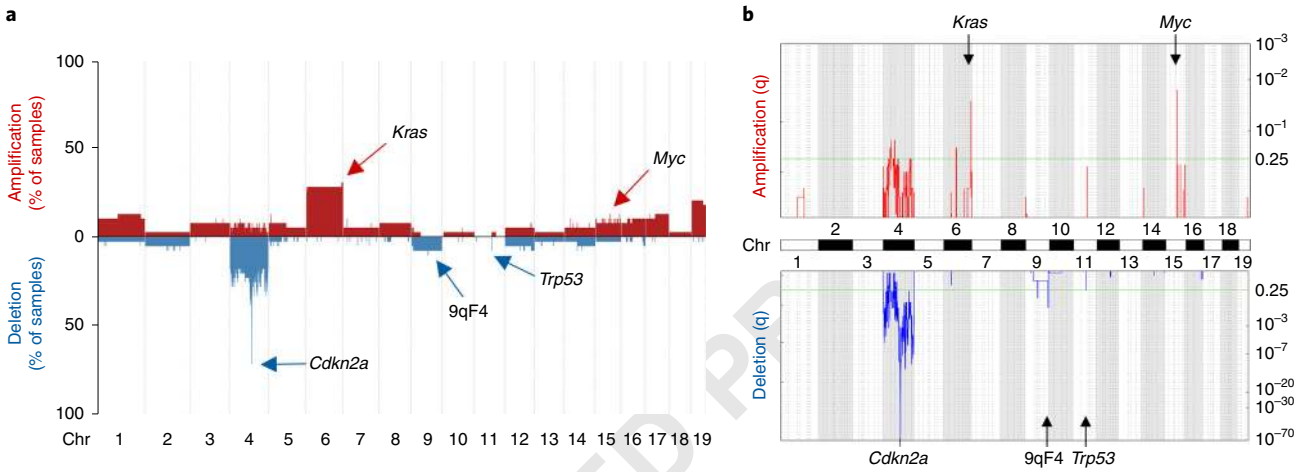


Fig. 6 | Analysis of copy-number changes across one cohort. a, Overlay of copy-number profiles from a cohort of primary pancreatic cancer cell cultures ($n = 38$). The *y* axis shows the frequency of amplifications (up) or deletions (down) in the cohort, with *Cdkn2a* and *Kras* loci being most frequently affected by copy-number alterations. **b**, GISTIC2 plot for the same cohort. The significance threshold is $q = 0.25$ (green line). Chr, chromosome.

For visualization of recurrent CNVs, multisample overlays showing changes at a cohort level can be generated (Fig. 6a). Moreover, to identify regions of the genome that are significantly amplified or deleted across samples, we use GISTIC2 (Fig. 6b)[53]. Because the GISTIC2 package does not include reference files for the analysis of mouse genomes, the respective files for GRCm38 are provided in the MoCaSeq repository ('Equipment setup' section).

Detection of LOH

LOH is a hallmark of cancer evolution. It can be studied by the analysis of common (SNPs) and rare SNVs in the genome using NGS. We tested a variety of LOH callers. Frequently, their output was erratic when using mouse samples, whereas the same tools robustly called LOH from human samples when compared with other methods.

Identification of heterozygous variant positions in the germline

For the detection of LOH, the first step is the identification of heterozygous variant positions in the germline of the respective animal ('informative' variants). A problem that arises when extracting these positions from mapped sequences is difficulty in differentiating between 'true variants' and sequencing artifacts, which typically requires extensive postprocessing (arrow in Fig. 7a).

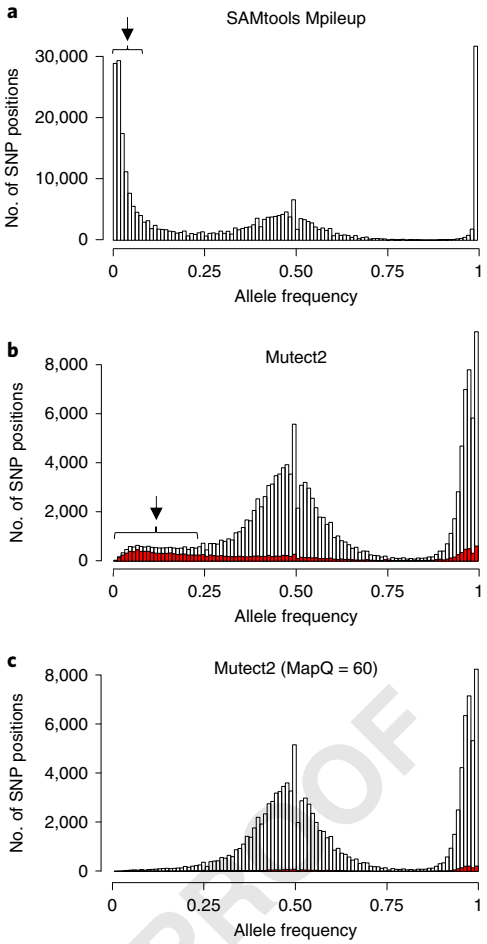


Fig. 7 | Identification of heterozygous variant positions in the mouse germline. a, Allele frequency distribution of all germline variants from WES data of a single mouse tail, extracted using SAMtools. Note the high rate of variants with an allele frequency <10% (arrow) and >98%. **b**, Distribution of allele frequencies of variants detected by Mutect2, run in single-sample mode, in the same tail sample. Most of the technical artifacts are removed. A number of variants between allele frequencies of 0 and 0.25 remain (arrow), which is unexpected in diploid genomes (peaks are expected at 0.5 and 1). These variants are not technical artifacts but true variants located almost exclusively in segmental duplications (number of positions in segmental duplications marked in red). **c**, Allele frequency distribution of all germline variants identified by Mutect2 after filtering for a mapping quality of 60. Here, only reads mapping uniquely to the reference genome remain, avoiding mapping in segmental duplications (marked in red), repetitive regions or pseudogenes. As expected, this results in a bimodal frequency distribution. MapQ, mapping quality.

Variant calling using Mutect2 in single-sample mode removes the vast majority of these sequencing artifacts (Fig. 7b).

After artifact removal, variant positions with allelic frequencies outside the expected range of a diploid genome are still present (arrow in Fig. 7b). We found that these variants originate from reads mapping to pseudogenes and regions of segmental duplication (overlap of these regions is marked red in Fig. 7b).

Duplicated sequences in mouse genomes have three to four times as many paralogs as compared with those of human genomes[8]. Our attempt to remove these regions using mouse databases of segmental duplications was unsuccessful, most likely because of incomplete annotation of these genomic features. We reasoned that filtering based on mapping quality could improve alleviate problem. Mappers such as bwa-mem assign each read a mapping quality score, which integrates different parameters, such as base quality, similarity (sequence identity) to the reference genome and 'uniqueness' of the mapped position. We tested various mapping quality thresholds on mouse data and found that reliable removal of ambiguous positions in segmental duplications and pseudogenes requires rigorous filtering using the maximum possible threshold (Fig. 7c).

Differences between humans and mice in the number of informative variant positions

Both the absolute number and the distribution of heterozygous SNVs in the germline are important criteria for the detection of LOH. In humans, these informative variants are distributed homogeneously throughout the genome/exome (Fig. 8a). The differences between individuals, even when including different ancestries, is negligible for the analysis of LOH.

By contrast, significant differences in the number and distribution of informative variants between individual mice can exist, depending on their genetic background and the level of inbreeding and backcrossing. In the mouse cohorts typically used in cancer research, these effects can be substantial. For example, in a cohort of mouse pancreatic cancers induced by pancreas-specific $Kras^{LSL\text{-}G12D}$ expression, some mice were kept on a mixed Sv/129;C57BL/6 background (Fig. 8b), whereas others were backcrossed to C57BL/6 to varying degrees (Fig. 8c,d).

In animals on mixed backgrounds or outbred mice, the distribution of germline variants allows LOH analysis at most genetic loci (Fig. 8b). By contrast, the frequent use of inbred genetic backgrounds in cancer research poses significant challenges to LOH analyses. Genomes of inbred mice have only a few nucleotides at which the maternal and paternal alleles differ. Figure 8c shows the variant allele frequency derived from WES for such an example: a knock-in mouse line that had been generated in Sv/129 embryonic stem cells and was subsequently extensively (13 generations) backcrossed to C57BL/6. LOH calling in cancer derived from this mouse is impossible for the majority of positions in the genome.

Figure 8d shows an example of only minimal backcrossing. Here, some chromosomes (e.g., Chr6) have high variant allele density, supporting LOH analyses. By contrast, for other chromosomes (e.g., Chr3), LOH analysis is impeded by the low variant allele density.

Visualization of LOH in mice

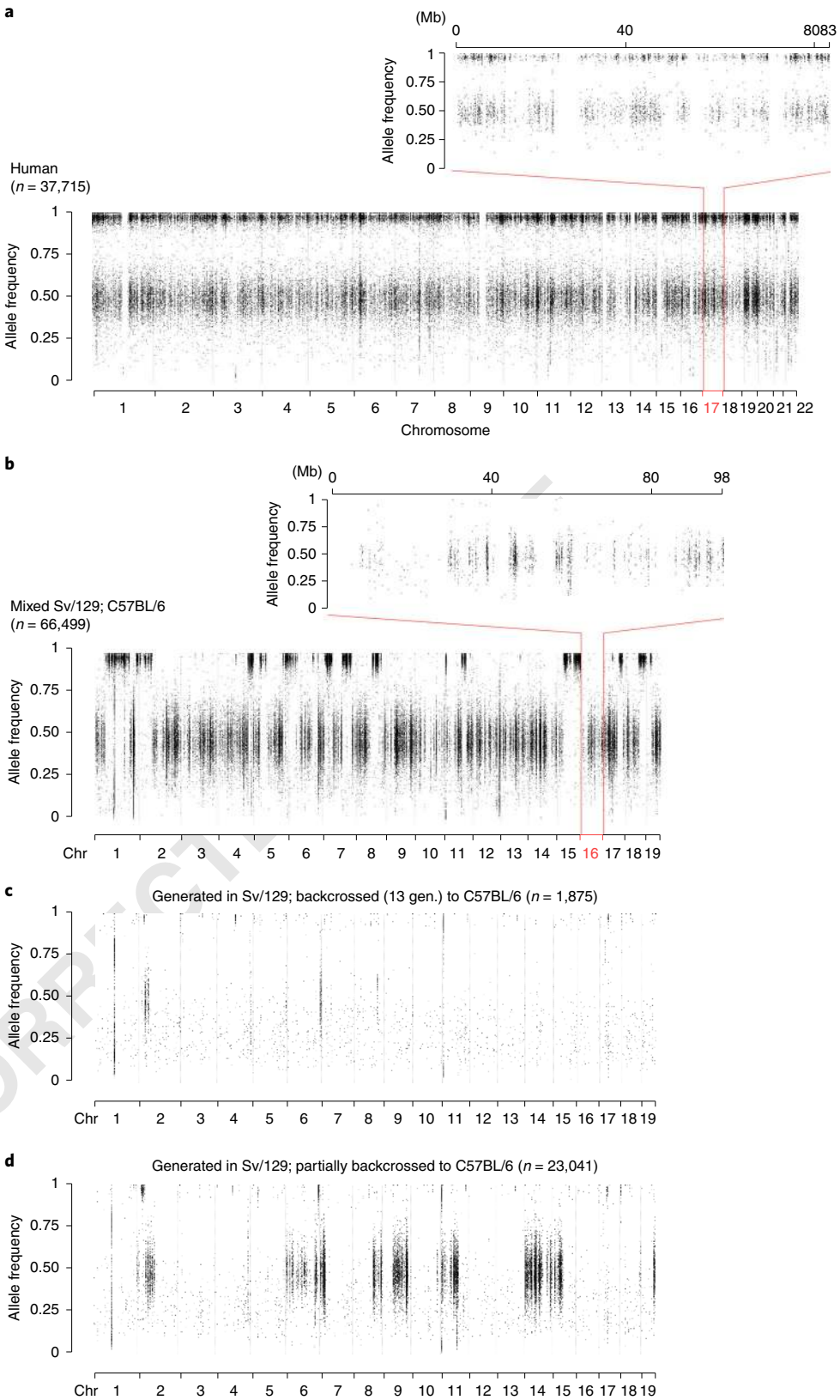
After the identification of heterozygous variant positions, the tumor variant allele frequency is plotted for these positions. When using human data from WES, this approach yields plots very similar to B-allele frequency plots derived from SNP arrays (Fig. 9a). In mice derived from crosses of different inbred strains, long stretches of variant alleles are often located on the same haplotype, even after many generations of interbreeding. In regions of LOH, this results in long blocks of continuous loss of either the variant or the reference allele (Fig. 9b,c). For inexperienced users, the visual interpretation of such data can be difficult. To simplify this, we developed a visualization tool creating LOH plots based on B allele frequency, as used earlier for visualization of SNP array data. This approach uses defined rules to designate each variant as the A or B allele. This leads to a symmetric representation of LOH regions on both sides of the LOH plot, similar to those for human data (see the transformation of the plot in Fig. 9c into the plot in Fig. 9d).

Complex genomic rearrangements

Complex genomic rearrangements can arise either through gradual acquisition or through a single catastrophic event. Breakage–fusion–bridge cycles, which are acquired during multiple cell cycles (progressive model), represent examples of the former. By contrast, chromothripsis is a single catastrophic event during which a localized region of a chromosome is shattered into multiple fragments. The chromosome is then reassembled through random re-joining of these fragments,

Fig. 8 | Mouse-specific limitations of LOH detection. a–d, Patterns of germline SNVs in healthy human and wild-type mouse genomes, on the basis of WES. Calls from Mutect2 were filtered for a mapping quality of 60 (Fig. 7). Each dot corresponds to the variant allele frequency of an individual SNP and its position in the mouse genome. For a diploid genome, the distribution of allele frequencies is expected to peak at 0.5 (heterozygous, variant allele inherited from one parent) and 1 (homozygous, inherited from both parents). Only heterozygous variants (informative positions) can be used for the detection of LOH. **a**, In the human germline, both hetero- and homozygous variants are distributed evenly throughout the genome. The plot on the right shows a zoom-in on Chr17. **b**, Mouse genome with mixed C57BL/6 and Sv/129 background. Although the absolute number of variants is comparable to those of human genomes, informative positions are not evenly distributed across the genome. Stretches of heterozygous variants are interrupted by blocks of homozygous variants, allowing the study of the LOH of most but not all genetic loci. The zoom-in on the right shows the distribution of germline variants on Chr16. **c**, Mouse genome with mixed C57BL/6 and Sv/129 background backcrossed to C56BL/6 background for 13 generations. Backcrossing resulted in extensive loss of informative germline variants, thus rendering LOH analysis impossible. **d**, Mouse genome with mixed C57BL/6 and Sv/129 background, partially backcrossed to C57BL/6. Note the strong variation of germline variant density at different chromosomes (e.g., Chr6 versus Chr3). Chr, chromosome; gen., generation.

during which genomic fragments can be lost[54]. Separate derivative or double minute chromosomes can be formed, which typically include oncogenes. Examples of chromothripsis affecting different chromosomes in mouse pancreatic cancers are shown in Fig. 10.



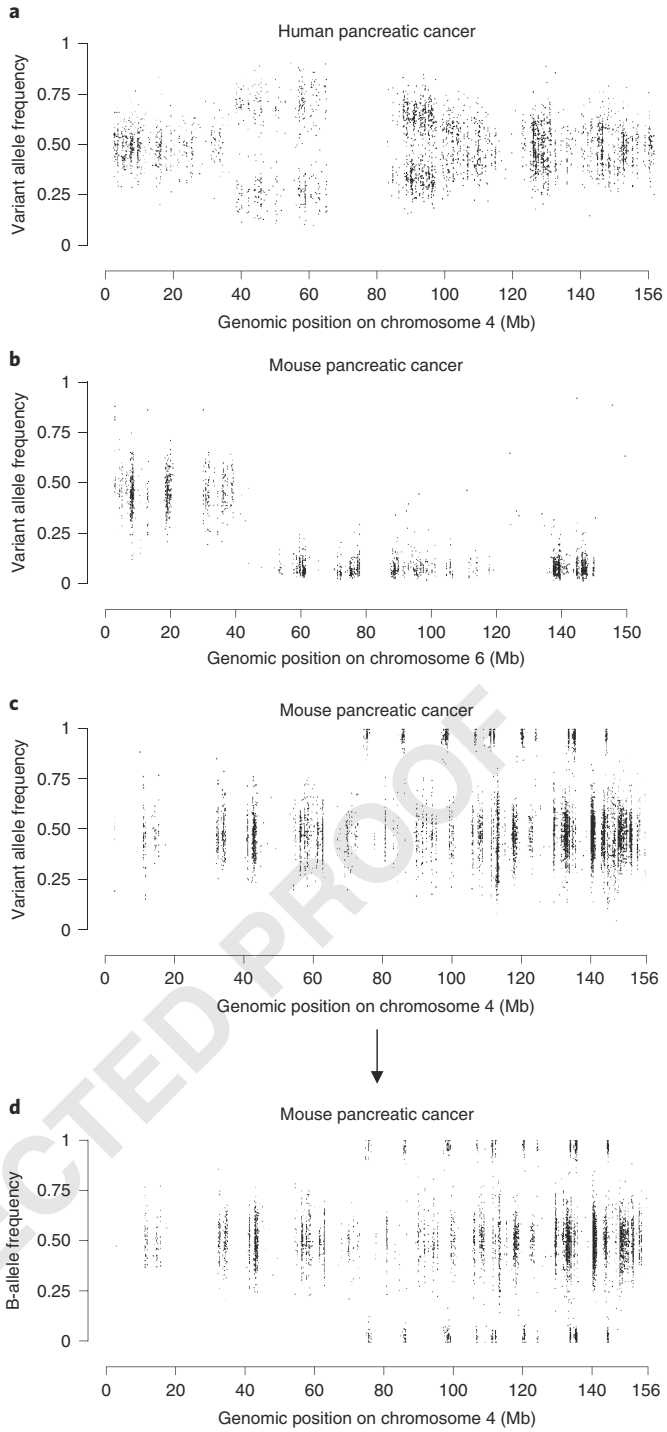


Fig. 9 | Visualization of LOH in human and mouse cancer genomes. a, Variant allele frequency plot of germline variants on Chr4 of a human pancreatic cancer genome on the basis of WES. **b**,**c**, Variant allele frequency plot for Chr6 (**b**, B590) and Chr4 (**c**, S821) in mouse pancreatic cancer cell cultures based on WES. In contrast to human cancer genomes, LOH in mouse cancer genomes results in long blocks with loss of either the variant (B590) or the reference allele (S821), evenly shifting all positions to one 'side' (toward 0 or 1) of the plot. **d**, B-allele Frequency (LOH plot) for Chr4 of mouse primary pancreatic cancer cell culture S821. The variant allele frequency was transformed into the corresponding B-allele frequency for each heterozygous germline variant. A- and B-alleles were defined following conventions developed by Illumina.

To differentiate chromothripsis from other forms of complex rearrangements, Korbel and Campbell proposed six hallmarks of chromothripsis[55]: (i) clustering of breakpoints, (ii) regularity of oscillating copy-number states, (iii) identical copy-number alteration and LOH patterns,

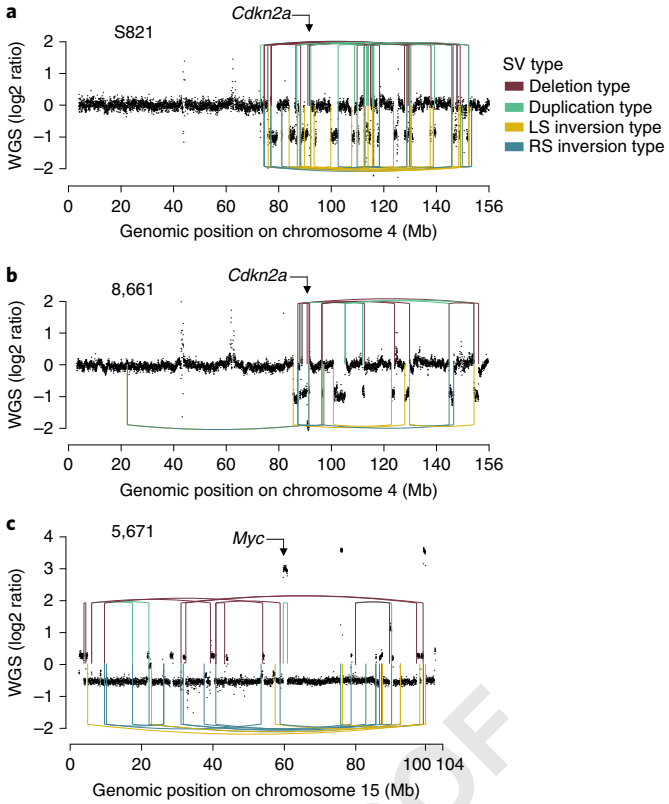


Fig. 10 | Examples of chromothripsis in mouse cancer genomes. a–c, Rearrangement graphs for chromothriptic chromosomes from three mouse pancreatic cancer cell cultures. In chromothripsis, a region of the chromosome is shattered into multiple fragments, which are then randomly rejoined. Fragments that are joined during chromothripsis are connected by lines that are superimposed on the copy-number profile (line color indicates fragment join type; see Fig. 12a for details). Note the association of chromothripsis with cancer-driving events such as deletion of *Cdkn2a* (**a**,**b**) and high-level amplification of *Myc* by double minute chromosome formation (**c**; double minute chromosome was excluded from rearrangement analysis). LS, xxxxxxxxx; RS, xxxxxxxxx; SV, xxxxxxxxx. Q28

(iv) rearrangement of only one haplotype, (v) randomness of DNA segment order/joints and (vi) the ability to walk the derivative chromosome (alternating head–tail sequences).

We implemented a pipeline for the systematic analysis and statistical testing of each hallmark in mouse cancer genomes. Input data for this pipeline are WGS-derived data describing structural variations (using Delly[56]), CNVs (HMMCopy[57]) and regions of LOH. Exemplary tests for these hallmarks for different mouse pancreatic cancers are shown in Fig. 11 and Supplementary Figs. 6 and 7. Note that although copy-number and LOH plots derived from WES can be used to 'screen' larger cohorts for potential chromothripsis, WGS is essential to detecting key hallmarks and providing definitive proof of chromothripsis.

Clustering of breakpoints. In contrast to a progressive model of rearrangement acquisition (sequential acquisition), in which breakpoints between fragments are distributed randomly across the chromosome, the breakpoints on a chromothriptic chromosome cluster together. This means that the observed distribution of distances between breakpoints after chromothripsis differs from a distribution of distances in which the breakpoints are randomly placed on a chromosome. An exponential distribution can describe the progressive model. The χ^2 test can be applied to test whether the observed breakpoint distances differ from this expected (exponential) distribution (Fig. 11a). Q29

Regularity of oscillating copy-number states. In a model of sequential acquisition of alterations, copy-number states of altered regions can change with the acquisition of new alterations, often resulting in multiple CNV states affecting a region of the chromosome (multi-stepped CNV plots). By contrast, chromothripsis is characterized by merely two to three distinct copy-number states (Fig. 11b).

For testing, a Monte Carlo approach can be used to simulate the sequential acquisition of all observed rearrangements affecting a chromosome. Our algorithm sequentially inserts a Q30

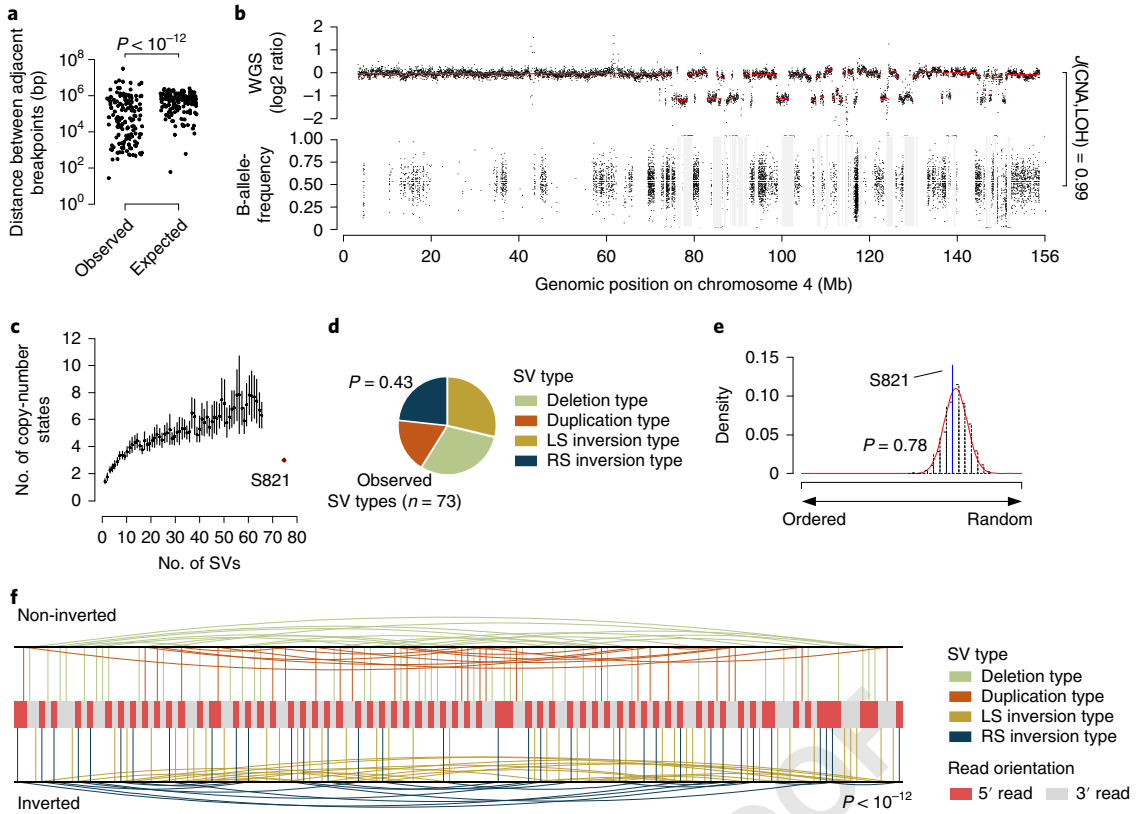


Fig. 11 | WGS-based inference of chromothripsis in mouse cancer genomes. a–f, Implementation of systematic analysis and statistical testing of chromothripsis hallmarks proposed by Korbel and Campbell[55]. WGS data from mouse pancreatic cancer primary cell culture S821 were used for analysis. **a**, Clustering of breakpoints: the distribution of observed distances between breakpoints ($n = 145$) differs significantly from an exponential distribution (Expected). $P < 10^{-12}$; χ^2 goodness-of-fit. **b**, Interspersed loss and retention of heterozygosity: comparison of CNV and LOH plots for Chr4. Copy-number and LOH events cluster in the second half of the chromosome. Only three distinct copy-number states (2, 1 and 0 copies) can be identified. Regions of loss and retention of heterozygosity alternate. There is near-perfect overlap between regions of LOH and copy-number loss (Jaccard index (J) = 0.99). **c**, Regularity of oscillating copy-number states: a Monte Carlo approach was used to simulate the sequential acquisition of observed rearrangements on Chr4 ($n = 1{,}000$ simulations per number of structural variations). Black dots represent the mean copy-number states. The associated 95% confidence intervals are shown as black lines. Chr4 showed fewer copy-number states than expected by sequential acquisition of observed rearrangements. **d**, Randomness of DNA fragment joins: all four types of structural variations are uniformly distributed in the chromothriptic chromosome. $P = 0.43$; χ^2 goodness-of-fit. **e**, Randomness of DNA fragment order: start and end positions of observed rearrangements ($n = 73$) were independently reordered and absolute rank differences were calculated to generate a random background distribution ($n = 1{,}000$ simulations). For sample S821, the observed value is located within the null model of random distribution, making it unlikely that the observed segment order arose in a progressive model. Two-sided $P = 0.78$. **f**, Ability to walk the derivative chromosome: rearrangement graph of Chr4 from sample S821 ($n = 146$ rearrangements). Each fragment is represented by two blocks, indicating the read orientations (3′ and 5′ in gray and red, respectively) for the start and the end of each chromosome segment when mapped to the reference genome (Fig. 12b). In a chromothriptic model, the read orientations will be alternating, resulting in a gray-red-gray-red pattern. The Wald–Wolfowitz test is used to test this alternating 3′-to-5′ pattern of paired-end read orientation ($P < 10^{-12}$). The connections between fragments are visualized above and below the blocks (line color indicates fragment join type; see Fig. 12a for details). See also Fig. 10a for visualization of the same connections superimposed on the copy-number profile. SV, structural variation. Adapted with permission from ref. [17], Springer Nature Limited.

number (n) of randomly chosen rearrangements from the input list of observed rearrangements into a chromosome and calculates the number of distinct copy-number states after each run. The simulation is re-iterated 1,000 times for each n between 1 and the total number of observed rearrangements.

In a sequential model, the number of distinct copy-number states increases with the absolute number of rearrangements, whereas in a chromothriptic model the number of distinct copy-number states is independent of the number of alterations (Fig. 11c).

Interspersed loss and retention of heterozygosity. In a diploid genome, loss of a chromosomal fragment leads to LOH of the corresponding region, which is irreversible. Therefore, in chromothripsis, there is high-level concordance between CNV and LOH patterns (Fig. 11b; concordance level reflected by Jaccard index).

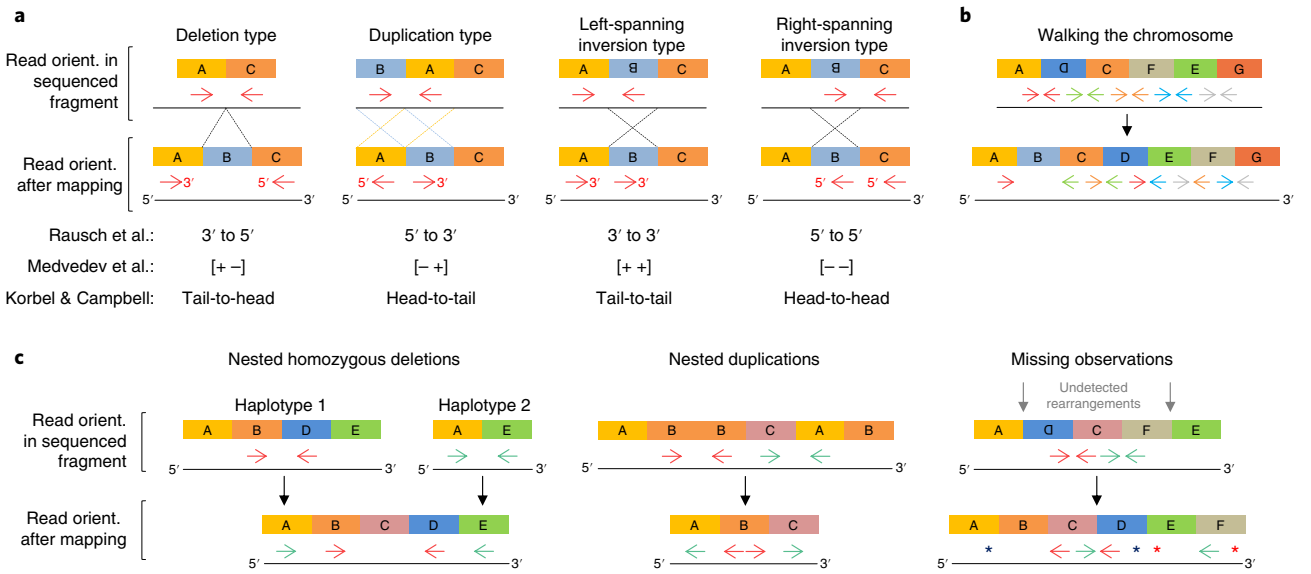


Fig. 12 | Features of chromothripsis. a, Nomenclatures for combinations of read orientations in paired-end sequencing used in the literature (Rausch et al.[56]; Medvedev et al.[59]; Korbel and Campbell[55]). The orientation of paired-end reads relative to the reference genome is altered by rearrangements and is specific for the rearrangement type (as proposed by Stephens et al.[54]). **b**, On a rearranged chromosome, each rearranged fragment contains a loose 3′ and 5′ ends that is covered by reads spanning the breakpoints in the 3′ and 5′ directions (colored arrows). In the case of a chromothriptic chromosome, mapping of breakpoint-flanking reads to the reference genome results in an alternating 3′-to-5′ read orientation pattern. **c**, The alternating pattern of 3′-to-5′ read orientations is disturbed by nested deletions or duplications originating from sequential accumulation of CNVs or by rearrangements, which are not detected by sequencing (Missing observations). Asterisks indicate missing read support for rearrangements, which remain undetected. orient., orientations.

As mentioned above, the analysis of LOH is highly dependent on the absolute number and distribution of heterozygous variants in the germline. Therefore, depending on the level of inbreeding, it can be difficult to evaluate this chromothripsis hallmark in mice.

Prevalence of rearrangements affecting one haplotype. During chromothripsis, a region of a single chromosome is shattered and reassembled, so that only one haplotype is affected by rearrangements. Testing of this hallmark therefore requires the reconstruction of the haplotypes for the affected chromosome. However, haplotype reconstruction from short-read paired-end WGS/NGS data (phasing) is possible only in combination with comprehensive databases of haplotype information, and even then results in only unconnected blocks of reconstructed haplotypes of ~2-Mbp length[58]. The precision of this reconstruction is determined by two factors: the quality and size of the haplotype databases and the density and distribution of heterozygous germline variants. In mice, the latter is a critical limiting factor because of the low variant number due to inbreeding. Therefore, testing for this hallmark is often not possible.

However, mouse crosses can be planned to overcome the necessity of haplotype reconstruction, facilitating the analysis of this hallmark; in crosses of two different inbred strains, the affected haplotype can be inferred directly from LOH plots (all LOH regions shift to one ‘side’ of the plot; Fig. 9b,c and Fig. 11b).

Randomness of DNA fragment joins and segment order. During chromothripsis, a region of the chromosome is shattered into multiple fragments and then randomly re-joined. Each join between two fragments, depending on the orientation of each fragment, can be classified into one of four categories: deletion type, duplication type and two different inversion types (Fig. 12a). Each of these categories is characterized by a unique pattern of read orientations between two paired-end reads when these are mapped onto the reference genome. In the literature, multiple different nomenclatures for these structural variants can be found[55,56,59].

The current assumption is that, during reassembly after chromothripsis, there is no preference for the type of join between two fragments. Therefore, each category should occur in 25% of the rearrangements. A χ^2 test can be used to test whether the observed distribution of joins significantly differs from the expected distribution. (Fig. 11d). In this test, a nonsignificant result supports the hypothesis of chromothripsis.

The order in which the fragments are reassembled after chromothripsis is random and independent of the types of joins between two segments. To this end, we order each segment according to its start position and assign ranks for both the start and the end positions. For a perfectly ordered chromosome, the difference between these ranks is 0. By contrast, for a chromothriptic chromosome, the difference in ranks is >0 and increases with the randomness of fragment order.

For statistical testing, we implemented a Monte Carlo approach by randomly reassigning the observed start and end positions 1,000 times and re-calculating the (absolute) mean rank difference for each simulation. The results of this test are shown in the histogram in Fig. 11e. If the observed mean rank difference is larger than the 5% percentile of this distribution, there is strong evidence that the observed fragment order originates from a random reassembly process.

Ability to walk the derivative chromosome. After a chromothriptic event, each chromosome fragment contains loose 3′ and 5′ ends, to which other fragments are joined during reassembly. In a paired-end sequencing approach, each breakpoint is supported by a read facing in the 3′ direction and a read facing in the 5′ direction (Fig. 12b). Mapping of these read orientations onto the reference chromosome results in an alternating 3′/5′ pattern, as shown in Fig. 11f. Statistically, this alternating pattern can be tested using the Wald–Wolfowitz test.

By contrast, in a progressive model with nested duplications or deletions, this pattern of read orientations is disturbed, leading to runs of segments with the same orientation (Fig. 12c). It should be noted that the test will also fail if breakpoint detection is insensitive, e.g., because of low sequence read coverage leading to missed observations.

Materials

Biological materials

Laboratory mouse strains can be obtained from external providers such as the Jackson Laboratory (e.g., *Kras*$^{tm4Tyj/J}$, stock no. 008180). Experimental mice are typically housed in isolated ventilated cages under specific pathogen-free conditions. The room temperature is set to 22 °C. Ambient lightning follows a 12 h/12 h dark/light cycle. Mice have free access to standard chow and water. Mouse handling is performed in a laminar flow cabinet. Mice are monitored daily by animal care staff. Ear clippings allow identification and genotyping of each animal **! CAUTION** All animal experiments must be approved by local authorities. They should be performed in accordance with the relevant local regulations and follow guidelines for the care of laboratory animals such as FELASA[60]. The experiments discussed in this paper were approved by the xxxxxxxx.

Reagents

- DNase-free water (Thermo Fisher Scientific, cat. no. 10977015)
- DNeasy Blood & Tissue Kit (Qiagen, cat. no. 69506) **▲ CRITICAL** We generally recommend this kit for the purification of genomic DNA. Yet the use of comparable genomic DNA purification kits or protocols would probably yield sequencing results similar to those shown here.
- Ethanol (absolute; Carl Roth, cat. no. 9065.2) **! CAUTION** Ethanol is flammable; use it while wearing appropriate personal protective equipment.
- Mayer's hematoxylin solution (Sigma-Aldrich, cat. no. MHS16-500ML)
- MinElute Reaction Cleanup Kit (Qiagen, cat. no. 28204)
- PBS (pH 7.4; Thermo Fisher Scientific, cat. no. 10010023)
- Proteinase K (Qiagen, cat. no. 19131)
- Qubit dsDNA BR Assay Kit (Thermo Fisher Scientific, cat. no. Q32850)
- Qubit dsDNA HS Assay Kit (Thermo Fisher Scientific, cat. no. Q32851)
- RNAlater (Thermo Fisher, cat. no. AM7020)
- Roti-Histofix (4%; Carl Roth, cat. no. P087) **! CAUTION** Avoid direct exposure; use under a fume hood.
- Xylene (Sigma-Aldrich, cat. no. 534056-500ML-D) **! CAUTION** Xylene is flammable, toxic upon inhalation, and a skin irritant. Use under a fume hood and wear appropriate safety equipment.
- HiSeq 3000/4000 PE Cluster kit (Illumina, cat. no. PE-410-1001)
- HiSeq 3000/4000 SBS kit (300 cycles; Illumina, cat. no. FC-410-1003)
- HiSeq X Ten Reagent Kit v2.5 (Illumina, cat. no. FC-501-2501)
- ddH_2O
- Paraffin

Equipment

Necropsy

- SafeLock Eppendorf tube (1.5 ml; Eppendorf, cat. no. 0030 108.051)
- Omnifix-F syringe (Braun, cat. no. 9161406V)
- Sterican needle, (18 gauge; Braun, cat. no. 4667123)
- Disposable scalpel (Swann-Morton)
- Surgical scissors (Fine Science Tools)
- Surgical forceps (Fine Science Tools) **! CAUTION** Handle with care; dispose of forceps in sharp containers.
- Tweezers
- Microscope (Carl Zeiss)
- Stereomicroscope (Carl Zeiss, model no. Stemi 508)
- Camera (Nikon, model no. D3400, cat. no. VBA490K001)
- Shaking heat block (Eppendorf, cat. no. 5383000019)
- Microtome (Thermo Fisher Scientific, cat. no. 902100)
- Centrifuge (Eppendorf, cat. no. 5401000010)
- Qubit fluorometer (Thermo Fisher Scientific, cat. no. Q33226) **▲ CRITICAL** Fluorescent dyes specific for dsDNA (e.g., Qubit dsDNA BR Assay Kit) do not overestimate DNA concentration and are therefore the method of choice for quantification.

Library preparation and sequencing

- 2100 Bioanalyzer instrument (Agilent, cat. no. G2939BA)
- Agilent DNA 1000 Kit (Agilent, cat. no. 5067-1504)
- Agilent SureSelect^{XT} Mouse All Exon kit (Agilent, cat. no. 5190-4641)
- TruSeq Nano DNA Low Throughput Library Prep Kit (Illumina, cat. no. 20015964) **▲ CRITICAL** The downstream analysis of the raw sequencing data shown here are optimized for the Illumina sequencing platform.
- cBot 2 instrument (Illumina, cat. no. SY-312-2001)
- HiSeq 4000 instrument (Illumina, cat. no. SY-401-4001)
- HiSeq 3000/4000 PE Cluster kit (Illumina, cat. no. PE-410-1001)
- HiSeq 3000/4000 SBS kit (300 cycles; Illumina, cat. no. FC-410-1003)
- HiSeq X Instrument (Illumina, cat. no. SY-412-1001)
- HiSeq X Ten Reagent kit v2.5 (Illumina, cat. no. FC-501-2501)

Hardware needed for data processing and analysis

- A workstation or computer cluster running a POSIX system (Unix, Linux or macOS) **▲ CRITICAL** Critical factors limiting the throughput of the pipeline are available RAM, number and performance of CPU threads, and speed of disk storage. The minimum requirements are an 8-core processor (48-core processor is preferred), 32 GB of RAM (256 GB is preferred) and 250 GB of disk space (a solid-state drive is preferred). In general, running an appropriate number of samples in parallel substantially increases the total throughput of the pipeline (Table 3). A comparison of runtimes for different systems is listed in Supplementary Table 2 **▲ CRITICAL** Both academic providers such as the European Open Science Cloud (https://www.eosc-portal.eu) and commercial cloud computing providers such as Amazon Web Services (AWS; https://aws.amazon.com) or Google Cloud (https://cloud.google.com) can be used to run this pipeline online. An overview for one provider (AWS) in regard to runtimes and associated costs is provided in Supplementary Table 3 **▲ CRITICAL** 15 GB of disk storage is needed for reference files. While running the WES (100×) analysis, ~170 GB of disk storage is needed for temporary files. The complete results of each tumor-normal pair can use up to ~30 GB of disk storage. For 30× WGS, ~1,000 GB of disk storage is needed for temporary files, whereas the results use ~300 GB of disk storage.

Software

! CAUTION When updating software tools, cross-compare results to older versions using test data.

- Docker (https://www.docker.com/) **▲ CRITICAL** Docker allows for packaging of software, including all dependencies, in containers. These containers can be run on most operating systems, including Windows, MacOS and Linux distributions. The Docker image provided online (see 'Equipment setup' section) includes the majority of tools listed below and makes separate installation of specific tools unnecessary. Versions of tools are listed as they are packaged in the Docker container.

Table 3 | Processing time for a cohort of 16 WES samples using different batch sizes

Samples running in parallel	Runtime per set of samples (h:min)	Total runtime for a cohort of 16 samples (h:min)
1	15:50	253:00
2	17:20	138:40
4	21:15	85:00
8	22:30	45:00

Matched tumor–normal data derived from WES for sample S821 were used. The pipeline was run on a Linux workstation, using 48 CPU threads, 256 GB of RAM and 2 TB of SSD storage. All steps were run sequentially.

- BAM-matcher (https://bitbucket.org/sacgf/bam-matcher)
- bcl2fastq v.2.20 (https://support.illumina.com/downloads/bcl2fastq-conversion-software-v2-20.html)
- bwa-mem v.0.7.17 (http://bio-bwa.sourceforge.net)
- bcftools v.1.9 (https://www.htslib.org)
- bedtools v.2.28.0 (https://github.com/arq5x/bedtools2)
- Delly2 v.0.8.1 (https://github.com/dellytools/delly)
- DNACopy v.1.57.0 (https://bioconductor.org/packages/release/bioc/html/DNAcopy.html)
- CNVKit v.0.9.6 (https://github.com/etal/cnvkit)
- CopywriteR v.2.15.2 (https://github.com/PeeperLab/CopywriteR)
- Fasta-to-Fastq (https://github.com/ekg/fasta-to-fastq)
- fastQC v.0.11.8 (https://www.bioinformatics.babraham.ac.uk/projects/fastqc)
- GATK v.4.1.2.0 (https://software.broadinstitute.org/gatk)
- GATK v.3.8.1.0 (https://software.broadinstitute.org/gatk)
- HMMCopy v.1.25.0 (http://bioconductor.org/packages/release/bioc/html/HMMcopy.html)
- HMMCopy Utils (https://github.com/shahcompbio/hmmcopy_utils)
- htslib v.1.9 (https://www.htslib.org)
- IGV v.2.4.16 (http://software.broadinstitute.org/software/igv)
- Java v.1.8 (https://java.com/download)
- Manta v.1.6.0 (https://github.com/Illumina/manta)
- MultiQC v.1.7 (https://github.com/ewels/MultiQC)
- msisensor v.0.5 (https://github.com/ding-lab/msisensor)
- Picard v.2.20.0 (https://broadinstitute.github.io/picard)
- R v.3.6.1 (https://www.r-project.org)
- samtools v.1.9 (https://www.htslib.org)
- SnpEff v.4.3T (http://snpeff.sourceforge.net)
- Strelka v.29.10 (https://github.com/Illumina/strelka)
- TitanCNA v.1.21.2 (https://github.com/gavinha/TitanCNA)
- Trimmomatic v.0.39 (http://www.usadellab.org/cms/index.php?page=trimmomatic)
- VCFtools v.0.1.16 (https://github.com/vcftools/vcftools)
- vcf2maf v.1.6.17 (https://github.com/mskcc/vcf2maf)
- VEP v.96 (https://github.com/Ensembl/ensembl-vep)
- (Optional) GISTIC2 v.2.0.23 (ftp://ftp.broadinstitute.org/pub/GISTIC2.0)
- (Optional) MuSiC2 v.0.2 (https://github.com/ding-lab/MuSiC2)
- (Optional) SomaticSignatures v.2.20.0 (http://bioconductor.org/packages/release/bioc/html/SomaticSignatures.html)

Equipment setup

▲ **CRITICAL** We recommend using the containerized version of this pipeline; you can either build it directly from a Dockerfile (available online: https://github.com/roland-rad-lab/MoCaSeq/blob/master/Dockerfile) or download an already-built version (detailed below). The steps detailed in this protocol can be used from inside the docker container, invoking BASH commands and executing scripts (Steps 6–48). This is very flexible; however, it can be cumbersome when processing large numbers of samples. Using the functionality provided by the Docker container to execute a scripted version of this pipeline greatly simplifies processing of a larger number of samples and increases reproducibility.

Initial setup

Define the location of a working directory and save it to a variable called `working_directory`, which will be used for testing and to hold downloaded reference files. Use a volume with at least 250 GB of free disk space.

```
working_directory=/PATH/TO/WORKING_DIRECTORY
```

Next, create the directory:

```
mkdir -p ${working_directory} \
&& cd ${working_directory}
```

Download and unzip the analysis workflow, available at https://github.com/roland-rad-lab/MoCaSeq:

```
wget https://github.com/roland-rad-lab/MoCaSeq/archive/master.zip \
&& unzip master.zip \
&& rm master.zip \
&& mv MoCaSeq-master ${working_directory}/MoCaSeq
```

Download the Docker image, available at https://cloud.docker.com/repository/docker/rolandradlab/mocaseq:

```
sudo docker pull rolandradlab/mocaseq:latest
```

The version of the pipeline can be specified by replacing `latest` with the corresponding release tag (listed at https://github.com/roland-rad-lab/MoCaSeq/releases). **! CAUTION** When downloading and unzipping the analysis workflow, do not change any file names or paths inside the downloaded folder.

Reference files

This pipeline requires several reference files, some of which can be downloaded directly, whereas others need to be generated before the first run of the pipeline. To facilitate these steps, the pipeline automatically checks whether the reference files have already been downloaded, and if not, will prepare them. For this, start the pipeline in test mode:

```
sudo docker run \
-v ${working_directory}:/var/pipeline/ \
rolandradlab/mocaseq:latest \
--test yes
```

A folder containing the necessary reference files (`ref`) will be created inside the current working directory. **! CAUTION** Owing to limits in downloading speed and computing-intensive steps during the generation of reference files needed for HMMCopy, this step can take up to 20 h. **▲ CRITICAL** To ensure comparability, make sure to use the same reference files for each sample of the experimental cohort.

Example dataset

We use exemplary sequencing data from a primary pancreatic cancer cell culture (sample S821), for which both WES (100× coverage) and WGS (30× coverage) are available. The raw data are available from the European Nucleotide Archive (https://www.ebi.ac.uk/ena) using the run accession numbers ERR2230828 (WES Tumor), ERR2230866 (WES Normal), ERR2210078 (WGS Tumor) and ERR2210079 (WGS Normal). A script, located in the repository folder, is provided to easily download data, resulting in eight files (WES and WGS data, separated into reverse and forward reads for the tumor and normal sample):

```
mkdir -p ${working_directory}/raw \
&& cd ${working_directory}/raw \
```

```
&& sh ${working_directory}/MoCaSeq/repository/Preparation_GetExemplary
Data.sh WES \
&& cd ${working_directory}
```

! CAUTION This step requires 100 GB of disk space.

Use of the Docker container

To illustrate the use of the Docker container, the following command will process the WES FASTQ files for Sample S821, a murine primary pancreatic cancer cell culture. This will automatically run through Steps 6–36 (for WES) of the Procedure. Additional options are displayed when the container is run without any options. Replace `<threads>` and `<RAM>` with appropriate values for your machine and then start the pipeline as follows:

```
sudo docker run \
-e USERID='id -u' -e GRPID='id -g' \
-v ${working_directory}:/var/pipeline/ \
rolandradlab/mocaseq:latest \
-nf '/var/pipeline/raw/S821-WES.Normal.R1.fastq.gz' \
-nr '/var/pipeline/raw/S821-WES.Normal.R2.fastq.gz' \
-tf '/var/pipeline/raw/S821-WES.Tumor.R1.fastq.gz' \
-tr '/var/pipeline/raw/S821-WES.Tumor.R2.fastq.gz' \
--name S821-WES \
--sequencing_type WES \
--threads <threads>\
--RAM <RAM>\
--artefact GT
```

Docker by design runs the container and its contents as user root (UID 1 and GID 1). Persistent directories mounted into the container with the option `-v` therefore are owned by root. Because this is often undesirable, the UID and GID of the current user can be passed into the container by specifying `-e USERID='id -u' -e GRPID='id -g'`.

By default, Docker containers cannot access files located on the machine on which they run. Therefore, local folders need to be mapped to folders inside the container using `-v local_folder:container_folder`. Importantly, the pipeline requires that the working directory be mapped to `/var/pipeline/`, the directory containing the reference folder (GRCm38) be mapped to `/var/pipeline/ref/` and the folder for temporary data be mapped to `/var/pipeline/temp/`. By default, both `ref` and `temp` are located inside the working directory.

Procedure

Sample collection ● Timing variable

1 Carefully extract the tumor from the euthanized mouse using surgical equipment. Remove excessive non-cancer tissue from the primary tumor. The aid of a dissection microscope is recommended. For large primary tumors (>0.5 cm), we perform regional sampling to extract material for use in Step 2. Cut out a central cross-section (>2.5 mm) from the tumor. To avoid degradation, immediately place the cross-section into a histology cassette and fixative for use in Step 2B. Both remaining tumor ends can be collected in RNAlater for direct DNA isolation (Step 2A) and/or used for the establishment of primary cultures (Step 2C). Check whether the primary tumor contains macroscopic heterogeneous regions, because this could point toward different cancer clones and should be addressed during regional sampling. For smaller primary tumors (<0.5 cm), isolate one part of the tumor in RNAlater (Step 2A) and (optionally) additional parts for histology (Step 2B) and/or for the establishment of primary cultures (Step 2C). For each tissue part, a diameter of at least 2.5 mm is recommended for Steps 2A and 2C. Metastatic lesions can be processed like primary cancer tissues; however, sample collection depends on the size, location and number of metastases. Count and describe the number of macroscopic metastatic nodes. Use scissors to cut a reference sample (>0.5 cm) from the tail and store it in RNAlater.
▲ CRITICAL STEP For each cancer sample, immediately continue with the corresponding extraction option in Step 2.

▲ CRITICAL STEP We recommend storing sample material in RNAlater because it greatly simplifies sample handling (e.g., no need to process samples immediately, no snap-freezing in liquid nitrogen necessary); preserves high-quality DNA and RNA, even though frequent freeze–thaw cycles; and yields reliable NGS data (e.g., in contrast to formalin fixation, which causes DNA sequence artifacts). Of course, conventional snap-freezing in liquid nitrogen is a good alternative for the isolation of high-quality DNA/RNA.

▲ CRITICAL STEP Samples must be stored in at least five volumes of RNAlater (e.g., 200 mg of tissue in 1 ml of RNAlater solution) and must be completely immersed. Use a SafeLock microcentrifuge tube because this prevents unintentional opening of the tubes during storage.

DNA extraction

2 Extract DNA, using option A for tissue stored in RNAlater, option B for microdissected FFPE-fixed material, or option C for cultured cells

(A) **Extract DNA from RNAlater tissue** ● Timing **1–2 d**

(i) Incubate the samples overnight (>12 h) at 4 °C. Subsequently, transfer the samples to −20 °C for long-term storage.

■ PAUSE POINT Tissue in RNAlater can be stored permanently at −20 °C. We recommend collecting all samples of a mouse cohort and continuing DNA extraction from this step.

(ii) For DNA extraction, remove tissue from the RNAlater solution; cut a sufficient, but not too large, piece (~25 mg of tumor tissue, ~0.5 cm of mouse tail) using a scalpel and tweezers in a clean Petri dish and chop it into fine pieces (<1 mm). Clean the instruments by washing in ddH_2O and 80% ethanol after each sample to avoid cross-contamination. Transfer the tissue to a 2-ml microcentrifuge tube and add 180 µl of ATL buffer (included in the Qiagen DNeasy Blood & Tissue Kit).

(iii) Add 20 µl of proteinase K solution and digest at 56 °C in a shaking heat block (~1,000 r.p.m.) until the tissue is completely lysed.

▲ CRITICAL STEP We strongly advise the use of fresh proteinase K or a stock solution stored at −20 °C, because proteinase K will degrade if improperly stored. Shaking in 2-ml tubes greatly enhances tissue disruption and speeds up lysis. If the sample is not lysed completely after 24 h, it was probably too large; in that case, we recommend adding another 180 µl of ATL buffer and 20 µl of proteinase K solution to complete the lysis. Take care to double the volumes of ATL buffer and ethanol in the next step as well and load the DNeasy mini spin column twice to bind all the DNA.

(iv) Proceed according to the Qiagen DNeasy Blood & Tissue Kit manufacturer's instructions.

(B) **Extract DNA from microdissected FFPE material** ● Timing **5–6 h**

▲ CRITICAL Depending on the size of the region of interest and cancer cell content, adjustments to the DNA extraction protocol are recommended for optimal results, as detailed below.

▲ CRITICAL Formalin covalently cross-links nucleic acids and proteins, and over-fixation can affect the integrity of the DNA; similarly, long-term or inappropriate storage can lead to DNA degradation; finally, carryover of organic solvents from de-paraffinization can also affect downstream reactions. Hence, we recommend using a DNA isolation procedure designed specifically for formalin-fixed sample material. The selection of an optimal protocol will produce amplifiable DNA and support sequencing quality.

(i) Remove the tissue from the tube containing fixative and embed the tissue in paraffin according to standard procedures[61].

■ PAUSE POINT FFPE material can be stored indefinitely protected from light at room temperature.

(ii) Cut the FFPE material into 10-µm-thick sections, mount specimens on a glass slide and air-dry the samples overnight at 37 °C as previously described[62].

▲ CRITICAL STEP Do not use sections <2 µm, because this will reduce the amount of extracted genomic DNA.

▲ CRITICAL STEP Overnight drying is recommended because the rehydration for microdissection can cause the whole specimen to detach from the glass slide.

(iii) Deparaffinize slides by immersion in fresh xylene twice for 10 min each.

! CAUTION Xylene is flammable, toxic when inhaled and a skin irritant. Use under a fume hood and wear appropriate safety equipment.

(iv) Rehydrate the slides by consecutive immersion in absolute ethanol (twice), 96% ethanol (twice) and 70% ethanol (once) for 2 min each.
! CAUTION Ethanol is highly flammable. Use under a fume hood and take appropriate care.

(v) Briefly submerge the slides in water and then stain them with Mayer's hematoxylin solution for 30–60 s. Wash the slides to remove excess staining solution.

(vi) Keep the specimen wet during the microdissection procedure. Use a microscope for magnification and scratch around the region of interest with a clean cannula. Either use the tip of the cannula to place the specimen into a Safe-Lock microcentrifuge tube pre-filled with ATL buffer (included in the Qiagen DNeasy Blood & Tissue Kit) or carefully pipette 20 µl of ATL buffer onto the region of interest, aspirate the material into the pipette tip and release the sample into a SafeLock microcentrifuge tube.
▲ CRITICAL STEP Avoid contamination of your cancer specimen with healthy wild-type surrounding tissue because this will affect downstream analyses.

(vii) Fill the sample tube to 180 µl with ATL buffer and add 20 µl of fresh proteinase K. Incubate at 56 °C and ~1,000 r.p.m. for 3 h in a shaking heat block.

(viii) Incubate specimens for 1 h at 90 °C without shaking to reverse cross-linking of DNA.

(ix) Proceed according to the Qiagen DNeasy Blood & Tissue Kit manufacturer's instructions. For very small sample amounts (<2-mm diameter), we recommend using QIAamp MinElute kit instead of DNeasy mini spin columns to yield higher DNA concentrations in a smaller elution volume.

(x) Finally, transfer the DNeasy mini spin column to a new Eppendorf LoBind microcentrifuge tube and add 100 µl of AE buffer (included in the Qiagen DNeasy Blood & Tissue Kit) to the center of the DNeasy mini spin column. Alternatively, for QIAamp MinElute spin columns, add 20–30 µl of AE buffer. Incubate for 3 min at room temperature. Elute by centrifugation at 8,000*g* at room temperature for 1 min.

(xi) Re-load the DNA-containing eluate onto the same DNeasy mini or QIAamp MinElute spin column and centrifuge at 8,000*g* at room temperature for 1 min. Store the DNA-containing eluate at −20 °C.

(C) **Extract DNA from cultured cells ● Timing 2 h**

(i) Apply cell isolation and culturing techniques appropriate to your cancerous tissue of interest.

(ii) Use a maximum of 5×10^6 cultured cells or frozen cells in 200 µl of PBS and extract DNA according to the Qiagen DNeasy Blood & Tissue Kit manufacturer's instructions.
▲ CRITICAL STEP Do not exceed the recommended cell numbers, because insufficient cell lysis will compromise DNA binding to the silica matrix of the DNeasy mini spin columns. If larger cell numbers need to be processed, scale up the buffer volumes accordingly and load the spin columns repeatedly.

DNA quantification ● Timing x x

3 Prepare the Qubit dsDNA BR fluorescent dye (from the Qubit dsDNA BR Assay Kit) in the reaction buffer according to the manufacturer's instructions and measure the DNA samples. Always perform a fresh standard curve for the measurement (included in the Qubit kit). Alternatively, for microdissected FFPE DNA samples, the Qubit dsDNA HS kit can be used.
▲ CRITICAL STEP Thaw the samples completely and vortex before pipetting. Vortex the samples as well as the standards after adding them to the Qubit buffer to ensure homogeneous fluorescence staining of the DNA.
▲ CRITICAL STEP UV spectrophotometry is precise and allows for the detection of contaminants such as protein or phenol. However, it relies on the wavelength-specific absorbance of nucleotides and therefore cannot distinguish between dsDNA, ssDNA, RNA and even dNTPs. This may lead to gross overestimation of DNA concentrations if the RNA is not removed during nucleic acid extraction; especially metabolically active tissues contain several times more mRNA than does genomic DNA. For these reasons, we strongly recommend quantification assays using dsDNA-specific fluorescent dyes.
▲ CRITICAL STEP Most sequencing facilities require ~1 µg of DNA at a concentration of at least 20 ng/µl for WES. For WGS, 250 ng of DNA is typically sufficient.
■ PAUSE POINT Purified DNA can be stored at −20 °C indefinitely.

Library preparation

4 Prepare DNA libraries for WES (option A) or WGS (option B) from extracted mouse genomic DNA.
▲ CRITICAL For cancer genome analyses, it is essential to prepare a separate library from the tail reference sample of each mouse.

(A) **Whole-exome library preparation** ● Timing **2 d**

(i) Prepare the exome DNA library from 1–2 µg of high-quality genomic DNA, using the Agilent SureSelect^{XT} Mouse All Exon kit according to the manufacturer's instructions.

(ii) Quantify individual sample libraries, using the 2100 Bioanalyzer in combination with an Agilent DNA 1000 Kit according to the manufacturer's instructions. Pool and quantify the final DNA library.
▲ CRITICAL STEP Quantify the pooled library to ensure optimal cluster density on the flow cell during the sequencing process.

(B) **Whole-genome library preparation** ● Timing **4–5 h**

(i) Prepare the whole-genome DNA library using the TruSeq Nano DNA Low Throughput Library Prep Kit from 250 ng of high-quality genomic DNA according to the manufacturer's instructions.

(ii) Quantify individual sample libraries using the 2100 Bioanalyzer in combination with the Agilent DNA 1000 Kit according to the manufacturer's instructions. Pool and quantify the final DNA library.
▲ CRITICAL STEP Quantify the pooled library to ensure optimal cluster density on the flow cell during the sequencing process.

Next-generation sequencing ● Timing Variable

5 Sequence libraries for WES by following option A and for WGS by following option B.

(A) **Whole-exome sequencing** ● Timing **2.5 d**

(i) Sequence the exome library on an Illumina HiSeq 4000 DNA sequencer in 2 × 100-bp paired-end sequencing mode to ~100× coverage with the HiSeq 4000 Reagent Kit according to the Illumina system guide. In general, 4–6 exomes per lane result in ~100× coverage per sample.

(B) **Whole-genome sequencing** ● Timing **3 d**

(i) Sequence the whole-genome DNA library on the Illumina HiSeq X in 2 × 150-bp paired-end sequencing mode to ~30× coverage with the HiSeq X Ten Reagent Kit according to the Illumina system guide. In general, one genome per lane results in ~30× coverage.

Bioinformatic analysis ● Timing 5 min

6 Steps 6–48 detail how the pipeline can be run manually, invoking BASH commands and executing scripts while running the Docker container in interactive mode. In this example, we will be using WES data for sample S821, a mouse primary pancreatic cancer cell culture sample. For this, choose option A if you have followed the steps detailed in the 'Equipment setup' section (creation of a working directory where the raw data and reference files are located); otherwise, choose option B to manually locate the working directory, the directories containing the reference and temporary files and the directory containing the raw data.

(A) **Interactive mode with default folders**

(i) Start the Docker container as follows:

```
sudo docker run \
-it --entrypoint=/bin/bash \
-e USERID='id -u' -e GRPID='id -g' \
-v ${working_directory}:/var/pipeline/ \
rolandradlab/mocaseq:latest
```

(B) **Interactive mode with custom folders**

(i) Map the local directories to the Docker container as follows:

```
working_directory=/PATH/TO/WORKING_DIRECTORY
reference_directory=/PATH/TO/REFERENCE_DIRECTORY
temp_directory=/PATH/TO/TEMP_DIRECTORY
rawdata_directory=/PATH/TO/RAWDATA_DIRECTORY
```

(ii) Start the Docker container as follows:

```
sudo docker run \
-it --entrypoint=/bin/bash \
-e USERID='id -u' -e GRPID='id -g' \
-v ${working_directory}:/var/pipeline/ \
-v ${reference_directory}:/var/pipeline/ref/ \
-v ${temp_directory}:/var/pipeline/temp/ \
-v ${rawdata_directory}:/var/pipeline/raw/ \
rolandradlab/mocaseq:latest
```

▲ **CRITICAL STEP** To improve reproducibility, use a MoCaSeq release tag (listed at https://github.com/roland-rad-lab/MoCaSeq/releases) instead of `latest`.

7 Now, define a set of basic variables that will be used throughout the pipeline. These include `name`, which identifies the samples and is prepended to all output files. Some specific aspects of this pipeline are different when used for WES versus WGS; therefore, define `sequencing_type` as either WES or WGS. Set `Species` to `Mouse`.

```
name=S821
sequencing_type=WES
species=Mouse
```

In the following commands, replace `<threads>` and `<RAM>` with values (in GB) appropriate for your machine

```
threads=<threads>
RAM=<RAM>
```

The configuration file is provided in the Docker image. Set `config_file` to the corresponding path and then load the configuration file using `source`.

```
config_file=/opt/MoCaSeq/config.sh
source $config_file
```

8 Now, create the folders for the output of the pipeline. Note that ~200 GB of data will be generated per pipeline run for WES, of which ~10 GB will be the raw read files, ~30 GB will be the results, including the mapped BAM files, and ~170 GB will be temporary files (located inside the `temp` folder), which can be deleted afterward.

```
mkdir -p $temp_dir/
mkdir -p $name/fastq/
mkdir -p $name/results/QC
mkdir -p $name/results/Genotype
mkdir -p $name/results/bam
mkdir -p $name/results/Mutect2
mkdir -p $name/results/LOH
if [$sequencing_type = 'WES']; then
mkdir -p $name/results/Copywriter
elif [$sequencing_type = 'WGS']; then
mkdir -p $name/results/Delly
mkdir -p $name/results/HMMCopy
mkdir -p $name/results/Chromothripsis
fi
```

▲ **CRITICAL STEP** This workflow expects that the folder structure created in this step and all generated files are neither renamed nor moved to other folders until all steps have been completed.

Formatting of raw data ● Timing 5–30 min, depending on initial format

9 The standard input format for this pipeline consists of gzipped FASTQ files for the tumor and normal control samples, separated by read mate (2 × 2 files in total) and named accordingly, in which case follow option A. Although this is commonly the case, sometimes raw reads are provided in multiple FASTQ files or as unmapped BAM files, which are handled differently, in which case use option B or C, respectively.

(A) **Handling standard input format files**

(i) In the case that gzipped FASTQ files are provided, simply copy these files to the working directory as follows:

```
cp  raw/S821-WES.Normal.R1.fastq.gz  $name/fastq/$name.Normal.
R1.fastq.gz
cp  raw/S821-WES.Normal.R2.fastq.gz  $name/fastq/$name.Normal.
R2.fastq.gz
cp raw/S821-WES.Tumor.R1.fastq.gz $name/fastq/$name.Tumor.R1.
fastq.gz
cp raw/S821-WES.Tumor.R2.fastq.gz $name/fastq/$name.Tumor.R2.
fastq.gz
```

(B) **Handling multiple FASTQ files**

(i) In some cases, raw reads are made available in multiple FASTQ files (e.g., when sequenced on multiple lanes). Simply merge these files using `cat` as follows:

```
cat  $name.Normal.Lane_1.R1.fastq.gz  $name.Normal.Lane_2.R1.
fastq.gz \
> $name/fastq/$name.Normal.R1.fastq.gz
```

(C) **Handling unmapped BAM files**

(i) If unmapped BAM files are provided, convert these to the FASTQ format beforehand as follows:

```
for type in Normal Tumor;
do
java -jar $picard_dir/picard.jar SamToFastq \
INPUT=$name.$type.bam \
FASTQ=$name/fastq/$name.$type.R1.fastq.gz \
SECOND_END_FASTQ=$name/fastq/$name.$type.R2.fastq.gz \
INCLUDE_NON_PF_READS=true VALIDATION_STRINGENCY=LENIENT
done
```

Quality control of raw data before trimming ● Timing 5 min

10 Generate basic quality checks of the raw data, using `fastqc` as follows. These, together with quality checks generated after trimming (Step 12), will be used in Step 18 for the evaluation of this run.

```
fastqc -t $threads \
$name/fastq/$name.Normal.R1.fastq.gz \
$name/fastq/$name.Normal.R2.fastq.gz \
$name/fastq/$name.Tumor.R1.fastq.gz \
$name/fastq/$name.Tumor.R2.fastq.gz \
--outdir=$name/results/QC
```

Read trimming ● Timing 10 min

11 Use `Trimmomatic` to discard short reads and reads with insufficient base qualities as follows. Depending on the version of the Illumina machine software used for sequencing the samples, different phred-scales (encoding the probability of an incorrectly sequenced base in the form of an ASCII character) are used to annotate base quality. Most modern sequencers provide quality information using Phred33, whereas older sequencers may use Phred64. This information is provided by the sequencing provider or can be found in the `fastqc` output generated in the

previous step (phred64 for Sanger/Illumina Encoding for versions 1.3 to 1.7; otherwise, phred33). Here, a custom script is implemented to automatically extract the phred-scale.

```
for type in Normal Tumor;
do
phred=$(sh $repository_dir/all_DeterminePhred.sh $name $type)
trimmomatic_file=$(basename $trimmomatic_dir)
java -Xmx${RAM}G -jar $trimmomatic_dir"/"$trimmomatic_file".jar" PE \
-threads $threads -$phred \
$name/fastq/$name.$type.R1.fastq.gz \
$name/fastq/$name.$type.R2.fastq.gz \
$temp_dir/$name.$type.R1.passed.fastq \
$temp_dir/$name.$type.R1.not_passed.fastq \
$temp_dir/$name.$type.R2.passed.fastq \
$temp_dir/$name.$type.R2.not_passed.fastq \
LEADING:25 TRAILING:25 MINLEN:50 \
SLIDINGWINDOW:10:25 \
ILLUMINACLIP:$trimmomatic_dir/adapters/TruSeq3-PE-2.fa:2:30:10
done
```

Quality control of raw reads after trimming ● Timing 5 min

12 Run `fastqc` to collect quality data using the trimmed reads as follows. The data from both pre- (Step 9) and post-trimming can be summarized using `multiqc` (Step 19). In this example, between 60 million and 70 million reads are available for the analysis of this sample. Inspect the distribution of quality scores of the raw reads, before and after trimming. The distribution of quality scores should be very similar for all reads.

```
fastqc -t $threads \
$temp_dir/$name.Normal.R1.passed.fastq \
$temp_dir/$name.Normal.R2.passed.fastq \
$temp_dir/$name.Tumor.R1.passed.fastq \
$temp_dir/$name.Tumor.R2.passed.fastq \
--outdir=$name/results/QC
```

Alignment to reference genome ● Timing 15 min

13 Align the trimmed reads to the reference genome as follows. The index files required by BWA need to be generated separately (see 'Equipment setup' section) to include alternative contigs, which are currently not placed on the auto- and allosomes.

```
for type in Normal Tumor;
do
bwa mem -t $threads $genomeindex_dir \
-Y -K 10000000 -v 1 \
$temp_dir/$name.$type.R1.passed.fastq \
$temp_dir/$name.$type.R2.passed.fastq \
> $temp_dir/$name.$type.sam
done
```

14 In each of following steps, processed files will be deleted once they are not needed anymore. Remove the trimmed raw files as follows:

```
for type in Normal Tumor;
do
rm $temp_dir/$name.$type.R1.passed.fastq
rm $temp_dir/$name.$type.R1.not_passed.fastq
rm $temp_dir/$name.$type.R2.passed.fastq
```

```
rm $temp_dir/$name.$type.R2.not_passed.fastq
done
```

Postprocessing of aligned reads ● Timing 2 h

15 Several postprocessing steps are required to prepare the files for use during SNV, LOH and CNV calling. Use `CleanSam` to provide information on soft-clipped reads, which are only partly aligned to the reference genome. Next, sort these files using `samtools`. Use `Picard Readgroups` to mark reads that have been sequenced together, as follows. Then duplicate reads (which possibly are PCR duplicates) are marked, which enables downstream tools to evaluate these reads differently.

```
for type in Normal Tumor;
do
MAX_RECORDS_IN_RAM=$(expr $RAM \* 250000)
java -Xmx${RAM}G -Dpicard.useLegacyParser=false \
-jar $picard_dir/picard.jar CleanSam \
-INPUT $temp_dir/$name.$type.sam \
-OUTPUT $temp_dir/$name.$type.cleaned.bam \
-VALIDATION_STRINGENCY LENIENT
rm $temp_dir/$name.$type.sam
samtools sort -@ $threads \
$temp_dir/$name.$type.cleaned.bam \
-o $temp_dir/$name.$type.cleaned.sorted.bam
rm $temp_dir/$name.$type.cleaned.bam
java -Xmx${RAM}G -Dpicard.useLegacyParser=false \
-jar $picard_dir/picard.jar AddOrReplaceReadGroups \
-I $temp_dir/$name.$type.cleaned.sorted.bam \
-O $temp_dir/$name.$type.cleaned.sorted.readgroups.bam \
-ID 1 -LB Lib1-Control -PL ILLUMINA -PU Run1 -SM $type \
-MAX_RECORDS_IN_RAM $MAX_RECORDS_IN_RAM
rm $temp_dir/$name.$type.cleaned.sorted.bam
java -Xmx${RAM}G -Dpicard.useLegacyParser=false \
-jar $picard_dir/picard.jar MarkDuplicates \
-INPUT $temp_dir/$name.$type.cleaned.sorted.readgroups.bam \
-OUTPUT $temp_dir/$name.$type.cleaned.sorted.readgroups.marked.bam \
-METRICS_FILE $name/results/QC/$name.$type.duplicate_metrics.txt \
-REMOVE_DUPLICATES false -ASSUME_SORTED true \
-VALIDATION_STRINGENCY LENIENT \
-MAX_RECORDS_IN_RAM $MAX_RECORDS_IN_RAM
rm $temp_dir/$name.$type.cleaned.sorted.readgroups.bam
done
```

Base recalibration ● Timing 2.5 h

16 Systematic errors, which can affect the base quality scores, can be introduced during sequencing. Therefore, recalibrate these scores in the final step of postprocessing as follows. Importantly, this step requires a VCF file of known germline variants, which should have been generated during the initial Equipment setup

```
(MGP.v5.snp_and_indels.exclude_wild.vcf.gz)
for type in Normal Tumor;
do
java -Xmx${RAM}G -jar $GATK_dir/gatk.jar BaseRecalibrator \
-R $genome_file \
-I $temp_dir/$name.$type.cleaned.sorted.readgroups.marked.bam \
--known-sites $snp_file \
--use-original-qualities \
-O $name/results/QC/$name.$type.GATK4.pre.recal.table
java -Xmx${RAM}G -jar $GATK_dir/gatk.jar ApplyBQSR \
```

```
-R $genome_file \
-I $temp_dir/$name.$type.cleaned.sorted.readgroups.marked.bam \
-O $name/results/bam/$name.$type.bam \
-bqsr $name/results/QC/$name.$type.GATK4.pre.recal.table
rm $temp_dir/$name.$type.cleaned.sorted.readgroups.marked.bam
java -Xmx${RAM}G -jar $GATK_dir/gatk.jar BaseRecalibrator \
-R $genome_file \
-I $name/results/bam/$name.$type.bam \
--known-sites $snp_file \
--use-original-qualities \
-O $name/results/QC/$name.$type.GATK4.post.recal.table
samtools index -@ $threads $name/results/bam/$name.$type.bam
rm $name/results/bam/$name.$type.bai
done
```

Quality control of the alignments ● Timing 30 min

17 Generate multiple quality controls for evaluating the mapped reads, which, together with other metrics, are summarized in Step 19, as follows.

```
for type in Normal Tumor;
do
java -Xmx${RAM}G -Dpicard.useLegacyParser=false \
-jar $picard_dir/picard.jar CollectSequencingArtifactMetrics \
-R $genome_file \
-I $name/results/bam/$name.$type.bam \
-O $name/results/QC/$name.$type.bam.artifacts
java -Xmx${RAM}G -Dpicard.useLegacyParser=false \
-jar $picard_dir/picard.jar CollectMultipleMetrics \
-R $genome_file \
-I $name/results/bam/$name.$type.bam \
-O $name/results/QC/$name.$type.bam.metrics
samtools idxstats $name/results/bam/$name.$type.bam \
> $name/results/QC/$name.$type.bam.idxstats
done
```

18 Collect metrics for coverage calculation. Here, WES and WGS are handled separately using options A and B, respectively, mainly for correct estimation of sequencing depth at each nucleotide (coverage).

(A) **Whole-exome sequencing**

(i) Calculate metrics for WES, including sequencing coverage, as follows. The correct estimation of coverage depth requires information about the target (baited) regions used in the exome extraction kit (Step X). This information is provided by the manufacturer. Q54

```
for type in Normal Tumor;
do
java -Xmx${RAM}G -Dpicard.useLegacyParser=false \
-jar $picard_dir/picard.jar CollectHsMetrics \
-SAMPLE_SIZE 100000 \
-R $genome_file \
-I $name/results/bam/$name.$type.bam \
-O $name/results/QC/$name.$type.bam.metrics \
-BAIT_INTERVALS $interval_file \
-TARGET_INTERVALS $interval_file
done
```

▲ **CRITICAL STEP** The corresponding file for Agilent SureSelect^{XT} Mouse All Exon, which was used for exome extraction for this sample, was generated for an older reference

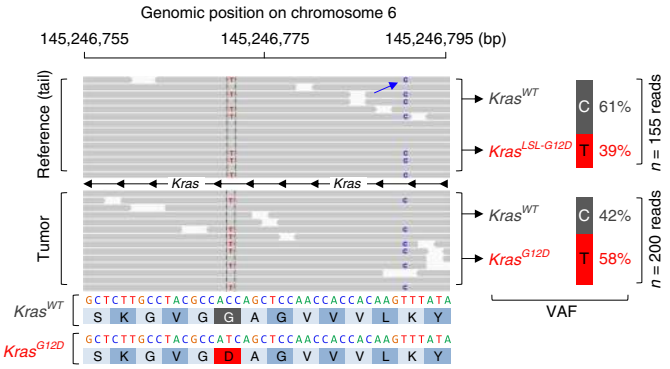


Fig. 13 | The mutant *Kras* allele is present in both tumor and matched normal tissue. Screenshot from IGV showing *Kras* exon 2 of mouse S821. The C>T mutation results in the $Kras^{G12D}$ allele. A T>C mutation 20 bp upstream was introduced during engineering of the locus. Approximately 50% of the reads (red T) carry the engineered mutation in both the tumor and normal matched tissue. VAF, variant allele frequency. WT, wild type.

genome (mm9). We generated the corresponding file for the current mouse reference genome, GRCm38; it is located in the `data` folder.

(B) **Whole-genome sequencing**

(i) Calculate metrics for WGS, including sequencing coverage, as follows.

```
for type in Normal Tumor;
do
java -Xmx${RAM}G -Dpicard.useLegacyParser=false \
-jar $picard_dir/picard.jar CollectWgsMetrics \
-R $genome_file \
-I $name/results/bam/$name.$type.bam \
-O $name/results/QC/$name.$type.bam.metrics \
-SAMPLE_SIZE 1000000
done.
```

? TROUBLESHOOTING

Visualize and check quality control metrics ● Timing 15 min

19 Use `multiqc` to summarize the output of all quality metric tools as follows.

```
multiqc $name/results/QC -n $name -o $name/results/QC/ --pdf
--interactive
```

Genotyping ● Timing 5 min

20 Some mouse models carry engineered mutations. Although specific polymerase chain reactions are typically used to determine these genotypes, in some cases they can be inferred from WES and WGS data. Here, we exemplarily detect the engineered *Kras* allele (see Fig. 13, one base pair change) using the following commands.

First, create a new TXT file containing the correct header line as follows:

```
echo -e 'Name\tAllele\tCHROM\tPOS\tREF\tALT\tCount_Ref\tCount_Alt
\tComment' \
> $name/results/Genotype/$name.Genotypes.txt
```

Next, define the specific genomic position for which the allele counts from the tumor and normal sample will be extracted:

```
allele=Kras-G12D
position=6:145246771-145246771
comment="GGT>GAT=G>D"
```

Then start the custom script, using the following command:

```
sh $repository_dir/SNV_GetGenotype.sh \
$name $allele $comment $config_file Mouse $position MS $types
```

21 In some genetically engineered mouse models, multiple exons and not single positions are affected. Here, we use the following commands to exemplarily test this for a *Trp53* knockout, in which, after recombination, exons 2–10 are lost:

```
allele=Trp53-fl
position=11:69580359-69591872
transcript=ENSMUST00000171247.7
wt_allele=1,11
del_allele=2,3,4,5,6,7,8,9,10
sh $repository_dir/CNV_GetGenotype.sh $name $position
Rscript $repository_dir/CNV_GetGenotype.R \
$name $genecode_file $transcript $allele $position $wt_allele
$del_allele
cat $name/results/Genotype/$name.Genotypes.temp.CNV.txt \
>> $name/results/Genotype/$name.Genotypes.txt
rm $name/results/Genotype/$name.Genotypes.temp.CNV.txt
```

SNVs and small indels ● Timing 5 h

22 Run `Mutect2` to call somatic point mutations and indels simultaneously and store the results as a VCF file as follows. Because the realignment step is directly implemented in Mutect2, we recommend storing the realigned reads in a separate BAM file (using `bamout`). This can be used to inspect callings afterwards that might not be explained by alignments generated in Step 13 of this protocol.

```
java -Xmx${RAM}G -jar $GATK_dir/gatk.jar Mutect2 \
--native-pair-hmm-threads $threads \
-R $genome_file \
-I $name/results/bam/$name.Tumor.bam \
-I $name/results/bam/$name.Normal.bam \
-tumor Tumor -normal Normal \
-O $name/results/Mutect2/$name.m2.vcf \
-bamout $name/results/Mutect2/$name.m2.bam
```

23 Use `FilterMutectCalls`, provided in the GATK package, to remove probable technical or germline artifacts as follows.

```
java -jar $GATK_dir/gatk.jar FilterMutectCalls \
--variant $name/results/Mutect2/"$name".m2.vcf \
--output $name/results/Mutect2/"$name".m2.filt.vcf \
--reference $genome_file
```

24 Filter Mutect2 calls for potential artifacts that arose through either oxidative DNA damage (G/T) during sample preparation or formaldehyde driven deamination of cytosines in FFPE samples (C/T) as follows. The tool makes use of the metrics file created in Step 17 of the protocol. Execute this step only if there is sufficient evidence that these samples are affected by one of these technical artifacts (using options B or C, respectively); otherwise, simply rename/copy the file (option A).

(A) **No artifact filtering**

(i) Rename/copy the files as follows:

```
cp $name/results/Mutect2/$name.m2.filt.vcf \
$name/results/Mutect2/$name.m2.filt.AM.vcf
cp $name/results/Mutect2/$name.m2.filt.AM.vcf \
$name/results/Mutect2/$name.m2.filt.AM.filtered.vcf
```

(B) **Filtering of FFPE artifacts**

(i) Filter FFPE artifacts as follows:

```
java -jar $GATK_dir/gatk.jar FilterByOrientationBias \
-V $name/results/Mutect2/$name.m2.filt.vcf -P \
$name/results/QC/$name.Tumor.bam.artifacts.pre_adapter_de-
tail_metrics \
--artifact-modes C/T --output $name/results/Mutect2/$name.
m2.filt.AM.vcf
cat $name/results/Mutect2/$name.m2.filt.AM.vcf \
| java -jar $snpeff_dir/SnpSift.jar filter \
"(((FILTER = 'PASS') & (exists GEN[Tumor].OBP) & \
(GEN[Tumor].OBP <= 0.05)) | ((FILTER = 'PASS')))" \
> $name/results/Mutect2/$name.m2.filt.AM.filtered.vcf
```

(C) **Filtering of oxidative DNA damage artifacts**

(i) Filter oxidative DNA damage artifacts as follows:

```
java -jar $GATK_dir/gatk.jar FilterByOrientationBias \
-V $name/results/Mutect2/$name.m2.filt.vcf -P \
$name/results/QC/$name.Tumor.bam.artifacts.pre_adapter_de-
tail_metrics \
--artifact-modes G/T --output $name/results/Mutect2/$name.
m2.filt.AM.vcf
cat $name/results/Mutect2/$name.m2.filt.AM.vcf \
| java -jar $snpeff_dir/SnpSift.jar filter \
"(((FILTER = 'PASS') & (exists GEN[Tumor].OBP) & \
(GEN[Tumor].OBP <= 0.05)) | ((FILTER = 'PASS')))" \
> $name/results/Mutect2/$name.m2.filt.AM.filtered.vcf
```

25 Use `SelectVariants` to filter out all indels >10 bp as follows:

```
java -jar $GATK_dir/gatk.jar SelectVariants --max-indel-size 10 \
-V $name/results/Mutect2/$name.m2.filt.AM.filtered.vcf \
-output $name/results/Mutect2/$name.m2.filt.AM.filtered.selected.vcf
```

26 Additional filters can be used to decrease the false-positive rate of reported mutations. We apply filters for mutant allele frequency (≥10%), coverage at particular positions in tumor and normal samples (≥10×) and supporting reads for mutation in tumor sample (at least two), as follows:

```
cat $name/results/Mutect2/$name.m2.filt.AM.filtered.selected.vcf \
| java -jar $snpeff_dir/SnpSift.jar filter \
"((FILTER = 'PASS') & (GEN[Tumor].AF >= 0.1) & \
((GEN[Tumor].AD[0] + GEN[Tumor].AD[1]) >= 10) & \
((GEN[Normal].AD[0] + GEN[Normal].AD[1]) >= 10) & \
(GEN[Tumor].AD[1] >= 3) & (GEN[Normal].AD[1] = 0))" \
> $name/results/Mutect2/$name.m2.postprocessed.vcf
```

27 To further reduce false-positive callings, compare the SNVs and indels to known polymorphisms as follows:

```
bgzip $name/results/Mutect2/$name.m2.postprocessed.vcf
tabix -p vcf $name/results/Mutect2/$name.m2.postprocessed.vcf.gz
bcftools isec -C -c none -O z -w 1 \
-o $name/results/Mutect2/$name.m2.postprocessed.snp_removed.vcf.gz \
$name/results/Mutect2/$name.m2.postprocessed.vcf.gz \
$alternate_snp_file
```

```
bcftools norm -m -any \
$name/results/Mutect2/$name.m2.postprocessed.snp_removed.vcf.gz \
-O z -o $name/results/Mutect2/$name.Mutect2.vcf.gz
gunzip -f $name/results/Mutect2/$name.Mutect2.vcf.gz.
```

? TROUBLESHOOTING

28 Use SNPeff to annotate the resulting set of SNVs and indels as follows:

```
java -Xmx${RAM}G -jar $snpeff_dir/snpEff.jar $snpeff_version -canon \
-csvStats $name/results/Mutect2/$name.Mutect2.annotated.vcf.stats \
$name/results/Mutect2/$name.Mutect2.vcf \
> $name/results/Mutect2/$name.Mutect2.annotated.vcf
```

29 To improve readability, split the effect of the same mutation on different transcripts into separate lines as follows:

```
cat $name/results/Mutect2/$name.Mutect2.annotated.vcf \
| $snpeff_dir/scripts/vcfEffOnePerLine.pl \
> $name/results/Mutect2/$name.Mutect2.annotated.one.vcf
```

30 Export the resulting file to a tab-separated TXT file as follows. The output format is explained in Box 1.

```
java -jar $snpeff_dir/SnpSift.jar extractFields \
$name/results/Mutect2/$name.Mutect2.annotated.one.vcf \
CHROM POS REF ALT "GEN[Tumor].AF" "GEN[Tumor].AD[0]" "GEN[Tumor].AD[1]" \
"GEN[Normal].AD[0]" "GEN[Normal].AD[1]" ANN[*].GENE ANN[*].EFFECT \
ANN[*].IMPACT ANN[*].FEATUREID ANN[*].HGVS_C ANN[*].HGVS_P \
> $name/results/Mutect2/$name.Mutect2.txt.
```

? TROUBLESHOOTING

31 Remove the intermediary files, if they are not needed for quality control, as follows:

```
sh $repository_dir/SNV_CleanUp.sh $name MS
```

Loss-of-heterozygosity ● Timing 4 h

32 As discussed in the 'Experimental design' section, use Mutect2 to extract positions for LOH analysis as follows:

```
for type in Normal Tumor;
do
java -Xmx${RAM}G -jar $GATK_dir/gatk.jar Mutect2 \
--native-pair-hmm-threads $threads \
-R $genome_file \
-I $name/results/bam/$name.$type.bam \
-tumor $type \
-O $name/results/Mutect2/$name."$type".m2.vcf \
-bamout $name/results/Mutect2/$name."$type".m2.bam
done
```

33 In Step 32, tumor and normal variants were called separately from the respective BAM files. Use the following commands to filter calls and extract positions that are evaluated for LOH plotting.

```
for type in Normal Tumor;
do
java -jar $GATK_dir/gatk.jar FilterMutectCalls \
--variant $name/results/Mutect2/$name.$type.m2.vcf \
```

Box 2 | Description of CNV output

The CNV section provides both a plot and an output file.
`Chrom` Chromosome name
`Start` Start position of the segment.
`End` End position of the segment.
`Mean` The mean log2 ratio between tumor and normal for the segment. In genomes of known ploidy, this can be converted to absolute copy-number change: ploidy × 2^{Mean}; e.g., $2 \times 2^{1.58} = 6$ (there are six copies of the affected region).

```
--output $name/results/Mutect2/$name.$type.m2.filt.vcf \
--reference $genome_file
java -jar $snpeff_dir/SnpSift.jar extractFields \
$name/results/Mutect2/"$name".$type.m2.filt.vcf \
CHROM POS REF ALT "GEN["$type"].AF" "GEN["$type"].AD[0]" \
"GEN["$type"].AD[1]" MMQ[1] MBQ[1] \
> $name/results/Mutect2/$name.$type.Mutect2.Positions.txt
done
```

34 Remove the intermediary files as follows:

```
for type in Normal Tumor;
do
sh $repository_dir/SNV_CleanUp.sh $name SS $type
done
```

35 Generate a list of variants to be used during the plotting procedure as follows. This custom script performs several steps sequentially; it filters out positions (i) with read coverage <10, (ii) with mapping quality <60, and (iii) that are potentially affected by strand artifacts. Positions that pass these filters in both the tumor and normal sample and for which the allele frequency is between 30 and 70% in the normal sample are used for plotting. Although the variant allele frequency can be used in LOH plots (Fig. 9a,b), we adapted the Illumina convention (https://www.illumina.com/documents/products/technotes/technote_topbot.pdf) of defining the A and B allele, which results in plots mirrored along the 0.5 axis (Fig. 9d)

```
Rscript $repository_dir/LOH_GenerateVariantTable.R \
$name $genome_dir/GRCm38.p6.fna $repository_dir
```

36 Plot the resulting list of heterozygous germline variants as follows:

```
Rscript $repository_dir/LOH_MakePlots.R \
$name $species $repository_dir
```

Copy-number variation

37 Use option A for WES data (to detect CNVs using CopywriteR) or option B for WGS data (to detect CNVs using HMMCopy). See Box 2 for an explanation of the output format. For the analysis of WES data, this step concludes the workflow.

(A) **CNVs from WES data** ● **Timing 1.5 h**

(i) Call CNVs from WES data using CopywriteR with 20-kB windows, as follows:

```
Rscript $repository_dir/CNV_RunCopywriter.R \
$name Mouse $threads MS $genome_dir $types
```

(ii) The called segments are located inside an Rdata object. Use the below command to extract the raw data generated by CopywriteR.

```
Rscript $repository_dir/CNV_CopywriterGetRawData.R $name MS
```

(iii) Re-center called segments using mode as location estimator as follows:

```
python $repository_dir/CNV_CopywriterGetModeCorrectionFactor.
py $name
Rscript $repository_dir/CNV_CopywriterGetModeCorrectionFactor.
R $name MS
```

(iv) Create CNV plots for both the MAD- and mode-centered segments as follows:

```
Rscript $repository_dir/CNV_PlotCopywriter.R $name Mouse
$repository_dir
```

(v) Extract exact copy-number state for each gene as follows:

```
Rscript $repository_dir/CNV_MapSegmentsToGenes.R $name Mouse
Copywriter
```

(vi) Clean up intermediary CNV files using the command below:

```
sh $repository_dir/CNV_CleanUp.sh $name
```

(B) **CNVs from WGS data** ● Timing **30 min**

(i) Run HMMCopy, using 20-kB windows, as follows:

```
sh $repository_dir/CNV_RunHMMCopy.sh $name Mouse $config_file
20000
```

(ii) Call segments and create CNV plots from the WIG file generated in the step above, as follows:

```
Rscript $repository_dir/CNV_PlotHMMCopy.R $name Mouse $repo-
sitory_dir
20000 $mapWig_file $gcWig_file $centromere_file $varregions_file
```

(iii) Extract exact copy-number state for each gene as follows:

```
Rscript $repository_dir/CNV_MapSegmentsToGenes.R $name Mouse
HMMCopy 20000
```

Structural variations and rearrangements ● Timing 2 h

38 Use the following command to call rearrangements using Delly:

```
delly call \
-o $name/results/Delly/$name.pre.bcf \
-g $genome_dir/GRCm38.p6.fna \
$name/results/bam/$name.Tumor.bam \
$name/results/bam/$name.Normal.bam
```

39 Filter the resulting structural variation calls as follows:

```
delly filter \
-f somatic -o $name/results/Delly/$name.bcf \
-s $genome_dir/Samples.tsv $name/results/Delly/$name.pre.bcf
```

40 Transform the Delly output to the VCF format, which is used in the chromothripsis workflow, as follows:

```
bcftools view $name/results/delly/$name.pre.bcf \
> $name/results/delly/$name.pre.vcf
```

Chromothripsis ● Timing 1 h

41 In our experience, Delly is very sensitive in detecting structural variations. In some cases, however, these variants are false positives. During benchmarking of these tests, we added several filter steps that are important to reducing these false-positive SV callings: (i) we exclude all short- and medium-length variants (<6 kb). (ii) Because it has been shown that chromothripsis happens very early during tumorigenesis, we exclude variants with allele frequencies <0.2. (iii) For all rearrangements not supported by 'split reads' (reads that span a specific breakpoint), we added additional filtering steps: Because highly repetitive regions are prone to false-positive callings from Delly, (i) we exclude all callings for which the read coverage significantly exceeds the mean coverage and (ii) the mapping quality score is <30.

Use the following commands to extract the mean coverage for each alignment file (using data generated in Step 18) and filter the Delly calls as explained above:

```
coverage=$(sh $repository_dir/Chromothripsis_GetCoverage.sh $name)
sh $repository_dir/Chromothripsis_FormatTable.sh $name
Rscript $repository_dir/Chromothripsis_AnnotateRatios.R \
-i $name/results/Delly/$name.breakpoints.tab \
 > $name/results/Delly/$name.breakpoints_annotated.tab
Rscript $repository_dir/Chromothripsis_FilterDelly.R \
-n $name -c $coverage \
-i $name/results/Delly/$name.breakpoints_annotated.tab
```

42 Using these results, as well as data from LOH and CNV calling, each hallmark of chromothripsis is tested separately on one chromosome at a time. For sample S821, the exemplary data used here, there is very strong suspicion that Chr4 was affected by chromothripsis. The output format (.tif or.emf) for all plots resulting from the chromothripsis workflow can be defined. Set up the test for Chr4 with the following commands:

```
chr=4
format="tif"
```

43 Test for the chromothripsis hallmark 'clustering of breakpoints' as follows. This results in Fig. 11a.

```
Rscript
$repository_dir/Chromothripsis_DetectBreakpointClustering.R \
-i $name/results/Delly/$name.breakpoints.filtered.tab \
-c $chr -n $name -f $format
```

44 Test for the chromothripsis hallmark 'regularity of oscillating copy-number states' as follows. This results in Fig. 11c.

```
Rscript $repository_dir/Chromothripsis_SimulateCopyNumberStates.R \
-i $name/results/Delly/$name.breakpoints.filtered.tab \
-o mouse -c $chr -n $name -s 1000 -a 1000 -f $format -v 1
```

45 Test for the chromothripsis hallmark 'interspersed loss and retention of heterozygosity' as follows. This results in Fig. 11b.

```
Rscript $repository_dir/Chromothripsis_PlotLOHPattern.R \
-s $name/results/HMMCopy/$name.HMMCopy.$resolution.segments.txt \
```

```
-d $name/results/HMMCopy/$name.HMMCopy.$resolution.log2RR.txt \
-v $name/results/LOH/$name.VariantsForLOH.txt \
-o mouse -c $chr -n $name -f $format
```

46 Test for the chromothripsis hallmark 'randomness of DNA fragment joins and segment order' as follows. This results in Fig. 11d,e.

```
Rscript $repository_dir/Chromothripsis_DetectRandomJoins.R \
-i $name/results/Delly/$name.breakpoints.filtered.tab \
-c $chr -n $name -f $format
```

47 Test for the chromothripsis hallmark 'ability to walk the derivative chromosome' as follows. This results in Fig. 11f.

```
Rscript $repository_dir/Chromothripsis_WalkDerivativeChromosome.R \
-i $name/results/Delly/$name.breakpoints.filtered.tab \
-c $chr -n $name -f $format
```

48 Visualize a combined rearrangement graph/copy-number plot using the following command. This results in Fig. 10a.

```
Rscript $repository_dir/Chromothripsis_PlotRearrangementGraph.R \
-i $name/results/Delly/$name.breakpoints.filtered.tab \
-d $name/results/HMMCopy/$name.HMMCopy.$resolution.log2RR.txt \
-c $chr -n $name -f $format
```

Troubleshooting

Troubleshooting advice can be found in Table 4.

Table 4 | Troubleshooting table

Step	Problem	Possible reason	Possible solution
18	Calculated sequencing coverage is lower than expected	Insert size is low. Picard discards reads for the calculation of sequencing coverage if forward and reverse reads overlap. This is most often a problem in FFPE-extracted DNA	Evaluate absolute number of mapped reads manually
27	Distribution of SNVs skewed to C>A/G>T	DNA was damaged through oxidative stress during library preparation	Filter artifacts in Step 23
30	Output contains an unexpectedly large number of SNVs	Tumor sample and normal sample are not from the same mouse	Use `BAM-matcher` (https://bitbucket.org/sacgf/bam-matcher) to check correct tumor-normal pairings for all animals (automatically included when running the Docker pipeline)

Timing

Timing estimates for the bioinformatic analysis are based on the analysis of one mouse cancer sample, using WES data (coverage ~100× for both tumor and normal) for Steps 6–37, and WGS data (coverage ~30× for both tumor and normal) for Steps 38–48. Table 3 shows runtime improvements when running multiple samples in parallel for a cohort of 16 matched WES tumor–normal pairs. Table 5 provides a comparison between runtimes for each step when using WES versus WGS. The hands-on time for Steps 6–48 is <10 min for either WES or WGS data.

In the analysis of WGS data, especially in very aneuploid tumors, somatic mutation calling using Mutect2 can be the limiting factor in overall throughput. The alternative use of Strelka2 can markedly improve the overall runtime (Table 5).

Table 5 | Comparison between runtimes for the analysis of WES and WGS.

Steps	WES runtime (h:min)	WGS runtime (h:min)	RAM (GB)
Trimming (Steps 9–12)	0:10	1:00	90
Alignment to the reference genome (Steps 13 and 14)	0:15	2:30	119
Postprocessing of aligned reads (Step 15)	1:45	8:00	201
Base recalibration (Step 16)	2:30	9:45	73
Quality control (Steps 17–19)	0:30	3:30	188
Genotyping (Steps 20 and 21)	0:05	0:05	4
SNV/indel (Mutect2, Steps 22–31)	4:45	19:00	90
SNV/indel (Strelka2, alternative for Steps 22–31)	0:20	1:00	47
LOH (Steps 32–36)	4:00	37:00	136
CNV (WES, Step 37)	1:30	—	66
CNV (WGS, Step 37)	—	0:30	45
SV (Steps 38–40)	—	2:00	9
Chromothripsis (Step 41–48)	—	1:00	22
Sum	15:50	85:20	—

Matched tumor-normal data derived from WES and WGS for sample S821 was used. The pipeline was run on a Linux workstation, using 48 CPU threads, 256 GB of RAM and 2 TB of SSD storage. RAM usage is similar for both WES and WGS but depends on the total capacity of available RAM. All steps were run sequentially.

Q62

Wet lab

Step 1, sample collection: variable
Step 2A, DNA extraction from tissue stored in RNAlater: 1–2 d
Step 2B, DNA extraction from microdissected FFPE material: 5–6 h
Step 2C, DNA extraction from cultured cells: 2 h
Step 3, DNA quantification: x x

Q64

Library preparation and sequencing

Step 4A, library preparation (WES): 2 d
Step 4B, library preparation (WGS): 4–5 h (WGS)
Step 5A, sequencing (WES): 2.5 d
Step 5B, sequencing (WGS): 3 d

Bioinformatic analysis

Steps 6–21, alignment and postprocessing: 6–6.5 h
Steps 22–31, SNV/indel calling: 5 h
Steps 32–36, LOH calling: 4 h
Step 37A, CNV calling (WES): 1.5 h
Step 37B, CNV calling (WGS): 30 min
Steps 38–40, SV calling (WGS only): 2 h
Steps 41–48, inference of chromothripsis (WGS only): 1 h

Anticipated results

Genetic alteration types and frequencies in an exemplary mouse cancer

Below we present results from the analysis of one individual cancer. The tumor was generated in a genetically engineered mouse model of pancreatic cancer (ID S821). The model is based on a heterozygous *Kras*$^{LSL\text{-}G12D}$ knock-in allele that was activated in a pancreas-specific manner using Cre recombination.

SNVs and indels

Step 30 of our protocol generates a list of 45 mutations (listed in Supplementary Table 4), of which 8 are mis- or nonsense mutations (listed in Table 6). An explanation of all columns in the file generated

Table 6 | Non-synonymous SNV calls for sample S821

Chrom	Pos	Ref	Alt	Allele freq.	Reads tumor (Ref)	Reads tumor (Alt)	Reads normal (Ref)	Read normal (Alt)	Gene	Effect	Impact	Transcript	HGVS_C	HGVS_P
2	13342476	C	A	0.132	44	6	59	0	*Cubn*	missense_variant	MODERATE	ENSMUST00000091436.6	c.6230G>T	p.Gly2077Val
2	85438826	C	A	0.11	54	6	102	0	*Olfr995*	missense_variant	MODERATE	ENSMUST00000099924.2	c.331G>T	p.Asp111Tyr
2	86690856	A	G	0.113	98	12	178	0	*Olfr1087*	missense_variant	MODERATE	ENSMUST00000099877.1	c.118T>C	p.Phe40Leu
3	96654785	C	A	0.114	52	6	48	0	*Itga10*	missense_variant	MODERATE	ENSMUST00000029744.5	c.1987C>A	p.Gln663Lys
5	25022007	C	T	0.27	66	24	92	0	*Prkag2*	missense_variant	MODERATE	ENSMUST00000030784.13	c.251G>A	p.Arg84Gln
10	70940567	C	A	0.107	56	6	73	0	*Bicc1*	missense_variant	MODERATE	ENSMUST00000143791.7	c.2301G>T	p.Lys767Asn
17	34034195	A	G	0.117	50	6	31	0	*Rxrb*	missense_variant	MODERATE	ENSMUST00000044858.14	c.775A>G	p.Arg259Gly
19	34021638	C	A	0.104	58	6	51	0	*Lipk*	missense_variant	MODERATE	ENSMUST00000054260.6	c.332C>A	p.Ala111Asp

Alt, variant (alternative) base; Chrom, chromosome; HGVS_C, nucleotide change; HGVS_P, amino acid change (for protein-coding genes); Pos, genomic position; Ref, reference base.

by the SNV/indel workflow is provided in Box 1. Note that the *Kras*$^{LSL\text{-}G12D}$ knock-in allele is a germline allele present in every cell, although it is expressed only in the pancreas (because of pancreas-specific recombination of the *LoxP*-flanked stop cassette). Thus, at the DNA level, the *Kras*$^{LSL\text{-}G12D}$ mutation is detectable in both the tumor and the control tissue. As a consequence, the final list of somatic tumor SNVs/indels will not contain this mutation. However, it can be visualized, for example, using IGV (Fig. 13) or can be extracted separately during the genotyping procedure in Steps 20 and 21.

CNV

In Step 37, a copy-number profile for the complete genome is generated (Fig. 14a). Multiple copy-number changes are located on Chr4 in an oscillating pattern very suggestive of chromothripsis (Figs. 5b,11 and 14c). Chr6, where *Kras* is located, is amplified (log2 ratio 0.42). Because sample S821 is tetraploid (as determined by M-FISH), this corresponds to five copies of Chr6. A table listing log2 ratios for all detected segments is also generated (Supplementary Table 5). An explanation of these columns is provided in Box 2.

LOH

The LOH plot in Fig. 14b is generated in Step 36. The animal from which this tumor originates is on a mixed background. However, several generations of backcrossing to C57BL/6 have been performed for this line. Therefore, not all regions of the genome can be adequately inspected for the occurrence of LOH.

Inference of chromothripsis

A comprehensive overview of the results generated by our chromothripsis pipeline (Steps 41–48) is shown in Fig. 11.

Integrative analyses of different alteration types affecting prototypic oncogenes and tumor suppressors

An individual genomic locus within a cancer cell can be affected by multiple alteration types. Integrative analysis of different alteration types affecting one locus is therefore essential for accurate interpretation of cancer genomic data. For example, tumor suppressors such as *Trp53* can be inactivated in multiple ways, either through somatic point mutations, larger copy-number changes or loss of wild-type alleles. Below we present exemplary data displaying common mechanisms of somatic alterations at prototypic tumor suppressors and oncogenes in mouse pancreatic ductal adenocarcinoma (Figs. 15 and 16). All tumors were derived in the abovementioned genetically engineered mouse model of *Kras*^{G12D} driven pancreatic cancer. Data were generated through WES of primary cancer cell cultures.

Allelic imbalance at the mutated Kras oncogene

A hallmark of pancreatic cancer evolution in humans and mice is allelic imbalance at the *Kras* locus, leading to *Kras*^{G12D} dosage gain (multiple copies of the mutant *Kras*^{G12D} allele). Examples of changes

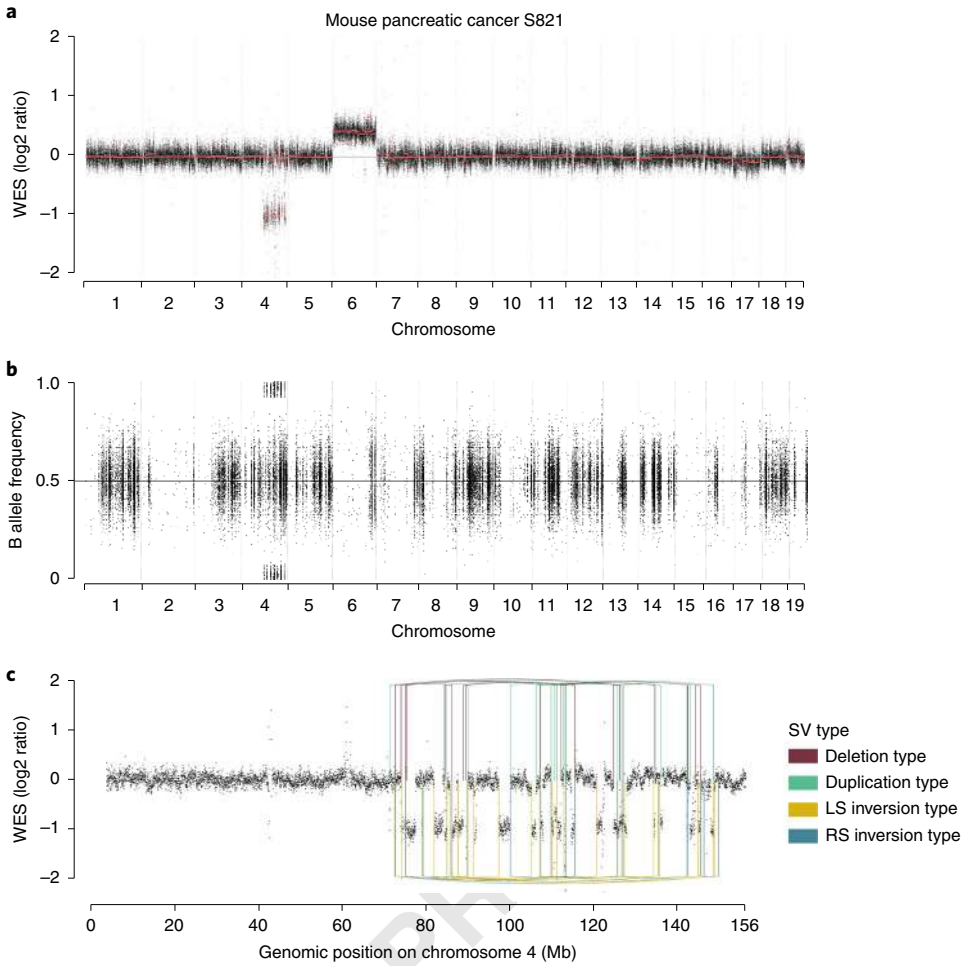


Fig. 14 | CNV and LOH profiles for sample S821. a,**b**, WES was used to infer copy-number (**a**) and loss-of-heterozygosity (**b**) profiles for mouse pancreatic cancer sample S821. Chr4 displays an oscillating pattern very suggestive of chromothripsis. Chr6, the location of *Kras*, is amplified. **c**, Rearrangement graph (from WGS) for the chromothriptic chromosome, Chr4. Fragments that are joined during chromothripsis are connected by lines superimposed on the copy-number profile.

affecting Chr6, the location of *Kras*, are shown for different cancers in Fig. 15. One cancer (Fig. 15a) displays arm-level gain of Chr6 (duplication of an entire chromosome; see CNV plot and M-FISH). The duplication affects the chromosome carrying the oncogenic $Kras^{G12D}$ point mutation, indicated by the elevated frequency of mutant $Kras^{G12D}$ reads (70% of reads are $Kras^{G12D}$ mutant; upper left panel). The LOH plot shows corresponding B allele frequency distribution peaks at 0.66 and 0.33 (lower right panel). Figure 15b shows a cancer with focal amplification (~6 copies) of the chromosomal region harboring the *Kras* locus. The amplification affects the chromosome carrying the mutant $Kras^{G12D}$ allele ($Kras^{G12D}$ and $Kras^{WT}$ allele frequencies are 89% and 11%, respectively). Because this region carries only a few heterozygous germline variants in this mouse, focal amplification cannot be easily seen in the LOH plot. Figure 15c and Fig. 15d show two cancers displaying $Kras^{G12D}$ dosage gain by copy-number-neutral (CN)-LOH ($Kras^{G12D}$ homozygosity, acquired uniparental disomy, loss of wild-type *Kras*). CN-LOH can affect either the whole chromosome (Fig. 15c; arising through chromosomal missegregation) or only parts of Chr6 (Fig. 15d; through mitotic recombination). Discriminating between these two scenarios is possible only through LOH analyses (bottom panels in Fig. 14c,d).

Alterations at prototype tumor suppressors

Examples of different types of tumor suppressor alterations are shown for *Trp53* (Fig. 16a) and *Cdkn2a* (Fig. 16b–d). One cancer has a somatic *Trp53* point mutation on Chr11 (Fig. 16a). Three

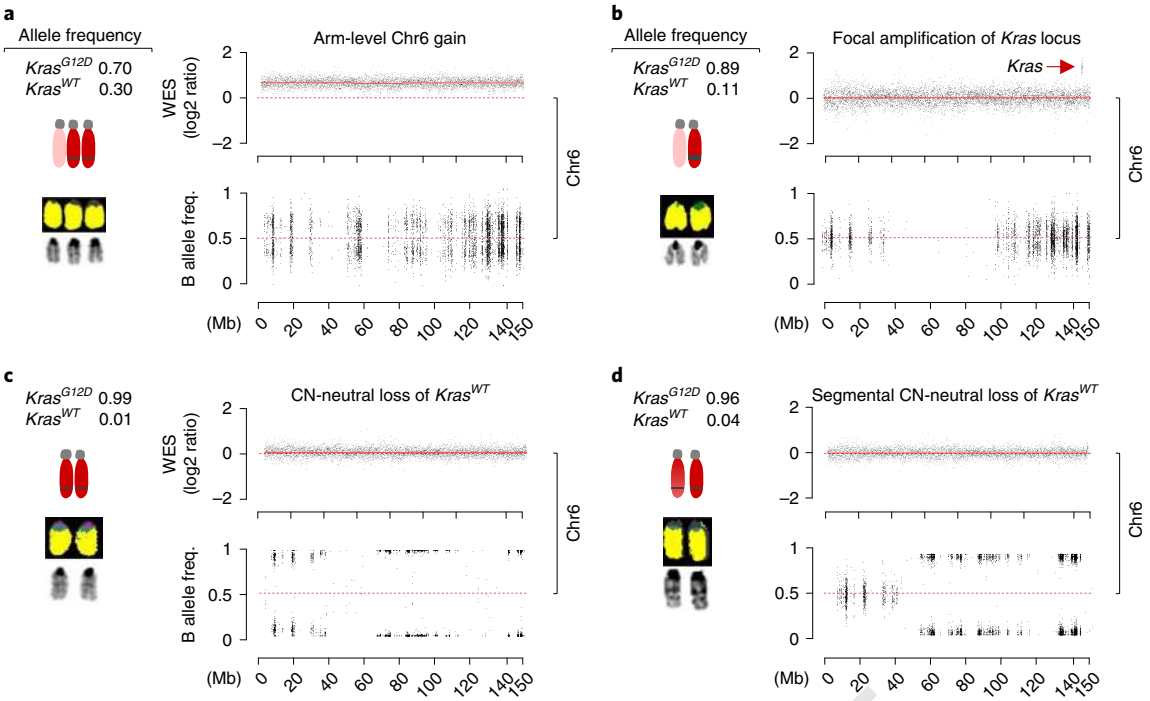


Fig. 15 | Patterns of genomic changes affecting oncogenes. a–d, Exemplary changes from a cohort of pancreatic cancer cell cultures, detected by the analysis workflow described (SNV, CNV, LOH) in this protocol and affecting the *Kras* locus (Chr6, **a**–**d**). In addition, representative M-FISH karyotypes are shown for each cancer. This pancreatic cancer mouse model is driven by an oncogenic $Kras^{G12D}$ mutation. Samples used in this figure are R1035 (**a**), S134 (**b**), 16992 (**c**), B590 (**d**). **a**,**b**, Chr6 trisomy (**a**) can be detected in both the CNV plot (top) and LOH plot (bottom), where a log2 ratio of 0.6 corresponds to the gain of one chromosome in a diploid genome. In the LOH plot, the gain of one copy results in shifts of allele frequencies of germline variants to 0.66 and 0.34. The focal amplification of the *Kras* locus (**b**) is only visible in the CNV plot, owing to insufficient numbers of heterozygous germline variants at this locus. In both cases, SNV calling is needed to determine whether the wild-type or the mutated *Kras* allele is amplified. **c**,**d**, Copy-number-neutral LOH losses can result from a two-step process: first, loss of the whole chromosome (**c**), and in a second step, amplification of the remaining chromosome to the diploid state. In **d**, mitotic recombination causes loss of only parts of the chromosome. These changes cannot be detected by CNV, but only by LOH analysis. Chr, chromosome; CN, copy number; WT, wild type.

copies of Chr11 are detectable in an otherwise tetraploid genome. All three carry the somatically acquired *Trp53* mutation. Owing to the low number of heterozygous germline variants (inbred mice), LOH analyses are impossible. Therefore, the exact evolution of these changes cannot be resolved in this cancer.

Figure 16b shows a heterozygous loss of Chr4, which harbors *Cdkn2a*, an important tumor suppressor locus in pancreatic cancer. In a different tumor (Fig. 16c), *Cdkn2a* is inactivated by two independent copy-number alterations: loss of one Chr4 and focal *Cdkn2a* deletion on the remaining chromosome. Finally, another cancer (Fig. 16d) displays a homozygous *Cdkn2a* loss. The genome of this cancer is tetraploid. Only two Chr4s are present, which are identical (identical focal deletion and haplotype). The data indicate that loss of one Chr4 and deletion of *Cdkn2a* on the remaining Chr4 happened before genome duplication. Not shown here are less frequent types of homozygous *Cdkn2a* inactivations, such as (i) two independent deletions on both Chr4s and (ii) deletion of *Cdkn2a* on one Chr4, followed by CN-LOH through mitotic recombination.

Reporting Summary

Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

NGS data from mouse pancreatic cancer cell cultures are available from the European Nucleotide Archive using study accession no. PRJEB23787. The validation datasets generated during the current study are available from the corresponding author upon request.

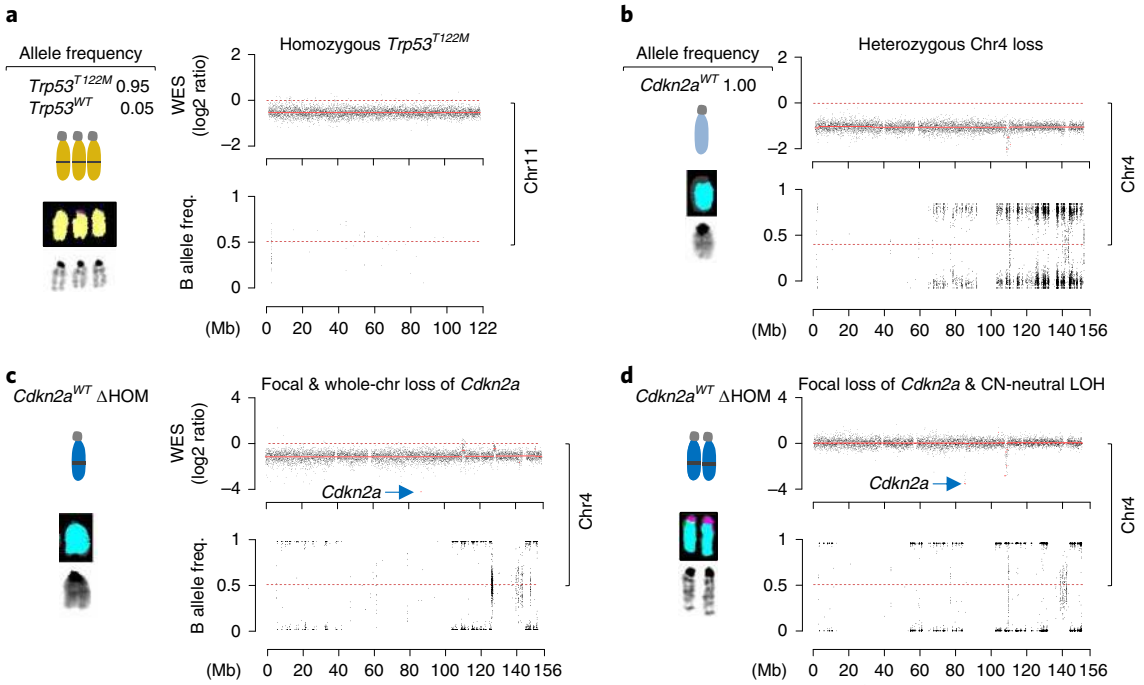


Fig. 16 | Patterns of genomic changes affecting tumor suppressor genes. a–d, Exemplary changes from a cohort of pancreatic cancer cell cultures, detected by the analysis workflow (SNV, CNV, LOH) described in this protocol, affecting the *Trp53* (Chr11, **a**) or *Cdkn2a* (Chr4, **b**–**d**) locus. Samples used in this figure are 9203 (**a**), 16990 (**b**), 5748 (**c**) and 53704 (**d**). **a**, SNV calling revealed a homozygous *Trp53* mutation. In the CNV plot, the whole chromosome is shifted to a log2 ratio of −0.4. This implies a tetraploid genome, where this log2 ratio corresponds to loss of one of four copies of Chr11. An insufficient number of germline variants for LOH analysis are available for this chromosome. **b**, Loss of one copy of Chr4, which can be visualized in both the CNV and LOH plots. **c,d**, Tumor suppressors are frequently inactivated homozygously either through (i) mutation or focal loss on one chromosome, followed by loss of the remaining wild-type allele through whole chromosome loss or chromosomal missegregation (**c**) or (ii) mutation or focal loss, followed by mitotic recombination (**d**). ΔHOM, homozygous deletion of *Cdkn2a* locus; Chr, chromosome; CN, copy number.

Code availability

The source code for all pipelines is available for public use at https://github.com/roland-rad-lab/MoCaSeq under the MIT license. In addition, the main workflow described in this protocol is packaged as a Docker container, available at https://cloud.docker.com/repository/docker/rolandradlab/mocaseq.

References

1. Morse, H. C. III. *Origins of Inbred Mice* (Elsevier Science, 2012).
2. van der Weyden, L., Adams, D. J. & Bradley, A. Tools for targeted manipulation of the mouse genome. *Physiol. Genomics* **11**, 133–164 (2002).
3. Jonkers, J. & Berns, A. Conditional mouse models of sporadic cancer. *Nat. Rev. Cancer* **2**, 251–265 (2002).
4. Weber, J. & Rad, R. Engineering CRISPR mouse models of cancer. *Curr. Opin. Genet. Dev.* **54**, 88–96 (2019).
5. Breschi, A., Gingeras, T. R. & Guigo, R. Comparative transcriptomics in human and mouse. *Nat. Rev. Genet.* **18**, 425–440 (2017).
6. Mouse Genome Sequencing, Consortium et al. Initial sequencing and comparative analysis of the mouse genome. *Nature* **420**, 520–562 (2002).
7. Lander, E. S. et al. Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).
8. She, X., Cheng, Z., Zollner, S., Church, D. M. & Eichler, E. E. Mouse segmental duplication and copy number variation. *Nat. Genet.* **40**, 909–914 (2008).
9. Egan, C. M., Sridhar, S., Wigler, M. & Hall, I. M. Recurrent DNA copy number variation in the laboratory mouse. *Nat. Genet.* **39**, 1384–1389 (2007).
10. Keane, T. M. et al. Mouse genomic variation and its effect on phenotypes and gene regulation. *Nature* **477**, 289–294 (2011).
11. Lek, M. et al. Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285–291 (2016).

12. Lee, G. H. et al. Strain specific sensitivity to diethylnitrosamine-induced carcinogenesis is maintained in hepatocytes of C3H/HeN in equilibrium with C57BL/6N chimeric mice. *Cancer Res.* **51**, 3257–3260 (1991).
13. Reilly, K. M., Loisel, D. A., Bronson, R. T., McLaughlin, M. E. & Jacks, T. Nf1;Trp53 mutant mice develop glioblastoma with evidence of strain-specific effects. *Nat. Genet.* **26**, 109–113 (2000).
14. Moser, A. R., Hegge, L. F. & Cardiff, R. D. Genetic background affects susceptibility to mammary hyperplasias and carcinomas in Apc(min)/+ mice. *Cancer Res.* **61**, 3480–3485 (2001).
15. Xu, X. et al. Induction of intrahepatic cholangiocellular carcinoma by liver-specific disruption of Smad4 and Pten in mice. *J. Clin. Invest.* **116**, 1843–1852 (2006).
16. Rad, R. et al. A genetic progression model of Braf(V600E)-induced intestinal tumorigenesis reveals targets for therapeutic intervention. *Cancer Cell* **24**, 15–29 (2013).
17. Mueller, S. et al. Evolutionary routes and KRAS dosage define pancreatic cancer phenotypes. *Nature* **554**, 62–68 (2018).
18. Cancer Genome Atlas Research Network. Electronic address: andrew_aguirre@dfci.harvard.edu; Cancer Genome Atlas Research, N. Integrated genomic characterization of pancreatic ductal adenocarcinoma. *Cancer Cell* **32**, 185–203 e113 (2017).
19. de Ruiter, J. R., Wessels, L. F. A. & Jonkers, J. Mouse models in the era of large human tumour sequencing studies. *Open Biol.* **8**, 180080 (2018).
20. McFadden, D. G. et al. Genetic and clonal dissection of murine small cell lung carcinoma progression by genome sequencing. *Cell* **156**, 1298–1311 (2014).
21. McFadden, D. G. et al. Mutational landscape of EGFR-, MYC-, and Kras-driven genetically engineered mouse models of lung adenocarcinoma. *Proc. Natl Acad. Sci. USA* **113**, E6409–E6417 (2016).
22. Koren, S. et al. PIK3CA(H1047R) induces multipotency and multi-lineage mammary tumours. *Nature* **525**, 114–118 (2015).
23. Ferreira, R. M. M. et al. Duct- and acinar-derived pancreatic ductal adenocarcinomas show distinct tumor progression and marker expression. *Cell Rep.* **21**, 966–978 (2017).
24. Chung, W. J. et al. Kras mutant genetically engineered mouse models of human cancers are genomically heterogeneous. *Proc. Natl Acad. Sci. USA* **114**, E10947–E10955 (2017).
25. Winters, I. P., Murray, C. W. & Winslow, M. M. Towards quantitative and multiplexed in vivo functional cancer genomics. *Nat. Rev. Genet.* **19**, 741–755 (2018).
26. Maronpot, R. R., Fox, T., Malarkey, D. E. & Goldsworthy, T. L. Mutations in the ras proto-oncogene: clues to etiology and molecular pathogenesis of mouse liver tumors. *Toxicology* **101**, 125–156 (1995).
27. Quintanilla, M., Brown, K., Ramsden, M. & Balmain, A. Carcinogen-specific mutation and amplification of Ha-ras during mouse skin carcinogenesis. *Nature* **322**, 78–80 (1986).
28. You, M., Candrian, U., Maronpot, R. R., Stoner, G. D. & Anderson, M. W. Activation of the Ki-ras protooncogene in spontaneously occurring and chemically induced lung tumors of the strain A mouse. *Proc. Natl Acad. Sci. USA* **86**, 3070–3074 (1989).
29. McCreery, M. Q. et al. Evolution of metastasis revealed by mutational landscapes of chemically induced skin cancers. *Nat. Med.* **21**, 1514–1520 (2015).
30. Nassar, D., Latil, M., Boeckx, B., Lambrechts, D. & Blanpain, C. Genomic landscape of carcinogen-induced and genetically induced mouse skin squamous cell carcinoma. *Nat. Med.* **21**, 946–954 (2015).
31. Westcott, P. M. et al. The mutational landscapes of genetic and chemical models of Kras-driven lung cancer. *Nature* **517**, 489–492 (2015).
32. Connor, F. et al. Mutational landscape of a chemically-induced mouse model of liver cancer. *J. Hepatol.* **69**, 840–850 (2018).
33. Arora, K. *et al.* Deep sequencing of 3 cancer cell lines on 2 sequencing platforms. *bioRxiv*, https://doi.org/10.1101/623702 (2019).
34. Weirather, J. L. et al. Comprehensive comparison of pacific Biosciences and Oxford Nanopore Technologies and their applications to transcriptome analysis. *F1000Res* **6**, 100 (2017).
35. Uchimura, A. et al. Germline mutation rates and the long-term phenotypic effects of mutation accumulation in wild-type laboratory mice and mutator mice. *Genome Res.* **25**, 1125–1134 (2015).
36. Milholland, B. et al. Differences between germline and somatic mutation rates in humans and mice. *Nat. Commun.* **8**, 15183 (2017).
37. Adewoye, A. B., Lindsay, S. J., Dubrova, Y. E. & Hurles, M. E. The genome-wide effects of ionizing radiation on mutation induction in the mammalian germline. *Nat. Commun.* **6**, 6684 (2015).
38. Einaga, N. et al. Assessment of the quality of DNA from various formalin-fixed paraffin-embedded (FFPE) tissues and the use of this DNA for next-generation sequencing (NGS) with no artifactual mutation. *PLoS One* **12**, e0176280 (2017).
39. Shi, W. et al. Reliability of whole-exome sequencing for assessing intratumor genetic heterogeneity. *Cell Rep.* **25**, 1446–1457 (2018).
40. Cibulskis, K. et al. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat. Biotechnol.* **31**, 213–219 (2013).
41. Francis, J. C. et al. Whole-exome DNA sequence analysis of Brca2- and Trp53-deficient mouse mammary gland tumours. *J. Pathol.* **236**, 186–200 (2015).
42. Ratnaparkhe, M. et al. Defective DNA damage repair leads to frequent catastrophic genomic events in murine and human tumors. *Nat. Commun.* **9**, 4760 (2018).

43. Kim, S. et al. Strelka2: fast and accurate calling of germline and somatic variants. *Nat. Methods* **15**, 591–594 (2018).
44. Koboldt, D. C. et al. VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res.* **22**, 568–576 (2012).
45. Poplin, R. *et al.* Scaling accurate genetic variant discovery to tens of thousands of samples. *bioRxiv*, https://doi.org/10.1101/201178 (2018).
46. Ye, K., Schulz, M. H., Long, Q., Apweiler, R. & Ning, Z. Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics* **25**, 2865–2871 (2009).
47. Costello, M. et al. Discovery and characterization of artifactual mutations in deep coverage targeted capture sequencing data due to oxidative DNA damage during sample preparation. *Nucleic Acids Res* **41**, e67 (2013).
48. Choi, Y. & Chan, A. P. PROVEAN web server: a tool to predict the functional effect of amino acid substitutions and indels. *Bioinformatics* **31**, 2745–2747 (2015).
49. Dees, N. D. et al. MuSiC: identifying mutational significance in cancer genomes. *Genome Res.* **22**, 1589–1598 (2012).
50. Gehring, J. S., Fischer, B., Lawrence, M. & Huber, W. SomaticSignatures: inferring mutational signatures from single-nucleotide variants. *Bioinformatics* **31**, 3673–3675 (2015).
51. Kuilman, T. et al. CopywriteR: DNA copy number detection from off-target sequence data. *Genome Biol.* **16**, 49 (2015).
52. Talevich, E., Shain, A. H., Botton, T. & Bastian, B. C. CNVkit: genome-wide copy number detection and visualization from targeted DNA sequencing. *PLoS Comput. Biol.* **12**, e1004873 (2016).
53. 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).
54. Stephens, P. J. et al. Massive genomic rearrangement acquired in a single catastrophic event during cancer development. *Cell* **144**, 27–40 (2011).
55. Korbel, J. O. & Campbell, P. J. Criteria for inference of chromothripsis in cancer genomes. *Cell* **152**, 1226–1236 (2013).
56. Rausch, T. et al. DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics* **28**, i333–i339 (2012).
57. Ha, G. et al. Integrative analysis of genome-wide loss of heterozygosity and monoallelic expression at nucleotide resolution reveals disrupted pathways in triple-negative breast cancer. *Genome Res.* **22**, 1995–2007 (2012).
58. Choi, Y., Chan, A. P., Kirkness, E., Telenti, A. & Schork, N. J. Comparison of phasing strategies for whole human genomes. *PLoS Genet.* **14**, e1007308 (2018).
59. Medvedev, P., Fiume, M., Dzamba, M., Smith, T. & Brudno, M. Detecting copy number variation with mated short reads. *Genome Res.* **20**, 1613–1622 (2010).
60. Guillen, J. FELASA guidelines and recommendations. *J. Am. Assoc. Lab Anim. Sci.* **51**, 311–321 (2012).
61. Slaoui, M. & Fiette, L. Histopathology procedures: from tissue sampling to histopathological evaluation. *Methods Mol. Biol.* **691**, 69–82 (2011).
62. Friedrich, M. J. et al. Genome-wide transposon screening and quantitative insertion site sequencing for cancer gene discovery in mice. *Nat Protoc.* **12**, 289–309 (2017).
63. Witkiewicz, A. K. et al. Whole-exome sequencing of pancreatic cancer defines genetic diversity and therapeutic targets. *Nat. Commun.* **6**, 6744 (2015).

Acknowledgements

R.R. was supported by the European Research Council (Consolidator Grants PACA-MET and MSCA-ITN-ETN PRECODE), the German Research Foundation (DFG RA1629/2-1; SFB1243; SFB1321; SFB1335), the German Cancer Consortium Joint Funding Program, and the Deutsche Krebshilfe (70112480).

Author contributions

S.L., T.E., M.Z., S.M., L.G.-S., I.V. and R.R. conceptualized, designed or developed analysis workflows, tools or procedures. S.L. integrated and validated bioinformatic workflows. S.M., R.M., M.J.F., R.B. and F.Y. performed wet-lab experiments. G.S., G.S.V. and D.S. provided biological resources and critical input during protocol development. S.L. and R.R. wrote the manuscript with input from T.E., S.M., R.M., M.J.F and I.V.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41596-019-0234-7.

Correspondence and requests for materials should be addressed to R.R.

Peer review information *Nature Protocols* thanks Malachi Griffith and other anonymous reviewer(s) for their contribution to the peer review of this work.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Received: 31 January 2019; Accepted: 27 August 2019;

Related links

Key references using this protocol

Mueller, S. et al. *Nature* **554**, 62–68 (2018) https://doi.org/10.1038/nature25459

Rad, R. et al. *Cancer Cell* **24**, 15–29 (2013) https://doi.org/10.1016/j.ccr.2013.05.014

Key data used in this protocol

Mueller, S. et al. Nature **554**, 62–68 (2018) https://doi.org/10.1038/nature25459

QUERY FORM

NPROT	
Manuscript ID	**[Art. Id: 234]**
Author	
Editor	
Publisher	

Journal: NPROT

Author :- The following queries have arisen during the editing of your manuscript. Please answer by making the requisite corrections directly in the e.proofing tool rather than marking them up on the PDF. This will ensure that your corrections are incorporated accurately and that your paper is published as quickly as possible.

Query No.	Description	Author's Response
AQ1	Please check the changes made to affiliation 9; note that per journal style, positions cannot be listed as affiliations, and Dr. Saur already has affiliation 3 listed.	
AQ2	Please check the changes made to the abstract. Please note that, per journal style, URLs are not allowed in the abstract, so we deleted it.	
AQ3	Please check the changes made to the Introduction.	
AQ4	Please check your article carefully, coordinate with any co-authors and enter all final edits clearly in the eproof, remembering to save frequently. Once corrections are submitted, we cannot routinely make further changes to the article.	
AQ5	Note that the eproof should be amended in only one browser window at any one time; otherwise changes will be overwritten.	
AQ6	Author surnames have been highlighted. Please check these carefully and adjust if the first name or surname is marked up incorrectly. Note that changes here will affect indexing of your article in public repositories such as PubMed. Also, carefully check the spelling and numbering of all author names and affiliations, and the corresponding email address(es).	
AQ7	You cannot alter accepted Supplementary Information files except for critical changes to scientific content. If you do resupply any files, please also provide a brief (but complete) list of changes.	

QUERY FORM

NPROT	
Manuscript ID	**[Art. Id: 234]**
Author	
Editor	
Publisher	

Journal: NPROT

Author :- The following queries have arisen during the editing of your manuscript. Please answer by making the requisite corrections directly in the e.proofing tool rather than marking them up on the PDF. This will ensure that your corrections are incorporated accurately and that your paper is published as quickly as possible.

Query No.	Description	Author's Response
AQ8	Please check the changes made to the figures and their captions, including expansions of abbreviations.	
AQ9	Insertion of minus signs in the y axis of Figure 2a OK?	
AQ10	Please spell out "PK".	
AQ11	Please provide the expansion of "n"; if it stands for "number", please change to "n = 4" and "n = 2", etc.	
AQ12	Change to "dimethylbenzanthracene" OK?	
AQ13	Please check the changes made to the sentence "We prefer Illumina systems for WGS or WES…"	
AQ14	Change to "Illumina HiSeq 3000/4000" OK?	
AQ15	Is the change to "megabase" OK? Or should it be "megabase pair"?	
AQ16	Please specify whether the percentage is "vol/vol", "wt/vol" or "wt/wt" for ethanol expressed in %.	
AQ17	Please provide a URL leading to the reference genome GRCm38 (mm10).	

QUERY FORM

NPROT	
Manuscript ID	**[Art. Id: 234]**
Author	
Editor	
Publisher	

Journal: NPROT

Author :- The following queries have arisen during the editing of your manuscript. Please answer by making the requisite corrections directly in the e.proofing tool rather than marking them up on the PDF. This will ensure that your corrections are incorporated accurately and that your paper is published as quickly as possible.

Query No.	Description	Author's Response
AQ18	Please check "685" number (Available for validation) in Figure 3a, the caption, and the main text; we added up the numbers and got 686.	
AQ19	Please provide the expansions of "F", "G", "I", and "L", if applicable.	
AQ20	Please check the changes made to all the tables.	
AQ21	In the Impact column, are both "Moderate" and "Modifier" correct?	
AQ22	Please check the changes made to the sentence "As an example, the total numbers…"	
AQ23	Change to "49.6 Mb" (megabases instead of megabytes)? Also, should "240K" be "240 kb"?	
AQ24	Please check the description of Figure 5e carefully and mention in the caption the difference between the top and bottom images.	
AQ25	Change to "owever, this advantage must be weighed against" OK?	
AQ26	Change to "marked with asterisks in Fig. 5a)" OK?	
AQ27	In The Figure 8 caption, please explain what the numbers at the top of the zoomed-in images in 8a and 8b represent.	

QUERY FORM

NPROT	
Manuscript ID	**[Art. Id: 234]**
Author	
Editor	
Publisher	

Journal: NPROT

Author :- The following queries have arisen during the editing of your manuscript. Please answer by making the requisite corrections directly in the e.proofing tool rather than marking them up on the PDF. This will ensure that your corrections are incorporated accurately and that your paper is published as quickly as possible.

Query No.	Description	Author's Response
AQ28	Please provide the expansions of the abbreviations used in Figure 10.	
AQ29	The format used to further describe the six hallmarks is incompatible with our numbering format; therefore, we have deleted the numbers; OK?	
AQ30	Please check the changes made to the sentence "Our algorithm sequentially inserts…"	
AQ31	Please check the changes made to the Materials section.	
AQ32	Usually in such a CAUTION note, the author names the specific IRB(s) or IAUAC(s) that approved the experiments discussed (and states the approval number(s), if applicable).	
AQ33	Note that the HiSeq kits were moved here from the Equipment list because they contain reagents.	
AQ34	Please note that "ddH2O" has been inserted here; please provide supplier and catalog number.	

QUERY FORM

NPROT	
Manuscript ID	**[Art. Id: 234]**
Author	
Editor	
Publisher	

Journal: NPROT

Author :- The following queries have arisen during the editing of your manuscript. Please answer by making the requisite corrections directly in the e.proofing tool rather than marking them up on the PDF. This will ensure that your corrections are incorporated accurately and that your paper is published as quickly as possible.

Query No.	Description	Author's Response
AQ35	Per journal requirements, all items listed in the "Biological materials", "Reagents" and "Equipment" sections (aside from materials such as regular water that would not be obtained from an outside supplier) should include the supplier name and catalog and/or model number or citation of a protocol for obtaining the item. Please check the lists carefully and add this information where applicable. Please check the protocol thoroughly to make sure that all materials mentioned are listed in the Materials section, along with supplier and catalog/model number. This will help to ensure that readers can successfully use your protocol.	
AQ36	Change to "Fine Science Tools" (twice) OK?	
AQ37	The "Tweezers" entry was inserted because tweezers were mentioned elsewhere; please provide supplier and catalog number.	
AQ38	Please check the URLs inserted for bcl2fastq and Trimmomatic; we could not find the software at the given web pages, so we did searches and found them at the inserted URLs.	
AQ39	Please check the sentence "This is very flexible…" and the following sentence. You seem to be saying that using the docker container "can be cumbersome when processing large numbers of samples", but then go on to say "Using the functionality provided by the Docker container to execute a scripted version of this pipeline greatly simplifies processing…"; the two statements seem to contradict each other.	

QUERY FORM

NPROT	
Manuscript ID	**[Art. Id: 234]**
Author	
Editor	
Publisher	

Journal: NPROT

Author :- The following queries have arisen during the editing of your manuscript. Please answer by making the requisite corrections directly in the e.proofing tool rather than marking them up on the PDF. This will ensure that your corrections are incorporated accurately and that your paper is published as quickly as possible.

Query No.	Description	Author's Response
AQ40	Please check all code throughout the paper to make sure it has been rendered properly. Please note that we have changed "curly" (slanted) quotation marks to straight quotation marks.	
AQ41	Per journal style, the CAUTION note must appear at the end of the entry, so we have moved it here and slightly reworded it; please check.	
AQ42	Per journal style, we have changed "artefact" to "artifact" in the text, but left it unchanged in the code. Please check whether it can be changed in the code as well or will cause it not to run.	
AQ43	Please check the changes made to the Procedure.	
AQ44	We inserted headings over Steps 2 and 3 so that they could be treated separately in the Timing section at the end of the paper; OK? Please insert or adjust timing as needed, i.e., can you provide timing for Step 3 and a range for Step 1 instead of "variable"?	
AQ45	We changed the CAUTION notes to CRITICAL STEP notes here because they did not seem to be related to personal safety or possible ethical violations; please check.	
AQ46	Please check the changes made to Step 2B(iv).	
AQ47	Please provide timing here for Step 3 and in the corresponding entry in the Timing section.	

QUERY FORM

NPROT	
Manuscript ID	**[Art. Id: 234]**
Author	
Editor	
Publisher	

Journal: NPROT

Author :- The following queries have arisen during the editing of your manuscript. Please answer by making the requisite corrections directly in the e.proofing tool rather than marking them up on the PDF. This will ensure that your corrections are incorporated accurately and that your paper is published as quickly as possible.

Query No.	Description	Author's Response
AQ48	"HiSeq 4000 Reagent Kit" is not in the Reagents list; please specify an item in the list, or add this to the list, along with supplier/catalog number.	
AQ49	Change to "HiSeq X Ten Reagent Kit" OK?	
AQ50	Change to "inside the temp folder" OK?	
AQ51	For Step 9C, does Step 9B have to be performed after Step 9C is performed?	
AQ52	Change to "found in the fastqc output" OK?	
AQ53	In the line of code in Step 11 that contains $trimmomatic_dir"/"$trimmomatic_file", should there be another set of quotation marks?	
AQ54	For "used in the exome extraction kit", please specify the step at which this was done.	
AQ55	Please check the code in Step 24 (and beyond) carefully, especially the parentheses and spaces.	
AQ56	Please check the text "and supporting reads for mutation in tumor sample" for clarity.	
AQ57	Change to "and (iii) that are potentially affected by strand artifacts" OK?	

QUERY FORM

NPROT	
Manuscript ID	**[Art. Id: 234]**
Author	
Editor	
Publisher	

Journal: NPROT

Author :- The following queries have arisen during the editing of your manuscript. Please answer by making the requisite corrections directly in the e.proofing tool rather than marking them up on the PDF. This will ensure that your corrections are incorporated accurately and that your paper is published as quickly as possible.

Query No.	Description	Author's Response
AQ58	Here and throughout the paper, if "mode" simply means "mode", please lowercase all instances; if it refers to a program or something similar, please change to "Mode" where necessary.	
AQ59	Change to "CNVs from WGS data" OK?	
AQ60	Change to "The output format (.tif or .emf)" OK? If not, please change to "The output format (TIF or EMF)".	
AQ61	Is it correct that Step 46 skips over the "Prevalence of rearrangements affecting one haplotype" (hallmark iv) test? If you need to insert an additional step here, please check all cross-references to step numbers to make sure that the revised step numbers are used.	
AQ62	Edit correct?	
AQ63	Please check the changes made to the Timing section to fit journal format; please check all timings given here against those given in the Procedure.	
AQ64	Changes to "5–6 h" for Step 2B, to "2 h" for Step 2C, and to 2.5 d for Step 5A (to match the Procedure) OK? If not, please adjust timing in the Procedure to match.	
AQ65	Change to "at prototypic tumor suppressors and oncogenes" OK?	

QUERY FORM

NPROT	
Manuscript ID	**[Art. Id: 234]**
Author	
Editor	
Publisher	

Journal: NPROT

Author :- The following queries have arisen during the editing of your manuscript. Please answer by making the requisite corrections directly in the e.proofing tool rather than marking them up on the PDF. This will ensure that your corrections are incorporated accurately and that your paper is published as quickly as possible.

Query No.	Description	Author's Response
AQ66	R.R. has 13 coauthors, some of them at different institutions and in different countries. Please make sure the Acknowledgements cover funding sources for all authors, and please check that all grant numbers are correct.	
AQ67	Please check that the Competing interests section is correct and complete.	

nature research

Corresponding author(s): Roland Rad

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistical parameters

When statistical analyses are reported, confirm that the following items are present in the relevant location (e.g. figure legend, table legend, main text, or Methods section).

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	An indication of whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistics including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☐	☒	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☐	☒	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated
☐	☒	Clearly defined error bars *State explicitly what error bars represent (e.g. SD, SE, CI)*

Our web collection on statistics for biologists may be useful.

Software and code

Policy information about availability of computer code

Data collection	The information is provided in the manuscript.
Data analysis	The information is included in the manuscript and available online (https://github.com/roland-rad-lab/MoCaSeq)

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

WES (n=38) and WGS (n=1) data is available at ENA (PRJEB23787)

Field-specific reporting

Please select the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences

For a reference copy of the document with all sections, see nature.com/authors/policies/ReportingSummary-flat.pdf

Life sciences

Study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	N/A
Data exclusions	N/A
Replication	N/A
Randomization	N/A
Blinding	N/A

Materials & experimental systems

Policy information about availability of materials

n/a	Involved in the study
☒	☐ Unique materials
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Research animals
☒	☐ Human research participants

Method-specific reporting

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ Magnetic resonance imaging