DeepClone, an end-to-end protocol to study somatic mutagenesis and selection at high resolution

Authors

Ferriol Calvet[1,2,3], Morena Pinheiro-Santin[1], Erika López-Arribillaga[1,3], Raquel Blanco Martinez-Illescas[1,2,3], Núria Samper[1], Miguel L. Grau[1], Ferran Muiños[1,3], Rocío Chamorro González[1], Maria Andrianova[1], Federica Brando[1], Stefano Pellegrini[1,2,3], Axel Rosendahl Huber[1,3], Marta Huertas[1], Elisabet Figuerola-Bou[1], Coohleen Coombes[4], Brendan F. Kohrn[4], Jeanne Fredrickson[4], Rosa Ana Risques[4], Nuria Lopez-Bigas[1,2,3,5,6,@], Abel Gonzalez-Perez[1,2,3,6,@]

[6] These authors jointly supervised this work: A. Gonzalez-Perez, N. Lopez-Bigas

[@] Correspondence should be addressed to Nuria Lopez-Bigas <nuria.lopez@irbbarcelona.org> and Abel Gonzalez-Perez <abel.gonzalez@irbbarcelona.org>

Affiliations

1. Institute for Research in Biomedicine (IRB Barcelona), The Barcelona Institute of Science and Technology, Baldiri Reixac, 10, 08028 Barcelona, Spain.
2. Department of Medicine and Life Sciences, Universitat Pompeu Fabra, Barcelona, Spain.
3. Centro de Investigación Biomédica en Red en Cáncer (CIBERONC), Instituto de Salud Carlos III, Madrid, Spain.
4. Department of Laboratory Medicine and Pathology, University of Washington, Seattle, WA.
5. Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Spain.

Abstract

The advent of Next Generation Sequencing (NGS) technologies opened up the door to the study of somatic mutagenesis and selection through the analysis of a salient clone in a sample, as in the case of cancer. This was made possible by the development of an ecosystem of computational methods that enabled mutation calling, the study of mutational processes, and the quantification of positive selection, among other research avenues. Recently, the introduction of DNA duplex sequencing technologies has allowed the detection of somatic mutations that occur in one or few cells in a sample. This has unlocked the possibility to study the interplay between mutagenesis and selection on hundreds or thousands of clones per sample, providing access to understanding somatic evolution in scenarios that were not previously easy to access, such as normal tissues. The adoption of DNA duplex sequencing technologies has been hindered, among other reasons, by the lack of a complete ecosystem of computational tools that support end-to-end analysis from the sequencing raw data to the quantification of multiple aspects of mutagenesis and selection. To overcome this problem, we

present DeepClone, a protocol that comprises an experimental DNA duplex sequencing library preparation solution, and two computational pipelines to readily identify somatic mutations, and carry out calculations of mutagenesis and selection in a cohort of samples.

Introduction

The advent of Next Generation Sequencing (NGS) technologies, in tandem with the completion of the first draft of the human genome, galvanized the study of germline and somatic variation in the last two decades. These technologies enabled the identification of genetic variants present across many (or all) cells in a sample, as their recurrent observation offset the sequencing error rate. They thus provided a clear advantage for the study of mutagenesis and selection in scenarios where a salient clone was present in a biological sample, as is the case in cancer[1–5], clonal hematopoiesis[6,7], in vitro clonal expansion of a single stem cell[8–10], and clonal or quasi-clonal structures microdissected from human normal tissues[11–14]. This led to important discoveries, such as a compendium of cancer driver genes, the possibility to interpret the mutations in an individual's tumor to enable precision cancer medicine, and the identification of endogenous and external mutational processes and their footprint in normal tissues and tumors. Such discoveries were supported by the development of a comprehensive arsenal of computational methods of analysis of NGS data, in particular in the realm of cancer genomics[3,5,15–17].

Nevertheless, the thorough and systematic identification of small clones in a sample (of a tumor or a normal tissue) is not possible through the use of standard NGS technologies, given their error rate. While the approaches to study microscopic clonal or quasi-clonal structures or in vitro expanded clones mentioned above provide a way to circumvent this problem, they all have very low throughput. To overcome this hurdle, in the last decade, a family of error-correcting sequencing technologies based on identifying mutations through independent consensus of the two strands of a DNA molecule (duplex DNA sequencing) have been developed and used in several studies[18–28]. While NGS technologies opened up the door to the study of somatic mutagenesis and selection on large expanded clones per sample sequenced, DNA duplex sequencing technologies allow the study of hundreds or thousands of small clones in one sample. Several experimental protocols for DNA duplex sequencing, and associated bioinformatics pipelines to call somatic mutations are public[18,19,21]. However, a fully developed ecosystem coupling experimental and bioinformatics tools for the study of mutagenesis and selection using DNA duplex sequencing is currently unavailable.

We present here an end-to-end experimental and bioinformatics protocol (DeepClone) to identify and analyze somatic mutations present at very low variant allele frequencies (i.e., present in 1/1000 cells or less) in a sample. DeepClone is composed of three separate elements. The first is an experimental protocol to build duplex DNA sequencing libraries. The other are two bioinformatics pipelines designed to accurately identify somatic mutations from duplex sequencing data (deepUMIcaller; https://github.com/bbglab/deepUMIcaller; Supp. Note), and to provide an ecosystem of computational methods to uncover different aspects of mutagenesis and selection –e.g., underlying mutational signatures, mutation density across genomic regions, strength of positive selection on different types of protein affecting mutations–, and their associations with the lifetime exposures of a group of individuals (deepCSA; https://github.com/bbglab/deepCSA; Supp. Note). These three elements can be used in tandem to go from the DNA to the final analysis of mutagenesis and selection. Alternatively, any of them can be combined with other available experimental and/or computational tools developed for similar aims. DeepClone can be applied to different research questions that require ultra-sensitive detection and characterization of mutations (see **Applications** section). These include human tissues collected from individuals with different lifestyle and history of exposures, before and after interventions that change exposures to exogenous agents (e.g., chemotherapy), or from *in vitro* or *in vivo* models exposed to known or suspected carcinogens

(Lopez-Bigas et al., *under review*). Furthermore, it can be applied to tumors to study selection under changing selective pressures, such as in response to a therapy. While DNA duplex sequencing (and the ecosystem of computational tools provided by DeepClone) can be applied also to clinical questions, such as tracking minimal residual disease upon cancer treatment, in this protocol, we will focus on applications concerning the study of mutagenesis and selection.

Development of the protocol

The development of DeepClone builds upon the experience of previous DNA duplex library preparation protocols[18,19,22,25,28] and existing pieces of software to analyze the resulting data and identify mutations. Ultradeep (more than 1,000x duplex depth) DNA duplex sequencing data is well suited to study mutagenesis and selection across small clones in polyclonal samples. Any method with that goal must provide enough statistical power for –and deal with– the calculation of mutation density and positive selection on the mutations in genes and subgenic elements (e.g., domains or exons) and their association with the lifetime exposure of individuals. To this end, we developed DeepClone, an end-to-end protocol composed of three parts: an experimental DNA duplex library construction method; deepUMIcaller[23], a computational pipeline to identify mutations from DNA duplex sequencing data; and deepCSA[23], a computational pipeline to analyze mutagenesis and selection from mutations called using this same data.

To construct DNA duplex sequencing libraries, we resorted to the adaptation of a commercial experimental protocol already in use[29], although other experimental approaches are available or can be readily implemented (e.g., refs[18–20,25] and Supp. Note) and used in its place. It is important, however, to understand the differences between duplex sequencing library preparation methods in order to select the most appropriate for each scientific question (see **Comparison to other methods** and **Experimental design** sections). We based the deepUMIcaller pipeline (Supp. Note) on the existing *fgbio Best Practices FASTQ to Consensus* pipeline[30,31] (https://nf-co.re/fastquorum/) and the VarDictJava[32] variant caller (https://github.com/AstraZeneca-NGS/VarDictJava) and added several quality control (QC) and monitoring steps along the process. To develop deepCSA, we collected a set of tools developed within cancer genomics for the analysis of mutagenesis (e.g., identifying mutational signatures[15,33] and calculating mutation density), and positive selection (e.g., OncodriveFML[28], Oncodrive3D[35]) and we adapted or developed new methods with the same purposes (e.g., omega, a dN/dS based method for positive selection[23]). It is important that, when running deepCSA, the output of every tool is carefully checked to guarantee that downstream steps receive inputs that correctly satisfy the analyses they implement. For example, it is important to verify that the magnitude of omega (dN/dS; see **Glossary**) calculated per sample is based upon a stable background mutation density before regressing them out on the information of exposures across donors (see **Experimental design** below).

Overview of the procedure

The preparation of libraries for DNA duplex sequencing uniquely tags both ends of each DNA molecule fragment in the sample with a double-stranded barcode, and amplifies by PCR both of its strands. Duplicates arising from each strand are then reunited computationally to form single strand consensus reads and the consensus of both strands arising from the same original fragment are combined to form a double strand (or duplex) consensus read. Nucleotide changes present in both strands of this duplex consensus read must precede the PCR, and are therefore not amplification or sequencing errors, but likely real mutations. The construction of a duplex consensus read from PCR duplicates to detect somatic mutations present in one or a

few such consensus DNA molecules requires a tailored computational pipeline. Furthermore, the analysis of mutagenesis and selection from these somatic mutations requires computational methods capable of dealing with the highly heterogeneous sequencing depth that is characteristic of DNA duplex sequencing. This heterogeneity implies that there is a different likelihood to detect mutations at regions sequenced at different duplex sequencing depth and, furthermore, different accuracy in the estimation of their variant allele frequency (VAF) in the sample. DeepClone covers these three steps: 1) the preparation of libraries for DNA duplex sequencing (Fig. 1a; Supp. Note), 2) the construction of double strand consensus sequences and the identification of somatic mutations, (Fig. 1b; Supp. Note) and 3) the analysis of mutagenesis and selection across samples (Fig. 1c,d; Supp. Note).

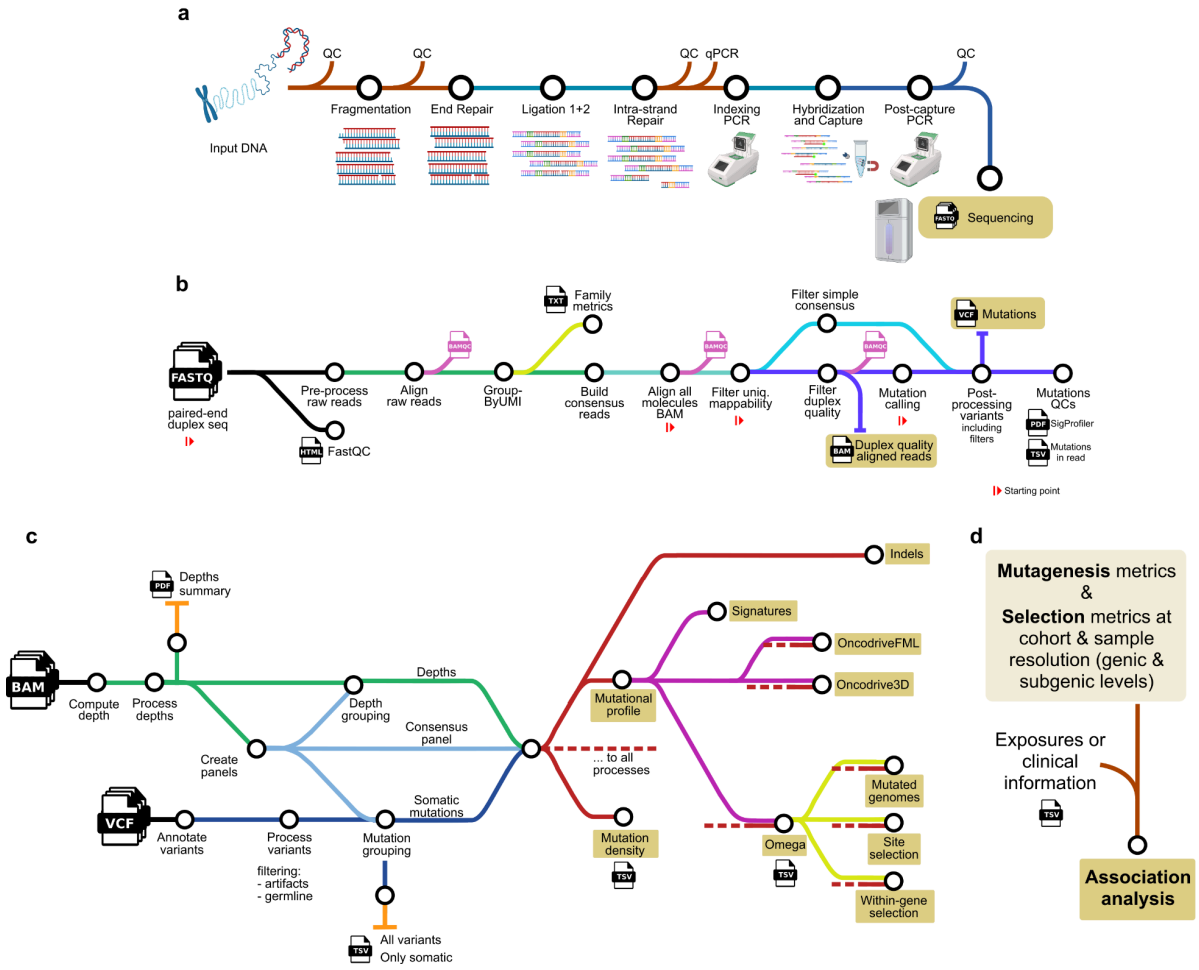

Figure 1. Schematic representation of the three parts of the DeepClone protocol.
a. DNA duplex library preparation protocol.
b. Computational pipeline to identify mutations in DNA duplex sequencing data (deepUMIcaller).
c. Computational pipeline to carry out analyses of mutagenesis and selection from DNA duplex sequencing data (deepCSA).
d. Different metrics of mutagenesis and selection across samples (such as the activity of mutational signatures and the value of omega of different genes) can then be tested for their association with information on the lifetime exposures of and other clinical information of donors.
Lines represent flow of DNA fragments (a) or data (b,c). The names represent independent steps in the DNA duplex library preparation or in the computational pipelines. The main outputs are highlighted within rectangular boxes. For a, b and c the shaded rectangular boxes at the end of each represent the input to the next set of steps (b, c, and d, respectively).

Applications

The main outcomes of this protocol are metrics of mutagenesis (such as active mutational signatures and mutation density) and selection obtained across samples and the association of these metrics with their history of exposures. When applied to polyclonal samples from *in vitro* or *in vivo* models exposed to known or suspected carcinogens, human samples obtained from individuals with different history of exposures, or before and after a potentially life-changing exposure (e.g., chemotherapy treatment, smoking cessation) the protocol can reveal which of these external agents act as mutagens or promoters of pre-existing mutated somatic cells. It can thus be used as a test of the mutagenic[22] or clonal promoting effect of any agent (Lopez-Bigas et al., *under review*). We recently demonstrated the application of this protocol to identify the associations of biological sex with the magnitude of positive selection on truncating mutations of three genes and of a history of smoking with the presence of activating mutations in the TERT promoter across normal urothelial samples[23]. Other recent articles have reported its use in the investigation of mutagenesis and selection in blood[20], buccal epithelium[20,29], or sperm[27], or across several tissues in autopsies[26].

We can envision that such a rapid test can be implemented using minimally invasive sampling (e.g., buccal swabs, blood, urine) to monitor specific population strata at higher risk of certain types of cancer, such as head and neck, esophageal or bladder carcinomas. DeepClone can also be applied to the study of subclonal evolution in tumors, in particular, in face of changing selective pressures, such as the exposure to an anti-cancer therapy[36]. Furthermore, it can be applied to the study of somatic mosaicism associated with cancer[37] or other diseases[38], or to understand the potential contribution of somatic mutations and clonal expansions to the aging soma[39,40].

The systematic application of the protocol to samples of the same tissue probing the mutations in a set of genes under positive selection can also be applied to the problem of saturation mutagenesis (understood here as the assessment of the potential functional impact of all possible mutations in a gene, especially in tumorigenesis). It is predicated on the idea that protein affecting mutations detected in a polyclonal sample reflect multiple expanded clones, which can be regarded as natural experiments revealing which mutations are under positive selection in the tissue in question. In theory, if a sufficiently large number of clones is interrogated (e.g., the cumulative depth resulting from pooling several samples), all mutations under positive selection will appear mutated over the expectation, unlike neutral or deleterious mutations. Two recent articles have demonstrated the power of this approach to build high resolution saturation mutagenesis maps in genes frequently mutated in the bladder urothelium, blood and the buccal epithelium[20,23].

Comparison to other methods

We briefly compare three similar protocols, DuplexSeq[18,25], Targeted NanoSeq[20] and UDSeq[21] with DeepClone (Supp. Table 1). All four use similar DNA duplex library preparation approaches that exploit the capabilities to uniquely tag DNA fragments in a sample using molecular barcodes (Supp. Table 1).

DuplexSeq, the first duplex sequencing protocol developed more than a decade ago[18,25,28], was based on *in house* duplex adaptors and originally employed sonication for DNA fragmentation. Sonication, shown to introduce a high number of sequencing artifacts, was later replaced with enzymatic fragmentation, to build a protocol that was commercialized by TwinStrand Biosciences. Recently, Targeted NanoSeq, which followed its untargeted version[19], was

developed on the basis of DuplexSeq by specifically blocking the repair of DNA fragments through the addition of ddBTPs. This way, the reduced error rate resulting from preventing the ligation of fragments with single-strand nicks, is obtained at the expense of reducing duplex depth and, as a consequence, the sensitivity to identify somatic mutations. Several newer duplex sequencing approaches (UDSeq, and DeepClone) are based on the original SaferSeqS[41]. DeepClone, however, introduces specific variations to obtain DNA blunt ends in fragmentation and to reduce the number of DNA bases damaged before ligation, while maintaining a number of ligated DNA molecules that is enough to obtain DNA duplex depth per sample above 3,000x (see description below). As explained above, Targeted NanoSeq differs from the other four in lacking a step of DNA end repair, which results in a smaller error rate ($<5x10^{-9}$ compared to $\sim10^{-8}$ of DeepClone) (Extended Data Fig. 1; Supp. Note).

Three of the protocols –UDseq, NanoSeq and DeepClone– provide similar computational pipelines to identify the somatic mutations from the DNA duplex sequencing data. DupCaller (UDseq) models the sample-specific error profiles resulting in a slightly more sensitive variant calling compared to NanoSeq[21]. Only DeepClone includes a computational pipeline that provides an ecosystem of tools to systematically analyze mutagenesis and selection across probed samples (deepCSA). This pipeline endows the user with the capability to run this ecosystem of tools to automatically compute a variety of outcomes, including the mutational signatures active across samples and the magnitude of positive selection on the mutations of different genes in each sample. More importantly, deepCSA furnishes an important number of quality control metrics of these outputs that allow the researcher to steer the analyses depending on their question and the results of intermediate steps (Extended Data Fig. 2-7). For example, deepCSA provides comparisons between the density of non-protein affecting mutations in every gene and the distribution observed across all genes, thus allowing the user to spot genes with extremely high or low values, which could be excluded from the calculation of positive selection with omega. We recommend that the first run of deepCSA in a cohort of samples be used primarily for quality control and not to directly obtain biologically meaningful results (see the **Procedure** section and Supp. Note).

DeepUMIcaller, the computational pipeline dedicated to identify somatic mutations in the DeepClone protocol can be easily coupled to DNA duplex data generated from different duplex library preparations (this was done in the original article presenting deepUMIcaller[23]). Similarly, deepCSA can be used on somatic mutations identified by other DNA duplex sequencing protocols and variant calling pipelines. This guarantees the flexibility needed to select different experimental approaches depending on specific research questions (see next section) at the convenience of the user. Both computational pipelines have been developed in Nextflow[42], increasing their portability and scalability.

Experimental design

The experimental design –starting DNA amount, specific DNA duplex library preparation (e.g., out of the four presented above), panel of genes to target, number of samples– needs to be guided by the research question (Supp. Note). Below, we describe the rationale to select the approach that is best suited for some typical research questions that can be tackled using DNA duplex sequencing.

If the aim of the experiment is to study mutagenesis (e.g., the mutational processes active in a tissue and the mutation density), the key element is to sample a representative set of all possible tri-nucleotides, as well as enough mutations to identify even mutational signatures with relatively low activity across samples. This depends on the number of sampled sites, which is

determined, ultimately by the product of the selected genomic region and the average depth of sequencing. As a result, if the region of the gene targeted for capture is sufficiently large, high sequencing depth is not required. All four protocols described in the previous section are well suited for this purpose.

When the researcher's goal includes the discovery of genes under positive selection in a tissue, the genomic region to target becomes more important, and so does the duplex sequencing depth. In this case, the panel only needs to include genic regions that may be under selection (relevant genes or non-coding regions with known regulatory function, such as the TERT promoter). Since the cost of sequencing the whole exome at sufficient depth across samples can scale very quickly, it is recommendable to target only the exons of a set of genes with reasonable probability to be under selection in the tissue in question, such as genes that drive tumors that arise from it, and/or genes already observed under selection in previous studies. In order to discover genes under positive selection, one requires enough statistical power to run computational driver discovery methods (such as those included in deepCSA; Supp. Note). This requires the identification of a sufficient number of protein and non-protein affecting mutations across samples. The latter critical requirement may be difficult to meet for short genes, with relatively few synonymous sites, such as TP53. If the duplex sequencing depth obtained per sample is not very high (e.g., below 1,000x), the experiment will require more samples to reach the minimum required number of mutations to run driver discovery methods. The same reasoning applies if the aim of the experiment is natural saturation mutagenesis (see above).

Finally, probing the association between the exposure to intrinsic (e.g., those associated with age or sex) or exogenous factors (e.g., tobacco smoking, alcohol drinking, or cytotoxic drugs) and the clonal landscape of a tissue requires to accurately infer the magnitude of positive selection shaping the mutational landscape of each gene of interest in each sample[23]. This type of analysis requires the interrogation of an even more reduced set of genes (only those with positive selection in the tissue in question, as others would be non-informative) and higher duplex depth per sample. Reaching the required depth per sample is difficult for Targeted NanoSeq, due to its DNA duplex library protocol, and for UDSeq, optimized for very low starting DNA quantities[21]. Overcoming this limitation to gain sufficient statistical power to do this analysis at scale requires either to prepare multiple duplex libraries using the DeepClone protocol, or the use of a more efficient and not limited DNA duplex library preparation (Supp. Note). We have demonstrated that both solutions work with DeepClone[23].

Sample requirements

It is critical to extract and manipulate the genomic DNA (gDNA) of interest in non-strand-denaturing conditions. We recommend using the DNeasy Blood & Tissue kit (Qiagen, cat. no. 69506) and if low DNA yield is expected, the QIAmp DNA micro kit (Qiagen, cat. no. 56304) is also suitable. The gDNA can be resuspended in a low-EDTA buffer such as low TE (10 mM Tris-Cl, 0.1mM EDTA, pH 8.0) or AE buffer (10 mM Tris-Cl, 0.5mM EDTA, pH 9.0). We recommend eluting the gDNA in a volume that will yield a suitable concentration range for optimal sample preparation. Upon quantification with fluorescence-based methods, such as Qubit, the desired gDNA input should be in a volume equal or less than 26 µL. If the gDNA is too diluted, bead-based or column-based methods are recommended for reconcentration. Evaporating methods (such as speed-vac) should be avoided to prevent salt concentration that can impair downstream steps. gDNA integrity should be assessed with a Genomic DNA ScreenTape Assay and run on an Agilent TapeStation 4200 or similar right before starting the DNA duplex library protocol, to account for any DNA integrity variations during sample shipping and/or storage. This protocol has been optimized for using high-molecular weight gDNA as input (DIN value above 6). Starting with lower-quality DNA is possible but the fragmentation step

needs to be tailored to avoid overfragmentation. Certain damage to the DNA introduced in the process of extraction or storage, or during fragmentation (e.g., single-stranded ends) could lead to a change in both strands in the DNA molecule as a consequence of the end repair process, which may then be wrongly identified as a somatic mutation. These changes of nucleotide in both strands that precede the ligation to duplex sequencing barcodes and are not due to mutational processes active in the tissue are known as pseudo-duplex nucleotide changes.

DNA fragmentation
Obtaining an optimal size distribution of fragments of gDNA is critical for library preparation and sequencing efficiency. To avoid single-stranded hanging ends in the fragmentation (see previous section), we chose to do an enzymatic fragmentation of the gDNA using the NEB UltraShear. The incubation time depends on sample characteristics (quantity, integrity of the starting DNA, elution buffer). Therefore a kinetics test is recommended to decide the optimal fragmentation conditions and maximize the number of DNA fragments with appropriate size for short-read NGS sequencing. The efficiency assessment of fragmentation can be performed by High Sensitivity DNA ScreenTape Assay and run on an Agilent TapeStation 4200 or similar, measuring the region ranging between 50 and 650 bp. The protocol has been designed to achieve optimal fragmentation status across samples being processed in parallel in one DNA duplex library preparation (see Figure 2 for an example case). However, the incubation step can be extended for specific samples in the batch, if needed. To decide the necessary sample-specific adjustment, all samples are kept on ice while running the High Sensitivity DNA ScreenTape Assay, and we recommend this time to be kept to a minimum, to prevent undesired overfragmentation.

Ligation
Most duplex sequencing methods use adapters with double-stranded molecular barcodes. Here we use a two step ligation method commercially available, which uses adapters with 8 bp molecular barcodes, without T overhangs. These are first ligated on the 3’ end of the insert, and in the second ligation the other adapter strand is added by gap filling to generate the double-stranded molecular barcodes.

Intra-strand repair
Damaged bases are susceptible to being fixed as mutations during the PCR amplification, and therefore detected as false-positive mutations. Consequently, the protocol incorporates a repair step before the indexing PCR to remove 8-oxo-7,8-dihydroguanine (8-oxoG) and uracil resulting from deaminated cytosines.

Quantification of unique ligated DNA fragments with qPCR
This technique allows us to specifically quantify the number of unique DNA fragments that have been successfully ligated with the duplex adapters. This number depends on the input DNA and the efficiency of the duplex library preparation steps, especially the ligation. Basing the calculation of required sequencing reads on the number of ligated molecules estimated in this step avoids oversequencing (too many PCR duplicates sequenced per original DNA fragment) or undersequencing (too few sequencing reads to form consensus duplex reads). However, if the efficiency of the duplex library preparation is constant, the sequencing reads can also be directly calculated based on the input DNA for each sample (Supp. Note).

Increasing duplex depth
The described protocol has a limitation of 250 ng of starting material. This input may not be enough to reach the desired duplex depth. In the cases where more starting material is available, it is possible to start with more than one DNA duplex library per sample. After the

libraries are amplified and uniquely tagged in the indexing PCR with Unique Dual Index (UDI) tags, libraries prepared from the same sample can be multiplexed for hybridization and capture with the desired targeted panel. We highly recommend multiplexing libraries that have a similar number of unique DNA fragments by qPCR and similar library sizes determined by TapeStation, since after pooling the different libraries they will be processed as a unique sample and will be sequenced together. In addition, we recommend only multiplexing up to 2 libraries from the same gDNA of origin.

Hybridization and capture

The minimum input of the amplified library recommended to go into the hybridization reaction is 100 ng. However, we recommend starting with 2-3 µg and do not exceed 6 µg as this could affect the capture's performance, as per manufacturer's instructions. Further recommendations regarding hybridization annealing time and number of post-capture PCR cycles are described in the procedure, although it is highly recommended to test those variables depending on the panel used to capture the target regions of interest. For panels <100kb we recommend two consecutive rounds of hybridization capture, as previously described[25], to obtain sufficient on-target reads.

Sequencing

The required sequencing output is estimated on the basis of the amount of unique DNA ligated fragments, calculated by qPCR (see above), the size of the capture panel, how many PCR duplicates are required on average to produce one consensus duplex read, and the efficiency of the hybridization capture (Supp. Note). As mentioned above, if too few sequencing reads are obtained, one may end up without enough PCR duplicates to form the expected number of duplex consensus reads (ideally, one per ligated DNA fragment). In this case the library is undersequenced, and this problem can be solved by requesting more sequencing reads. On the other hand, if too many sequencing reads are obtained, this will result in more PCR copies per ligated DNA fragment but duplex depth will not increase since the number of possible duplex consensus reads is limited. This problem, known as oversequencing, implies that more resources than necessary have been spent in sequencing the library in question. We have carried out the sequencing of DNA duplex libraries on Illumina NovaSeq 6000 and Illumina NovaSeq X instruments[23].

Building duplex consensus sequences

The deepUMIcaller pipeline receives the reads produced by the sequencer from a library and outputs a list of the somatic mutations present in it. It can be run serially or in parallel (see Supp. Note for how this is achieved in each step of the pipeline). Parallel running is recommended in the case of ultradeep sequencing libraries captured with large sequencing panels (e.g. more than 100 genes).

The deepUMIcaller pipeline requires demultiplexed pairs of FASTQ files from each sample as input. If this is split in multiple pairs of FASTQs (e.g., if the sample has been sequenced on multiple lanes), the pipeline will run in parallel mode, thus speeding up the first steps of raw read processing and alignment. After extracting the duplex tags, we recommend to set an automatic clipping of the first 10 bps of the reads to reduce potential artifacts usually seen at the beginning of the reads. Once the sequencing reads are aligned, the pipeline proceeds to build single and double strand consensus using the *fgbio* suite of tools, and calibrated duplex quality thresholds (see Supp. Note). This process can be executed separately by chromosome to reduce the execution time.

As a final step of the duplex building process, deepUMIcaller writes the BAM files with single strand and duplex consensus reads mapped to the reference genome passing an unambiguity threshold (AS-XS > 50; Supp. Note). This file contains the maximum number of original DNA fragments that are available to identify somatic mutations. If multiple libraries have been prepared from the same sample (see **Increasing duplex depth** section) these can be merged at this step in order to proceed to the variant calling. This merging is implemented in deepUMIcaller and it can be triggered by assigning the same name in the parent_dna column of the input file (see: https://github.com/bbglab/deepUMIcaller/tree/master/docs for more information).

Variant calling

The BAM files containing all unique molecules generated in the consensus building steps can next be filtered to keep only duplex consensus reads with at least 2 copies supporting each of its single strand consensus to proceed to the variant calling. We use VarDictJava in pileup mode for calling all the potential genetic variants and this process is followed by the application of successive filters based on different criteria. These filters add flags to the FILTER column of the VCF file, which can then be used to discard specific sets of potential artifacts (see Supp. Note). These filters include genomic regions of low complexity, problematic mappability, polymorphic positions or genomic sites with recurrent sequencing errors, which are applied using a set of BED and mask files[20].

Variant allele frequency calculation

The quotient between the number of duplex reads with a mutation and the total number of duplex reads covering the genomic position yields the variant allele frequency (VAF) of the mutation on DNA duplex reads, or duplex VAF. Despite calling mutations only if they are supported by both single strand consensus in a duplex consensus read, we can leverage the information from the rest of unique molecules sequenced to refine (with more reference and alternate reads) the calculation of the variant allele frequency of a mutation. Thus, we compute a second VAF value taking into account all unique molecules (including those that do not form a duplex consensus), which we call the all molecules VAF. We compute a third VAF value using exclusively the unique molecules that do not form a duplex consensus. This third metric is called no duplex VAF. The three values are, of course, highly correlated. The all molecules or no duplex VAF metrics always include more reads and are therefore preferred to avoid low-depth related artifacts (Supp. Note).

Quality controls

DeepUMIcaller generates multiple QCs that allow evaluation of key library preparation and sequencing parameters. These include the rate of duplicates of original DNA fragments, the percentage of raw sequencing and duplex consensus reads on target, the overall sequencing depth, and others. In addition, a final quality check for each sample consists of plotting the distribution of the mutations along the positions of the reads to identify potential enrichment on the first bases due to DNA damaged bases at read ends (Figure 4). If the increase of variants remains despite clipping the first 10 bases of reads, the clipping should be extended further.

DeepUMIcaller additional run modes

Several additional parameters can be tuned according to the user's preferences, and there are also multiple possible starting points depending on the input data available. For more information on all possibilities, see the documentation at the GitHub repository: github.com/bbglab/deepUMIcaller/blob/dev/docs. Only starting from the mapping step allows the generation of all QC metrics required for downstream compilation.

Integrated duplex metrics

These scripts compile the QCs from the library preparation with the results of data processing to allow quality assessment of the duplex data generated. The value of these metrics (Supp. Table 2) increases with the availability of more samples so it is critical to compile all the runs together in the templates provided. All the code required for the computation of this and other metrics is available here: https://github.com/bbglab/wetdry-metrics and see Supplementary Note for more information.

Assemble a cohort

For the applications of DNA duplex sequencing described above, ideally, one would assemble a cohort of polyclonal samples with enough DNA to obtain a duplex sequencing depth that is enough to answer the desired research questions. At least 250 ng of DNA are needed to achieve a mean duplex depth of ~3,000x; if deeper DNA duplex sequencing is required, more input DNA should be used. If the aim of the project is to understand the effect of internal or endogenous exposures on mutagenesis and selection in a cohort, samples with different exposures in numbers sufficient to achieve the necessary statistical power, given the expected effect sizes, must be included in the cohort. Groups of samples (for example depending on exposures) can be defined when executing deepCSA, so that analyses are run on each sample, on each group of samples, and across all samples.

Select mutations in well-covered regions across samples

To avoid biases introduced by areas with different coverage across samples, only mutations in well-covered regions across samples are selected for analyses (at least ~200x, but dependent on average depth of the cohort). This is one of the first steps of deepCSA, which generates a file with the selected genomic regions that we refer to as the consensus panel. To carry out this step, the user needs to provide both a VCF file with mutations and a BAM file with the duplex quality reads for each sample. Both files can be directly obtained from the output of deepUMIcaller. If another computational pipeline is used to identify mutations, these files need to be extracted from their outputs and fed to deepCSA.

Annotate and filter mutations

All the mutations provided to deepCSA are annotated using Ensembl VEP and are assigned a consequence type based on their impact on the corresponding MANE transcripts. The sequencing coverage information is also added for each mutation, as well as the flags warning about potentially artifactual mutations (cohort n-rich, repetitive, potential SNP[23]). After all the preprocessing steps, the pipeline applies the desired set of filters (filters_criteria, filter_criteria_somatic; Supp. Note). This includes VAF thresholds to discard rare or private germline variants and flags generated by prior steps of the pipeline. These filtered mutations, together with the depths per position and the consensus panel, constitute the main inputs for all downstream analysis of mutagenesis and selection implemented by deepCSA.

Analysis of mutagenesis

The basic mutagenesis metrics implemented within deepCSA are mutation densities and mutational profiles. The mutation density of a gene in a sample (or across groups of samples) is computed by dividing the total number of somatic mutations with a given consequence (e.g., protein-affecting or non-protein-affecting mutations) that were sequenced in the gene. This total number of sites is calculated as the product of the total number of sites with potentially the same consequence type in the gene, multiplied by the number of times each of them is covered by a duplex consensus read (in Mbps). The mutational profile computed by deepCSA is the classical tri-nucleotide context single base substitution (SBS) profile of 96 channels, resulting from the quotient of the number of tri-nucleotide changes observed by the number of times the given tri-nucleotide was sequenced and re-scaled to the whole genome trinucleotide counts (https://github.com/bbglab/deepCSA/blob/main/docs/output.md). DeepCSA uses the SBS

mutational profile across samples to identify the mutational signatures active in the cohort through HDP[33] and SigProfilerAssignment[15], on the basis of a reference set of signatures provided by the user (e.g., the COSMIC[43] reference).

Selection analyses
The analysis of selection requires tools specifically adapted for duplex sequencing data. To this end, we developed a new method to compute dN/dS-based positive selection signals called omega (ref.[23], Supp. Note) and we adapted OncodriveFML[34] and Oncodrive3D[35] to use position specific weights that ensure a proper correction for the uneven depth per site in the calculation of their background models. The ultradeep sequencing of multiple polyclonal samples results in hundreds of mutations for some genes enabling the possibility of computing selection metrics at subgenic or even site resolution. The omega method can compute the selection signals for each exon, known protein domains or other custom regions. Additionally, we provide estimates for site specific selection, particularly useful for quantifying selection in known hotspots and for natural saturation mutagenesis, where we can estimate the effect of every possible mutation in a given gene. All these measurements of mutagenesis and selection define the mutational and clonal landscape of the tissue/cohort of interest.

Interindividual variability and associations with clinical variables
DeepCSA also includes a regression-based approach to test the association between the exposure to endogenous or exogenous factors, or any other clinical data, and the clonal landscape of samples. This can serve as a first pass exploration of potential relationships (Supp. Note). For a more fine-tuned analysis, a more comprehensive module of data exploration and regressions can be used (https://github.com/bbglab/bbgregressions). This part of the pipeline is, of course, dependent on the user providing metadata of the samples that contain relevant exposures to test (Fig. 1d).

Glossary

Read family: sequencing raw reads derived that are PCR duplicates of the same DNA fragment; it can encompass the duplicates of only one strand or both strands of the fragment, if available.

Single strand consensus reads: read family derived encompassing only one strand of the original DNA fragment, also referred to as single strand consensus (SSC).

Duplex consensus reads: read family encompassing both strands of an original DNA fragment, also referred to as double strand consensus (DSC).

Unique molecules: DNA fragments successfully ligated during duplex library preparation and recovered through computational analyses as duplex consensus reads.

All molecules: set of single strand consensus reads and duplex strand consensus reads representing all the unique DNA fragments that were detected when sequencing a duplex library.

Consensus panel: genomic regions with sufficient duplex sequencing coverage across the majority of samples in a cohort the specific thresholds for the definition of this panel are cohort dependent and can be defined by the user.

Protein affecting mutations: variants that change the sequence of a protein (non-synonymous, truncating, etc), and are potentially subject to selection, as opposed to those (as synonymous or intronic) that are non-protein affecting.

Mutation density: number of mutations identified per megabase sequenced of any relevant genomic region (exon, gene, consensus panel, etc).

Omega: dN/dS-based method to calculate positive selection on the protein affecting mutations of a gene in one sample or in a group of samples; also used for referring to the metric itself.

Clone: Group of cells derived from the same stem cell.

Polyclonal sample: biological sample (e.g., from a human or animal model tissue, organoids, etc) containing several clones.

Pseudo-duplex artifact: variant present in both strands in a DNA fragment prior to ligation which did not arise from a mutational process active in the sample, but as a consequence of DNA damage caused by DNA extraction, fragmentation or other technical process.

Error rate: number of falsely identified mutations (pseudo-duplex artifacts) per base with duplex coverage.

Saturation mutagenesis: high throughput evaluation of the effect of all (or virtually all) mutations in a gene or gene fragment in the development of a phenotype, for example, tumorigenesis.

Natural saturation mutagenesis: approximation to saturation mutagenesis using mutations observed across naturally evolving polyclonal samples (e.g., from human normal tissues).

Materials

Reagents

- DNA of interest

    CRITICAL STEP: It must be purified under non-strand-denaturing conditions.

- Agilent Genomic DNA TapeStation kit (Agilent, cat. no. 5067-5365 and 5067-5366)
- Agilent DNA high-sensitivity D5000 TapeStation kit (Agilent, cat. no. 5067-5592 and 5067-5593)
- Agencourt AMPure XP beads (Beckman Coulter, cat. no. A63880)
- Nuclease-free water (Invitrogen, cat. no. 10977035)
- EB buffer (Qiagen, cat. no. 19086)
- Ethanol absolute (Reag. USP, Ph. Eur.) for analysis, ACS, ISO (PanReac AppliChem, cat. no. 131086)

    Caution: Ethanol is flammable. Keep it away from flames.

- Qubit reagents: Broad Range (Invitrogen, cat. no. Q33265) and High Sensitivity (Invitrogen, cat. no. Q33230)
- NEBNext UltraShear® (NEB, cat. no. M7634S)
- xGen™ cfDNA & FFPE DNA Library Prep v2 MC (IDT, cat. no. 10010206 for 16rxn or 10010207 for 96rxn)
- Fpg (NEB, cat. no. M0240S)
- UDG (NEB, cat. no. M0280S)
- KAPA Library Quantification Kits - Complete kit (Universal) (Roche, cat. no. KK4824)
- qPCR primers, HPLC purified (Sigma-Aldrich, see below for sequences)

    Forward: 5′-ACACTCTTTCCCTACACGAC-3′
    Reverse: 5′-GTGACTGGAGTTCAGACGTG-3′

- xGen™ UDI Primers (IDT, cat. no. 10005975 for 16rxn or 10005922 for 96rxn)
- xGen™ Hybridization and Wash Kit, 16 rxn (IDT, cat. no. 1080577)
- xGen™ Universal Blockers TS, 16 rxn (IDT, cat. no. 1075474)
- xGen™ Custom Hyb Panel (IDT) or equivalent (see Supp. Note)
- xGen™ Library Amplification Primer Mix, 16 rxn (IDT, cat. no. 1077675)

Equipment

- Eight-well PCR strip tubes 0.2 mL (NerbePlus, cat. no. 04-032-0500)
- Eight-well PCR strip dome caps (NerbePlus, cat. no. 04-042-0500)
- Microcentrifuge tubes (1.5 mL, Eppendorf)
- Filtered pipette tips
- DynaMag-96 side magnetic plate separator (Invitrogen, cat. no. 12331D)
- Thermocycler (BioRad T100 or Applied Biosystems MicroAmp)
- qPCR thermocycler (Life Technologies QS6Flex or QS6Pro or similar)
- Agilent TapeStation 4200 (Agilent, cat. no. G2991BA) or similar
- BioShake iQ Thermal Mixer (Bulldog Bio)
- Illumina NovaSeqX or NovaSeq6000 and associated equipment (Illumina)

Hardware

All computational steps described between 183 to 201 of the protocol were executed using a high-performance computing cluster. A multi-threaded processor is required to run most of the steps, particularly for deepUMIcaller. The total amount of required resources is highly dependent on the amount of sequencing reads and the size of the targeted regions, both of which have a big impact on the total size of the temporary directory generated during the deepUMIcaller run. The temporary files generated by deepUMIcaller and stored in the work directory (see Nextflow instructions) typically occupy 3 to 5 times the size of the input FASTQs, and the final output directory will have a size corresponding to 10 to 20 % of the initial FASTQ size. The temporary directories for deepCSA are dependent on the number of samples and panel size, but the expected sizes are within a few GBs (up to 20 max).

Software

- Nextflow (version >=25.04)
- Containerization software (Docker, singularity, apptainer, …)
- Conda

Data files

- Reference genome file for the species of interest. For the human genome GRCh38 download it from: ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/000/001/405/GCA_000001405.15_GRCh38/seqs_for_alignment_pipelines.ucsc_ids/GCA_000001405.15_GRCh38_no_alt_analysis_set.fna.gz
- BED file (BED4-5-6 format) with the targeted regions of the genome. For optimal on target metrics move the ends of all intervals to cover 250bp flanking regions. See Equipment setup.

- (optional) Additional files can be downloaded for adding filters to mutation calls.
    - **Low complexity:** This file identifies repetitive genomic regions from RepeatMasker annotations that can cause alignment artifacts and variant calling errors. RepeatMasker output files for commonly used human reference genomes can be downloaded from the RepeatMasker human pages.
    - **Low mappability:** This file comes from an article describing regions with mappability problems. The ENCODE Blacklist: Identification of Problematic Regions of the Genome. Use the appropriate file for your genome version.
    - (only human) **Nanoseq genomic masks:** These files identify sites overlapping common SNPs and noisy or variable genomic regions, as described in Abascal et al, 2021 and used in the Nanoseq pipeline. Both files are available for GRCh37 and GRCh38 at the shared folder from the Martincorena Group, at the Wellcome Sanger Institute. Two BED files are available to be used:
        - Nanoseq SNP: Common SNP positions that should be excluded from analysis
        - Nanoseq Noise: Regions with high noise or variability
- See here for more information on additional datasets to be downloaded for deepUMIcaller and deepCSA as convenient:
  https://github.com/bbglab/deepUMIcaller/tree/master/docs
  https://github.com/bbglab/deepCSA/tree/main/docs

Reagent setup

Ethanol solution, 80% (vol/vol)

Add 400 µL of ddH2O to 1.6 mL of 100% ethanol. Prepare this solution fresh before each bead cleanup. Solution can be stored at room temperature.

qPCR primers

Dissolve each respective PCR primer in ddH2O to a final concentration of 100 µM each. Prepare a dilution to the final working concentration of 10 µM and store at -20 ºC until further use. Primers can be stored at this temperature for 6–12 months.

AMPure XP magnetic beads

The beads are normally stored at 4 °C. Allow the beads to warm to room temperature before use.

CRITICAL STEP: The beads will not function properly if they are not at room temperature.

When removing the supernatant from the last ethanol wash, do not allow beads to over-dry. This phenomenon can manifest with the appearance of cracks in the bead pellet.

CRITICAL STEP: The DNA recovery will be directly affected if beads are over-dried.

Equipment setup

Collection of experimental protocol metrics

Before starting the experimental protocol, download the excel file wetlab_qc_metrics.xlsx available in the templates directory of the GitHub repository (https://github.com/bbglab/wetdry-metrics). This will allow you to collect the main metrics of the experimental protocol and have them ready to be provided to the downstream metrics compilation step.

Installation of Nextflow

Follow the Nextflow installation instructions here: https://www.nextflow.io/docs/latest/install.html#install-nextflow

```
curl -s https://get.nextflow.io | bash
chmod +x nextflow
mkdir -p $HOME/.local/bin/
mv nextflow $HOME/.local/bin/
```

Preparation of the Nextflow config file

Configure your available HPC server Nextflow executor profile. Prepare an `executor.config` file with the following information.

```
process {
   executor = 'slurm'
   queue = '<your queues>'

   errorStrategy = 'retry'
   maxRetries = 3

}

executor {
   queueSize = 100
}
```

Preparation of the targets BED file

For optimal computation of on target metrics extend the ends of all intervals to cover the 250 bp flanking regions. You can use the following command:

```
awk -v OFS='\t' '{$2=$2-2;$3=$3+2; print $0}' panel.bed6.bed > panel.expanded250.bed6.bed
```

deepCSA datasets creation

This process with examples is documented in : https://github.com/bbglab/deepCSA/blob/main/docs/usage.md

- Download Ensembl VEP cache from version 111 (see: Ensembl VEP docs).
- Download of additional specific datasets:

    - Generation of Oncodrive3D datasets
      (see: Oncodrive3D repo datasets building process)
    - CADD scores
      (see: CADD downloads page "All possible SNVs of GRCh38/hg38" file)
    - COSMIC signatures
      (i.e. COSMIC signatures downloads page (select context size = 96 and your desired species of interest))

- Protein domain definition file
    - This file must contain the following columns:
        - Ens_Transcr_ID: Ensembl Transcript ID
        - Begin: Starting protein position of the domain or subgenic region of interest.
        - End: Last protein position of the domain or subgenic region of interest.
        - NAME: Name of the subgenic region of interest (without spaces).
    - This information for the human genome is available at: https://github.com/bbglab/bbgdomains

Procedure

Fragmentation and cleanup

Timing 1 h 30 min

1. Ensure that the NEBNext UltraShear Reaction Buffer is completely thawed and quickly vortex to mix. Place on ice until use.
2. Vortex the NEBNext UltraShear for 5–10 s prior to use and place on ice.

   CRITICAL STEP: It is important to vortex the enzyme mix prior to use for optimal performance.
3. Add the following components to a 0.2 mL thin wall PCR tube on ice:

Reagents	Volume
DNA (up to 250 ng)	1–26 µL
NEBNext UltraShear Reaction Buffer	14 µL
NEBNext UltraShear	4 µL
Sterile Water	variable
Final Volume	44 µL

   CRITICAL STEP: If processing more than one sample, make a mastermix with 10% overage.
4. Vortex the reaction for 5–10 s and briefly spin in a microcentrifuge.
5. In a thermal cycler, preheated and with the heated lid set to 75 °C, run the following program:

Fragmentation program		
Step	Temperature (ºC)	Time (min)
Fragmentation	37	5-45

6. Transfer samples immediately to ice, and assess fragment size in TapeStation using 1 µL diluted 1:5 in sterile water and following manufacturers instructions.
7. If the percentage of DNA fragments in the 50-650bp region assessed by TapeStation is below 30%, add extra time on the Fragmentation program and check again.

   TROUBLESHOOTING

8. Once samples have reached this 30% threshold of DNA fragments within 50-650bp, transfer samples to a thermal cycler, preheated and with the heated lid set to 75 °C, and run the following program:

Inactivation program		
Step	Temperature (ºC)	Time (min)
Inactivation	65	15
Hold	4	∞

CRITICAL STEP: Make sure the thermocycler block is already at 65 ºC when you place the samples.

9. For easier handling of the sample and faster collection of the beads to the magnet, it is recommended to dilute the sample with sterile water. Add 6 µL nuclease-free water (or up to 50 µL) to the sample.

10. Thoroughly resuspend AMPure XP beads before use, then add 90 µL of AMPure beads (1.8X volume) to each tube and pipette 10 times to thoroughly mix.

11. Incubate the samples at room temperature for 10 min.

12. Place the samples on a magnet and wait for the liquid to clear completely or at least for 2 min.

13. Remove and discard the cleared supernatant making sure not to remove any beads.

14. Keeping the samples on the magnet, add 160 µL of 80% ethanol and incubate for 30 s.

15. Remove and discard the supernatant.

16. Repeat steps 14 and 15 for a total of 2 washes.

17. Use a P10 pipette tip to remove any residual ethanol.

18. Dry the beads at room temperature for 1–3 min.

19. Remove the samples from the magnet, then add 52 µL of Buffer EB and resuspend beads fully.

20. Allow the samples to incubate at room temperature for 5 min to elute DNA off the beads.

21. Place the samples on a magnet and wait for the beads to be cleared from the liquid (approximately 1–2 min).

22. Carefully transfer 52 µL of the cleared liquid containing the eluted DNA into a new well.

23. Quantify 1 µL the fragmented product with a Qubit HS assay and assess its size distribution using 1 µL in an Agilent TapeStation 4200 or similar. Fragmented products at this point should be around 250-350 bp. Proceed to the next step with the remaining 50 µL.

End repair and cleanup

Timing 1 h

24. For each sample, make the following End Repair Master Mix:

Reagents	Volume
End Repair Buffer	6 µL
End Repair Enzyme	3 µL
Final Volume	9 µL

CRITICAL STEP: If there is precipitate in the End Repair Buffer, vortex until the precipitate becomes clear in solution.

CRITICAL STEP: The resulting Master Mix is viscous and requires careful pipetting.

25. Pulse-vortex the master mix for 10 s, then briefly centrifuge. Keep the master mix on ice.

26. Add 9 µL of End Repair Master Mix to each well. Using a pipette set to 40 µL, pipette 10 times to mix.

27. In a thermal cycler, with the heated lid set to OFF, or to 40 °C, run the following program:

End Repair program		
Step	Temperature (ºC)	Time (min)
End repair	20	30
Hold	4	∞

While the end repair program runs, make the Ligation 1 Master Mix in preparation for the Post-End Repair cleanup steps.

Reagents	Volume
Ligation 1 Buffer	25 µL
Ligation 1 Adapter	2 µL
Ligation 1 Enzyme	3 µL
Final Volume	30 µL

Caution: Use extra caution when handling Ligation 1 and Ligation 2 adapter tubes: never handle Ligation 2 adapter before or during Ligation 1 Master Mix setup and handling. Trace

contamination of Ligation 2 adapter into Ligation 1 adapter has been shown to induce adapter-dimer formation.

28. Pulse vortex the master mix for 10 s, then briefly centrifuge. Keep the master mix on ice until ready to use.

29. After the End Repair program reaches 4 °C, proceed immediately to the bead cleanup.

    CRITICAL STEP: Before starting cleanup, make sure the Ligation 1 Master Mix has already been prepared.

30. Thoroughly resuspend AMPure XP beads before use, then add 147.5 µL of AMPure beads (2.5X volume) to each well and pipette 10 times to thoroughly mix.

31. Incubate the samples at room temperature for 10 min.

32. Place the samples on a magnet and wait for the liquid to clear completely or at least for 2 min.

33. Remove and discard the cleared supernatant making sure not to remove any beads.

34. Keeping the samples on the magnet, add 160 µL of 80% ethanol and incubate for 30 s.

35. Remove and discard the supernatant.

36. Use a P10 pipette tip to remove any residual ethanol.

37. Dry the beads at room temperature for 1–3 min.

Ligation 1

Timing 30 min

38. Remove the samples from the magnet, then add 30 µL Ligation 1 Master Mix.

39. Pipette mix a minimum of 10 times.

    CRITICAL STEP: Make sure the samples are thoroughly mixed and that the beads are fully resuspended before proceeding.

40. In a thermal cycler, preheated and with the heated lid set to 70 °C, run the following program:

Ligation 1 program		
Step	Temperature (ºC)	Time (min)
Ligation	20	15
Inactivation	65	15
Hold	4	∞

PAUSE POINT: The samples can temporarily remain at 4 °C (no more than 2 hours). It is normal for beads to settle during this reaction.

Ligation 2 and cleanup

Timing 1 h

41. For each sample, prepare the Ligation 2 Master Mix.

Reagents	Volume
Ligation 2 Buffer	4.5 µL
Ligation 2 Adapter	4 µL
Ligation 2 Enzyme A	0.5 µL
Ligation 2 Enzyme B	1 µL
Final Volume	10 µL

42. Pulse-vortex the master mix for 10 s, then briefly centrifuge. Keep the master mix on ice until ready to use.

43. Add 10 µL of the Ligation 2 Master Mix to each well.

44. Using a pipette set to 35 µL, pipette 10 times to mix.

    CRITICAL STEP: Make sure the samples are thoroughly mixed and that the beads are fully resuspended before proceeding.

45. In a thermal cycler, preheated and with the heated lid set to 70 °C, run the following program:

Ligation 2 program		
Step	Temperature (ºC)	Time (min)
Ligation	65	30
Hold	4	∞

46. Add 100 µL of PEG/NaCl (2.5X volume) to each well, then pipette 10 times to mix.

47. Incubate the samples at room temperature for 10 min.

48. Place the samples on a magnet and wait for the liquid to clear completely or at least for 2 min.

49. Keeping the samples on the magnet, add 160 µL of 80% ethanol and incubate for 30 s.

50. Remove and discard the supernatant.
51. Repeat steps 49 and 50 for a total of 2 washes.
52. Use a P10 pipette tip to remove any residual ethanol.
53. Dry the beads at room temperature for 1–3 min.
54. Remove the samples from the magnet, then add 25 µL of Buffer EB and resuspend beads.
55. Allow the samples to incubate at room temperature for 5 min to elute DNA off the beads.
56. Place the samples on the magnet and wait for the beads to be cleared from the liquid (approximately 1–2 min).
57. Carefully transfer 25 µL of the cleared liquid containing the eluted DNA into a new tube.

PAUSE POINT: Samples can be stored at –20 °C overnight.

Intra-strand repair and cleanup

Timing 1 h 30 min

58. For each sample, prepare the Intra-strand repair Master Mix.

Reagents	Volume
NEB1 buffer	3 µL
Fpg	1 µL
UDG	1 µL
Final Volume	5 µL

59. Pulse-vortex the master mix for 10 s, then briefly centrifuge. Keep the master mix on ice until ready to use.
60. Add 5 µL of the Intra-strand repair Master Mix to each sample.
61. Using a pipette set to 25 µL, pipette 10 times to mix.
62. If necessary, briefly centrifuge to collect contents to the bottom of the wells.
63. In a thermal cycler, preheated and with the heated lid set to 40 °C, run the following program:

Intra-strand repair program		
Step	Temperature (ºC)	Time (min)

Intra-strand repair	37	60
Hold	4	∞

64. Once incubation is over, add 20uL of nuclease-free water to each reaction and proceed to the bead cleanup.

65. Thoroughly resuspend AMPure XP beads before use, then add 45 µL (0.9X volume) of Agencourt AMPure XP beads to each sample.

66. After adding the beads, mix thoroughly and incubate for 10 min.

67. Place the samples on a magnet until the supernatant is clear (2–5 min).

68. Remove supernatant without disturbing the beads.

69. While keeping the samples on the magnet, add 125 µL of 80% ethanol, then incubate for 30 s.

70. Remove and discard the supernatant.

71. Repeat steps 69 and 70 for a total of 2 washes.

72. Use a P10 pipette tip to remove any residual ethanol.

73. Allow the beads to air dry for 1–3 min. Do not over-dry the beads.

74. Remove the samples from the magnet and elute in 23 µL of Buffer EB. Mix thoroughly.

75. Incubate for 5 min at room temperature.

76. Place the samples on a magnet until the supernatant is clear (1–2 min).

77. Transfer 23 µL of eluate to a fresh tube making sure that no beads are carried over.

78. Quantify 1 µL the product with a Qubit HS assay and assess its size distribution using 1 µL in an Agilent TapeStation 4200 or similar. Fragmented products at this point should be around 350-450 bp. Proceed to the next step (Indexing PCR) with the remaining 20 µL.

Quantification of unique ligated DNA fragments with qPCR

Timing 3 h

79. Taking into account the previous quantification, dilute samples accordingly by doing serial dilutions in nuclease-free water, knowing that the six standards are in the range of 5.5 – 0.000055 pg/µL. The recommendation would be to dilute samples at least to around ~1 pg/µL and test two different dilutions. This quantification is not strictly required before proceeding to the next step (Indexing PCR). It is performed to inform about the number of ligated molecules to properly allocate raw sequencing reads when the duplex library is ready for sequencing (see Supp. Note).

80. For each reaction (which will be in triplicate), make the following mastermix adding an overage of 2 reactions:

Reagents	Volume
KAPA MasterMix 2X	5 µL
Illumina primer mix 10X	1 µL
10 µM qPCR primer Fw	0.2 µL
10 µM qPCR primer Rv	0.2 µL
Nuclease-free water	1.6 µL
Final Volume	8 µL

81. Set up triplicate 10 µL qPCR reactions (8 µL master mix, 2 µL sample/standard) in a 384 well plate. Add also a negative control with nuclease-free water.

    PAUSE POINT: Samples can be stored at –20 ºC overnight.

82. Run samples on a qPCR thermocycler and run the following programme:

qPCR program for Absolute Quantification analysis					
Step	Ramp rate (ºC/s)	Temperature (ºC)	Time	Acquisition	Cycles
Hold stage	1.6	50	2 min		1
Initial denaturation	1.6	95	5 min		1
Denaturation	1.6	95	30 s		40
Annealing/Extension/ Data acquisition	1.6	60	45 s	Yes	
Melt curve analysis (Continuous)	1.6	95	15 s		1
	1.6	60	1 min		
	0.05	95	15 s	Yes	

TROUBLESHOOTING

Indexing PCR and cleanup

Timing 1 h 30 min

CRITICAL STEP: Sample index barcodes are introduced during PCR; double-check that a unique primer pair is used for each sample.

83. Add 5 µL of xGen UDI Primer Pairs to each well.
84. Add 25 µL of xGen 2x HiFi PCR Mix to each well, then pipette 10 times to mix.
85. Then briefly centrifuge to recover any volume from the walls into the bottom of the wells.
86. In a thermal cycler, preheated and with the heated lid set to 105 °C, run the following program:

Indexing PCR program			
Step	Temperature (ºC)	Time	Cycles
Polymerase activation	98	45 s	1
Denature, Anneal, Extend	98	15 s	10
	60	30 s	
	72	1 min	
Final Extension	72	3 min	1
Hold	4	∞	

87. Thoroughly resuspend AMPure XP beads before use, then add 45 µL of AMPure beads (0.9X ratio), to each well, then pipette 10 times to thoroughly mix.
88. Incubate the samples at room temperature for 5 min.
89. Place the samples on a magnet and wait for the liquid to clear completely or at least for 2 min.
90. Remove and discard the cleared supernatant; make sure not to remove any beads.
91. Keeping the samples on the magnet, add 160 µL of 80% ethanol, then incubate for 30 s.
92. Remove and discard the supernatant.
93. Repeat steps 91 and 92 for a total of 2 washes.
94. Use a P10 pipette tip to remove any residual ethanol.
95. Dry the beads at room temperature for 1–3 min.
96. Remove the samples from the magnet, then add 32 µL of Buffer EB. Gently vortex (use 70% vortex capacity) to resuspend beads.

97. Allow the samples to incubate at room temperature for 5 min to elute DNA off beads. Then place the samples on a magnet and wait for the liquid to clear completely for 1–2 min.

98. Carefully transfer 32 µL of eluted DNA into a new tube.

99. Quantify 1 µL the PCR product with a Qubit HS assay and assess its size distribution using 1 µL in an Agilent TapeStation 4200 or similar. Fragmented products at this point should be around 350-650 bp.

    TROUBLESHOOTING

Hybridization

Timing 1 h (handling)

100. Use half of the Indexing PCR product for capture and keep the other half at -20 ºC as a back up sample. If capture fails, the second half can be used for recapture. If samples are being multiplexed, pool half of each Indexing PCR product combined into a single tube.

101. Add 7.5 µL of Human Cot DNA.

102. Add 1.8X volume of AMPure XP beads.

103. Vortex thoroughly to mix. Adjust the settings to prevent any splashing onto the seal or cap.

104. Incubate for 10 min at room temperature.

105. Incubate the samples on the magnet for at least 2 min or until supernatant is clear.

106. Remove and discard the supernatant. Keeping the samples on the magnet, add 80% ethanol to cover the surface of the beads. Incubate for 30 s without disturbing the beads.

107. Remove and discard the supernatant, then repeat another ethanol wash for a total of two washes.

108. Allow the beads to air dry for approximately 3 min.

    CRITICAL STEP: Do not over-dry.

109. Add these components to the tube to make the Hybridization Reaction Mix:

Reagents	Volume
xGen 2X Hybridization Buffer	9.5 µL
xGen Hybridization Buffer Enhancer	3 µL
xGen Universal Blockers	2 µL
xGen Hyb Panel	4 µL

Final Volume	18.5 µL

110. Remove the samples from the magnet and vortex to mix. Ensure that the beads are fully resuspended.

111. Incubate for 5 min at room temperature.

112. After incubation, place the samples on a magnet for 5–10 min or until the supernatant is clear.

113. Transfer 18 µL (or all the volume you are able to recover without taking any beads) of the supernatant to a new well, where the hybridization will occur. Make sure to avoid bead carryover during the transfer process.

114. Vortex briefly to mix and spin down.

115. Incubate the samples in a thermal cycler, preheated and with the heated lid set to 100 °C, run the following program:

Hybridization program		
Step	Temperature (ºC)	Time
Denaturing	95	30 s
Annealing	65	16-20 h
Hold	65	∞

CRITICAL STEP: The annealing time should be optimized per panel size, for reference, for relatively big panels (>17,000 probes) we recommend doing an overnight annealing (equivalent to a minimum of 16 h incubation).

Streptavidin bead wash

Timing 30 min

CRITICAL STEP: Make sure the Dynabeads M270 Streptavidin beads, which are stored at 4 °C, have been at RT for a minimum of 30 min before performing the washes.

116. Prepare the following buffer to create a 1X working solution:

Reagents	Volume
xGen 2X Bead Wash Buffer	160 µL
Nuclease-free water	160 µL
Final Volume	320 µL

117. Prepare the following Bead Resuspension Mix in a low-bind tube:

Reagents	Volume
xGen 2X Hybridization Buffer	8.5 µL
xGen Hybridization Buffer Enhancer	2.7 µL
Nuclease-free water	5.8 µL
Final Volume	17 µL

118. Add 50 µL of capture bead-mix per sample to 1.5 mL tube.

119. Collect the beads on a magnet and remove the supernatant.

120. Add twice the volume of bead-mix of 1X Bead Wash Buffer to the tube. For example, if 500 µL of bead-mix is being washed, use 1 mL of Bead Wash Buffer.

121. Remove from the magnet and briefly vortex the beads to resuspend, then spin down.

122. Collect the beads on a magnet and remove the supernatant.

123. Repeat steps 120−122 twice more for a total of 3 washes.

124. Following the last wash, collect the beads on the magnet and remove the supernatant.

125. Remove from the magnet and resuspend the beads in 17 µL of the Bead Resuspension Mix per reaction. Mix solution thoroughly pipetting until homogenous. If necessary, spin down briefly at low speed to avoid bead pelleting.

Capture

Timing 1 h

CRITICAL STEP: Before adding the beads to your samples, vortex or pipette mix the washed bead-mix to ensure uniformity.

126. Remove samples from the thermal cycler and stop the Hybridization program. Immediately start the Wash program.

127. Spin down the tubes briefly to ensure no sample or condensation is left on the seal or tube cap.

128. Ensure washed capture beads are homogeneously resuspended immediately before aliquoting.

129. Add 17 µL of resuspended washed capture beads to each sample at room temperature.

130. Gently vortex the samples to resuspend the mixture.

131. Incubate the samples in a thermal cycler, preheated and with the heated lid set to 70 °C, and run the following program:

CRITICAL STEP: Reduce the lid temperature to 70 °C for the Wash program.

CRITICAL STEP: Every 10–12 min, remove the tube from the thermal cycler and gently vortex to ensure the sample is fully resuspended.

Wash program	
Temperature (ºC)	Time
65	45 min

132. At the end of the 45 min, take the sample off the thermal cycler and proceed immediately to Heated washes.

Heated washes

Timing 30 min

CRITICAL STEP: Inspect the xGen 10X Wash Buffer I. If it appears cloudy, warm solution to 65 °C and mix until homogeneous. Allow it to cool to RT after heating. Use all Wash buffers at RT.

133. Dilute the following xGen buffers to create 1X working solutions:

Reagents	Volume	Nuclease-free water	Final volume
xGen 10X Wash Buffer 1	252 µL	28 µL	280 µL
xGen 10X Wash Buffer 2	144 µL	16 µL	160 µL
xGen 10X Wash Buffer 3	144 µL	16 µL	160 µL
xGen 10X Stringent Wash Buffer	288 µL	32 µL	320 µL

134. Aliquot 110 µL per sample of the 1X Wash Buffer 1 into a separate tube. Heat tubes to 65°C in the thermocycler with the Wash program. The remaining solution should be kept at room temperature.

135. Aliquot 160 µL each per sample into two tubes (with 160 µL each) of the 1X Stringent Wash Buffer. Heat tubes to 65°C in the thermocycler with the Wash program.

CRITICAL STEP: The 1X Wash Buffer 1 (110 µL aliquot) and the 1X Stringent Wash Buffer (both aliquots) should be in the 65 °C thermocycler with the Wash program for at least 15 min. We recommend starting this incubation at the same time as the bead capture, so that the buffers will be at the correct temperature when needed.

136. After capture, spin down and place the samples on a magnet to collect the beads.

137. Transfer 100 µL of heated Wash Buffer 1 to the sample, then pipette mix 10 times, being careful to minimize bubble formation.

138. Place the tube on a magnetic rack for 1 min. Remove the supernatant.

139. Remove the tube from the magnet and add 150 µL of heated Stringent Wash Buffer to the sample. Pipette mix 10 times, being careful to not introduce bubbles.

140. Incubate in the thermocycler at 65 °C for 5 min.

141. Place the tube on a magnetic rack for 1 min. Remove the supernatant.

142. Remove the tube from the magnet and add 150 µL of heated Stringent Wash Buffer to the sample. Pipette mix 10 times, being careful to not introduce bubbles.

143. Incubate in the thermocycler at 65 °C for 5 min.

144. Place the tube on a magnetic rack for 1 min. Remove the supernatant. Use a P10 to remove any residual liquid.

Room temperature washes

Timing 30 min

145. Add 150 µL of Wash Buffer 1 equilibrated to room temperature.

146. Vortex thoroughly until fully resuspended.

147. Incubate for 2 min while alternating between vortexing for 30 s and resting for 30 s, to ensure the mixture remains homogenous.

148. At the end of the incubation, briefly centrifuge the tube.

149. Place on the magnet for 1 min.

150. Remove the supernatant. Add 150 µL of Wash Buffer 2.

151. Vortex thoroughly until fully resuspended.

152. Incubate for 2 min while alternating between vortexing for 30 s and resting for 30 s, to ensure the mixture remains homogenous.

153. At the end of the incubation, briefly centrifuge the tube.

154. Place on the magnet for 1 min.

155. Remove the supernatant. Add 150 µL of Wash Buffer 3.

156. Vortex thoroughly until fully resuspended.

157. Incubate for 2 min while alternating between vortexing for 30 s and resting for 30 s, to ensure the mixture remains homogenous.

158. At the end of the incubation, briefly centrifuge the tube.

159. Place the sample tube on the magnet for 1 min.

160. Remove and discard the supernatant.

161. Use a P10 to remove any residual liquid.
162. Add 20 µL of Nuclease-Free Water to each capture.
163. Pipette mix 10 times to resuspend any beads stuck to the side of the tube.

Post-capture PCR and cleanup

Timing 1 h

164. In a tube, prepare the Amplification Reaction Mix:

Reagents	Volume
xGen 2x HiFi PCR Mix	25 µL
xGen Library Amplification Primer Mix	1.25 µL
Nuclease-Free Water	3.75 µL
Final Volume	30 µL

165. Add 30 µL of Amplification Master Mix to each sample for a final reaction volume of 50 µL.
166. Then gently vortex to thoroughly mix the reaction.
167. If necessary, briefly centrifuge to collect contents to the bottom of the wells.
168. Place the samples in a thermal cycler, and run the following program with the lid temperature set to 105 °C:

Post-capture PCR program			
Step	Temperature (ºC)	Time	Cycles
Polymerase activation	98	45 s	1
Denature, Anneal, Extend	98	15 s	variable
	60	30 s	
	72	30 s	
Final Extension	72	1 min	1
Hold	4	∞	

CRITICAL STEP: The number of PCR cycles should be optimized per panel size. For reference for a panel size of >17,000 probes we recommend to start with 13 cycles of amplification.

169. Add 45 µL of AMPure beads (0.9X ratio), to each sample, then pipette 10 times to thoroughly mix.

170. Incubate the samples at room temperature for 5 min.

171. Place the samples on a magnet and wait for the liquid to clear completely or at least for 2 min.

172. Remove and discard the cleared supernatant, and make sure not to remove any beads.

173. Keeping the samples on the magnet, add 160 µL of 80% ethanol, then incubate for 30 s.

174. Remove and discard the supernatant.

175. Repeat steps 173 and 174 for a total of 2 washes.

176. Use a P10 pipette tip to remove any residual ethanol.

177. Dry the beads at room temperature for 1–3 min.

178. Remove the samples from the magnet, then add 22 µL of Buffer EB.

179. Gently vortex (use 70% vortex capacity) to resuspend beads.

180. Allow the samples to incubate at room temperature for 5 min to elute DNA off beads. Then place the samples on a magnet and wait for the liquid to clear completely for 1–2 min.

181. Carefully transfer 22 µL of eluted DNA into a new tube.

182. Quantify 1 µL the PCR product with a Qubit HS assay and assess its size distribution using 1 µL in an Agilent TapeStation 4200 or similar. Final libraries should yield >15-20 ng/µL and fragment distribution should be between 350-650 bp. At this point libraries are ready for sequencing in an Illumina platform or equivalent.

    TROUBLESHOOTING

Sequencing data processing: consensus building, QCs and variant calling

There is no limit on the number of samples to be processed together and these do not necessarily need to come from a single experiment. Our recommendation is that for libraries larger than ~800kb targeted panel at >2,000x target depth no more than 10 samples are run at once. A detailed list of all the internal processing steps of deepUMIcaller can be found in the extended protocol in the Supplementary Note.

183. The reference genome needs to be uncompressed and indexed, which can be achieved using the following commands:

```
gunzip <genome>.fa.gz
bwa index <genome>.fa
```

```
samtools faidx <genome>.fa
gatk-launch CreateSequenceDictionary -R <genome>.fa
```

184. Clone the GitHub repository: bbglab/deepUMIcaller and move inside the directory.

```
git clone https://github.com/bbglab/deepUMIcaller.git
cd deepUMIcaller
```

185. Generate the low complexity file from the RepeatMasker information. The script accepts RepeatMasker .out files (compressed or uncompressed) and generates a BED format file containing filtered repetitive regions suitable for variant calling exclusion. This can be done for any species.

```
python assets/generate_low_complex_rep_bed.py hg38.fa.out.gz
low_complexity_regions.bed
```

186. Prepare your input.csv file. For each sample provide: sample (sample name), fastq_1, fastq_2 and read_structure. The sample name can be composed of alphanumerical characters, '-' and '_'. The read structure has to follow the *fgbio* guidelines (https://github.com/fulcrumgenomics/fgbio/wiki/Read-Structures). See example of a file below:

```
sample,fastq_1,fastq_2,read_structure
sample1,sample1_R1.fastq.gz,sample1_R2.fastq.gz,8M1S+T 8M1S+T
sample2,sample2_R1.fastq.gz,sample2_R2.fastq.gz,8M1S+T 8M1S+T
sample3,sample3_R1.fastq.gz,sample3_R2.fastq.gz,8M1S+T 8M1S+T
…
```

187. Create a params.deepUMIcaller.yml file with all the customized parameters. Special attention should be given to the required input parameters, the desired outputs and duplex quality standards. Full documentation of the parameters can be found in the extended protocol and in the deepUMIcaller repository.

```
input                   : <input.csv>
ref_fasta               : <reference Fasta file>
targetsfile             : <BED: targeted regions>
filter_min_reads_duplex : 4 2 2
perform_qcs             : true
```

188. If available, add the BED files required for filtering the mutations in the post-processing steps to the params.deepUMIcaller.yml file.

```
low_complex_file        : null
low_mappability_file    : null
nanoseq_snp_file        : null
nanoseq_noise_file      : null
```

189. You should now proceed to launch deepUMIcaller via the following Nextflow command:

```
nextflow run bbglab/deepUMIcaller \
  -profile singularity \
  -c executor.config \
```

```
-params-file params.deepUMIcaller.yml
```

TROUBLESHOOTING

Sequencing metrics compilation

190. (optional; highly recommended) Collect the library prep metrics in the standardized format provided via the template (*wetlab_qc_metrics.xlsx*) available in the GitHub repository. (https://github.com/bbglab/wetdry-metrics)

191. Clone the GitHub repository: bbglab/WetDryMetrics and move inside the directory.

```
git clone https://github.com/bbglab/wetdry-metrics.git
cd wetdry-metrics
```

192. Update the `runs_list.json` file, also provided as a template, with the new information from your last run. Add a name of the run and the path to its deepUMIcaller output directory.

193. Create an environment with the required dependencies:

```
conda env create -f environment.yml
conda activate metrics-env
```

194. Run the metrics compilation script with the following command:

```
cd scripts
python BuildWetDryMetrics.py --runs_list runs_list.json \
      --output_data_dir <output_directory> \
      --wetlab_qc_metrics_file wetlab_qc_metrics.xlsx \
      --plot_dir <path_to_plot_folder>
```

`--wetlab_qc_metrics_file` flag is optional depending on the availability of library prep. metrics.

`--plot_dir` flag automatically triggers the plotting of some summary metrics by batch and is also optional.

Check the printed progress for details on the steps executed.

TROUBLESHOOTING

195. Check the main metrics as described in Supplementary Table 2 to ensure that the run was successful.

Mutations data analysis

There are multiple input options for running deepCSA and these are further documented in the repository. Here we will cover the continuation from the deepUMIcaller outputs generated in the

previous run. A detailed list of all the internal processing steps of deepCSA can be found in the extended protocol in the Supplementary Note.

196. Prepare the input csv file for deepCSA by taking the vcf files in mutations_vcf and the BAM files in sortbamduplexcons. See example of a file below:

```
sample,vcf,bam
donor1_BDO,donor1_BDO.filtered.vcf,donor1_BDO.sorted.bam
donor1_BTR,donor1_BTR.filtered.vcf,donor1_BTR.sorted.bam
donor2_BDO,donor2_BDO.filtered.vcf,donor2_BDO.sorted.bam
donor2_BTR,donor2_BTR.filtered.vcf,donor2_BTR.sorted.bam
...
```

    In case all the samples of a cohort were run in a single run of deepUMIcaller, these files can be obtained at: pipeline_info/deepCSA_input_template.csv .

197. Create a params.deepCSA.yml file with all the customized parameters.
    Special attention should be put to the required input parameters and the sequencing depth dependent parameters. The consensus panel depth should be big enough to include only regions that are properly sequenced across samples. This definition depends on each cohort, but it should not be smaller than 100 (Supp. Note).
    Full documentation of the parameters can be found here: https://github.com/bbglab/deepCSA/blob/main/docs/README.md.

```
input                : <input.csv>
fasta                : <reference Fasta file>

cosmic_ref_signatures      : <COSMIC mutational signatures TSV>
wgs_trinuc_counts                                                    :
https://raw.githubusercontent.com/bbglab/deepCSA/refs/heads/main/asse
ts/trinucleotide_counts/trinuc_counts.<homo_sapiens|mus_musculus.mm39
>.tsv

vep_cache         : <Ensembl VEP cache location>
vep_genome        : <GRCh38|GRCm39>
vep_species       : <homo_sapiens|mus_musculus>

consensus_panel_min_depth : <custom depth threshold>
```

198. (optional) Prepare an input data table to use for grouping samples into specific groups. i.e. defined based on clinical data or metadata. If provided, only the samples included in this table will be included in the analysis. All samples in this table must be also in the input csv. The unique identifier needs to match the sample names provided in the input csv. See example below.

    features_example.csv:

```
SAMPLE_ID,BLADDER_LOCATION,SEX,SMOKING_STATUS
donor1_BDO,dome,M,never
donor1_BTR,trigone,M,never
donor2_BDO,dome,F,former
donor2_BTR,trigone,F,former
```

…

```
features_table                : features_example.csv
features_table_separator      : 'comma'
features_unique_identifier    : 'SAMPLE_ID'
features_groups_list                    :  [  [BLADDER_LOCATION],
[BLADDER_LOCATION,      SEX],       [BLADDER_LOCATION,      SEX,
SMOKING_STATUS] ]
```

199. Choose the desired outputs by selecting one of the following profiles: basic, get_signatures or clonal_structure. We recommend starting with the **basic** run to be able to assess the quality of the samples and the simple metrics before starting to look at more complex ones.

    Proceed to launch deepCSA via the following Nextflow command:

```
nextflow run bbglab/deepCSA \
  -profile singularity,basic \
  -c executor.config \
  -params-file params.deepCSA.yml
```

    TROUBLESHOOTING

200. Check the results available in the following directories part of deepCSA output and confirm that your cohort of samples is suitable for proceeding to additional analysis. Part of these QCs are described in depth in the Supplementary Note.

    i. depthssummary: check the outputs in this folder to explore the values of depth per gene and per sample and identify or discard the presence of any big differences in sequencing coverage. (Extended Data Fig. 2a-c,3a-c)

    ii. germline_somatic vs clean_somatic: these two folders contain the mutation files before and after applying all the filters. The final set of mutations is used for all the analysis so it is important to check that no likely relevant mutation is missed, nor any likely artifactual mutation is kept.
        i. plotmaf & plotsomaticmaf: use the plots generated in these two directories for a high level comparison

    iii. concatprofiles: check if there are major differences between the mutational profiles of all the samples, this is a basic mutagenesis analysis but will inform on the uniformity of the mutational processes in the cohort. (Extended Data Fig. 4a,b)

    iv. plotmutdensityqc: explore this directory to identify specific genes (Extended Data Fig. 5a) or samples (Extended Data Fig. 5b) whose background mutation density falls outside the overall distribution of background mutation densities.
        i. mutdensity: the specific values of all mutation densities that can be compared to identify outliers are available here. The user can also assess if these values are inflated due to poor sequencing depth (Extended Data Fig. 3c).

ii. plotinterindividualvariability: additional interindividual variability in the values of mutation density can be observed with these plots

v. plotcontamination: measure the proportion of cases where a specific sample contains several if not all germline mutations of another sample. If this were the case, there would have likely been contamination of DNA from one sample to another. The consequences are that it could invalidate some preliminary results and samples.

If all the checks have been successful, proceed with the last step.

201. Proceed to resume your previous run with an updated configuration. Below (a,b) are two different predefined sets of analysis. Optionally, custom sets of analysis can be selected by tuning the parameters directly in the params.deepCSA.yml file (c).
In either case, when activating new analysis, most of them have associated QCs (Extended Data Fig. 6) that should be checked, find more information at:

a. Running with the get_signatures profile can provide additional insights on the mutagenesis analysis.

```
nextflow run bbglab/deepCSA \
  -profile singularity,get_signatures \
  -c executor.config \
  -params-file params.deepCSA.yml
```

b. If the main goal is the analysis of selection of somatic mutations, you should use the clonal_structure predefined profile.

```
nextflow run bbglab/deepCSA \
  -profile singularity,clonal_structure\
  -c executor.config \
  -params-file params.deepCSA.yml
```

c. Custom configuration.

```
nextflow run bbglab/deepCSA \
  -profile singularity \
  -c executor.config \
  -params-file params_custom.deepCSA.yml
```

Timing

Steps 1–23, Fragmentation and cleanup (1 h 30 min)
Steps 24–37, End repair and cleanup (1 h)
Steps 38–40, Ligation 1 (30 min)
Steps 41–57, Ligation 2 and cleanup (1 h)
Steps 58–78, Intra-strand repair and cleanup and cleanup (1 h 30 min)
Steps 79–82, Quantification of unique ligated DNA fragments with qPCR (3 h)
Steps 83–00, Indexing PCR and cleanup (1 h 30 min)
Steps 100–115, Hybridization (1 h handling and 16h incubation)
Steps 116–125, Streptavidin bead wash (30 min)
Steps 126–132, Capture (1 h)
Steps 133–144, Heated washes (30 min)
Steps 145–163, Room temperature washes (30 min)
Steps 164–182, Post-capture PCR and cleanup (1 h)
Steps 183–189, Sequencing data processing: consensus building, QCs and variant calling.
Variable and dependent on input size and computational resources. Up to ~3 days for whole-exome libraries at high depth.
Steps 190–195, Sequencing metrics compilation (5 min)
Steps 196–201, Mutations data analysis (1-10 h)
Variable and dependent on input size, analysis requested and computational resources.

Troubleshooting

Step (step number)	Problem	Possible reason	Solution
Fragmentation (7)	Percentage of library fragments from 50-650bp is below 30%	Library is underfragmented	Add more incubation time at 37ºC in the thermocycler. If the problem persists, consider adding 1 uL more enzyme and more incubation time. High EDTA (1mM) concentrations might result in lower fragmentation kinetics, consider eluting the gDNA in lower EDTA buffers.
	Percentage of library fragments from 50-650bp is above 50%	Library is overfragmented	Some overfragmentation (31-40% fragments in 50-650bp region) may be acceptable. However, if this percentage is above 50%, it could directly impact the insert size of the final library. If more genomic DNA material is available, consider re-starting the fragmentation and shortening the incubation time.
Quantification of unique ligated DNA fragments with qPCR (82)	Femtomol yield is lower than expected	Dilutions were not properly prepared and accounted for when calculating molarity	Prepare dilutions again, and re-do the qPCR. If the qPCR repetition gives the same results, DNA may have been lost in previous steps; consider re-starting the library if more gDNA is available.
Indexing PCR (99)	Overamplification	The number of PCR cycles is in excess	Some overamplification can be observed at the

			Indexing PCR step, while still resulting in successful library preparations. If this is suspected to have an impact on library performance, consider adjusting the number of PCR cycles.
Post-capture PCR (182)	Less than 10-15 ng/uL after cleanup	Not enough post-capture PCR cycles	While setting up the correct post-capture PCR cycle number per panel, it is recommended to measure directly the PCR product before final elongation step and cleanup. If concentration is lower than 8-10 ng/uL, add an additional 2-3 cycles before continuing with the cleanup.
		Failed capture	Repeat capture from the second half of the Indexing PCR, which was stored at -20ºC.
Integrated duplex metrics (194)	Wetlab information missing for some columns	IDs not matching during merging	Verify that the sample IDs in the deepUMI input csv match the IDs listed under the ‘DryLab ID’ column in the wetlab qc metrics table.
deepUMIcaller / deepCSA (189, 199, 201)	Path or file not found	File system is not mounted	Request the mounting of the required file systems in the config. `singularity { autoMounts = true runOptions = '-B /data/bbg'`

			`}`
deepUMIcaller / deepCSA (189, 199, 201)	Long execution time pulling containers	Containers are pulled to a default location that is not cached across runs	Indicate a directory to store the cached container images. `singularity { cacheDir = '<custom_location>/nextflow_containers' libraryDir = '<custom_location>/nextflow_containers' }`
deepCSA (200)	Missing sample in output	Missing sample in the features_table file or no mutations detected in that sample	Revise the content of the features_table and the minimum mutations per sample threshold.
deepUMIcaller (189)	Requested resources are not available and very long execution times		Check Nextflow troubleshooting, and adjust the max_cpus, max_memory, max_time. If the input file is too big and resources are already at max. take advantage of the parallelization capabilities with the splitfastq options

Anticipated Results

Duplex library preparation

The aim of fragmentation is to yield DNA fragments of size 300-400 bps for ligation. Starting with 250 ng DNA, a fragmentation for 15 minutes at 37ºC is normally required when starting with genomic DNA with integrity values between 6 and 10 (see **Experimental Design** section). If the fragments of sizes between 50 and 650 bp represent less than 30% of the total (Fig. 2a,b), it is possible to continue the fragmentation. It should be stopped when between 30% and 50% of fragments fall in this range, which predicts more than 80% of fragments in this range immediately before ligation (Fig. 2c) and optimal fragment size distribution before sequencing (Fig. 2d). The qPCR used to estimate ligated DNA fragments, yields values that are usually in the range of 20-80 fmol (Supp. Note; 1.7-6% recovery). The expected yield of the indexing PCR is between 2-8 µg with an average DNA fragment size of 415 bps. The overnight hybridization (16-20 h) and capture of half of the indexing PCR product (target panel of 737kb), followed by post-capture PCR leads to ~750 ng of DNA with average fragment size of 470 bps. The required number of sequencing reads to attain efficient duplex consensus read building is estimated at 170-680 billion basepairs.

Building consensus and variant calling

The main outputs of deepUMIcaller are: mutations (single VCF file), duplex-quality consensus reads and metrics for quality control purposes. Each mutation row in the VCF contains the number of duplex consensus reads, unique molecules, and unique molecules that do not form duplex consensus reads supporting it. Aligned duplex-quality consensus reads, required for downstream analysis, are stored in the `sortbamduplexcons` directory. (see Supp. Note)

Sequencing metrics compilation

The metrics compilation (collecting deepUMIcaller QC outputs and, optionally, the QCs from the library preparation steps) are stored in the WetDryMetrics file (Supp. Table 2; Fig. 3 and Supp. Note).

For example, the distribution of family size of the two libraries (of separate samples) exemplified in Figure 3a,b have been correctly sequenced; each initial molecule was sequenced on average between 5-7 times, and single and double strand consensus reads were correctly generated, and the number of unique molecules was very similar to the estimation from the qPCR (Supp. Note). The library with the family size distribution shown in Figure 3c is undersequenced; more raw reads are required to produce more duplex consensus reads, and that of Figure 3d has been oversequenced.

An incorrect estimation in the number of raw reads required for optimal sequencing may be caused by decreased efficiency of the hybridization and capture (Extended Data Fig. 7a-c). Figure 3e presents the fraction of on-target sequencing reads, all unique molecules and duplex consensus reads across libraries prepared and sequenced across different batches. Ideally, one would like to obtain an on-target rate of unique molecules above 80%.

Four examples of the distribution of the six types of nucleotide changes along the sequence of consensus duplex reads are shown in Figure 4a-d. The library in Figure 4a presents an even

(expected) distribution of different nucleotide changes along the sequence of reads, while the other three exhibit a clear enrichment in the first positions of the reads. This would prompt to increase the number of nucleotides that are clipped at the start of the reads. Nevertheless, if the enrichment persists despite this extended clipping (which could point to other sources of pseudoduplex artefacts, such as nucleotide damage incurred in the process of DNA extraction or during storage), the inclusion of the sample in the analysis cohort should be reconsidered.

Mutations data analysis

The expected outputs of any deepCSA run include a summary of the sequencing depth across each sample in the cohort, as well as within each sequenced genomic element. It also contains several files with the somatic mutations annotated using the Ensembl Variant Effect Predictor (VEP) and their variant allele frequencies. These files should be explored to revise that the automated filtering is not excluding any relevant mutations. DeepCSA also provides several metrics of mutagenesis and selection across samples, and pre-defined groups of samples (Fig. 5a-f). The mutagenesis metrics include the activity of mutational signatures across samples (Fig. 5a,b) and the density of mutations (mutations/Mbps) of each gene (and all sequenced genomic regions) across samples and at the cohort level (Fig. 5c). The metrics of selection include values computed by several methods that detect signals of positive selection in the pattern of mutations in genes, including a dN/dS-based method (omega; Fig. 5d) and three others (Fig. 5e) as well as metrics of selection per site (Fig. 5f). Regressions to test the association between several of these metrics and the history of exposure of the donors of the samples can also be produced (Fig. 5g). All outputs of deepCSA are described in detail in the Supplementary Note and https://github.com/bbglab/deepCSA/blob/main/docs/output.md.

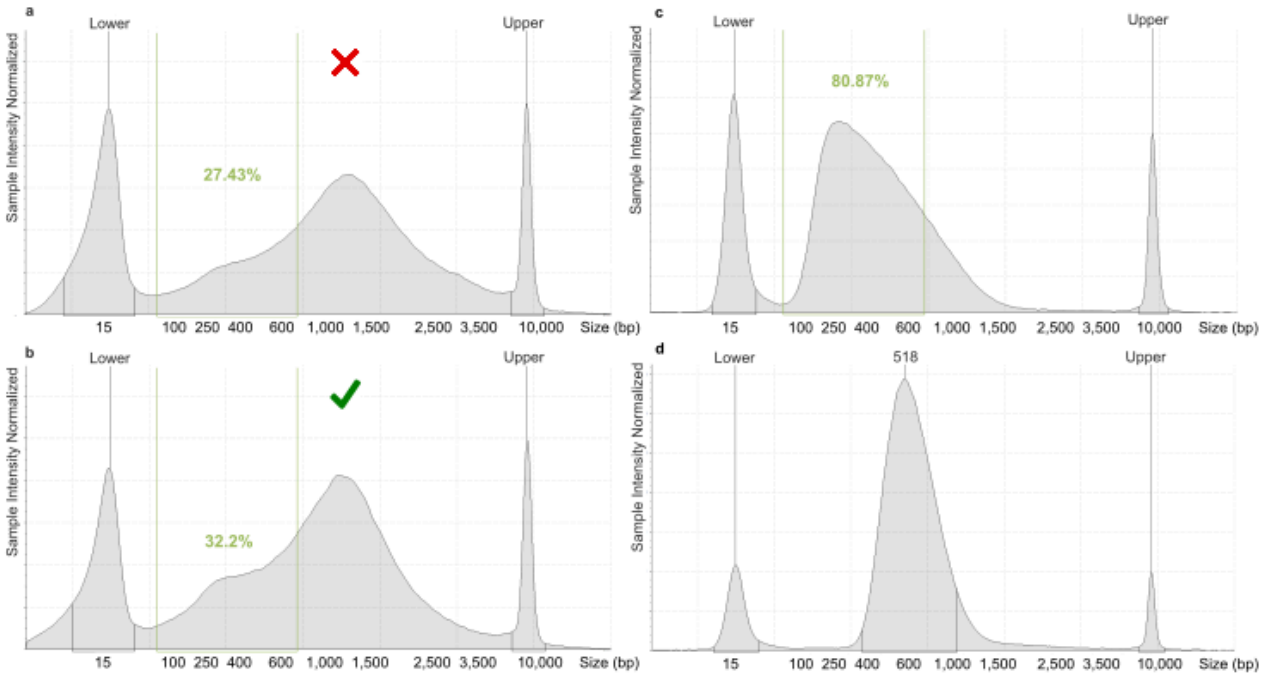

Figure 2. DNA Duplex library preparation expected size distribution examples.
Histograms showing the distribution of DNA fragment size in mixtures of DNA fragments,
a. Underfragmented DNA mixture, showing the percentage of DNA fragments between 50 and 650bp is under 30%, in this case 27.43%.
b. A case of optimal fragmentation profile after additional fragmentation of the mixture in (a), showing percentage of DNA fragments above 30% within the same range.
c. Distribution of DNA fragments of the mixture in (b) after fragmentation inactivation and clean-up. The percentage of DNA fragments between 50 and 650bp in this case is 80.87%.
d. Final DNA duplex library fragment size distribution before sequencing.

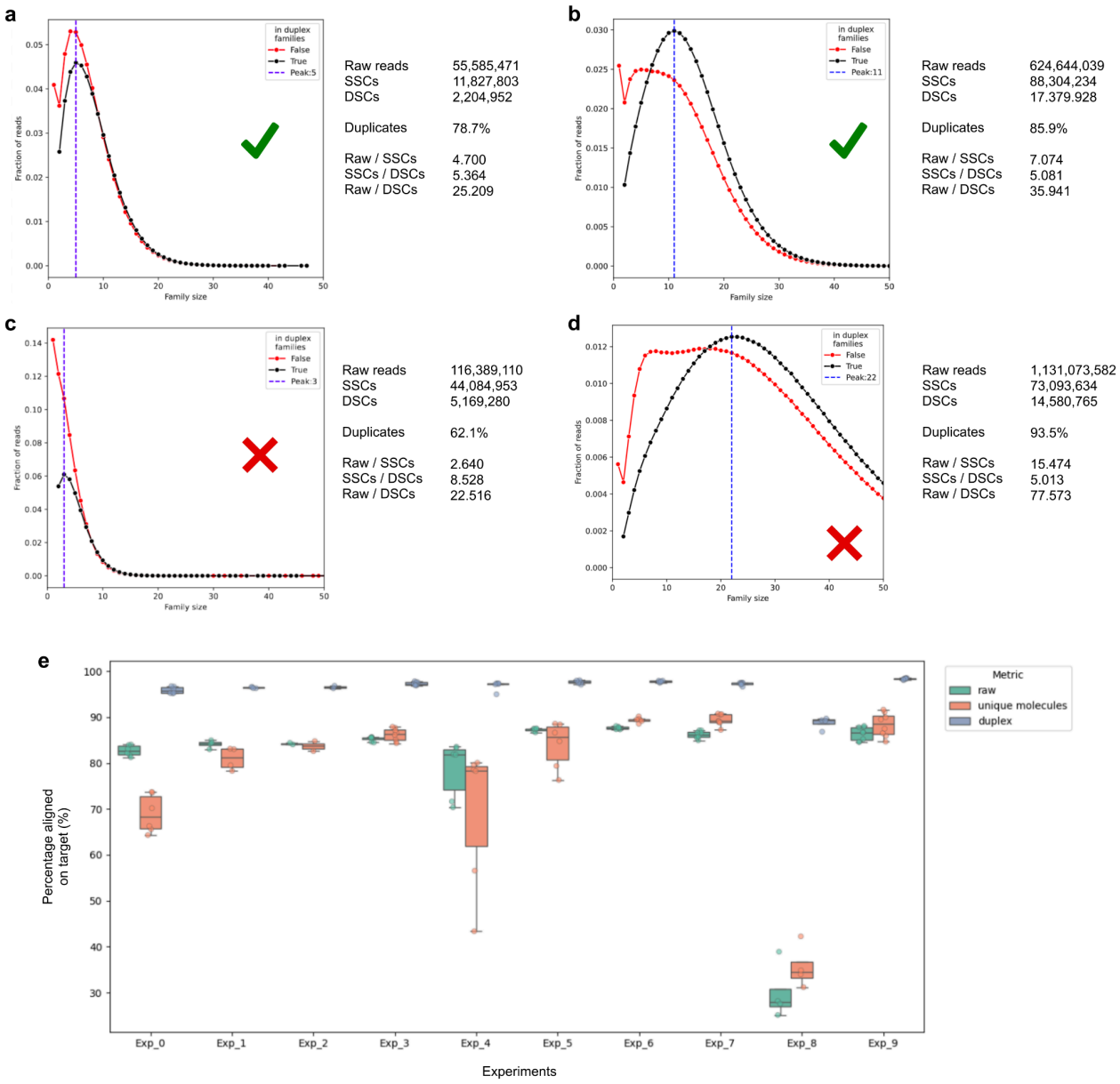

Figure 3. Building duplex consensus metrics and metrics compilation outputs.
a-d. Distribution of family sizes of unique molecules in 4 DNA duplex libraries prepared from different samples. Unique molecules are separated depending on whether (red) or not (black) they form duplex consensus reads. A vertical blue line represents the mode of the SSCs family sizes. The value of this mode (family size peak) is presented within the legend box. Numbers at the right side of the plots present key metrics of the duplex consensus building process. These plots are generated by the deepUMIcaller family metrics module (see FAMILYMETRICS deepUMIcaller step 11b,d in the Supp. Note). e. Percentage of molecules aligned within the targeted regions for 10 different experiments, generated by the Integrated duplex metrics. SSC: Single Strand Consensus. DSC: Double Strand Consensus (also duplex consensus).

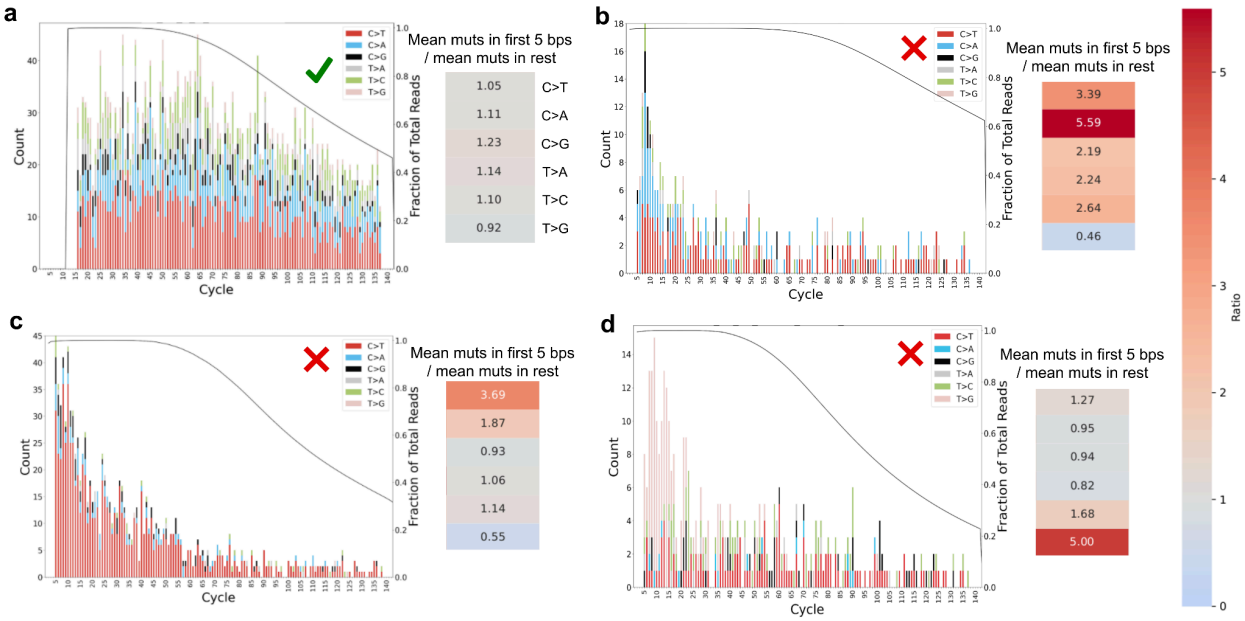

Figure 4. Mutations per duplex consensus read position.
a-d. Counts of mutations identified at each position of their supporting reads, colored by nucleotide change type. The black line represents the fraction of mutations identified at each position.The heatmaps represent the ratio of the mean number of mutations in the first 5 bps of the read divided by the mean number of mutations per bp in the rest of the positions of the read. These are computed separately for each possible mutation type and are plotted in the same order as the legend of the plot. Values greater than 2 indicate the need to carry out further clipping. These plots are generated in the MUTSPERPOS and COHORTMUTSPERPOS steps of deepUMIcaller (see deepUMIcaller Supp. Note steps 42 and 43). Sample on panel a has already been clipped to remove a possible enrichment of mutations on the read ends.

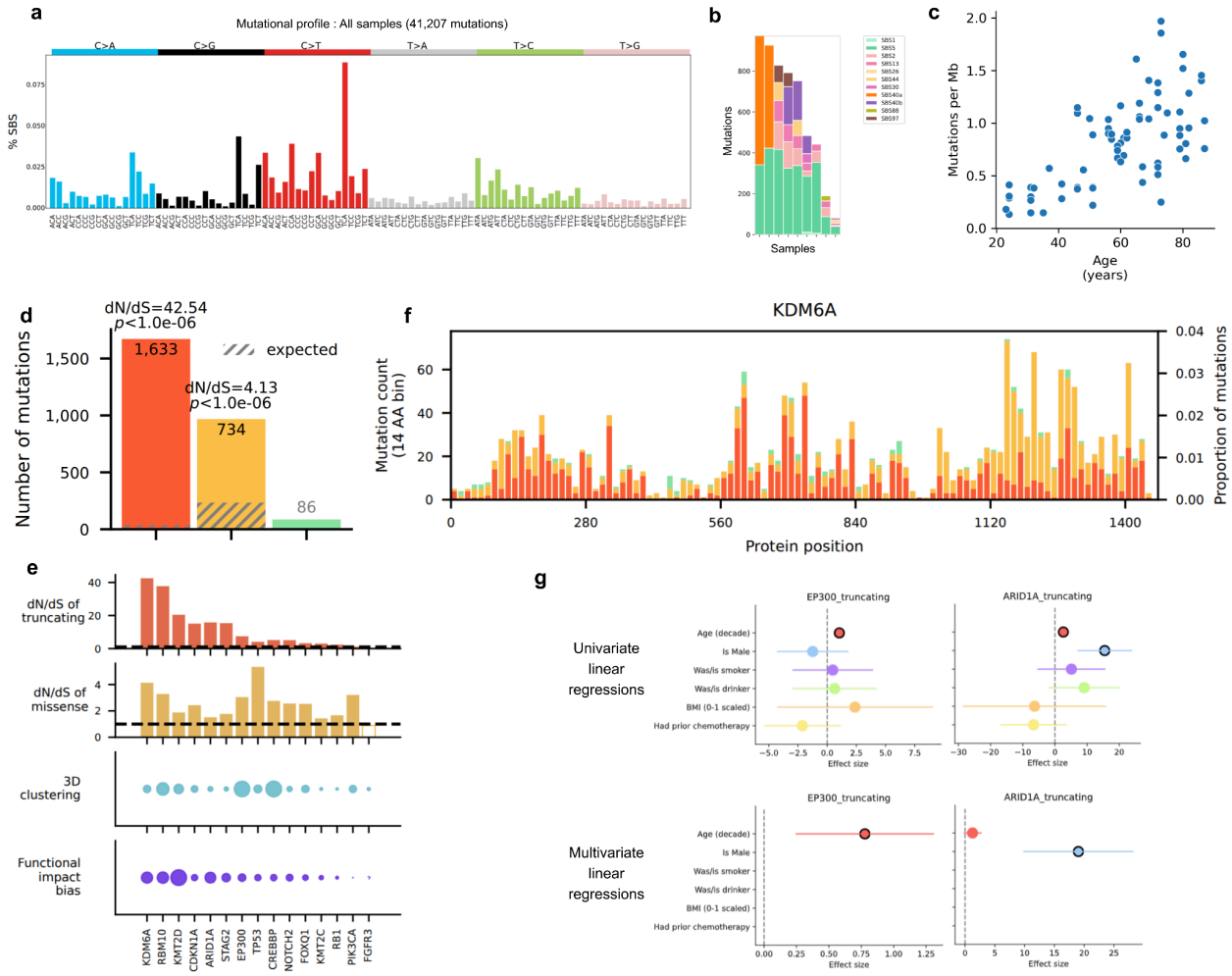

Figure 5. Overview of deepCSA results.

a. Exemplary tri-nucleotide mutational profile obtained from all mutations in a cohort of samples.

b. Activity of mutational signatures identified in a cohort across its samples.

c. Density of mutations of different samples obtained from donors of varying age.

d. Magnitude of positive selection (computed using omega, which implements a dN/dS approach) on missense and truncating mutations observed in KDM6A across 71 normal urothelium samples.

e. Magnitude of positive selection on mutations of 15 genes across 71 normal urothelium samples, calculated using three independent methods.

f. Frequency of mutations observed at different sites of the KDM6A in 71 normal urothelium samples.

g. Association of the magnitude of positive selection (dN/dS of truncating mutations) observed for EP300 and ARID1A genes across donors of normal urothelium samples. Showing the results of univariate and multivariate linear regressions.

References

1. Alexandrov, L. B. *et al.* The repertoire of mutational signatures in human cancer. *Nature* **578**, 94–101 (2020).
2. Campbell, P. J. *et al.* Pan-cancer analysis of whole genomes. *Nature* **578**, 82–93 (2020).
3. Martínez-Jiménez, F. *et al.* A compendium of mutational cancer driver genes. *Nat. Rev. Cancer* **20**, 555–572 (2020).
4. Sabarinathan, R. *et al.* The whole-genome panorama of cancer drivers. *bioRxiv* https://doi.org/10.1101/190330 (2017) doi:10.1101/190330.
5. Gerstung, M. *et al.* The evolutionary history of 2,658 cancers. *Nature* **578**, 122–128 (2020).
6. Jaiswal, S. & Ebert, B. L. Clonal hematopoiesis in human aging and disease. *Science* **366**, eaan4673 (2019).
7. Pich, O., Reyes-Salazar, I., Gonzalez-Perez, A. & Lopez-Bigas, N. Discovering the drivers of clonal hematopoiesis. *Nat. Commun.* **13**, 4267 (2022).
8. Lee-Six, H. *et al.* Population dynamics of normal human blood inferred from somatic mutations. *Nature* **561**, 473–478 (2018).
9. Fabre, M. A. *et al.* The longitudinal dynamics and natural history of clonal haematopoiesis. *Nature* **606**, 335–342 (2022).
10. Yoshida, K. *et al.* Tobacco smoking and somatic mutations in human bronchial epithelium. *Nature* **578**, 266–272 (2020).
11. Lee-Six, H. *et al.* The landscape of somatic mutation in normal colorectal epithelial cells. *Nature* **574**, 532–537 (2019).
12. Lawson, A. R. J. *et al.* Extensive heterogeneity in somatic mutation and selection in the human bladder. *Science* **370**, 75–82 (2020).
13. Martincorena, I. *et al.* High burden and pervasive positive selection of somatic mutations in normal human skin. *Science* **348**, 880–886 (2015).

14. Martincorena, I. *et al.* Somatic mutant clones colonize the human esophagus with age. *Science* **362**, 911–917 (2018).
15. Islam, S. M. A. *et al.* Uncovering novel mutational signatures by de novo extraction with SigProfilerExtractor. 2020.12.13.422570 Preprint at https://doi.org/10.1101/2020.12.13.422570 (2022).
16. Muiños, F., Martínez-Jiménez, F., Pich, O., Gonzalez-Perez, A. & Lopez-Bigas, N. In silico saturation mutagenesis of cancer genes. *Nature* **596**, 428–432 (2021).
17. 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).
18. Kennedy, S. R. *et al.* Detecting ultralow-frequency mutations by Duplex Sequencing. *Nat. Protoc.* **9**, 2586–2606 (2014).
19. Abascal, F. *et al.* Somatic mutation landscapes at single-molecule resolution. *Nature* **593**, 405–410 (2021).
20. Lawson, A. R. J. *et al.* Somatic mutation and selection at population scale. *Nature* **647**, 411–420 (2025).
21. Nandi, S. P. *et al.* A Universal Duplex Sequencing Approach for Accurate Detection of Somatic Mutations. 2025.09.14.676103 Preprint at https://doi.org/10.1101/2025.09.14.676103 (2025).
22. Valentine, C. C. *et al.* Direct quantification of in vivo mutagenesis and carcinogenesis using duplex sequencing. *Proc. Natl. Acad. Sci.* **117**, 33414–33425 (2020).
23. Calvet, F. *et al.* Sex and smoking bias in the selection of somatic mutations in human bladder. *Nature* **647**, 436–444 (2025).
24. Liu, M. H. *et al.* DNA mismatch and damage patterns revealed by single-molecule sequencing. *Nature* **630**, 752–761 (2024).
25. Schmitt, M. W. *et al.* Detection of ultra-rare mutations by next-generation sequencing. *Proc. Natl. Acad. Sci.* **109**, 14508–14513 (2012).

26. Pich, O. *et al.* Somatic evolution following cancer treatment in normal tissue. *Nature* 1–11 (2025) doi:10.1038/s41586-025-09792-4.
27. Neville, M. D. C. *et al.* Sperm sequencing reveals extensive positive selection in the male germline. *Nature* **647**, 421–428 (2025).
28. Salk, J. J., Schmitt, M. W. & Loeb, L. A. Enhancing the accuracy of next-generation sequencing for detecting rare and subclonal mutations. *Nat. Rev. Genet.* **19**, 269–285 (2018).
29. Yokoyama, A. *et al.* Somatic mosaicism in the buccal mucosa reflects lifestyle and germline risk factors for esophageal squamous cell carcinoma. *Sci. Transl. Med.* **17**, eadq6740 (2025).
30. Norgaard, Z., Homer, N., Pearce, S., bot, nf-core & Pedersen, B. nf-core/fastquorum: 1.2.0. Zenodo https://doi.org/10.5281/zenodo.15198598 (2025).
31. Langer, B. E. *et al.* Empowering bioinformatics communities with Nextflow and nf-core. *Genome Biol.* **26**, 228 (2025).
32. Lai, Z. *et al.* VarDict: a novel and versatile variant caller for next-generation sequencing in cancer research. *Nucleic Acids Res.* **44**, e108 (2016).
33. Liu, M., Wu, Y., Jiang, N., Boot, A. & Rozen, S. G. mSigHdp: hierarchical Dirichlet process mixture modeling for mutational signature discovery. *NAR Genomics Bioinforma.* **5**, lqad005 (2023).
34. Mularoni, L., Sabarinathan, R., Deu-Pons, J., Gonzalez-Perez, A. & López-Bigas, N. OncodriveFML: a general framework to identify coding and non-coding regions with cancer driver mutations. *Genome Biol.* **17**, 128 (2016).
35. Pellegrini, S., Dove-Estrella, O., Muiños, F., Lopez-Bigas, N. & Gonzalez-Perez, A. Oncodrive3D: fast and accurate detection of structural clusters of somatic mutations under positive selection. *Nucleic Acids Res.* **53**, gkaf776 (2025).
36. Woolston, D. W. *et al.* Ultra-deep mutational landscape in chronic lymphocytic leukemia

uncovers dynamics of resistance to targeted therapies. *Haematologica* **109**, 835–845 (2024).

37. Pareja, F. *et al.* Cancer-Causative Mutations Occurring in Early Embryogenesis. *Cancer Discov.* **12**, 949–957 (2022).
38. Schmitz, E. G., Griffith, M., Griffith, O. L. & Cooper, M. A. Identifying genetic errors of immunity due to mosaicism. *J. Exp. Med.* **222**, e20241045 (2025).
39. Ren, P., Zhang, J. & Vijg, J. Somatic mutations in aging and disease. *GeroScience* **46**, 5171–5189 (2024).
40. Koch, Z., Li, A., Evans, D. S., Cummings, S. & Ideker, T. Somatic mutation as an explanation for epigenetic aging. *Nat. Aging* **5**, 709–719 (2025).
41. Cohen, J. D. *et al.* Detection of low-frequency DNA variants by targeted sequencing of the Watson and Crick strands. *Nat. Biotechnol.* **39**, 1220–1227 (2021).
42. Di Tommaso, P. *et al.* Nextflow enables reproducible computational workflows. *Nat. Biotechnol.* **35**, 316–319 (2017).
43. Forbes, S. A. *et al.* {COSMIC} (the Catalogue of Somatic Mutations in Cancer): a resource to investigate acquired mutations in human cancer. *Nucleic Acids Res.* **38**, D652—-657 (2010).