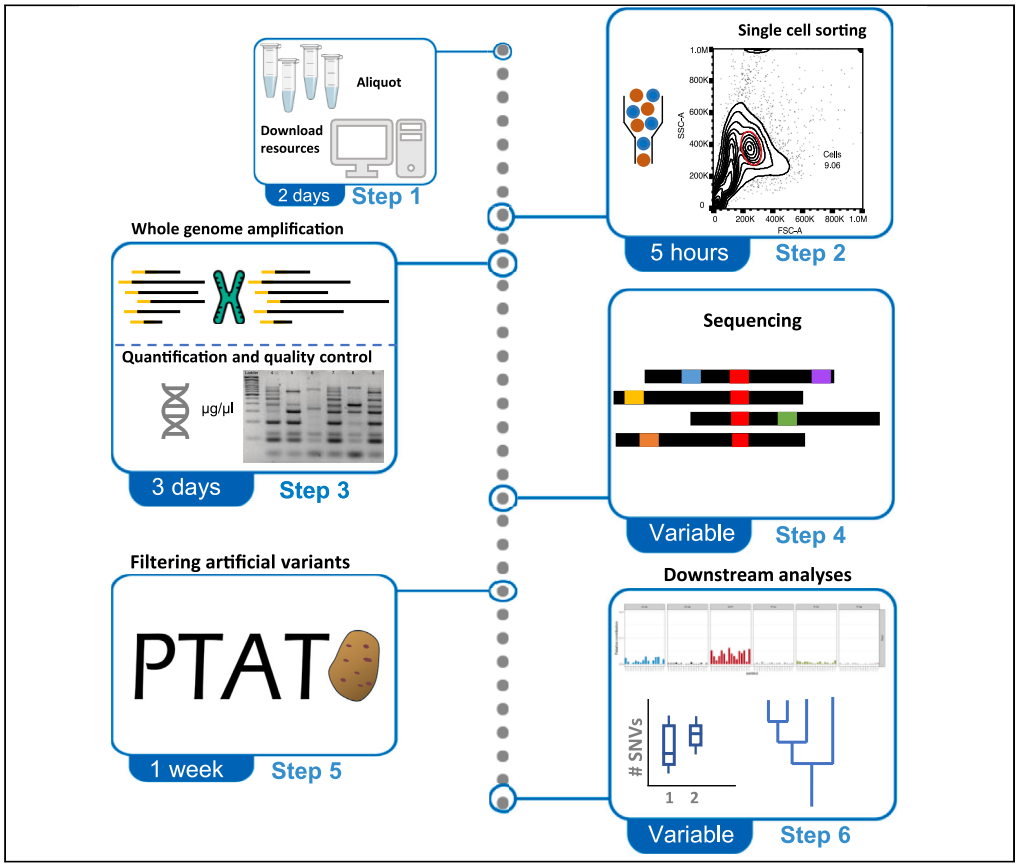
STAR Protocols

CellPress
OPEN ACCESS

Protocol

Protocol for genome-wide analysis of somatic variants at single-cell resolution using primary template-directed DNA amplification



The study of somatic mutations in single cells provides insights into aging and carcinogenesis, which is complicated by the dependency on whole-genome amplification (WGA). Here, we describe a detailed workflow starting from single-cell isolation to WGA by primary template-directed amplification (PTA), sequencing, quality control, and downstream analyses. A machine learning approach, the PTA Analysis Toolkit (PTATO), is used to filter the hundreds to thousands of artificial variants induced by WGA from true mutations at high sensitivity and accuracy.

Publisher's note: Undertaking any experimental protocol requires adherence to local institutional guidelines for laboratory safety and ethics.

Lucca L.M. Derks, Anaïs J.C.N. van Leeuwen, Alexander S. Steemers, ..., Mark Verheul, Sjors Middelkamp, Ruben van Boxtel

r.vanboxtel@prinsesmaximacentrum.nl

Highlights

Isolation, whole-genome amplification, and sequencing of single cells

Removal of artificial variants from somatic mutations using a machine learning approach

Implementation of quality control steps in experimental and computational workflows

Derks et al., STAR Protocols 6, 103499
March 21, 2025
https://doi.org/10.1016/j.xpro.2024.103499

STAR Protocols

CellPress
OPEN ACCESS

Protocol

Protocol for genome-wide analysis of somatic variants at single-cell resolution using primary template-directed DNA amplification

Lucca L.M. Derks,[1,2,4] Anaïs J.C.N. van Leeuwen,[1,2,4] Alexander S. Steemers,[1,2] Laurianne Trabut,[1,2] Markus J. van Roosmalen,[1,2] Vera M. Poort,[1,2] Rico Hagelaar,[1,2] Mark Verheul,[1,2] Sjors Middelkamp,[2,3] and Ruben van Boxtel[1,2,5,6,*]

[1]Princess Máxima Center for Pediatric Oncology, Utrecht 3584 CS, the Netherlands

[2]Oncode Institute, Utrecht 3521 AL, the Netherlands

[3]Center for Molecular Medicine, University Medical Center Utrecht, Utrecht 3584 CG, the Netherlands

[4]These authors contributed equally

[5]Technical contact

[6]Lead contact

*Correspondence: r.vanboxtel@prinsesmaximacentrum.nl
https://doi.org/10.1016/j.xpro.2024.103499

SUMMARY

The study of somatic mutations in single cells provides insights into aging and carcinogenesis, which is complicated by the dependency on whole-genome amplification (WGA). Here, we describe a detailed workflow starting from single-cell isolation to WGA by primary template-directed amplification (PTA), sequencing, quality control, and downstream analyses. A machine learning approach, the PTA Analysis Toolkit (PTATO), is used to filter the hundreds to thousands of artificial variants induced by WGA from true mutations at high sensitivity and accuracy.

For complete details on the use and execution of this protocol, please refer to Middelkamp et al.[1]

BEFORE YOU BEGIN

Characterization of somatic mutations at single-cell resolution using whole-genome sequencing (WGS) has improved our understanding of genetic heterogeneity, clonal relatedness, cellular plasticity, and mutation burden. However, single cell (sc) whole-genome amplification (WGA) often leads to disproportional distribution of maternal and paternal alleles, resulting in allelic imbalance, bias or even dropout in the sequencing reads. In addition, the sequencing data is associated with low coverage uniformity and elevated false positive as well as false negative variant calling.[2–4] Primary template-directed amplification (PTA) enables a more uniform and reproducible profiling of whole genomes. However, PTA still generates hundreds to thousands of artificial variants during amplification.[5] Existing bioinformatic approaches have a low detection sensitivity ($\pm$40%) and are thus unable to detect all true variants.[6]

Here, we describe a streamlined workflow to generate high quality scWGS data. Our approach leverages additional levels of quality control on top of the standardized PTA protocol, including indexed flow cytometry sorting and a multiplexed polymerase chain reaction (PCR) to assess the amplification of all chromosomes. Finally, we employ a machine-learning approach, the PTA Analysis Toolkit (PTATO), to remove false positive calls from the data. This results in data that is highly accurate and reaches a variant calling sensitivity up to 90%.[1] In the protocol below, we applied

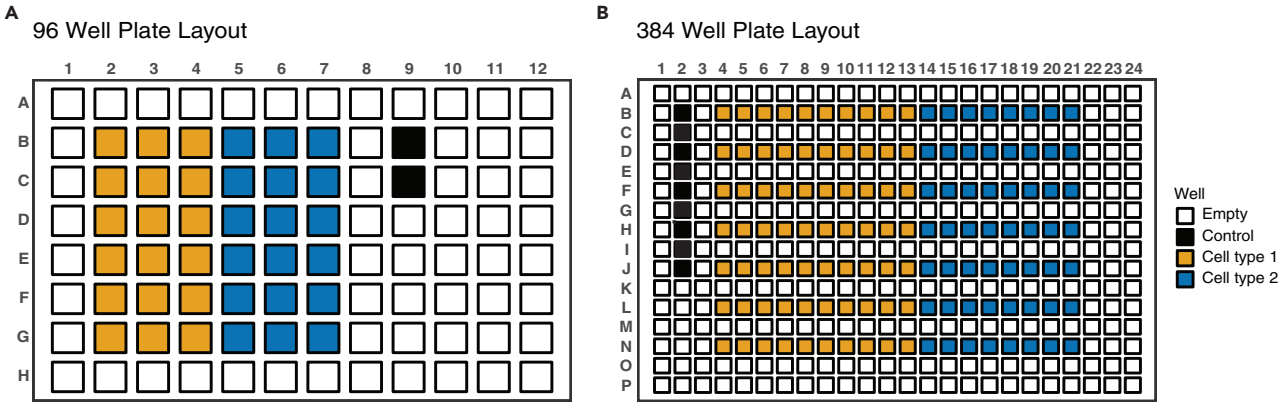


Figure 1. Example of plate layouts for single-cell sorting

(A) Example plate layout for the 96 well format.

(B) Example plate layout for the 384 well format. In each plate, wells are reserved for controls (black).

our workflow on different cell types and states, including malignant and non-malignant B-cells, immature T-cells and hematopoietic stem and progenitor cells (HSPCs), illustrating the robustness of our approach.

Institutional permissions

When performing whole-genome sequencing on primary human cells, institutional permission should be obtained in addition to informed consent from the individuals included in the study. The samples included in the experimental results in the step-by-step method details and expected outcomes of this protocol were obtained from the Biobank and Data Access Committee of the Princess Máxima Center for pediatric oncology with ethical approval under proposals PMCLAB2022.303 (Figures 2 and 3B), PMCLAB2021.249 (Figures 3A, 5, 6, and 7), and PMCLAB2018.026 (Figures 3C and 4). Written informed consents from the included individuals, or their parents or legal guardians, were obtained by the Princess Máxima Center for pediatric oncology (B-cells, Burkitt lymphoma cells, and hematopoietic stem and progenitor cells) or the institutional review board of the Erasmus MC (T-cells/T-cell acute lymphoblastic leukemia cells). Samples were obtained in accordance with the Declaration of Helsinki. The WGS data included in the Expected Outcomes of this protocol (Figures 5–10) were obtained from Poort et al.[7]

Choosing an appropriate whole-genome amplification method

Timing: 15 min

Various versions of WGA kits are available using PTA technology. Initially, ResolveDNA Whole Genome Amplification (96 reactions) (Bioskryb) was released, which allows the manual processing of up to 96 cells at a time. When using the updated version of this method, ResolveDNA Whole Genome Amplification kit (v2.0 and newer, Bioskryb), the user can choose between a high-throughput, automated workflow (384 reactions) or a high-yield, manual workflow (96 reactions). Although up to 384 samples can be processed simultaneously in the 384 reactions format and sufficient DNA is produced for WGS, the method is accompanied by a reduced DNA yield after WGA due to the lower reaction volume (see Figure 3 in expected outcomes). The manual workflow is more suitable when the collection volume of single cells exceeds 1 μL, when higher yields of DNA are required, or when dispensing equipment compatible with the automated workflow is not available. In addition, the ResolveOME Whole Genome and Transcriptome Single-Cell Core Kit (Bioskryb) enables the amplification of both the whole DNA and RNA contents of a single cell.

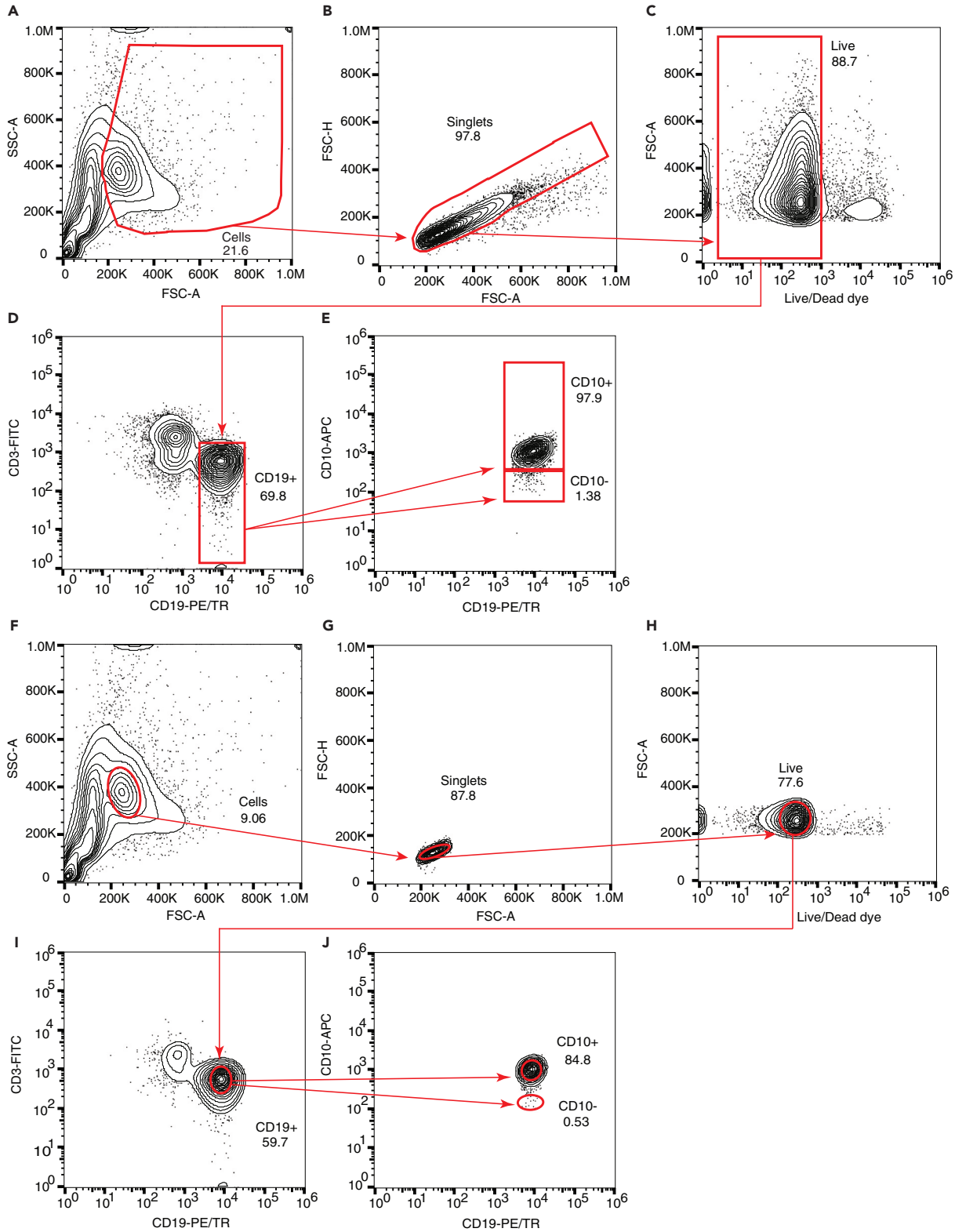


Figure 2. Gating strategies for lenient and strict sorting of single cells

(A) Exclusion of debris based on the FSC-A and SSC-A channels.

(B) Exclusion of doublets using the FSC-A and FSC-H channels.

Figure 2. *Continued*
(C) Exclusion of dead cells using a dead cell dye.
(D) Selection of B cells using B-cell (progenitor) marker CD19 and exclusion of T-cells using T-cell marker CD3.
(E) Selection of malignant and non-malignant cells using CD10.
(F–J) Same as in (A)–(E), but with stricter gating to ensure that sorted cells are of the populations of interest.

Throughout this protocol, a distinction is made between the manual processing method (96 reactions) and the automated processing method (384 reactions). This applies to both ResolveDNA Whole Genome Amplification (ResolveDNA) and ResolveOME Whole Genome and Transcriptome Single-Cell Core (ResolveOME) kits (Bioskryb).

Whole-genome amplification reagents aliquot preparation

Timing: 1 h

It is critical to restrict the number of freeze-thaw cycles of the WGA reagents and thus, single-use aliquots are highly recommended. The ResolveDNA and ResolveOME kit contents (Bioskryb) can be stored until one year after date of manufacturing (as indicated by the manufacturer). Here, we describe aliquot sizes suitable for manual workup of cells in sets of six using the ResolveDNA Whole Genome Amplification Kit - 96 Reactions; for other processing workflows, adjust the aliquot sizes as needed.

1. Thaw the reagents in the ResolveDNA Whole Genome Amplification Kit - 96 Reactions on ice.

   Note: The kit contains the following reagents: SS2, SM2, SN1, SDX, SB4, SEZ1, and SEZ2.

   Note: Aliquoting the reagents can also be performed after the first thaw and use of the kit, the remainder of the reagents can then be aliquoted.

2. Choose the number of reactions per aliquot (1, 6, 12, or 96).

   Note: The size of the aliquots is determined by the number of cells that will be processed for whole-genome amplification in one go. The surplus present in each aliquot is generally enough to add a single positive control amplification reaction to the 6, 12, or 96 reactions performed.

3. Working in a DNA-free pre-PCR hood if possible, prepare the reagent aliquots in 0.5 mL Safe-Lock Tubes (Eppendorf) according to table below.

	Supplied (µL)	Needed 1x (µL)	Aliquot 1x[a] (µL)	Aliquot 6x[a] (µL)	Aliquot 12x[a] (µL)	Aliquot 96x[a] (µL)
12x SS2	500	0.31	2	12	24	192
SM2[b]	500	1.67	2.5	15	30	240
SN1	500	3.00	4	24	48	384
SDX	360	3.00	3.5	21	42	336
SB4	550	5.50	5.5	33	66	528
SEZ1[c]	96	0.88	0.95	5.7	11.4	91.2
SEZ2[c]	144	1.32	1.4	8.4	16.8	134.4

[a]Volume per aliquot includes overages.
[b]The SM2 buffer should never be kept at −80°C or directly on dry ice.
[c]Ensure careful pipetting of the enzymes (SEZ1, SEZ2), which are the limiting factor for the number of reactions per kit. The enzyme solutions are viscous; slow pipetting is required.

⚠ CRITICAL: Keep reagents on ice or in a PCR-cooler throughout the aliquoting process.

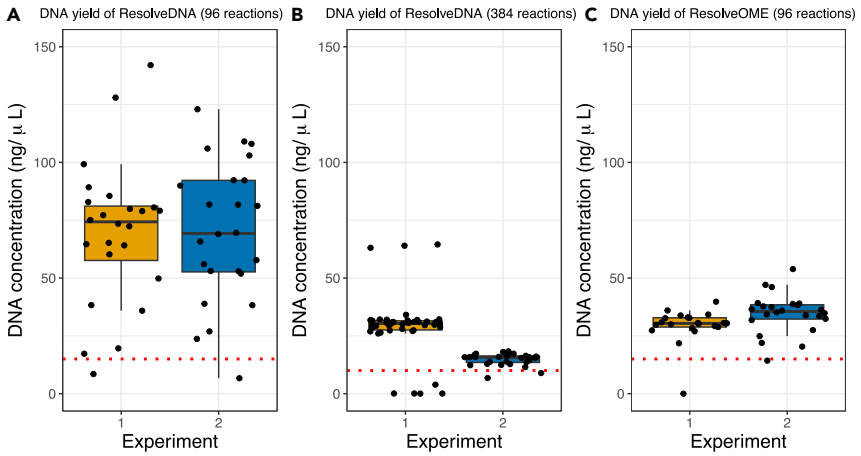


Figure 3. DNA yield expectations across ResolveDNA/ResolveOME kit versions

(A) The DNA yields of two experiments in which single human T-cells and T-cell acute lymphoblastic leukemia (T-ALL) blasts were processed using the ResolveDNA Whole Genome Amplification kit (96 well format) (n = 24, 24 respectively).

(B) The DNA yields of two experiments in which single human B-cells and Burkitt lymphoma cells were processed using the ResolveDNA Whole Genome Amplification kit (384 well format) (n = 42, 28 respectively).

(C) The DNA yields of two experiments in which single human HSPCs were processed using the ResolveOME Whole Genome and Transcriptome Amplification kit (96 well format) (n = 21, 24 respectively). The boxplots depict the median (center line), 25th and 75th percentiles (box), and the largest values no more than 1.5* the interquartile range (whiskers). The dotted red lines show the 15 ng/μL and 10 ng/μL minimum yield thresholds for the 96 reactions and 384 reactions, respectively.

4. Store aliquots at −20°C until use.

Positive control DNA aliquot preparation

Timing: 30 min

The positive control that is included in the ResolveDNA/ResolveOME kits (Bioskryb) consists of DNA provided as a stock of 50 ng/μL. To provide an appropriate control for the DNA in a single cell, dilution is required to 10 and 100 pg/μL as indicated in the protocol included with the kits. In our experience, the use of a single positive control reaction with 100 pg of DNA as input is sufficient to assess DNA amplification efficiency in the 96-reaction format.

5. Thaw the Control Genomic DNA (50 ng/μL) (positive control DNA) at 15°C–25°C.

Note: Control Genomic DNA (50 ng/μL) is part of the ResolveDNA/ResolveOME kits (Bioskryb).

6. Working in a DNA-free pre-PCR hood, if possible, dilute the positive control DNA.
   a. Dilute 50-fold using Elution Buffer (Bioskryb) or TE buffer, Low EDTA (G-Biosciences).

   Note: Elution Buffer is part of the ResolveDNA / ResolveOME kits (Bioskryb).

   b. Measure the DNA concentration using high sensitivity fluorometry, e.g., a Qubit fluorometer, dsDNA Quantification Assay, and Assay Tubes (Invitrogen) according to the manufacturer's instructions available here.
   c. Dilute the positive control DNA to 100 pg/μL using Elution Buffer (Bioskryb) or TE buffer, Low EDTA (G-Biosciences), based on the DNA concentration obtained in step 6b.

   Note: The dilution factor for this step is approximately 10-fold.

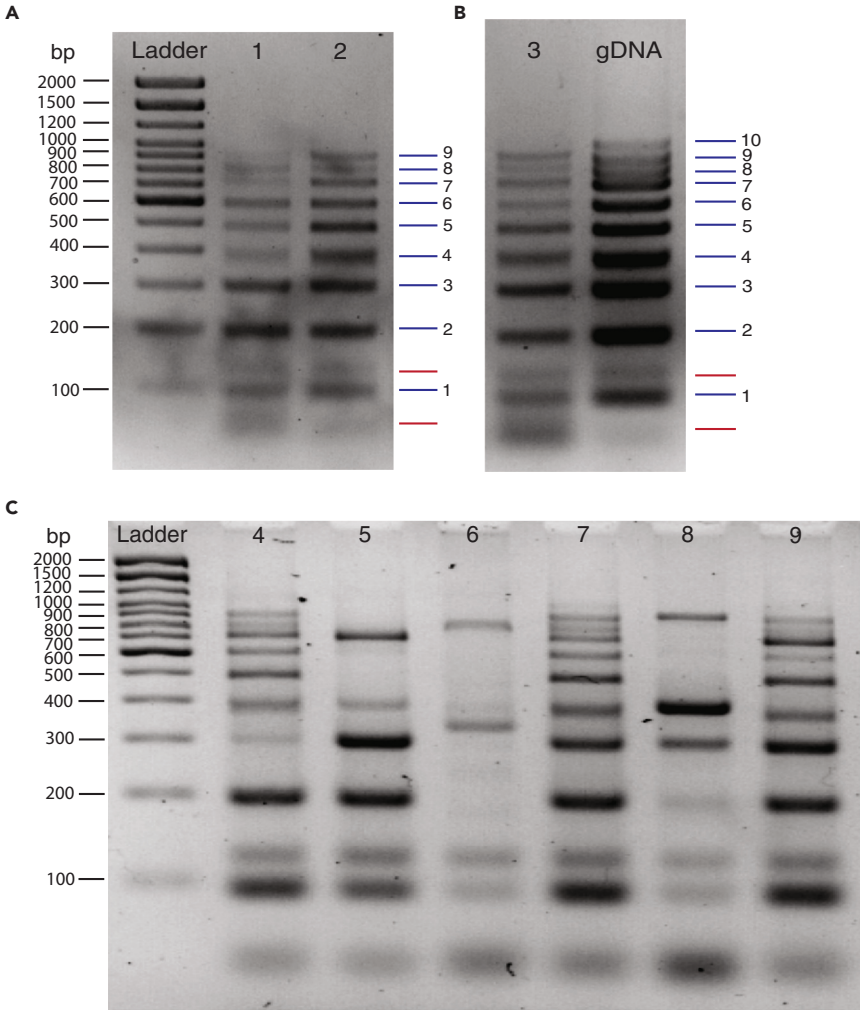


Figure 4. Gel results from the multiplexed QC PCR of single HSPCs that were processed using the ResolveDNA Whole Genome Amplification kit (96 well format)

(A) Example of high-quality amplified genomes. In both samples (cells 1 and 2), nine bands are visible which are indicated with blue lines. Bands indicated with red lines are off-target bands and not used for quality assessment. In the first lane, a 100 bp DNA ladder (Invitrogen) was used. Right lanes removed (additional samples, not needed for representation).

(B) As in (A), showing a WGA-sample (cell 3, nine bands) and a positive control (genomic DNA, ten bands). Left lanes removed (additional samples, not needed for representation).

(C) Representative gel result of an experiment containing both high- and low-quality samples. In lane 1, a 100 bp DNA ladder was used. High-quality cells 4, 7 and 9 all have 9 bands while low-quality cells 5, 6 and 8 have 5, 3 and 5 bands, respectively. Right lanes removed (additional samples, not needed for representation).

d. Re-measure the DNA concentration using high sensitivity fluorometry, e.g., a Qubit fluorometer, dsDNA Quantification Assay, and Assay Tubes (Invitrogen) according to the manufacturer's instructions available here.

7. Prepare aliquots of 30 μL in Safe-Lock Tubes (0.5 mL, Eppendorf).
8. Store aliquots at −20°C until use.

Hardware specifications

The complete pipeline described here runs on a high-performance computing cluster. For one single cell processed by WGA and sequenced at 15× coverage, along with a bulk germline control sample sequenced at 30× coverage, expect to use a maximum of 100 gigabytes (GB) of random

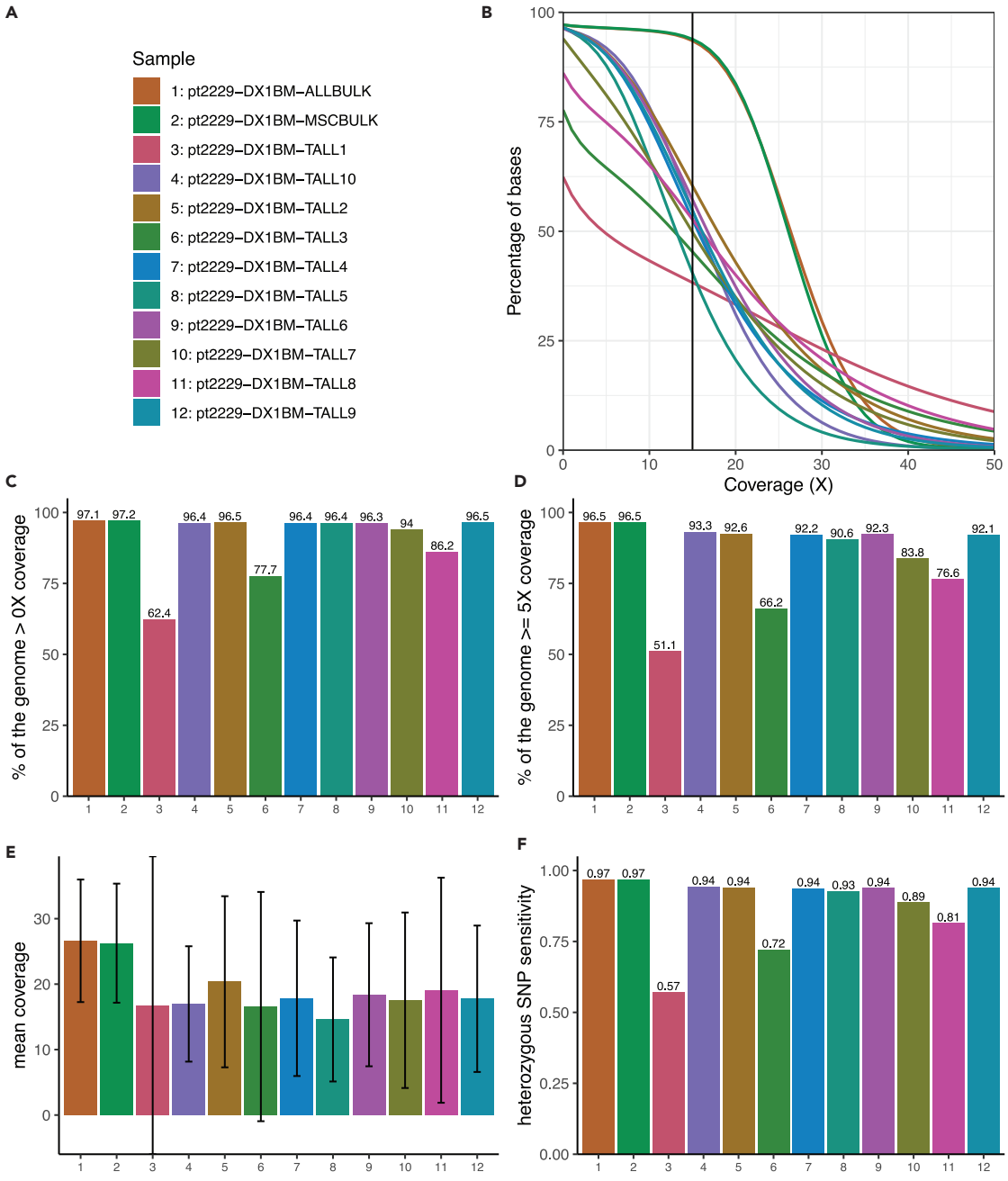


Figure 5. Coverage and heterozygous SNP detection sensitivity outputs from pre-PTATO QC report

(A) Sample information of a single PTATO run. In this example, one bulk tumor sample, one bulk MSC sample, and ten WGA cells were sequenced from the bone marrow (BM) of T-ALL patient pt2229 taken at diagnosis (DX1). WGA samples were sequenced at 15X coverage, while bulk samples were sequenced at 30X.

(B) The percentage of bases with at least X coverage is plotted against the coverage per sample.

(C) Bar graph illustrating the quantification of the genomes with more than 0X coverage.

(D) Same as (C), but more than or equal to 5X coverage.

(E) Bar plot depicting the mean read coverage of the samples with error bars showing the standard deviation.

(F) Bar plot depicting the heterozygous SNP detection sensitivity per sample which is a theoretical estimate of the sensitivity to detect heterozygous SNPs based on the coverage and base quality distributions.

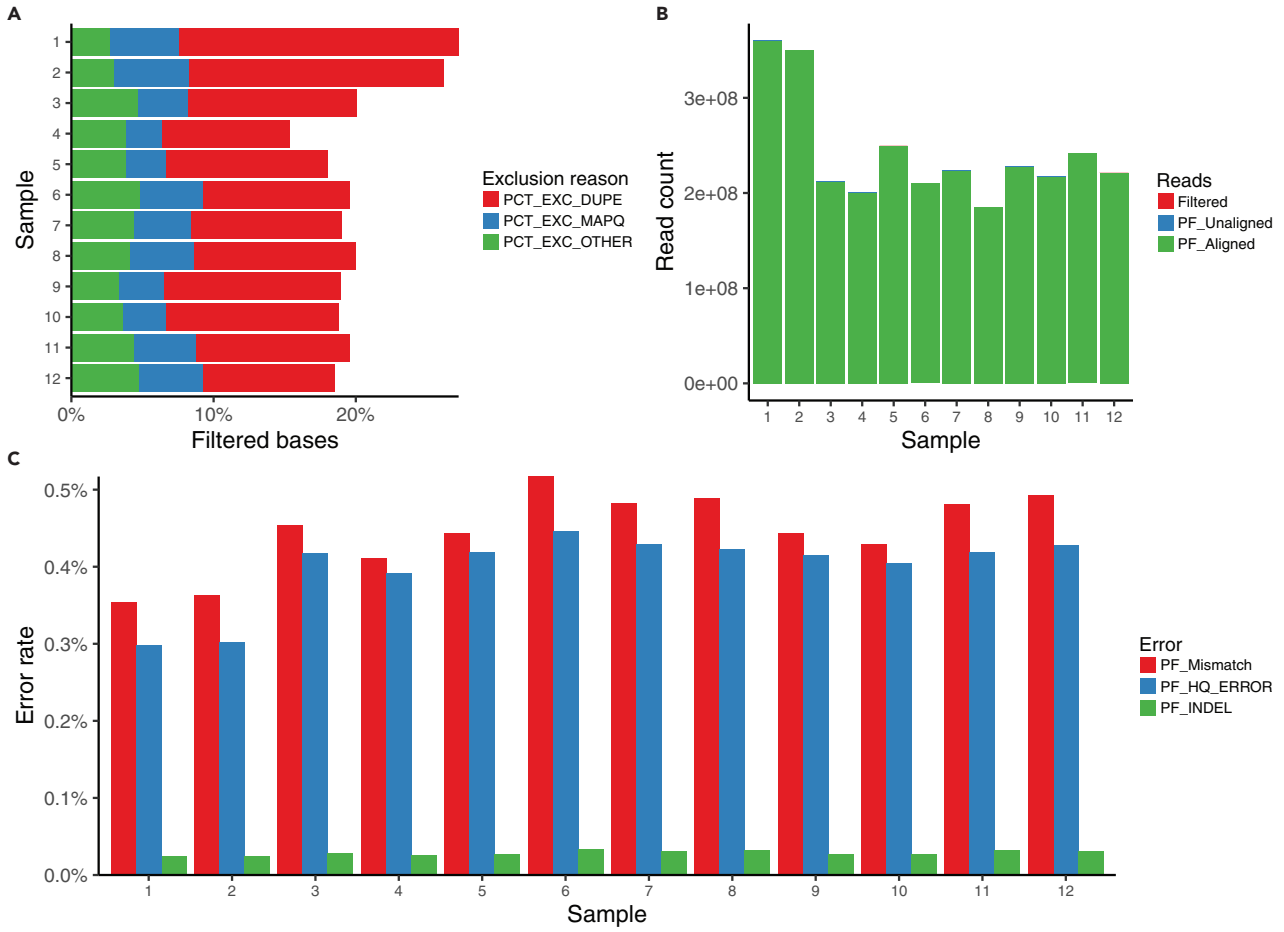


Figure 6. Sequencing read quality outputs from pre-PTATO QC report

(A) Bar plot depicting the percentage of bases that were filtered. The color indicates the reason a base is excluded which can be because a base is found in a read marked as a duplicate ("PCT_EXC_DUPE"), low mapping quality ("PCT_EXC_MAPQ"), or due to other reasons ("PCT_EXC_OTHER"). The other reasons are further delineated in the quality control table generated by PTATO.

(B) Bar plot depicting the number of reads per sample. The color indicates if the read was filtered out ("Filtered"), unaligned ("PF_Unaligned") or aligned ("PF_Aligned"). PF stands for reads that have passed Illumina's filters.

(C) Bar plot depicting the error rate per sample. This is shown separately and in different colors for the percentage of mismatched bases in aligned reads ("PF_Mismatch"), the percentage of mismatched bases in aligned reads with a mapping quality of at least 20 ("PF_HQ_ERROR"), and the number of indels per 100 aligned bases ("PF_INDEL").

access memory (RAM) and approximately 1500 GB of hard disk space during data processing. PTATO requires 100–200 central processing unit (CPU) hours per 15× WGA sample.

Download and install software and tools

Timing: 5 h

The PTATO workflow is implemented in Nextflow, enabling flexible use of pipelines on multiple scientific platforms i.e., a high-performance computing cluster. Additionally, it requires the installation of R and Apptainer/Singularity as dependencies.

9. Install Nextflow v21.10.6.5661 following the official documentation.
10. Install Apptainer/Singularity v3.5 following the instructions.
11. Install the container SnpSift v4.3.1 available in PTATO's Demo Dataset in folder "containers".

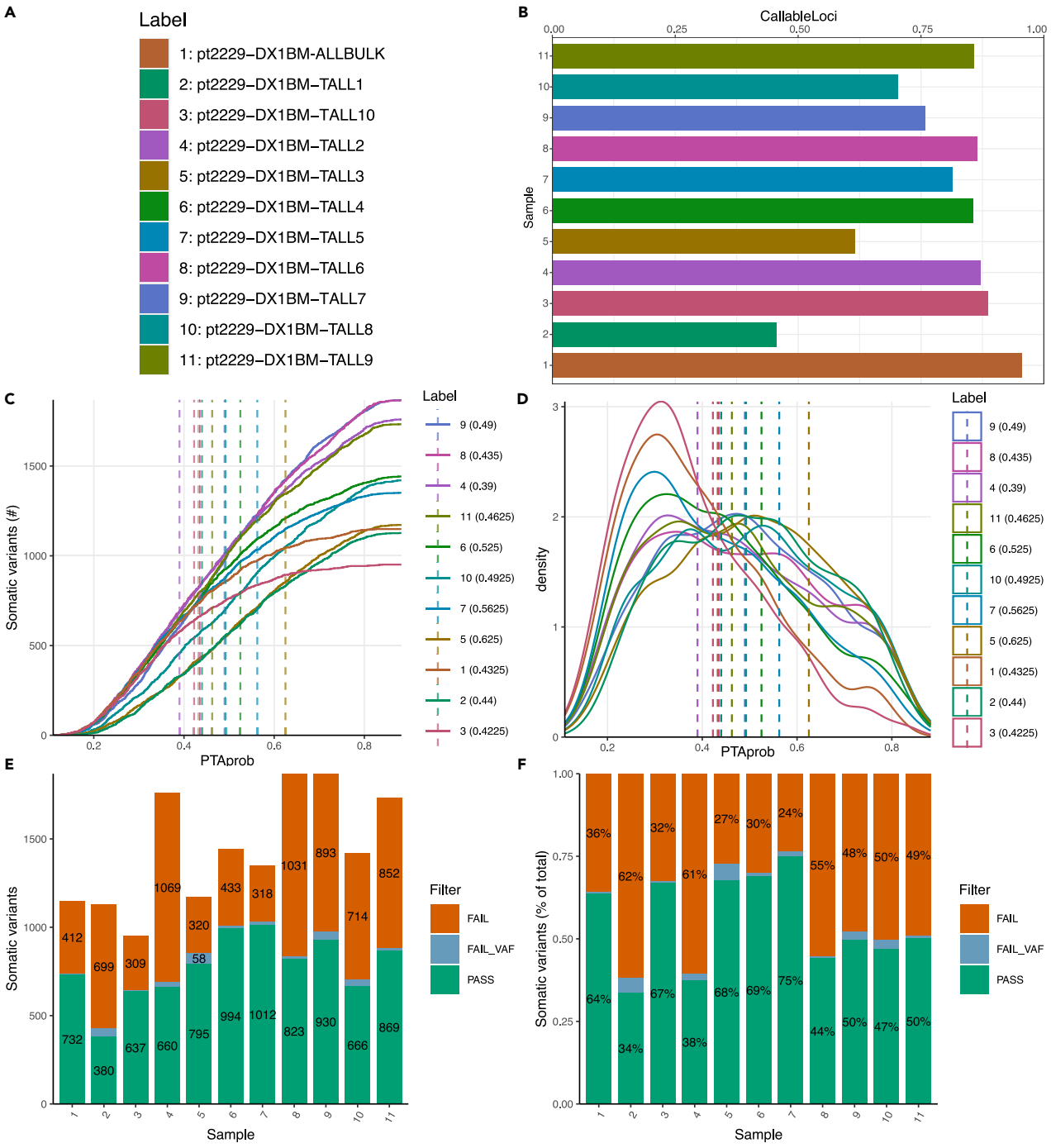


Figure 7. Callable loci and PTA probability outputs from post-PTATO QC report

(A) Sample information. These results were obtained from the same PTATO run as in Figures 5 and 6.

(B) Bar plot depicting the fraction of the genome where sufficient data has been collected to confidently call genetic variants (CallableLoci), per sample.

(C) Cumulative distribution plot depicting the PTA probability of somatic variants and cutoff (dotted line), per sample. The PTA probability cutoff here is the mean of the PTA probability cutoffs generated by Walker and Cossim (cosine similarity with PTA signature).

(D) Density plot of PTA probabilities, per sample.

(E) Bar plot depicting the absolute number of somatic variants that have passed or failed filtering step. Variants with VAF<VAF_threshold will be flagged as "FAIL_VAF", variants with PTAprob>PTAprobCutoff will be flagged as "FAIL", while variants with PTAprob<=PTAprobCutoff will be flagged as "PASS" in the FILTER field.

(F) Same as (E), but the percentage of passed and failed variants relative to total number of variants.

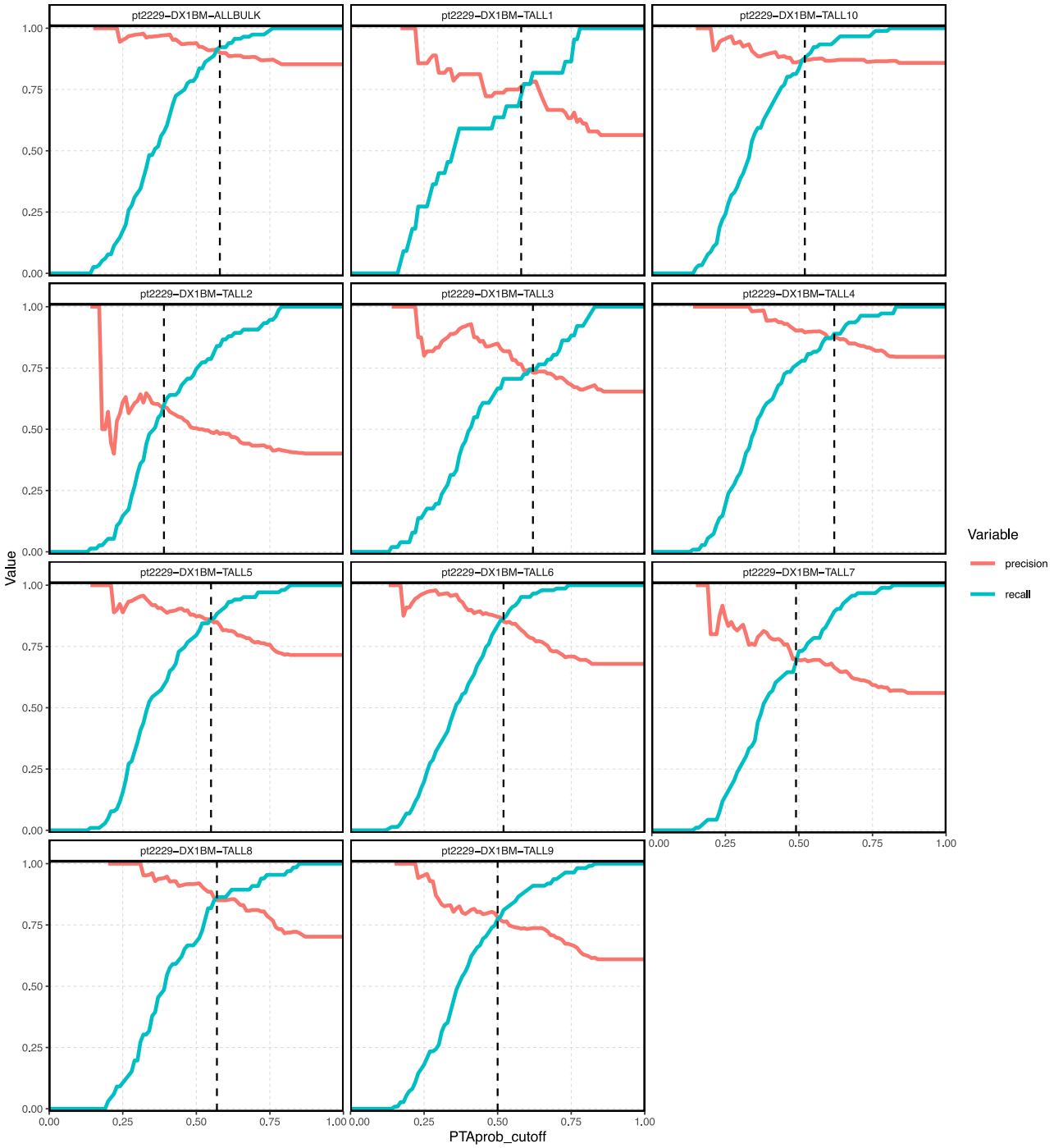


Figure 8. Precision-Recall plot outputs from post-PTATO QC report

Precision-Recall plot based on Walker, per sample. The PTA probability cutoff is where precision and recall intersect.

a. Move this file to the Singularity cache directory for your cluster.

12. Install R v4.1.2 or higher following the instructions.
13. Install the necessary R libraries following the installation guides:
    a. ggplot2 v3.4.1.
    b. MutationalPatterns v3.6.0.

c. VariantAnnotation v1.42.1.
d. StructuralVariantAnnotation v1.12.
e. BSgenome.Hsapiens.UCSC.hg38 v1.4.4.
f. BSgenome v1.72.0.
g. copynumber v3.19.
h. cowplot v1.1.3.
i. gtools v3.9.5.
j. randomForest v4.7-1.1.
k. scales v1.3.0.

14. Use git to obtain a copy of the latest version of the PTATO repository:

```
git clone git@github.com:ToolsVanBox/PTATO.git
```

Note: Alternatively, download the latest release from the GitHub page here: https://github.com/ToolsVanBox/PTATO.

Note: For issues installing PTATO, please see troubleshooting problem 7.

Download and install resources for PTATO

Timing: 3 h

We provide most of the resources required for PTATO using GRCh38 as reference genome, this is available through the PTATO repository. However, some files need to be extracted or downloaded separately, these are available in PTATO's Demo Dataset.

15. Download the following files from the Demo Dataset:
    a. The reference genome files in the folder "resources".

Note: Be sure to download *.dict; *.fasta and *fasta.fai files. In addition, we recommend downloading the *.fasta.amb, *.fasta.ann, *.fasta.bwt, *.fasta.dict, *.fasta.gridsscache, *.fasta.pac and *.fasta.sa files. However, the latter files will be created by the pipeline if not provided.

    b. Move the reference genome resource files to your PTATO directory `[PTATO_dir]/resources/hg38/`.
16. Download the Segmented HAPlotype Estimation and Imputation Tools (SHAPEIT) v4 resource files following the official guidelines.
    a. Phasing_reference files.
    b. shapeit_maps files.
    c. Unzip the reference genome file.

```
tar -zxf genetic_maps.b38.tar.gz
```

    d. Create the following SHAPEIT directories: `[PTATO_dir]/resources/hg38/shapeit/Phasing_reference/` and `[PTATO_dir]/resources/hg38/shapeit/shapeit_maps/`
    e. Move the SHAPEIT resource files from step 16a and 16b to the newly generated folders (16d). Respectively, `[PTATO_dir]/resources/hg38/shapeit/Phasing_reference/` (step 16a) and `[PTATO_dir]/resources/hg38/shapeit/shapeit_maps/` (step 16b).
    f. Remove "chr" prefix in the "phasing reference files" and "shapeit_maps" files if the output of your mapping will be without it.

Note: If using a different reference mapping tool than the one we recommend (See "preprocessing: Map WGS reads to the reference genome, variant calling and annotation and quality control"), this may not be necessary.

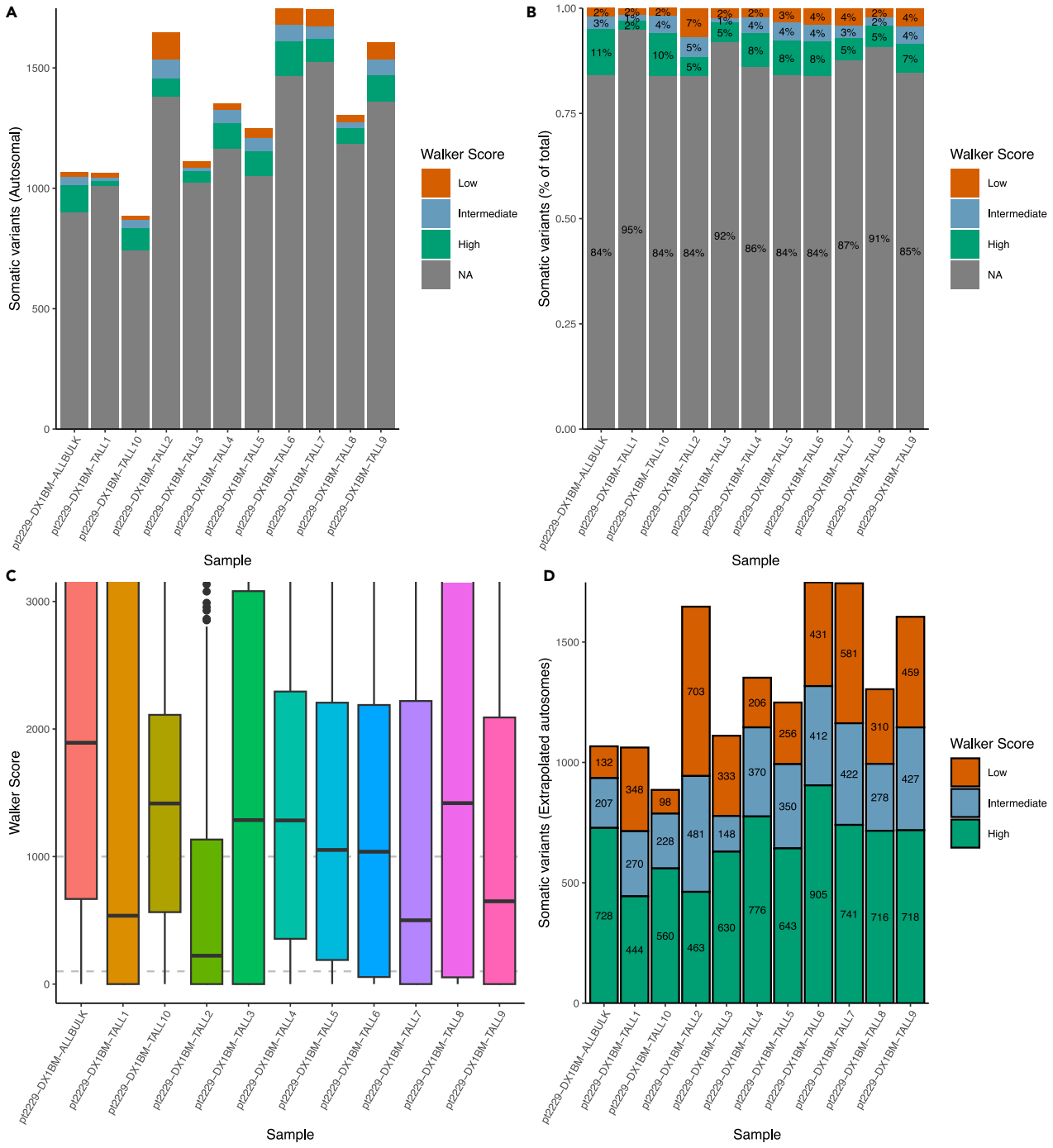


Figure 9. Walker score outputs from post-PTATO QC report

(A) Bar plot of Walker scores across different samples, in absolute counts. Autosomal somatic variants with Walker score <1 will be flagged as "Low", variants with score >1000 as "High", variants with score >=1 and =<1000 as "Intermediate", whereas variants with no score as "NA".

(B) Same as in (A), but as relative proportions.

(C) Boxplot depicting the distribution of Walker scores across different samples. The boxplots show the median (center line), 25th and 75th percentiles (box), the largest values no more than 1.5* the interquartile range (whiskers), and potential outliers.

(D) Bar plot depicting the extrapolated Walker scores across different samples. Here, the distribution of uncalled variants (those initially categorized as "NA") is estimated by using the frequencies of the called variants ("Low", "Intermediate", "High"). It then adds these extrapolated counts to the original counts, providing an estimated total for each Walker score category.

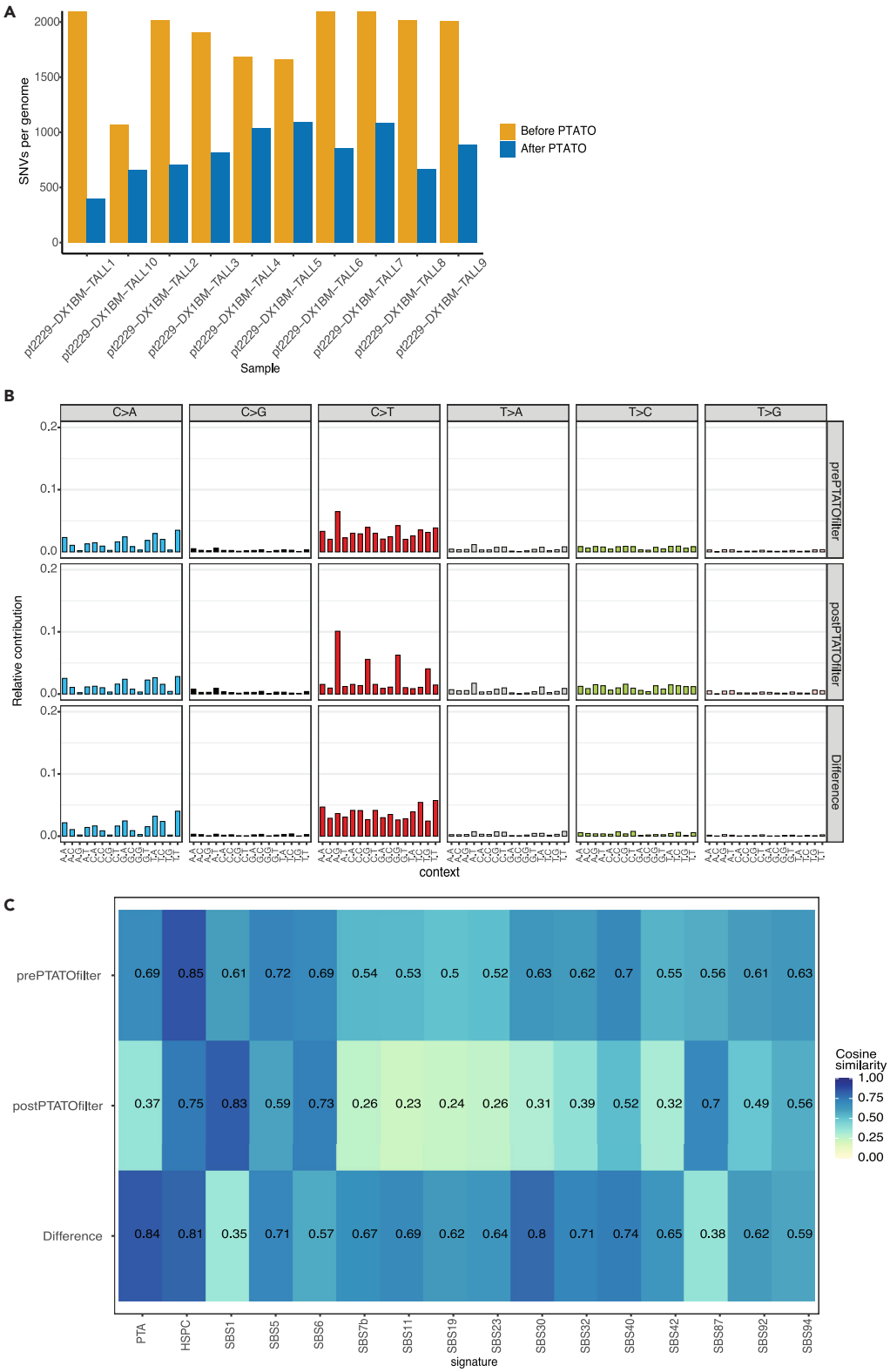


Figure 10. Mutational load and signature before and after PTATO

(A) SNV load before and after PTATO per PTA-amplified single cell from T-ALL patient pt2229. The SNV load per genome is calculated by normalizing the SNV load to the GATK CallableLoci's CALLABLE length.

Figure 10. *Continued*

(B) Mutational signatures in 96-trinucleotide profiles when all SNVs from PTA-amplified single cells are merged. Before PTATO, there is no obvious mutational signature (top profile). After PTATO, the mutational signature SBS1 is clearly present (middle profile). Most mutations filtered out by PTATO are PTA-artefact C>T mutations (bottom profile).

(C) Heatmap depicting the cosine similarities between the "prePTATOfilter", "postPTATOfilter" and "Difference" profiles and individual signatures. For this analysis, 60 COSMIC signatures, the HSPC signature and the PTA signature were included. Signatures with <0.6 cosine similarity in all three profiles were removed.

17. Extract the following reference genome resource files from the PTATO repository here.
    a. [PTATO_dir]/resources/hg38/gripss/gridss_pon_breakpoint.tar.gz.
    b. [PTATO_dir]/resources/hg38/cobalt/COBALT_PTA_Normalized_Full.tar.gz.
    c. [PTATO_dir]/resources/hg38/smurf/Mutational_blacklists/Fetal_15x_raw_variants_hg38.tar.gz.
    d. [PTATO_dir]/resources/hg38/smurf/Mutational_blacklists/MSC_healthyBM_raw_variants_hg38.tar.gz.

Verify whether PTATO works using the demo dataset

Timing: 3 h

Before running the PTATO pipeline with your own samples, check whether the software is installed properly and if the pipeline works on your computing cluster using the input files in the Demo Dataset. The test dataset contains bulk WGS data of a clonal cell line, to be used as the germline control, in addition to a WGA data from a single cell derived from this clone. The following steps are specific for SLURM as an executor for Nextflow. Adjust the steps for your specific executor if needed.

18. Edit the `nextflow.config`in the `[PTATO_dir]/configs/`by specifying the base and cache directories for your cluster.
    a. Change the directory to the base directory of your cluster in `runOptions`.
    b. Change the path in `cacheDir`to the cache directory on your cluster.

```
singularity {
  enabled = true
  autoMounts = true
  runOptions = '-B /path/to/your/base_dir-B $TMPDIR:$TMPDIR'
  cacheDir = '/path/to/your/cache_dir'
}
```

19. Download the SHAPEIT resources specifically for the Demo Dataset, available here.
    a. Move these files according to the following path structure.

```
[PTATO_dir]/demo/resources/hg38/shapeit
  ./Phasing_reference
    ./.vcf.gz
    ./.vzf.gz.csi
    ./.vzf.gz_tmp
  ./Shapeit_maps/
./.gz
  ./Chromosomes.txt
```

20. Download the `resources.config` file from the Demo Dataset, available here.
    a. Move this config file to the `[PTATO_dir]/demo/configs/` directory.
21. Download BAMS and corresponding index files (BAI) and VCF input file from the Demo Dataset respectively in the folders "`input/bams/donor`" and "`input/vcf/donor`".
22. Move the input files to your PTATO directory according to the following input folder structure.

⚠ CRITICAL: The structure of the input folder specified below is required to run PTATO.

```
/path/to/bams_dir
  ./donor
    ./mycontrol.bam
    ./mycontrol.bai
    ./mysample1.bam
    ./mysample1.bai
    ./mysample2.bam
    ./mysample2.bai
    ..
/path/to/vcfs_dir
  ./donor
    ./myfile.vcf(.gz)
```

Here.

```
/[PTATO_dir]/demo/input/bams/
  ./donor
    ./PMCAHH1-WT-C6_dedup.ch20.bam
    ./PMCAHH1-WT-C6_dedup.ch20.bai
    ./PMCAHH1-WT-C6SC1_dedup.ch20.bam
    ./PMCAHH1-WT-C6SC1_dedup.ch20.bam
[PTATO_dir]/demo/input/vcf/
  ./donor
    ./demo.chr20.vcf
```

23. Edit the `run-template.config` file in the `[PTATO_dir]/configs/` directory.
    a. Adjust the path to the `resources.config` file.

```
includeConfig "${projectDir}/demo/configs/resources.config"
```

    b. Specify which part(s) of PTATO to run under header `run {}` set to `'=true'` or `'=false'`.

Note: Adjust this for each new PTATO run as it contains the paths to the input files and name(s) of the sample(s).

Note: We recommend starting only with the QC and once this has finished successfully, moving on to variant calling. It is also possible to set all parameters to `'=true'` if you want to run the workflow in at once.

```
run {
    snvs = false
    QC = true
    svs = false
    indels = false
    cnvs = false
    postqc= false
  }
```

c. Specify the paths to the input directories containing the multi-sample VCF and BAM files under header `// TESTING`.

```
// TESTING
input_vcfs_dir = '[PTATO_dir]/demo/input/vcf/'
bams_dir = '[PTATO_dir]/demo/input/bams/'
// END TESTING
```

⚠ CRITICAL: The ID of the donor should not be included in the path, i.e. not `input_vcfs_dir='/path/to/vcf_dir/donor/'`

d. Specify the path to the output directory.

```
out_dir= '/path/to/output_dir'
```

e. Specify the donor ID of the sample and the name of the germline control sample under header `'bulk_names'`.

```
bulk_names = [
    ['donor', 'PMCAHH1-WT-C6'
  ]
```

Note: Here you can only put two samples per line. The first position is for the donor ID, so define a string that is representative for all samples included in this run. Moreover, the donor ID should match with the folder name. The second position should always be your germline control sample (e.g. bulk MSCs, T cells or B cells).

The remaining fields in the `run-template.config` file are optional and can be left empty as PTATO will generate these files. If you need to rerun parts of PTATO at a later point, you can specify the paths to the files previously generated by PTATO. This allows the old files to be re-used, saving time and resources.

Note: Your `run-template.config` should look identical to `the run_demo.config` file in the "output" folder in the Demo Dataset.

24. Start the PTATO pipeline, for example on SLURM workload manager.

```
/path/to/nextflow run [PTATO_dir]/ptato.nf -c [PTATO_dir]/configs/run-template.config --out_
dir /path/to/output_dir -profile slurm -resume
```

Note: A shell template is provided on the PTATO Github page, "`start_pipeline.sh`", in addition to a shell template for the singularity, "`start_pipeline_singularity.sh`".

PTATO will now generate the QC report which can be found in your newly generated QC folder, e.g., `/path/to/output_dir/QC/reports/donor_ID`.

After running PTATO for QC, re-run PTATO for the calling of somatic nucleotide variants (SNV), insertions and deletions (indels), copy number variants (CNVs), structural variants (SVs) and post QC (postqc).

25. Under the header `run {}` in the `run-template.config` file in `[PTATO_dir]/configs/` directory set the options for snvs, indels, svs and cnvs to true.

```
run {
  snvs = true
  QC = false
  svs = true
  indels = true
  cnvs = true
  postqc= true
}
```

Note: The "`snvs = true`" and "`cnvs = true`" parts of PTATO are required to run the "`svs = true`" part.

26. Start the PTATO pipeline following step 24.
27. Check whether PTATO ran successfully by comparing your output with the output in the Demo dataset "output" folder. Once you have determined that you are able to run PTATO, proceed with preparing your samples.

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Biological samples		
Genomic DNA positive control	Human	N/A
Chemicals, peptides, and recombinant proteins		
Ethanol 100% absolute analytical reagent, ACS, ISO	Boom	84028185.2500
TE buffer, 1X solution, pH 8.0, low EDTA[a]	G-Biosciences	786-150
PBS (Ca2+/Mg2+ free) (one tablet dissolved in 500 mL distilled water)	Gibco	18912014
EDTA, UltraPure, 0.5 M, pH 8.0	Invitrogen	15575020
Bovine serum albumin	Sigma-Aldrich	A3311

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
UltraPure agarose	Invitrogen	16500500
Midori Green Advance	NIPPON Genetics	MG04
Gel loading dye, purple (6X)	NEB	B7025S
Critical commercial assays		
Qubit dsDNA quantification assay, high sensitivity	Invitrogen	Q32854
Qubit assay tubes	Invitrogen	Q32856
Qubit 4 fluorometer	Invitrogen	Q33238
ResolveDNA whole genome amplification kit - 96 reactions[b]	BIOKÉ	BSG 100136
ResolveDNA whole genome amplification kit v2.0[b]	BIOKÉ	BSG 100545
ResolveOME whole genome and transcriptome single-cell core kit, 96 reactions, with adapter sets A, B[b]	BIOKÉ	BSG 100956
WGA QC mix (50 rx)	VyCAP	WGA QC-50
QIAGEN Multiplex PCR *Plus* kit (100)	QIAGEN	206152
QIAamp DNA micro kit	QIAGEN	56304
Software and algorithms		
Nextflow v21.10.6.5661	Di Tommaso et al.[8]	https://www.nextflow.io/
Apptainer/Singularity v3.5	Kurtzer et al.[9]	https://apptainer.org/
SnpSift v4.3.1	Cingolani et al.[10]	https://surfdrive.surf.nl/files/index.php/s/l3FX6eLnTtuVK1g?path=%2F.
R v4.1.2	N/A	https://www.r-project.org/
PTATO v1.3.3	Middelkamp et al.[1]	https://github.com/ToolsVanBox/PTATO
NF-IAP v1.3.1	Princess Máxima Center	https://github.com/ToolsVanBox/NF-IAP
ggplot2 v3.4.1	Wickham et al[11]	https://ggplot2.tidyverse.org/
MutationalPatterns v3.6.0	Manders et al.[12]	https://bioconductor.org/packages/release/bioc/html/MutationalPatterns.html
VariantAnnotation v1.42.1	Obenchain et al.[13]	https://bioconductor.org/packages/release/bioc/html/VariantAnnotation.html
StructuralVariantAnnotation v1.12	Cameron and Dong[14]	https://www.bioconductor.org/packages/release/bioc/html/StructuralVariantAnnotation.html
BSgenome.Hsapiens.UCSC.hg38 v1.4.4	The Bioconductor Development Team[15]	https://bioconductor.org/packages/release/data/annotation/html/BSgenome.Hsapiens.UCSC.hg38.html
Bsgenome v1.72.0	Pagès[16]	https://bioconductor.org/packages/release/bioc/html/BSgenome.html
copynumber v3.19	Nilsen, Liestoel, and Lingjaerde[17]	https://bioconductor.org/packages/3.14/bioc/html/copynumber.html
cowplot v1.1.3	Wilke et al.[18]	https://cran.r-project.org/web/packages/cowplot/index.html
gtools v3.9.5	Warnes et al.[19]	https://cran.r-project.org/web/packages/gtools/index.html
randomForest v4.7-1.1	Wiener et al.[20]	https://cran.r-project.org/web/packages/randomForest/index.html
scales v1.3.0	Wickham[21]	https://scales.r-lib.org
SHAPEIT v4	Delaneau et al.[22]	https://github.com/odelaneau/shapeit4?tab=readme-ov-file
GATK's HaplotypeCaller v4.1.4.1	Poplin et al.[23]	https://gatk.broadinstitute.org/hc/en-us/articles/360037225632-HaplotypeCaller
Burrow-Wheeler Aligner v0.7.17	Li et al.	https://github.com/lh3/bwa
Other		
Safe-Lock tubes, 0.5 mL	Eppendorf	0030121023
PCR-Cooler, 0.2 mL (for 96 reaction format)	Eppendorf	3881000015
twin.tec PCR plate 96 LoBind (for 96 reaction format)	Eppendorf	0030129504
Hard-Shell PCR plates 384-well, thin-wall (for 384 reaction format)	Bio-Rad	HSP3805
Microseal 'F' PCR plate seal, foil, pierceable	Bio-Rad	MSF1001
Cell sorter	Sony	SH800

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
SPRIPlate 96R ring super magnet plate (only for 96 reaction format)	Beckman Coulter	A32782
Thermoshaker	Grant-Bio	PHMT
Block for 1 × 96-well PCR microplate[c]	Grant-Bio	PSC96
Plate spinner	Eppendorf	Centrifuge 5920 R
100 bp DNA ladder	Invitrogen	15628050
Sapphire PCR 8-tube strips, 0.2 mL, PP, natural (alternative to PCR plate)	Greiner Bio-One	673210
Sapphire 8-cap strip, PP, natural, domed (alternative to PCR plate)	Greiner Bio-One	373270
Hard-Shell 96-well PCR plates, high profile, semi skirted, clear/clear (alternative to PCR strip)	Bio-Rad	HSS9601
AMPure XP bead-based reagent[d]	Beckman Coulter	A63882
DNA LoBind tube, 1.5 mL	Eppendorf	0030108051
Corning 96-well black polystyrene microplate (only for plate-based DNA quantitation)	Merck	CLS3603
Plate reader (only for plate-based DNA quantitation)	BMG LABTECH	FLUOstar Omega
Single-cell dispenser (only for 384 reaction format)	Tecan	HP D100
Uno D1 Dispensehead Cassettes (only for 384 reaction format)	Tecan	30230843
Uno T1 Dispensehead Cassettes (only for 384 reaction format)	Tecan	30230842
Thermal cycler	Bio-Rad	T100, CFX384

[a]Alternative: Elution Buffer (Part of all WGA kits supplied by Bioskryb).
[b]Choose one of the available versions. See section 'choosing an appropriate whole-genome amplification method' (before you begin) for details.
[c]The same block can be used for the 96 and 384 reaction formats. Use tape to secure the 384-well plate.
[d]Alternative: ResolveDNA Bead Purification Kit – 96 Reactions (BIOKÉ, BSG100182).

MATERIALS AND EQUIPMENT

FACS buffer

Reagent	Final concentration	Amount
PBS (Ca^{2+}/Mg^{2+} free)[a]	1X	50 mL
Bovine Serum Albumin	0.05% (w/v)	25 mg
EDTA (0.5 M)	1 mM	100 µL
Total	N/A	50 mL

[a]Before use, sterilize through autoclaving at 121°C for 15 min.

After sterile filtration through a 0.2 µM filter, store the FACS buffer at 4°C for at least one year, or until protein aggregates appear.

STEP-BY-STEP METHOD DETAILS

Preparation of a reference DNA sample

Timing: Variable (up to 4 h)

For the execution of this protocol, bulk WGS data of the donor is required to remove germline variants in the single-cell WGS data. In this step, the availability of a pre-existing dataset is checked, or if unavailable, a suitable DNA sample is prepared for WGS. In our experience, the filtering of germline variants performs best when the germline normal sample is sequenced at 30X coverage (100 Gb).

1. Determine whether a pre-existing WGS dataset is available.

⚠ CRITICAL: The WGS dataset needs to meet the following requirements: obtained from the same individual for which single cells will be processed, a base coverage of at least 30X coverage for a bulk sample, and derived from a different developmental lineage or at least different cell type as the assessed single cells. For example, use mesenchymal stromal cells (MSCs) as a reference for bone marrow-derived blood cells or use a bulk myeloid cell population as reference when investigating lymphoid cells at single-cell level.[7,24]

Note: A different developmental lineage or cell type is recommended to reduce the clonal relatedness between the single cells and the reference as much as possible. Early acquired somatic variants that overlap between the reference and the single cells will be removed from the dataset, as they will be falsely categorized as germline variants.

2. If no dataset is available, prepare a bulk DNA sample for WGS.
   a. Collect at least 30,000 cells of a different developmental lineage or at least different cell type of the individual studied at single-cell level.

   Note: If possible, collect up to 1 million cells. For blood lineages derived from bone marrow, we recommend using *in vitro*-expanded mesenchymal stromal cells. When using peripheral blood samples, a different cell type can be isolated in bulk through fluorescence-activated cell sorting (FACS) (e.g. monocytes as reference when studying T-cells at a single-cell level.[7]

   b. Extract genomic DNA using a QIAamp DNA Micro Kit according to manufacturer's instructions available here.
   c. Include this sample in step 34 (step-by-step method details) and perform WGS to a depth of 30X (100 Gb) using 200–500 ng of DNA as input.

   Note: At least 50 ng of DNA is required for low-input WGS.

Preparation of cell collection plates

Timing: 1 h

In this step, the plates in which the cells will be deposited during single cell sorting are prepared. Working in a DNA-free pre-PCR hood is recommended to prevent contamination.

3. Design the cell collection plate layout.
   a. Determine the number of cells that can be worked up at the same time.

   Note: The number of cells per plate should be based on the number of desired cells, the success rate of the WGA, and the size of the single-use reagent aliquots. Before WGA, the best cells can be selected through the index sorting data, and less cells on the edge or outside of the desired gates can be excluded. In our hands, the WGA process generally has a success rate of ±80%. This percentage may vary between users and/or institutes. Experienced users can reliably process more than 24 cells in parallel for manual use of the kits. For automated dispensing of reagents (e.g. using a D100 Single Cell Dispenser (HP)), full 384 wells plates can be processed at once.

   b. Include controls wells (Figures 1A and 1B).
   c. Fill the 96 or 384 wells-plate in a column-based way (Figures 1A and 1B).

   Note: The type of plate that is used is dependent on the kit version that is used, i.e. the 96 well format or the 384 well format. We recommend leaving the outer wells empty. For the 384 well

format, downstream steps by multi-channel pipette (8 channels) are facilitated by alternating filled and empty rows for experiments where plates are only partially used.

Note: When sorting multiple cell types, it is important to think about the order of sorting. For example, when sorting abundant and rare cell types, such as leukemic blasts and HSPCs, sort the abundant cell types first before starting to sort the rare cell types. In this way, the abundant cell type(s) can be sorted quickly, and the remaining sample can be used to sort the rare cell type(s).

4. Thaw the Cell Buffer (Bioskryb) on ice for 20 min.

Note: The Cell Buffer is part of the ResolveDNA / ResolveOME kits (Bioskryb).

5. If applicable: start the UV sterilization of a DNA-free pre-PCR hood.
6. For 96 well plates, pipette 3 μL of Cell Buffer per well into a twin.tec PCR Plate 96 LoBind (cell collection plate) using reverse pipetting. For 384 well plates, dispense 1 μL per well using a D100 Single Cell Dispenser (HP) or equivalent.

⚠ CRITICAL: Keep all reagents and plates cold by placing them on ice or in cooling blocks.

7. Seal the wells using a Microseal 'F' PCR Plate Seal and centrifuge at 1,000 × *g* for 1 min at 4°C.

Note: When filling up to five columns, it is possible to cut the foil in half before removing the non-adhesive layer to reduce waste.

8. Store plates on ice/4°C or at −20°C until use.

Note: Plates can be stored on ice for same-day experiments. Plates can be stored at −20°C for several weeks without loss of whole-genome amplification efficiency.

Isolation of single cells prior to whole-genome amplification

Timing: 3–6 h

The isolation of single cells through cell sorting is usually required due to the small volumes used during PTA. For the isolation of specific cell types, FACS is required. Depending on the experimental set-up, cells or nuclei can be isolated directly from primary tissue, live cell culture, or viably frozen cells.

9. Perform the start-up procedure of the cell sorter.

Note: Depending on facility availability, sorting equipment may vary. We have performed all experiments here using a Sony SH800S cells sorter in combination with a 100 μM filter chip.

10. Perform plate calibration using an empty cell collection plate, covered with foil.

⚠ CRITICAL: Plate calibration needs to be performed carefully, as the cell collection surface is small for 96- and 384-well plates. The sorted droplet needs to be in the exact center of the well.

11. Create single-cell suspensions of the sample of interest.

Optional: Thaw viably frozen cells and/or trypsinize the sample into a single-cell suspension according to routine procedures.

Optional: Perform antibody staining on ice with appropriate lineage markers to identify cell populations on FACS.

⚠ CRITICAL: Viable cells can be positively selected through Calcein AM staining and negative selection through propidium iodide staining. When sorting nuclei, it is important to not use DNA-binding stains (e.g. propidium iodide, DAPI) as these may interfere with the WGA.

12. Wash the single-cell suspension using FACS buffer (materials and equipment).
    a. Centrifuge the sample at 350 × *g* for 5 min at 4°C.
    b. Add 5 mL of ice-cold FACS buffer.
    c. Centrifuge the sample at 350 × *g* for 5 min at 4°C.
    d. Resuspend the cells in ice-cold FACS buffer, to maximum 10 million cells per mL.
13. If the cell collection plates were stored at −20°C, place them on ice to thaw.

Note: This step can also be performed during the optional antibody staining.

14. Acquire 20,000–50,000 events on the cell sorter and set gating strategy.
    a. Exclude debris and dead cells using the forward-scatter-area (FSC-A) and side-scatter-area (SSC-A) channels (Figure 2A).
    b. Select single cells using the forward-scatter-area (FSC-A) and forward-scatter-height (FSC-H) channels (Figure 2B).

Optional: Create gates with positive and negative selection of additional dyes or fluorescent antibodies used in staining of the cells (Figures 2C–2E).

    c. Create a sorting gate which is at least 10% smaller than the population of interest.

Note: Setting the sorting gates stringently is recommended, as cell sorting can be too lenient for cells on the border of or just outside of the gated population (Figures 2F–2J).

Optional: For rare cell populations, or when limited cell numbers are available, the selection and sorting gates can be placed more leniently (Figure 2E). However, the cells to be selected for WGA need to be selected stringently post-sorting using index data.

15. Place a PCR-cooler, stored in −20°C for at least 2 h, in the plate holder of the sorting device.

Note: A PCR cooling block is recommended to ensure cell viability of the sorted cells. If a PCR cooler does not fit in the cell collection chamber, make sure that the cell collection plate fits securely in the plate holder. If not, place the cell collection plate in a supporting plate with V-bottom wells.

16. Sort one cell per well of the desired population(s).
    a. Centrifuge a foil-covered collection plate at 1,000 × *g* for 1 min at 4°C.
    b. Remove the foil and place it on the PCR cooling block in the cell collection chamber.
    c. Sort one cell per well of the desired population(s), while recording the index of the sorted events.

⚠ CRITICAL: It is important to record the index data of the sorted events in the software of the sorting device to check the lineage identity of the sorted cells after sorting.

Optional: For small cell numbers and/or rare populations, device-specific settings can be adjusted to be more lenient (e.g. adjusting the sort mask to 40 on a Sony SH800).

d. Once sorting is finished, remove the plate from the cell collection chamber and close off the wells using a Microseal 'F' PCR Plate Seal.
e. Centrifuge the collection plate at 1,000 × *g* for 1 min at 4°C.
f. Place the plate on dry ice to snap freeze.

17. Check the index of the sorted events and adjust sort gates if necessary.

Note: If all sorted events are on the edges of the sort gates, decrease the size of the sort gate to exclude these events, as they are often of bad quality.

18. Repeat steps 15–17 for the remainder of the plates, or until the sample is finished.
19. Transfer the cell collection plates from dry ice to −80°C.

Pause point: The cells can be processed for WGA immediately (steps 20–24). If not processed immediately, cell collection plates containing cells can be stored at −80°C for four to five months.

⚠ CRITICAL: Perform the WGA within four to five months after cell sorting, preferably as soon as possible. In our hands, the amplification efficiency is reduced with long-term storage resulting in a higher fraction of cells with reduced yield of the amplification and/or uneven amplification of the genome.

Single-cell whole-genome amplification through primary template-directed amplification

Timing: Variable

In this step, the sorted single cells are processed for WGA. To ensure sufficient yield and an even amplification of the genome, follow the protocol closely with special attention to reagent temperature and timing of steps. The time required for this step is dependent on the ResolveDNA/ResolveOME kit version that is used (see manufacturer's instructions, available upon request here).

20. Choose the plate(s) and well(s) to be processed by checking the index data of the sorted cells.
    a. Exclude cells too close to the debris cloud or which may be doublets based on the forward- and side-scatter channels (Figures 2A, 2B, 2F, and 2G).
    b. Exclude cells outside of the (optional) lineage gates (Figures 2C–2E and 2H–2J).

Note: When more lenient sorting strategies are used to include as many events as possible, stringently exclude all cells which may be of the wrong identity. Cells closer to the debris cloud can be worked up, but this may result in reduced efficiency and therefore increased cost of reagents.

21. Perform the single-cell WGA using the ResolveDNA/ResolveOME kit according to manufacturer's instructions (available upon request here).

Note: Cell lysis may need optimization depending on the cell type that is used (see troubleshooting problem 2).

22. Perform the cleanup of amplified DNA according to the manufacturer's instructions (available upon request here).

Optional: The cleanup protocol may be adapted in the following way: Use of alternative DNA cleanup beads, such as AMPure XP beads. Elute using Low EDTA TE buffer (G-Biosciences) instead of the provided elution buffer.

Note: Working in a DNA-free pre-PCR hood is recommended to prevent contamination of samples.

Note: Steps 21 and 22 can be altered when using alternative versions of ResolveDNA or ResolveOME kits according to the provided protocols. For ResolveDNA Whole Genome Amplification v2.0 (Bioskryb), no cleanup of the amplified DNA is required before proceeding to WGS. For ResolveOME kits (Bioskryb), the next-generation sequencing library preparation is performed with reagents included in the kit.

23. Measure the DNA amplification yield using sensitive fluorometric quantification, e.g., a Qubit fluorometer, dsDNA Quantification Assay, and Assay Tubes (Invitrogen) according to the manufacturer's instructions available here.

Note: Use 1–2 μL of amplified DNA as input for Qubit measurements.

Note: For the quantification of many samples in parallel, a fluorescence-based plate reader can be used to perform the readout and quantification of the samples (for example using 96 Well Black Polystyrene Microplates (Merck) and the FLUOstar Omega (BMG Labtech)).

24. Discard samples with a DNA concentration below 15 ng/μL (96 reactions) or 10 ng/μL (384 reactions) to reduce the used reagents in the next major step.

Note: WGA samples with a very low concentration generally contain unevenly amplified genomes (see Figure 3 in expected outcomes). See troubleshooting section Problem 1–3 for reasons and solutions for a lack of sufficient DNA output, or a DNA output of over 100 ng for the negative control reaction.

Pause point: WGA samples can be stored at −20°C after amplification until further processing.

Quality control through multiplexed PCR of amplified genomic DNA

Timing: 1 day

Before proceeding to WGS, the quality of the WGA step can be checked using shallow WGS or through amplification of ten representative loci of the genome to visualize on an agarose gel. Here, the latter method (referred to as QC PCR) is explained in detail, which is suitable for both euploid and aneuploid human cells (see Figure 4 in expected outcomes).

25. Create 1 ng/μL suspensions of the post-amplification DNA samples and a positive control (human genomic DNA).
    a. Calculate the required dilution volume (sample concentration in ng/μL minus 1).
    b. Pipette the dilution volume of MQ or TE buffer, Low-EDTA (G-Biosciences) into wells of a PCR strip or plate.

Note: Depending on the number of reactions, use PCR strips (8-Tube/8-Cap PCR Strips, Sapphire (Greiner bio-one)) or PCR plates (Hard-Shell 96-Well PCR Plates (Bio-Rad)).

    c. Add to each well 1 μL of the amplified DNA product.

d. Centrifuge the plate at 1,000 × *g* at 15°C–25°C for 30 s - 1 min.

26. Prepare the PCR strip(s)/plate(s).
    a. Prepare the PCR master mix as follows:

Reagent	Amount (μL)
Multiplex PCR *Plus* mix (QIAGEN)	5
ddH2O	3.5
WGA QC mix (Vycap)	1
Total	**9.5**

    b. Add 9 μL PCR master mix per tube/well of the PCR strip/plate.
    c. Add 1 μL of 1 ng/μL sample dilution to each tube/well.
    d. Centrifuge the strip/plate at 1,000 × *g* at 15°C–25°C for 30 s - 1 min.

Note: Depending on the number of reactions, use PCR strips (8-Tube/8-Cap PCR Strips, Sapphire (Greiner bio-one)) or PCR plates (Hard-Shell 96-Well PCR Plates (Bio-Rad)).

27. Run the reaction according to the following PCR program:

Steps	Temperature	Time	Cycles
Initial denaturation	95°C	5 min	1
Denaturation	95°C	30 s	35 cycles
Annealing + extension	72°C	3 min	
Final extension	68°C	10 min	1
Hold	12°C	∞	N/A

Note: A holding temperature of 12°C at the end of the PCR program does not impact stability of the amplification products compared to lower temperatures.

28. Run 5 μL of the QC PCR products on a 3% agarose gel with a DNA dye for 1.5–2 h on 100 V in TAE buffer.
29. Image the gel and export for analysis.
30. Discard samples showing no bands or more than two missing bands (see Figure 4 in expected outcomes).

Note: See troubleshooting problem 4 for reasons and solutions when many samples show more than two missing bands.

Optional: It is possible to genotype samples prior to WGS. For all variants of interest, perform a PCR encompassing the location of the variant using the 1 ng/μL sample dilutions created in this step. Next, perform Sanger sequencing and determine the genotype of each sample.

Pause point: WGA samples can be stored at −20°C until further processing.

Preparing and sending DNA samples for whole-genome sequencing

Timing: 1 h, depending on the number of samples

The following procedure describes how to prepare your WGA samples for WGS. The samples with the highest DNA yield and the most bands in the QC PCR are diluted to an appropriate final concentration. For ResolveDNA samples, you have the choice to perform next-generation sequencing

library preparation in-house or to outsource this step. For ResolveOME samples, library preparation must be performed in-house using reagents included in the kit and indexed using ResolveDNA Multi-Use Library Adapter Plates. Discuss the details with your sequencing facility.

31. Dilute your PTA DNA samples with Low EDTA TE buffer (G-Biosciences) (or similar standard elution buffer) to the final concentration in accordance with your sequencing provider.

Note: Dilute your sample plus an additional 2 μL, allowing you to use 2 μL for the final DNA quantification, see step 32.

Note: We recommend using a total amount between 200 – 500 ng for WGS.

32. Remeasure the DNA concentration using high sensitivity dsDNA fluorometry, e.g., a Qubit HS DNA fluorometer (Invitrogen).

Note: Use 2 μL of eluted DNA for each sample.

33. Construct whole genome libraries using a TruSeq Nano kit (500 bp insert size) (Illumina).

Note: For complete instructions, refer to the TruSeq(R) Nano DNA Library Prep Reference guide by Illumina.

34. Sequence each library with the following method:

Platform options	Illumina NovaSeq 6000 sequencing platform Illumina NovaSeqX (plus) sequencing platform	N/A
Read length	2 × 150 bp (paired end)	N/A
Mean depth of coverage	15× = 50 Gb	WGA samples
	30× = 100 Gb	Bulk germline control
Analysis readout	FASTQ	N/A

Preprocessing: Map WGS reads to the reference genome, variant calling and annotation and quality control

Timing: Variable (up to 5 days)

Before running the PTATO pipeline, map the raw sequencing reads to the reference genome (GRCh38 for human samples) using Burrows Wheeler Aligner (BWA) v0.7.17. Next, perform raw variant calling, filtration, and annotation in multi-sample mode using Genome Analysis Toolkit's (GATK) HaplotypeCaller v4.1.4.1. PTATO requires the single-sample binary alignment map (BAM) and multi-sample variant call format (VCF) files as input. In addition, we advise assessing the quality of your sequencing run, however, this output is not required for running PTATO. Our approach utilizes NF-IAP v1.3.1, which encompasses all steps described above, and starts with FASTQ files as input. Directions to run this pipeline can be found here.

Filtering WGA artificial variants using PTATO

Timing: Variable (up to 3 days)

PTATO has been optimized for implementation in Nextflow v21.10.6.5661 using human samples and SLURM workload manager as the computing cluster. The methodology described below contains instructions for running the workflow with this in consideration. Single base substitutions

caused by WGA-based DNA amplification are removed using PTATO's built-in machine learning approach. In addition, PTATO filters indels and SVs. The PTATO pipeline requires a multi-sample VCF file along with (an) individual BAM file(s), including the germline control sample for each donor as input.

35. Follow the instructions as per steps 18 and 22–24 in the "before you begin" section with the following adjustments.
    a. Edit the `process.config`in the`[PTATO_dir]/configs/`directory to define the time and memory settings for each job.

Note: These settings depend on the number of samples in the PTATO run. The provided template specifies settings for approximately 5 samples. It is critical to scale up time and resources in the `process.config` if running more than 5 samples (see troubleshooting problem 6).

Note: Tailor these settings to your computing cluster. This configuration can be reused for future runs but may require some tweaking for your specific setup.

Note: It is not recommended to run PTATO for more than 20 samples at once. If you have more samples, you can split them into batches (e.g. by cell type) and run GATK's HaplotypeCaller after PTATO, as detailed below in steps 37–39. If running in multiple batches, make sure to include the germline control file for each run.

    b. Edit the run-template.config file in the [PTATO_dir]/configs/ directory.

Note: Adjust this for each new PTATO run as it contains the paths to the input files and name(s) of the sample(s).

Note: When analyzing multiple samples from different donors add an additional row specifying the ID of the donor in the first position and the name of the germline control sample in the second position. This has not been tested extensively by us and recommend executing one PTATO run per donor. Also, keep in mind that you will most likely need to scale up the number of resources you define for each job.

```
bulk_names = [
  ['donor1', 'germline_control_sample1'],
  ['donor2', 'germline_control_sample2'],
]
```

Note: It is also possible to include more than one germline control sample for one donor, for example MSCs and T cells. The output VCF will specify variants from both germline samples in the same file.

```
bulk_names = [
  ['donor', 'germline_control_sample1'],
  ['donor', 'germline_control_sample2'],
]
```

Note: The `nextflow.config` will most likely not need to be modified.

c. After running PTATO for only QC, remove samples that do not pass the QC from further analyses (see Figures 5 and 6 in expected outcomes). The QC report can be found in the directory specified below.

```
QC
  ./reports
    ./donor/
      ./donor.postqcreport.pdf
      ./donor.postqcreport.txt
      ./donor.qcreport.pdf
      ./donor.qcreport.txt
```

d. Run PTATO for indels, SNVs, CNVs and SVs and postqc according to steps 25 and 26 in the "before you begin section".

Note: For PTATO versions up to v1.3.3, SVs and CNVs can only be run on R version 4.1.2, while indels and SNVs can be run using R version 4.3.0.

36. After PTATO has finished, check whether the following directories were generated: QC, indels, intermediate, log, ptato_vcfs and snvs.

Further examine the output carefully and familiarize yourself with the results, ensuring a thorough understanding of the data for downstream analyses. Below, we have included a blueprint for the files we utilize most often.

Within the `snvs` or `indels` folder, you will find the annotated `.ptato.vcf.gz` files with corresponding index files, in addition to the vcf files filtered for the WGA artificial variants in the subfolders. The directories are structured as follows.

```
snvs (or indels)
  ./donor/
    ./myvcf.snvs.ptato.vcf.gz
    ./myvcf.snvs.ptato.vcf.gz.tbi
    ./myvcf
      ./myvcf.ptatotable.txt
      ./myvcf.snvs.ptato.filtered.vcf.gz
      ./myvcf.snvs.ptato.filtered.vcf.gz.tbi
```

The merged annotated (but not filtered) vcf file can be found in the ptato_vcfs/donor folder.

```
ptato_vcfs
  ./donor/
    ./myvcf.ptato.merged.vcf.gz
    ./myvcf.ptato.merged.vcf.gz.tbi
```

In the `intermediate` folder, you can find the copy number and B allele frequency (BAF) plots of the entire genome for all samples. You may also use the Circos plots to visualize genomic rearrangements, this may be especially helpful for fusions or translocations.

```
intermediate
  ./svs
    ./Plots
      ./donor
        ./germline_control_sample
          ./baf
          ./copynumber
          ./karyogram
    ./Circos
      ./plots
        ./donor
```

Calling shared mutations between cells processed in separate batches using GATK's HaplotypeCaller

Timing: Variable (up to 2 days)

If you analyzed a large number of single cells using PTATO in separate runs, you may want to determine whether somatic variants are found (and thus shared) across all samples. Here, we do this by running GATK's HaplotypeCaller v4.1.4.1 on all samples simultaneously, but specifying only the regions of the genome where variants have been identified.

37. Create a merged interval list (for example a BED file) containing all the genomic regions where somatic mutations are found after running PTATO on your samples.

Note: Use the "`.snvs.filtered.ptato.vcf.gz`" file.

38. Run GATK's HaplotypeCaller v4.1.4.1 using this merged interval list.
39. Use this file for downstream analyses as it includes the shared mutations across all samples.

Mutation load and downstream analyses

Timing: 5 h

To obtain the mutation burden after PTATO filtering, determine the fraction of the sequenced genome that has sufficient coverage and quality for variant calling using GATK's v3.8.1 CallableLoci, (with parameters –minBaseQuality 10 –minMappingQuality 10 –minDepth 8 –minDepthForLowMAPQ 10 –maxDepth 100). This script is incorporated in the PTATO package, and the BED files specifying which regions are callable, low coverage, have poor mapping quality or no coverage can be found in `QC/CallableLoci`folder, according to the directory structure below.

```
QC
  ./donor/
    ./CallableLoci
      ./donor/
```

40. Determine the fraction of overlapping callable regions between your sample and your germline control and take the sum for each sample, this is the total number of callable bases.

Note: We do this using bedtools, with the BED files in the "CallableLoci" directory as input.

41. Normalize the variant (for example SNVs, or indels) load to the callable bases for each sample.
    a. Exclude variants that do not overlap with the callable regions.
    b. Count all remaining variants on autosomal chromosomes and extrapolate this by dividing it by the fraction of the genome that was surveyed (=total number of callable bases, 2745186691 for GRCh38).

Mutational patterns

After determining the true number of SNVs and indels per genome, visualize the 96-trinucleotide context of the artificial variants using MutationalPatterns v3.6.0.[12] In-depth somatic mutation analysis can shed light on operative mutational processes in a single cell, which act as so-called genomic footprints. We regularly employ this analysis to extract relevant mutational patterns from our sequencing data.

42. First, load the unfiltered and filtered data in R and filter on "snv."

Note: All analyses in the code blocks below are performed in R.

```
> vcf_files = list.files('/path/to/out_dir/donor/snvs/donor/', pattern =
'*.snvs.ptato.vcf.gz$', full.names = TRUE)
> sample_names = gsub('*sorted_*', '\\1', vcf_files)
> sample_names = gsub('*.snvs.ptato.vcf.gz', '\\1', sample_names)
> grl = read_vcfs_as_granges(vcf_files, sample_names, ref_genome, type = 'all')
> grl_snv = get_mut_type(grl, type = 'snv')
> mut_mat = mut_matrix(grl_snv, ref_genome)
> filtered_vcf_files= list.files('//path/to/out_dir/donor/snvs/donor/', pattern =
'*.snvs.ptato.filtered.vcf.gz$', full.names = TRUE,
recursive = T)
> filtered_grl = read_vcfs_as_granges(filtered_vcf_files, sample_names,
ref_genome, type = 'all')
> filtered_grl_snv = get_mut_type(filtered_grl, type = 'snv')
> filtered_grl = read_vcfs_as_granges(filtered_vcf_files, sample_names,
ref_genome, type = 'all')
> filtered_grl_snv = get_mut_type(filtered_grl, type = 'snv')
> filtered_mut_mat = mut_matrix(filtered_grl_snv, ref_genome)
```

43. Make a mutational matrix for the filtered and the unfiltered data and subtract them.
44. Visualize the mutational matrix in a 96-trinucleotide context (see Figure 10 in expected outcomes).

```
> mut_mat_merge<-as.matrix(rowSums(mut_mat))

> mut_mat_merge<-cbind(mut_mat_merge, as.matrix(rowSums(filtered_mut_mat)))

> colnames(mut_mat_merge)<-c("prePTATOfilter","postPTATOfilter")

> mut_mat_merge <- as.data.frame(mut_mat_merge)

> mut_mat_merge$Difference = mut_mat_merge$prePTATOfilter ->

mut_mat_merge$postPTATOfilter

> mut_mat_merge <- as.matrix(mut_mat_merge)

> plot_96_profile(mut_mat_merge)
```

Note: This analysis can also be performed for double base substitutions (dbs) and indels, change the type = in "get_mut_type () to respectively "dbs" or "indel." For more details, see the MutationalPatterns vignette here.

EXPECTED OUTCOMES

Yield of whole-genome amplification

In general, the yield of DNA after WGA depends on the kit version used and varies between experiments. For the ResolveDNA Whole Genome Amplification kit (96 reactions), a yield of above 2 μg for the positive control (100 pg input) is expected. After successful amplification of single cells, over 1 μg of DNA is expected, which equals to approximately 25 ng/μL in a total volume of 38–40 μL after elution. For experiments in which the cells of interest were highly viable prior to sorting, over 85% of cells will result in DNA yields higher than 25 ng/μL (Figure 3A).

The yield of DNA amplification is generally lower for the ResolveDNA Whole Genome Amplification kit (384 reactions), due to the lower reaction volume. For this kit, successful amplification will result in a DNA yield of approximately 15–20 ng/μL or higher in a volume of 18 μL (Figure 3B). In these two example experiments, 90% of the cells had a concentration of more than 10 ng/μL and could therefore be sent for WGS. When using a dilution series of positive control DNA, the output should correlate to the input that is used, producing up to 0.5–1 μg of DNA in the highest concentration (600 pg). The yield of the DNA fraction of the ResolveOME Whole Genome and Transcriptome Amplification Kit is similar to that of the ResolveDNA kit (Figure 3C). In conclusion, the genome of cells with an amplification yield of less than 15 (96 reactions) or 10 ng/μL (384 reactions) is in our experience not evenly amplified and thus not suitable for WGS analysis. In all kit versions, the negative control should result in a total yield of less than 100 ng of DNA.

Quality of whole-genome amplification based on multiplexed PCR

A high-quality amplified genome from a single cell is defined as having a sufficient DNA yield after whole-genome amplification and showing 8 to 10 homogeneous bands on the PCR-based quality control gel (Figure 4A; nine bands). The top/largest band, which is visible when looking at genomic DNA (gDNA) control, is often not present in high-quality amplified genomes (Figure 4B; cell 3 versus genomic gDNA). This is because a PTA reaction results in smaller fragments that often do not span the full targeted locus. Overall, experiments on highly viable samples and sorting of viable cells as described in this protocol will result in high-quality amplified genomes for 80%–100% of cells.

Lower quality amplified genomes are characterized by multiple missing bands following the multiplexed QC PCR, also in cases where the DNA yield is within the expected range of good quality samples. If WGS is performed on these samples, our experience is that the allelic dropout rate will increase substantially and the proportion of the genome which can be reliably analyzed will be reduced. Especially in cases where even coverage of the whole genome is important, for example when estimating the mutational burden per cell, samples with fewer than 8 bands should be

discarded. When performing experiments using samples with lower viability (e.g., sensitive primary samples), a higher fraction of cells may show reduced DNA yield after amplification and/or multiple missing bands on the QC gel (Figure 4C; cells 5, 6 and 8). In all experiments, stringently select the cells for whole-genome sequencing with DNA yields in the expected range and the best quality based on the multiplexed PCR gels. In this example, cells 1 and 2 from Figure 4A and cells 4, 7 and 9 from Figure 4C are high quality and can be sent for WGS.

Mapping QC report

The QC reports generated following "filtering WGA artificial variants using PTATO" step 35C, shows the overall sequencing coverage, in addition to the distribution of the mean coverage per sample. Use this report to determine which samples to continue through the PTATO analysis.

In our experience, we obtain the best results when including samples where more than 75% of bases are covered with a sequencing coverage $\geq$5X. In the example below, genomic DNA from bulk MSCs, bulk tumor cells, and 10 WGA single cells were sequenced from a T-ALL patient (Figure 5A).[7] All WGA samples were sequenced at 15X coverage, while bulk samples were sequenced at 30X (Figure 5B). Here, samples "pt2229-TALL-1" and "pt2229-TALL-3" have a low percentage of genome coverage (Figures 5C and 5D). These samples can still be included in downstream analysis, but any biological conclusions need to be carefully considered. Other QC parameters can also be considered before including or excluding samples from downstream analyses, including mean coverage, heterozygous single nucleotide polymorphism (SNP) sensitivity, percentage of filtered bases, read counts, and error rates (Figures 5E, 5F, and 6A–6C). In this case, samples "pt2229-TALL-1" and "pt2229-TALL-3" did not have significantly higher filtered bases (Figure 6A), the read counts (Figure 6B) and error rate (Figure 6C) fit in line with the rest of the samples and were thus not excluded for further analyses.

postPTATO quality report

The QC report post-PTATO ("`postqc`"), see "filtering WGA artificial variants using PTATO" step 35D, gives an overview of the callable loci and the estimation of the PTA-induced artificial variant probability of each sample (Figures 7A and 7B). As WGA samples generally have a lower fraction of the genome covered compared to bulk samples, we want to ensure which loci can be considered. In addition, the distribution of the somatic variants and the artificial variant probability is estimated per sample (Figures 7C and 7D) and is expected to be around 0.5. To assess PTATO's ability to estimate the PTA probability cutoff (PTAprob_Cutoff), we employ the linked-read analysis called Walker. The final PTA probability cutoff is determined when most true positives (recall) and least false positives (precision) are called (Figure 8). Higher mutation loads generally tend to result in higher precision-recall values as the presence of more variants reduces the impact of individual variants, thereby increasing precision. In addition, the cosine similarity of the PTA-induced artificial variants is also taken into account.

The complete overview of passed and failed variants, including the failed variants as a result of the user-specified VAF (default setting 0.15), informs on the level of noise (Figures 9A and 9B). In a single cell, sequencing artefacts tend to have a VAF of value below 0.5, while true variants are between 0.5 and 1. The lower you set the VAF cutoff value (i.e., <0.15), the higher the level of artificial variants retained in the dataset. Increasing the VAF cutoff value (i.e., >0.25) removes more false positive variants. These false positive variants can be the result of WGA artificial variants or sequencing artefacts. Keep in mind that increasing the VAF cutoff value towards 0.5 leads to the exclusion of true variants. Lastly, the post-PTATO QC report shows the results of the Walker analysis (Figures 9C and 9D), which distinguishes between true and false positives. A low number of callable loci can lead to PTA probability cutoffs that deviate significantly from 0.5 (i.e., <0.4 or >0.6). A high number of variants with low Walker scores are an indication of low-quality samples.

Mutation characteristics of PTA

After running PTATO, the SNV load, corrected for callable loci, can be compared (Figure 10A). The artificial variants generated by PTA follow a specific 96-trinucleotide mutational profile and mostly consist of C to T mutations (Figure 10B).[5] Indeed, the artificial variant pattern shows the highest cosine similarities with COSMIC signatures that mostly include C to T mutations, such as SBS30 (Figure 10C). Once these are filtered out, the true COSMIC signatures SBS1 and SBS6 become more apparent.

LIMITATIONS

Although the workflow described here results in the detection of base substitutions and indels at high sensitivity (70%–89%) and high precision (70%–92%), several limitations of this protocol remain. First, the quality of the whole-genome sequencing data is highly dependent on the quality of the WGA. Uneven or incomplete amplification of the genome will result in allelic drop-out and a reduction of the callable region of the genome. This will result in regions with loss of heterozygosity and noisy copy-number profiles, which complicates the downstream analysis of e.g., structural variations. The loss of (some) samples throughout the process cannot be avoided, despite the key steps for the generation of high-quality amplified genomes highlighted in this protocol. In addition, the use of PTATO for removing artificial variants may remove true indel mutations in homopolymer stretches and is less accurate for samples with an extremely low mutational burden (<200 mutations). Furthermore, the accuracy for removal of artefact structural variations is unknown, as this requires the analysis of a larger cohort of matched single cells and bulk samples containing the same structural variations. Finally, the methods described in this protocol may not be applicable to all eukaryotic cell types. We have successfully applied these methods to several primary human cell types, such as intestinal stem cells and hematopoietic progenitor and mature lineages, as well as leukemic blasts and lymphoma cells. For other cell types, the lysis step of the WGA may require optimization. Finally, the use of different cell type(s), model organisms, and other PTA-based WGA methods may result in artificial mutations with different characteristics. This will require the user to train a new random forest model based on true and false positive variant calls, an option which is implemented in the PTATO package.

TROUBLESHOOTING

Problem 1

Entire or almost full plate(s) show(s) no or extremely low yield after whole-genome amplification, **as does the positive control** (related to step 23).

Potential solutions

- Check the expiration date of the kit that is used. The kits can be used up to one year after the date of manufacturing (related to step 21).
- Use fresh aliquots of the kit reagents, including the positive control (related to step 21). Potentially, the reagents have been exposed to too many freeze-thaw cycles or to the wrong temperature. For example, the SM2 buffer should never be kept at −80°C or directly on dry ice.
- Thaw and use a new kit, if all aliquots of the initial kit do not work with an aliquot of the positive control (related to before you begin steps 1–8).

Problem 2

Entire or almost full plate(s) show(s) no or extremely low yield after whole-genome amplification, **despite the positive control working as expected** (related to step 23). There are several critical steps that can lead to a lack of successful amplification, such as incomplete cell lysis, sorting of dead/dying cells or debris, or inaccurate sorting into the wells.

Potential solution

Before performing additional troubleshooting experiments, several steps can be checked to assess the quality of the cells that failed to show sufficient DNA amplification, and the reagents that were used.

- The sorting equipment was not configured appropriately for the cell type(s) of interest (related to step 10). Check the size of the sorting chip or nozzle and sorting mask used compared to the size of the cells of interest (e.g., larger cells may require a larger nozzle and mask).
- The cells of interest were stored for too long at −80°C (related to step 19). Check the time between sorting and sample workup to ascertain that the samples were not stored for longer than five months. This can only be solved by processing samples within five months after sorting.
- The sorted events consisted of debris or dead cells (related to steps 14–17). Check the index data of the sorting experiment to ascertain that the sorted events were part of the live cell gate in the scatter channels (Figures 2A and 2B). For future experiments, set the gates stricter (Figures 2F–2J), and check the index data of the sorted cells prior to WGA.

Complete functionality of the reagents, sorting equipment, and storage methods can be checked through single-cell sorting and whole-genome amplification of a highly viable cell line (e.g., lymphoblastic cell lines). If this does not result in good DNA yields, the following points should be considered.

- The (cell lysis) reagents of the kit are expired or exposed to too many freeze-thaw cycles or the wrong temperature (related to step 21). See problem 1 for potential solutions.
- Inaccurate sorting due to unstable instruments or incorrect calibration (related to steps 9–17). For future experiments, re-perform plate calibration in between the sorting of consecutive plates to check the accuracy of the cell sorting throughout the sort. For additional troubleshooting, cells of any kind can be stained fluorescently, sorted into a glass-bottomed plate, and checked under a fluorescent microscope.

If, in contrast, the use of a viable cell line results in well-amplified DNA, the problem may have arisen due to cell type-specific lysis requirements, or the sorting of bad quality cells (e.g., debris/dead).

- The lysis of the cells may be incomplete and optimization of the lysis for this specific cell type is required (related to step 21). For cells derived from bone marrow or peripheral blood samples, we have successfully used the following incubation method:
  - Per protocol:
    - Add 3 μL of MS mix to each well.
    - Seal the plate with sealing film.
    - Spin for 10 s, mix at 15°C–25°C for 1 min at 1,400 rpm (plate mixer), spin for 10 s and place plate back on ice.
  - Incubate on ice for 5 min.
  - Incubate at 15°C–25°C for 15 min, shaking on a plate mixer (1,400 RPM).
  - Per protocol:
    - Add 3 μL of SN1 Reagent.
    - Bad quality cells were sorted (e.g., debris/dead cells, related to steps 14–17). In future experiments, use a live/dead cell dye (see step 11 of step-by-step method details), in addition to stricter sorting gates and the checking of index data prior to WGA. This is especially important for samples with a lower viability, e.g., sensitive primary material.

Problem 3

The negative control reaction has a high DNA yield (>100 ng in total), related to step 23 of step-by-step method details.

Potential solution

Use new aliquots of all enzyme mixes and buffers, as they may be contaminated with DNA. Make sure to use filter tips throughout the protocol and if possible, work in a DNA-free pre-PCR hood.

Problem 4

Many samples show multiple missing bands based on the multiplexed QC PCR, despite sufficient DNA yield after whole-genome amplification (step 30).

Potential solution

The viability, storage, quality, lysis, or DNA amplification of the cells is inadequate. For potential causes and solutions, see problem 2.

Problem 5

Your sequencing mapping quality report shows less than 75% of bases with ≥5× coverage, see expected outcomes Figures 5B and 5C (related to step 35C).

Potential solution

The viability, storage, quality, lysis, or DNA amplification of the cells is inadequate. This can also be related to problem 4. For potential causes and solutions, see problem 2.

Problem 6

PTATO workflow was interrupted due to an out-of-time or out-of-memory error (related to step 35).

Potential solution

Scale memory and/or time parameters in the `process.config`, see step-by-step method details step 35A for the specified job in the `PTATO.error` file.

Problem 7

PTATO cannot be used due to issues during the cloning of the repository and subsequent installation, related to 'before you begin' step 24.

Potential solution

Make sure you installed everything properly following the official guidelines as described here ('before you begin' steps 9–17). For additional computational issues, please report to the Issues section in the PTATO GitHub https://github.com/ToolsVanBox/PTATO/issues.

RESOURCE AVAILABILITY

Lead contact

Further information and requests for reagents and resources should be directed to and will be fulfilled by the lead contact, Ruben van Boxtel (R.vanBoxtel@prinsesmaximacentrum.nl).

Technical contact

Issues regarding the installation and use of PTATO can be raised at the GitHub page (https://github.com/ToolsVanBox/PTATO). Additional requests should be directed to Ruben van Boxtel (R.vanBoxtel@prinsesmaximacentrum.nl).

Materials availability

This study did not generate new unique reagents.

Data and code availability

This protocol did not generate new datasets. The WGS data used in this protocol is available through the European Genome-phenome Archive (EGA): EGAS00001007446. No new data was generated for this study. This protocol makes use of a nextflow-based bioinformatic pipeline to process raw sequencing data. The code for the full mapping and variant calling pipeline can be retrieved from https://github.com/ToolsVanBox/NF-IAP; Zenodo: https://doi.org/10.5281/zenodo.13903645. The PTATO pipeline is available at https://github.com/ToolsVanBox/PTATO; Zenodo: https://doi.org/10.5281/zenodo.8424848. Mutational signature analysis can be performed using the R package MutationalPatterns available at Bioconductor: https://bioconductor.org/packages/release/bioc/html/MutationalPatterns.html.

ACKNOWLEDGMENTS

This work was funded by an ERC Consolidator grant from the European Research Council (ERC; no. 864499) and the Foundation Kids Cancer Free (KiKa; no. 424) to R.v.B. In addition, this work was supported by funding from the Oncode Institute. R.v.B. is a New York Stem Cell Foundation – Robertson Investigator. This research was supported by The New

York Stem Cell Foundation. The authors want to thank all patients and parents at the Princess Máxima Center for Pediatric Oncology for their contribution to this work.

AUTHOR CONTRIBUTIONS

Conceptualization, L.L.M.D., A.J.C.N.v.L., A.S.S., and V.M.P.; software, S.M., M.J.v.R., and R.H.; investigation, L.L.M.D., A.J.C.N.v.L., A.S.S., L.T., V.M.P., M.V., and S.M.; formal analysis, M.J.v.R. and R.H.; resources, S.M., L.L.M.D., A.S.S., and A.J.C.N.v.L; writing – original draft, L.L.M.D., A.J.C.N.v.L., and L.T.; writing – review and editing, L.L.M.D., A.J.C.N.v.L., A.S.S., L.T., S.M., M.J.v.R., and R.H.; visualization, L.L.M.D., A.J.C.N.v.L., and A.S.S.; supervision, R.v.B.; funding acquisition, R.v.B.

DECLARATION OF INTERESTS

The authors declare no competing interests.

REFERENCES

1. Middelkamp, S., Manders, F., Peci, F., van Roosmalen, M.J., González, D.M., Bertrums, E.J.M., van der Werf, I., Derks, L.L.M., Groenen, N.M., Verheul, M., et al. (2023). Comprehensive single-cell genome analysis at nucleotide resolution using the PTA Analysis Toolbox. Cell Genom. *3*, 100389. https://doi.org/10.1016/j.xgen.2023.100389.

2. Biezuner, T., Raz, O., Amir, S., Milo, L., Adar, R., Fried, Y., Ainbinder, E., and Shapiro, E. (2021). Comparison of seven single cell whole genome amplification commercial kits using targeted sequencing. Sci. Rep. *11*, 17171. https://doi.org/10.1038/s41598-021-96045-9.

3. Raz, O., Tao, L., Biezuner, T., Marx, T., Neumeier, Y., Tumanyan, N., and Shapiro, E. (2022). Whole-Genome Amplification—Surveying Yield, Reproducibility, and Heterozygous Balance. Int. J. Mol. Sci. *23*, 6161. https://doi.org/10.3390/ijms23116161.

4. Luquette, L.J., Bohrson, C.L., Sherman, M.A., and Park, P.J. (2019). Identification of somatic mutations in single cell DNA-seq using a spatial model of allelic imbalance. Nat. Commun. *10*, 3908. https://doi.org/10.1038/s41467-019-11857-8.

5. Gonzalez-Pena, V., Natarajan, S., Xia, Y., Klein, D., Carter, R., Pang, Y., Shaner, B., Annu, K., Putnam, D., Chen, W., et al. (2021). Accurate genomic variant detection in single cells with primary template-directed amplification. Proc. Natl. Acad. Sci. USA *118*, e2024176118. https://doi.org/10.1073/pnas.2024176118.

6. Luquette, L.J., Miller, M.B., Zhou, Z., Bohrson, C.L., Zhao, Y., Jin, H., Gulhan, D., Ganz, J., Bizzotto, S., Kirkham, S., et al. (2022). Single-cell genome sequencing of human neurons identifies somatic point mutation and indel enrichment in regulatory elements. Nat. Genet. *54*, 1564–1571. https://doi.org/10.1038/s41588-022-01180-2.

7. Poort, V.M., Hagelaar, R., van Roosmalen, M.J., Trabut, L., Buijs-Gladdines, J.G.C.A.M., van Wijk, B., Meijerink, J., and van Boxtel, R. (2024). Transient Differentiation-State Plasticity Occurs during Acute Lymphoblastic Leukemia Initiation. Cancer Res. *84*, 2720–2733. https://doi.org/10.1158/0008-5472.can-24-1090.

8. Di Tommaso, P., Chatzou, M., Floden, E.W., Barja, P.P., Palumbo, E., and Notredame, C. (2017). Nextflow enables reproducible computational workflows. Nat. Biotechnol. *35*, 316–319. https://doi.org/10.1038/nbt.3820.

9. Kurtzer, G.M., Bauer, M., Kaneshiro, I., Trudgian, D., and Godlove, D. (2021). hpcng/singularity: Singularity 3.7.3. Zenodo. https://doi.org/10.5281/zenodo.1310023.

10. Cingolani, P., Platts, A., Wang, L.L., Coon, M., Nguyen, T., Wang, L., Land, S.J., Lu, X., and Ruden, D.M. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff. Fly *6*, 80–92. https://doi.org/10.4161/fly.19695.

11. Wickham, H. (2016). ggplot2: Elegant Graphics for Data Analysis (Springer-Verlag). https://ggplot2.tidyverse.org.

12. Manders, F., Brandsma, A.M., de Kanter, J., Verheul, M., Oka, R., van Roosmalen, M.J., van der Roest, B., van Hoeck, A., Cuppen, E., and van Boxtel, R. (2022). MutationalPatterns: the one stop shop for the analysis of mutational processes. BMC Genom. *23*, 134. https://doi.org/10.1186/s12864-022-08357-3.

13. Obenchain, V., Lawrence, M., Carey, V., Gogarten, S., Shannon, P., and Morgan, M. (2014). VariantAnnotation: a Bioconductor package for exploration and annotation of genetic variants. Bioinformatics *30*, 2076–2078. https://doi.org/10.1093/bioinformatics/btu168.

14. Cameron, D., and Dong, R. (2024). StructuralVariantAnnotation: Variant annotations for structural variants. R package version 1.22.0. https://www.bioconductor.org/packages/release/bioc/html/StructuralVariantAnnotation.html.

15. The Bioconductor Development Team (2023). BSgenome.Hsapiens.UCSC.hg38: Full genomic sequences for Homo sapiens (UCSC genome hg38). R package version 1.4.5. https://www.bioconductor.org/packages/release/data/annotation/html/BSgenome.Hsapiens.UCSC.hg38.html.

16. Pagès, H. (2024). BSgenome: Software infrastructure for efficient representation of full genomes and their SNPs. R package version 1.72.0. https://bioconductor.org/packages/release/bioc/html/BSgenome.html.

17. Nilsen, G., Liestøl, K., Van Loo, P., Vollan, H.K.M., Eide, M.B., Rueda, O.M., Chin, S.-F., Russell, R., Baumbusch, L.O., Caldas, C., et al. (2012). Copynumber: Efficient algorithms for single- and multi-track copy number segmentation. BMC Genomics *13*, 591. https://doi.org/10.1186/1471-2164-13-591.

18. Wilke, C.O. (2024). cowplot: Streamlined Plot Theme and Plot Annotations for "ggplot2". R package version 1.1.3. https://doi.org/10.32614/cran.package.cowplot.

19. Warnes, G., Bolker, B., Lumley, T., Magnusson, A., Venables, B., Ryodan, G., and Moeller, S. (2023). gtools: Various R Programming Tools. https://doi.org/10.32614/cran.package.gtools.

20. Wiener, M., Breiman, L., Cutler, A., and Liaw, A. (2024). randomForest: Breiman and Cutler's Random Forests for Classification and Regression. https://doi.org/10.32614/cran.package.randomforest.

21. Wickham, H., Pedersen, T.L., and Seidel, D. (2023). scales: Scale Functions for Visualization. R package version 1.3.0. https://doi.org/10.32614/cran.package.scales.

22. Delaneau, O., Zagury, J.-F., Robinson, M.R., Marchini, J.L., and Dermitzakis, E.T. (2019). Accurate, scalable and integrative haplotype estimation. Nat. Commun. *10*, 5436. https://doi.org/10.1038/s41467-019-13225-y.

23. Poplin, R., Ruano-Rubio, V., DePristo, M.A., Fennell, T.J., Carneiro, M.O., Van der Auwera, G.A., Kling, D.E., Gauthier, L.D., Levy-Moonshine, A., Roazen, D., et al. (2018). Scaling accurate genetic variant discovery to tens of thousands of samples. Preprint at bioRxiv. https://doi.org/10.1101/201178.

24. Bertrums, E.J.M., de Kanter, J.K., Derks, L.L.M., Verheul, M., Trabut, L., van Roosmalen, M.J., Hasle, H., Antoniou, E., Reinhardt, D., Dworzak, M.N., et al. (2024). Selective pressures of platinum compounds shape the evolution of therapy-related myeloid neoplasms. Nat. Commun. *15*, 6025. https://doi.org/10.1038/s41467-024-50384-z.