Long-read whole genome analysis of human single cells

Joanna Hård[1,*], Jeff E Mold[1], Jesper Eisfeldt[2,3], Christian Tellgren-Roth[4], Susana Häggqvist[4], Ignas Bunikis[4], Orlando Contreras-Lopez[5], Chen-Shan Chin[6], Carl-Johan Rubin[5], Lars Feuk[4], Jakob Michaëlsson[7,†], Adam Ameur[4,8,†,*]

[1] Department of Cell and Molecular Biology, Karolinska Institutet, Stockholm, Sweden

[2] Department of Molecular Medicine and Surgery, Karolinska Institutet

[3] Department of Clinical Genetics, Karolinska University Hospital, Stockholm, Sweden

[4] Science for Life Laboratory, Department of Immunology, Genetics and Pathology, Uppsala University, Uppsala, Sweden

[5] Science for Life Laboratory, Royal Institute of Technology (KTH), Stockholm, Sweden

[6] Foundation for Biological Data Science, Belmont, CA, USA

[7] Center for Infectious Medicine, Department of Medicine, Karolinska Institutet, Stockholm, Sweden

[8] Department of Epidemiology and Preventive Medicine, Monash University, Melbourne, Australia

[†] Equal contributions

[*] Corresponding authors

Abstract

With long-read sequencing we have entered an era where individual genomes are routinely assembled to near-completion and where complex genetic variation can efficiently be resolved. Here we demonstrate that long reads can be applied also to study the genomic architecture of individual human cells. Clonally expanded CD8+ T-cells from a human donor were used as starting material for a droplet-based multiple displacement amplification (dMDA) method designed to ensure long molecule lengths and minimal amplification bias. Sequencing of two single cells was performed on the PacBio Sequel II system, generating over 2.5 million reads and ~20Gb HiFi data (>QV20) per cell, achieving up to 40% genome coverage. This data allowed for single nucleotide variant (SNV) detection, including in genomic regions inaccessible by short reads. Over 1000 high-confidence structural variants (SVs) per cell were discovered in the PacBio data, which is four times more than the number of SVs detected in Illumina dMDA data from clonally related cells. In addition, several putative clone-specific somatic SV events could be identified. Single-cell *de novo* assembly resulted in 454-598 Mb assembly sizes and 35-42 kb contig N50 values. 1762 (12.8%) of expected gene models were found to be complete in the best single-cell assembly. The *de novo* constructed mitochondrial genomes were 100% identical for the two single cells subjected to PacBio sequencing, although mitochondrial heteroplasmy was also observed. In summary, the work presented here demonstrates the utility of long-read sequencing towards understanding the extent and distribution of complex genetic variation at the single cell level.

Keywords

Single-cell sequencing, long-read sequencing, structural variation, somatic variation, de novo assembly, droplet amplification, dMDA, Xdrop, HiFi sequencing

Background

During the last few years, long-read sequencing technologies have made remarkable progress in terms of throughput and data quality. Due to their capability to read through repetitive and high GC-content regions, these technologies are essential for the ambitious plans to generate reference genomes for virtually all of Earth's eukaryotic biodiversity[1, 2], as well as complete telomere-to-telomere maps of human chromosomes[3, 4]. A further advantage of long-read sequencing is that it facilitates genotyping of complex structural variation (SVs) and repeat elements, which can be difficult or impossible to identify with other genomic sequencing approaches[5-7]. Although clinical long-read sequencing is still in its infancy[8, 9], several studies have already demonstrated the potential to discover novel disease-causing human genetic variation. Long sequencing reads can also enable the detection of clinically relevant genetic variation in 'dark DNA', representing regions of the human genome that cannot be analyzed with standard short-read technologies[10].

Long-read sequencing holds many promises, but one intriguing research area that remains unexplored is single-cell genomics. Human single-cell whole genome sequencing (WGS) emerged about a decade ago[11-15], and has become an active field of research with potential to answer fundamental questions in several areas of cell biology, such as somatic genetic variation[16], tumor evolution[11], de novo mutation rates[14], meiotic recombination of germ cells[14, 17], or neurogenetics[18-20]. Until now, single-cell WGS projects have focused on characterizing genetic variation detectable from short-read Illumina sequencing protocols[21-25], including single nucleotide variants (SNVs)[19, 21, 26-31], large-scale copy number variation[30, 32-34] and retrotransposon elements[12, 18, 35, 36]. To our knowledge, there are today no published reports of long-read WGS of individual human cells. In part, this can be explained by the throughput of long-read instruments, which until recently has been relatively modest. In addition, single-cell

genome sequencing is associated with technical challenges[37]. In a diploid cell, only two DNA molecules exist at each locus in the genome, and every molecule that is lost during sample preparation, or fails to be sequenced, inevitably leads to allelic drop-out and missing data. Moreover, the long-read sequencing protocols require large amounts, typically several micrograms, of input DNA. This is about a million times more DNA than what is contained within a single human cell, which implies that a substantial DNA amplification is required.

Whole genome amplification has a profound detrimental effect on the sequencing results and should be avoided when possible, since it introduces amplification bias, chimeric molecules and allelic dropout. Several different amplification protocols have been developed[15, 38, 39] and it is crucial to choose a method that minimizes artefacts and biases, while at the same time being compatible with the downstream sequencing technology. Multiple displacement amplification (MDA)[39] has capacity to amplify kilobase-length molecules and could therefore a suitable approach for long-read sequencing. With regards to amplification bias, it has been proposed that a droplet-based MDA (dMDA) reaction, performed on DNA fragments contained within nano- or picoliter droplets, can minimize differences in amplification gain among the fragments[40-42]. Such a droplet-based amplification could also be an efficient approach to remove inter-molecular chimeras, since MDA chimeras only can be formed between molecules contained within the same droplet.

Single-cell DNA fragments amplified by MDA methods are well-suited for PacBio high-fidelity (HiFi) sequencing[43], since this protocol enables to read molecules of at least 20kb length. Moreover, the resulting PacBio HiFi reads have very high accuracy (>QV20), and not only allows identification of complex genetic variation such as SVs and repeat elements, but also SNVs at an accuracy that matches the ability of short-read sequencing[43]. PacBio HiFi sequencing has also proven to be an excellent method for high-quality genome assembly[4, 44-46], thereby raising the prospect of long-read de novo assembly of genomic DNA from individual

cells. Taken together, a more detailed analysis of single cell genomes using highly accurate long reads could allow detection of new classes of somatic variation, including for example SVs and repeats, which have not been possible to study in single cells before. Eventually, this could lead to a better understanding of somatic variation, mutation rates and the functional impact of these elements. The potential applications are not limited to human cells. Long-read WGS could also potentially generate improved genome assemblies also for other types of cells, such as single cellular organisms that are difficult to culture.

In this study, we established a long-read based approach for single-cell whole genome sequencing using a new automated dMDA technique for single-cell whole genome amplification coupled with PacBio HiFi whole genome sequencing. The method was evaluated on two clonally expanded CD8+ T-cells from a human donor, and in parallel other cells from the same T-cell clones were sequenced with short-read Illumina WGS. Our data demonstrates that SV discovery in single cells is substantially improved by long-read sequencing, and that genetic variation can be discovered also in regions inaccessible by short reads. We further performed de novo assembly of each of the two human single cells. Albeit fragmented due to dropout, these assemblies represent the first step towards reference-free analysis of the genomes in individual cells. Taken together, these findings open up new possibilities to characterize the landscape of complex genetic variation and genome organization at unprecedented resolution.

Results

Amplification of single-cell DNA in droplets

We first aimed to develop a DNA amplification method that preserves molecule lengths and reduces amplification bias (Figure 1A). Briefly, one single cell is isolated by fluorescence-activated cell sorting (FACS) and placed into a well containing lysis buffer, so that the DNA fragments are released. The DNA molecules are then encapsulated in approximately 50,000 droplets, after which a dMDA reaction takes place within each droplet. The droplets have a diameter of <100 µm and are generated using the Xdrop system[47] (Figure 1B). Only one or a few DNA fragments will be located in each droplet, and since the amplification takes place in a small volume containing limited reagents this will prevent molecules from being heavily over-amplified. Moreover, the risk of forming inter-molecular chimeras during the dMDA reaction is greatly reduced, and completely eliminated in droplets harboring a single DNA fragment. Once the dMDA reaction is complete, the amplified DNA can be used for preparation of short- and long-read sequencing libraries. For our experiments, two individual CD8+ T-cells (A and B) from the same human donor were clonally expanded *in vitro*, and the resulting cell collections were used as starting material for whole genome amplification and sequencing (Figure 1C). In addition, bulk DNA isolated from peripheral blood mononuclear cells (PBMC) obtained from the same individual was analyzed by short-read WGS.

dMDA increases whole-genome sequencing coverage uniformity

Sixteen single-cell DNA samples from the two T-cell clones A and B were investigated using Illumina WGS. Eight of the samples were amplified using dMDA, while the remaining eight samples were subjected to regular MDA. The sequencing resulted in 100 to 200 million read pairs per sample, and these were aligned to the GRCh38 human reference build. To facilitate direct comparisons between the samples, all Illumina datasets were randomly subsampled to

contain about 100 million read pairs. As expected, the eight dMDA samples displayed a more uniform coverage across the genome as compared to the eight MDA samples (Figure 2A-C). The reduced bias in dMDA can also be seen in the mitochondrial genome, where dMDA resulted in more than 10-fold higher coverage as compared to regular MDA. Furthermore, our results reveal that the uneven coverage in the MDA samples originates from a limited number of fragments that are being amplified to extreme coverage (Figure 2D). For the MDA samples, on average 68.9% of the reads align to regions with ≥200x coverage, while the corresponding percentage for dMDA is only 16.0%. In these downsampled datasets, 33.8% of bases were covered by at least one read in dMDA as compared to 23.4% for MDA (Figure 2E). Based on these results, we conclude that dMDA gives increased sequencing coverage uniformity as compared to regular MDA, thereby corroborating previous evaluations of droplet-based MDA methods[40, 41].

Long-read whole-genome sequencing of two individual T-cells

Two dMDA single-cell samples, one from T-cell clone A and one from T-cell clone B, were selected for PacBio long-read sequencing. The dMDA reactions generated 3160 ng (T-cell A) and 1850 ng (T-cell B) amplified DNA and the fragment size distributions displayed peaks around 9 kb. The PacBio HiFi sequencing protocol for the Sequel II instrument requires 10 µg of input DNA, and to enable library preparation from smaller DNA amounts SMRT bell size selection was performed using beads instead of using a gel-based system (see Methods). The resulting SMRT bell libraries were run on two separate 8M cells with 30h movie time. Over 2.5 million reads and ~20Gb HiFi data (>QV20) was obtained for each of the two samples (Table 1). Virtually all reads could be aligned to GRCh38 and the average alignment concordance was over 99%. More than 6 million alignments were produced per sample, indicating that chimeric artifacts from the dMDA are found in many reads, since each read

gives rise to between 2-3 separate alignments on average. The aligned read length is a good indicator of the non-chimeric part of a read, since it corresponds to the longest subsequence that can be continuously matched to the GRCh38 reference. The N50 aligned read length was 5.4 kb for T-cell A and 6.4 kb for T-cell B, and the maximum read alignment was 43.4 kb (T-cell A) and 48.7 kb (T-cell B). An average of 6x coverage was obtained from both samples. However, just like in the Illumina data there is a high level of allelic drop out. For T-cell A, 40% of the genome was covered, while T-cell B had an even lower genome coverage of 28%.

Single nucleotide variants can be detected in single-cell long read data

Having generated single-cell whole genome data both using short- and long-read technologies, we were interested to analyze single nucleotide variants (SNVs) in the different datasets. More than twice the amount of sequencing data was generated for the Illumina single-cell samples (average 48.7 Gb) as compared to PacBio (average 20.0 Gb). (Figure 3A). In the Illumina data, between 0.3 to 2.1 million SNVs were detected in each sample, and an average of 992k SNVs/sample were found to be overlapping with SNVs called in the PBMC bulk sequencing data. In the PacBio data, a total of 1.7M SNVs (T-cell A) and 1.2M SNVs (T-cell B) were detected by the software DeepVariant[48]. Of these, an average of 900k SNVs/sample were found to be overlapping with SNVs called in the PBMC bulk DNA sample (Figure 3B). This means that a similar number of germline SNVs were detected using PacBio as compared to Illumina, despite the much lower total data amount for PacBio. 78,775 of the PacBio SNVs that failed to be identified in the PBMC bulk sample sequenced on the Illumina platform were found to be located within previously reported "dark" genic regions of potential importance for human health[10]. One such region comprises introns and exons of *NBPF8* (Figure 3C). Another example is *CDC73*, where a repeat resolved in the PacBio single-cell data is represented as an alignment gap in the Illumina bulk data (Figure 3D).

Single-cell analysis of structural variation

We performed SV calling using Manta[49] and TIDDIT[50] in the Illumina datasets, while the PacBio SVs were called using PBSV[51]. To compare the number of detected true SVs called in single cells, we focused on germline SVs that overlapped with SVs called in the PBMC bulk sample. In the eight Illumina dMDA samples, an average of 326.5 SV events were overlapping with SVs detected in the PBMC bulk sample (Figure 4A). The corresponding number for the eight Illumina MDA samples was 46.4 SVs. By far, the highest numbers of SVs overlapping with the PBMC bulk sample were found in the PacBio data; 1620 for T-cell A and 1126 for T-cell B. This finding demonstrates that PacBio sequencing outperforms Illumina in SV detection in single cells. As seen in Figure 4B, this pattern holds true for deletions, insertions and tandem duplications. We further developed a computational strategy to screen for somatic SV differences between the two T-cell clones A and B, which resulted in three candidate events. One of these is a 50 bp deletion on chromosome 1, clearly visible in the PacBio data for T-cell B (Figure 4C). The Illumina data for T-cell clone B also has support for a genomic aberration in this region, even though the exact break points are difficult to see in the short-read alignments. However, there is no visible support for this deletion either in the bulk sequencing data or in the single cell data for T-cell clone A. Due to the presence of heterozygous SNVs a few kb downstream of the 50bp deletion it is clear that both alleles have been sequenced in the bulk sequencing data, and the allele harboring the deletion event can be determined through phasing.

De novo assembly of single-cell long-read data

PacBio HiFi reads are ideal for generating high-quality assemblies of human genomes[4, 43-46], and we were interested to see whether some pieces of the single-cell genomes could be

reconstructed de novo. Since assembly of single-cell PacBio data is challenging due to allelic dropout and chimeric reads, we developed a filtering method to remove chimeric reads from the dataset prior to assembly. Because of the dMDA, chimeras are mainly formed within the same molecule, and by screening each read for inverted or duplicated elements, chimeric reads could be identified and removed ab initio (see Methods). For T-cells A and B, 44.2% and 46.6% of PacBio reads, respectively, passed our filtering criteria. However, this filtering is very stringent and does not only remove chimeras, but also many correct reads harboring repeat elements. Hifiasm[52] generated primary assemblies of size 598.3 Mb for T-cell A and 454.1 Mb for T-cell B, corresponding to approximately 19% and 15% of the human reference (Table 2). The contig N50 values were 35 kb (T-cell A) and 42 kb (T-cell B), and the largest contig of 578.3 kb was detected in the T-cell B assembly. In addition, approximately 40 Mb of alternative contigs were found in each sample. These alternative contigs correspond to regions where hifiasm reported two distinct haplotypes. We further performed an analysis of BUSCO gene models[53] and could conclude that 12.8% of genes (n=1762) were completely assembled for T-cell A, and 9.0% of genes (n=1236) for T-cell B. Complete mitochondrial genomes were obtained and these were identical for T-cells A and B. Looking closer at the mtDNA data, there is one location (chrM:16,218) where a C>T nucleotide substitution occurs in 42% of PacBio reads for T-cell B, while being completely absent from PacBio reads for T-cell A as well as from the Illumina bulk DNA sample. By further analyzing the Illumina dMDA data for the two single cell clones, we validate that the nucleotide substitution is present in T-cell B, but not in T-cell A, consistent with mitochondrial heteroplasmy in T-cell clone B.

Discussion

By a combination of methods for single-cell isolation, whole-genome amplification and PacBio HiFi sequencing, we were able to sequence long DNA fragments from two human T-cells. The long sequencing reads give improved analyses of genetic variants as compared to short-read technologies, including single nucleotide variation in "dark" regions of the human genome, larger structural variants, and even enables de novo assembly of single cell genomes. The single cells used as starting point for this study were obtained through *in vitro* expansion of CD8+ T-cells from a healthy human donor. These T-cells are more challenging to obtain as compared to cells from an established cell line, but have the advantages of being representative of healthy somatic cells from a human being, and additionally enable screening for somatic variation between clones based on known lineage relationships.

By short-read sequencing we could demonstrate that dMDA of single-cell DNA results in more uniform coverage and improved SV calling as compared to regular MDA. However, the best performance is obtained when coupling dMDA with PacBio HiFi sequencing. Despite that the HiFi sequencing yield was below 50% of the average data amount generated for the Illumina single-cells, a similar number of SNVs and 3-5 times as many SVs could be detected. Most likely, there are many additional true events among the remaining PacBio SNVs and SVs calls, although only high-confidence events overlapping with variants called in an unamplified bulk sample were considered in the current report. Furthermore, our data allowed us to identify somatic SVs that distinguish the two expanded T cell clones.

A human cell contains about six picograms of DNA, and this is a major challenge for PacBio HiFi sequencing which typically requires several micrograms of input material. Recently, an ultra-low input HiFi protocol was released[54], but this still requires five nanograms DNA. Thus,

we have ventured far beyond the limits of available protocols, with about 1000-fold less DNA, and with a large (3 Gb) genome size. The single cell assemblies presented here represent a reduced overall completeness relative to published human long-read studies[44-46, 55]. This is not surprising since HiFi sequencing has previously only been performed on bulk DNA isolated from millions of cells, which typically yield reads of 15-20 kb length, uniformly distributed across the genome, and without amplification errors and chimeras. It is, however, encouraging that genome assembly can also be achieved at the single cell level, despite the errors and biases present in single cell data that can be expected in all available technologies for whole genome amplification. Taken together, the results presented here represent the very first benchmark for reference-free analysis of single human cells.

Allelic dropout is always a challenge for single-cell WGS, and our results could be improved by having a higher proportion of DNA fragments encapsulated and amplified in the droplets. Several factors could lead to allelic dropout, such as DNA molecules that gets stuck in plastic, problems with getting sufficient reagents into all droplets, or DNA fragments that are either are too short or too long to be efficiently encapsulated. Another important issue is that the whole genome amplification introduces chimeras and errors. To some extent this might be helped by alternative amplification methods or modified experimental conditions, but regardless of such optimizations, there will likely remain a significant proportion of amplification errors in the resulting reads. This opens up for new bioinformatics tools, specifically designed for single-cell long-read WGS data, which are able to resolve amplification errors and maximize the utility of the data.

In this project, we opted for PacBio HiFi sequencing since it currently offers the highest per-read accuracy[56]. Although nanopore WGS[55] could be an alternative, it would likely be more

challenging to study SNVs and identify chimeric artefacts from nanopore reads because of their higher error rate. However, it is still an open question which platform would be best suited for this application in the future. This will depend on factors such as the sequencing yield, quality, and cost per sample, for coming versions of instruments. Due to the rapid developments of the long-read technologies, we anticipate that several of these parameters can be radically improved over the coming years.

In conclusion, we demonstrate that long-read genome analysis can be performed not only at a species, population, or individual-level, but also for a single human cell. Ultimately, new innovations and technical advances may in the future enable near-complete genome assemblies and full haplotype reconstructions from individual cells. Our work presented here is a first step in that direction.

Acknowledgements

We thank Thorarinn Blondal at Samplix for assistance in setting up the Xdrop system for dMDA experiments. The sequencing was performed at the SciLifeLab National Genomics Infrastructure in Stockholm and Uppsala. Computations were performed on resources provided by SNIC through Uppsala Multidisciplinary Center for Advanced Computational Science (UPPMAX) under Project SNIC sens2016003.

Methods

Single-cell samples

T cell samples were isolated from peripheral blood mononuclear cells taken from a living healthy human donor. The donor was previously vaccinated with the live, attenuated Yellow Fever Virus (YFV) vaccine (YFV-17D) as part of an ongoing study to investigate the dynamics of adaptive immunity to YFV vaccination (approved by the Regional Ethical Review Board in Stockholm, Sweden: 2008/1881-31/4, 2013/216-32, and 2104/1890-32). To expand CD8+ T cell clones from single YFV-specific memory CD8+ T-cells, mononuclear cells were isolated from peripheral blood by density centrifugation, and were first stained with HLA-A2/YFV(LLWNGPMAV)-dextramer FITC (Immudex, Denmark) for 15min at 4°C, followed by staining with anti-CD8a-BV570 (clone RPA-T8, Biolegend), anti-CD3-PE/Cy5 (clone UCHT1), anti-CD14-V500 (clone MφP9), anti-CD19-V500 (clone HIB19) (all from BD Biosciences), and LIVE/DEAD™ Fixable Aqua Dead Cell Stain (ThermoFisher) for 20 min at 4°C. After washing, single live $CD14^{-}CD19^{-}CD8^{+}CD3^{+}$HLA-A2/YFV-dextramer^{+} cells were sorted directly into 96 well U-bottom plates containing 500ng/ml HLA-A2/YFV peptide (LLWNGPMAV), 40U/ml human recombinant IL-2, and 40.000 irradiated (25Gy) CD3-depleted autologous PBMCs in T-cell media (RPMI1640 with 10% heat inactivated human AB sera, 1mM sodium pyruvate, 10mM Hepes, 50µM 2-mercaptoethanol, 1mM L-glutamine, 100U/ml penicillin and 50µg/ml streptomycin) and were cultured for 20 days. Every 7 days half of the media was replaced with fresh T-cell media containing 50U/ml IL-2, 500ng/ml peptide and 40.000 irradiated CD3-depleted autologous PBMCs, and the wells were visually inspected for proliferation. Clonal expansions of single HLA-A2/YFV-specific $CD8^{+}$ T-cells clones was confirmed by flow cytometry by using the same staining protocol as described above. Clones with sufficient number of clonal progeny were subsequently cryopreserved in fetal bovine serum with 10% DMSO and stored in liquid nitrogen until sorting for DNA/RNA

sequencing analysis. To isolate single cells from two selected YFV-specific CD8+ T cell clones (A and B), the clones were thawed, washed twice in RPMI1640 supplemented with 10% fetal bovine serum, and stained with as described above for initial sort and index-sorted into 96 well PCR plates (Thermo Fisher) or dMDA cartridge containing lysis buffer as described in the following sections.

Whole-genome amplification by droplet MDA (dMDA)

The single T-cells were sorted in a FACS instrument equipped with a custom 3D printed adapter holding a dMDA cartridge (cat# CA20100-16, Samplix ApS, Herlev, Denmark) and deposited directly into 2.8 µL lysis buffer (200 mM KOH, 5 mM EDTA (pH 8) and 40 mM 1.4 DTT) positioned at the dMDA cartridge's Inlet site. Single cells were lysed, and DNA denatured for 5 minutes at room temperature followed by addition of 1.4 µL neutralization buffer (400 mM HCl and 600 mM Tris HCl (pH 7.5)) and incubated for 5 min at room temperature. Then, 15.8 µL MDA amplification mixture including polymerase, primers, dNTP and reaction buffer (Samplix dMDA kit item# RE20300, Samplix ApS, Herlev, Denmark), was added, by injecting it into the dMDA cartridge Inlet site using a wide bore pipette. Finally, 75 µL dMDA oil (Samplix dMDA kit item# RE20300, Samplix ApS, Herlev, Denmark) was added into the inlet well (general cavity). The dMDA cartridge was moved into the Xdrop™ droplet generator (item# IN00100-SF002 Samplix ApS, Herlev, Denmark) to create single emulsion dMDA droplets. Droplets were collected into low bind 0.2 ml PCR vials from the Collection container of the dMDA cartridge and excess oil was removed from the bottom. The MDA droplets were incubated in a thermal block at 30°C for 16 hours and then heat inactivated at 65°C for 10 minutes and then cooled down to 4°C. Droplets were broken by adding 20 µL Break solution (Samplix dMDA kit item# RE20300, Samplix ApS, Herlev, Denmark) and the aqueous phase collected containing the amplified DNA. DNA material from Xdrop™ droplet MDA reactions

were quantified using Qubit ™ Fluorometer (ThermoFisher Inc., Waltham, MA, USA) and the DNA integrity investigated using Fragment Analyzer (Agilent Inc., Santa Clara, CA, USA) according to the manufacturer's instructions.

Whole-genome amplification by regular MDA

For comparison to the Xdrop™ droplet MDA process, single T-cells were sorted in the FACS and singly deposited directly into 2.8 µL lysis buffer (200 mM KOH, 5 mM EDTA (pH 8) and 40 mM 1.4 DTT) at the bottom of a 0.2 ml PCR vial or 96 well plate. Single cells were lysed, and DNA denatured for 5 minutes at room temperature followed by addition of 1.4 µL neutralization buffer (400 mM HCl and 600 mM Tris HCl (pH 7.5)) and incubation for 5 min at room temperature. The MDA reactions were prepared using RepliPHI Phi29 DNA polymerase and Reagent set (Epicentre, Illumina, Madison, WI, USA) according to the manufacturer's instructions. The reactions were carried out at 30°C for 8-16 hours and then heat inactivated at 65°C for 10 minutes.

Illumina whole genome sequencing

Illumina libraries were prepared using an automated version of the TruSeq DNA PCR-Free kit. Briefly, DNA was quantified using Qubit HS DNA and 1ug of DNA was used as input. The samples were then fragmented using Covaris E220 system, aiming for a fragment size of 350bp. Fragmented DNA was end-repaired, followed by size selection using Dynabeads MyOne Carboxylic Acid beads. Illumina TruSeq DNA CD Indexes with sample-specific barcode sequences were ligated and the final product was cleaned up using AMPure XP beads. Finished libraries were normalized based on their concentration and pooled for clustering. Clustering was done by 'cBot' and samples were sequenced on NovaSeq6000 (NovaSeq Control Software 1.6.0/RTA v3.4.4) with a 2x151 setup using 'NovaSeqXp' workflow in 'S4' mode flowcell. Bcl

to FastQ conversion was performed using bcl2fastq_v2.20.0.422 from the CASAVA software suite.

Mapping and variant detection in Illumina data

Illumina data was aligned to GRCh38 using BWA mem (0.7.17-r1188)[57]. The aligned data was sorted using Samtools sort (1.10)[58], and deduplicated using Picard MarkDuplicates (2.20.4-SNAPSHOT) (https://broadinstitute.github.io/picard/). Quality control was performed using Picard CollectGCMetrics and Picard WGSMetrics, as well as Samtools flagstats. The analysis was performed on the PCR-free bulk WGS and on each of the single cell samples. The subsequent bam files were searched for SNV and SV. The SNV calling was performed using Bcftools call (1.10+htslib-1.10) and the resulting SNV were decomposed and normalized using Vt[59]. SV detection was performed using TIDDIT (2.11.0)[50] and Manta (1.6.0)[49]. Briefly, the TIDDIT calls were filtered based on the Filter column – keeping only PASS variants. Next, the SV calls were combined using SVDB merge (2.4.0), combining calls positioned within 200 bp form each other, and sharing an overlap of at least 10% bases.

Downsampling and quality control of Illumina data

Downsampling to 100M read pairs was performed for each Illumina dataset using Samtools view. Thereafter, the coverage was analysed using TIDDIT cov, computing the coverage in bins sized 5 kbp and 500 kbp across the entire genome. The 500 kbp analysis was visualized using Circos[60], displaying coverage levels as a heatmap. The 5 kbp analysis was used to estimate the fraction of reads within high (>200X) coverage regions; the fraction of reads in such regions were computed using Samtools view, searching for reads overlapping high coverage regions as reported by TIDDIT cov.

PacBio whole genome sequencing

Two MDA samples, one from clone A and one from clone B, were chosen for sequencing based on input fragment length and DNA amount. The samples were fragmented to 10 kb using Megaruptor 2 (Diagenode). For each fragmented sample, SMRTbell construction was performed using the Express Template prep kit 2.0 and incomplete SMRTbells were removed using the SMRTbell Enzyme Clean up Kit. SMRTbells were size selected using AMPure beads to remove fragments shorter than 3kb. The library preparation procedure is described in the protocol "Preparing HiFi Libraries from Low DNA Input Using SMRTbell Express Template Prep Kit 2.0" from PacBio. The SMRTbell library sizes and profiles were evaluated using the Agilent DNA 12000 kit on the Bioanalyzer system. PacBio sequencing was performed on the Sequel II instrument with 30h movie time.

Mapping and variant detection in PacBio data

The PacBio data was analyzed using tools available in the SMRTLink v8 GUI. HiFi reads were generated using the circular consensus sequencing (CCS) tool. The HiFi reads were aligned to hg38 using Minimap2[61]. Structural variants were detected using PBSV. DeepVariant[48] (v 1.0.0) was used for PacBio SNVs calling.

Detection of somatic SV events

Somatic SV were found by removing the previously mentioned germline SVs from the single-cell sequencing WGS data. Candidate somatic SV were discovered in a variety of settings: through the previously mentioned PacBio analysis, through Illumina WGS analysis, or through the combination of Illumina and PacBio data. The combined analysis was initiated by removing any germline calls from the Illumina single-cell WGS lists. Next the remaining calls were merged with the quality controlled PacBio calls, and the intersection was considered potential

somatic SV. The Illumina-only somatic analysis focused on the intersection between Manta and TIDDIT, searching for shared calls not present in the Germline SV list. In all these cases, the intersection of callers/technologies, were found through SVDB merge. Somatic SV of interest were manually inspected using[62].

Assembly of PacBio single-cell data

The PacBio HiFi reads were first filtered to remove intramolecular chimeras. This filtering was done by aligning each read to its own sequence using BLAST[63] and removing all reads that have a secondary blast hit against themselves, at an identity higher than 90%. In this way, reads containing intramolecular chimeras such as inversions and duplications are efficiently removed. The HiFi reads that pass the chimera filtering were then assembled using hifiasm[52] (v. 0.7-dirty-r255).

Tables

Table 1. PacBio Sequel II run statistics and alignment results for two human single T-cells

	Single-cell A	**Single-cell B**
≥ Q20 reads	2,750,802	2,547,184
≥ Q20 yield (bp)	19,880,131,345	20,169,954,798
≥ Q20 read length (mean, bp)	7,227	7,918
≥ Q20 read quality (median)	Q36	Q36
Number aligned reads	2,739,035 (99.57%)	2,517,588 (98.83%)
Number of alignments	6,508,237	6,235,924
Aligned read mean concordance	99.18%	99.11%
Aligned read length (mean)	2,994	3,149
Aligned read length N50	5,429	6,476
Aligned read length 95%	8,691	9,739
Aligned read length Max	43,386	48,731
Mean coverage	6	6
Covered bases	39.60%	27.71%

Table 2: *De novo* assembly results for the two T-cells A and B

	Single-cell A	**Single-cell B**
Assembly statistics:		
Filtered CCS reads (bp)	8,794,585,174 (44.2%)	9,405,139,162 (46.6%)
Assembly size, primary (bp)	598,293,718	454,096,399
Assembly completeness	19.4%	14.7%
Contig N50	34,883	41,528
Max contig size	206,875	578,275
Assembly size, alternative (bp)	44,706,740	36,132,542
Contig N50, alternative	18,969	20,976
Max contig size, alternative	79,718	94,865
BUSCO gene models:		
complete	1762 (12.8%)	1236 (9.0%)
duplicated	17 (0.1%)	14 (0.1%)
fragmented	58 (0.4%)	250 (1.8%)
missing	11960 (86.8%)	12294 (89.2%)
Mitochondrion:	Complete	Complete

Figures

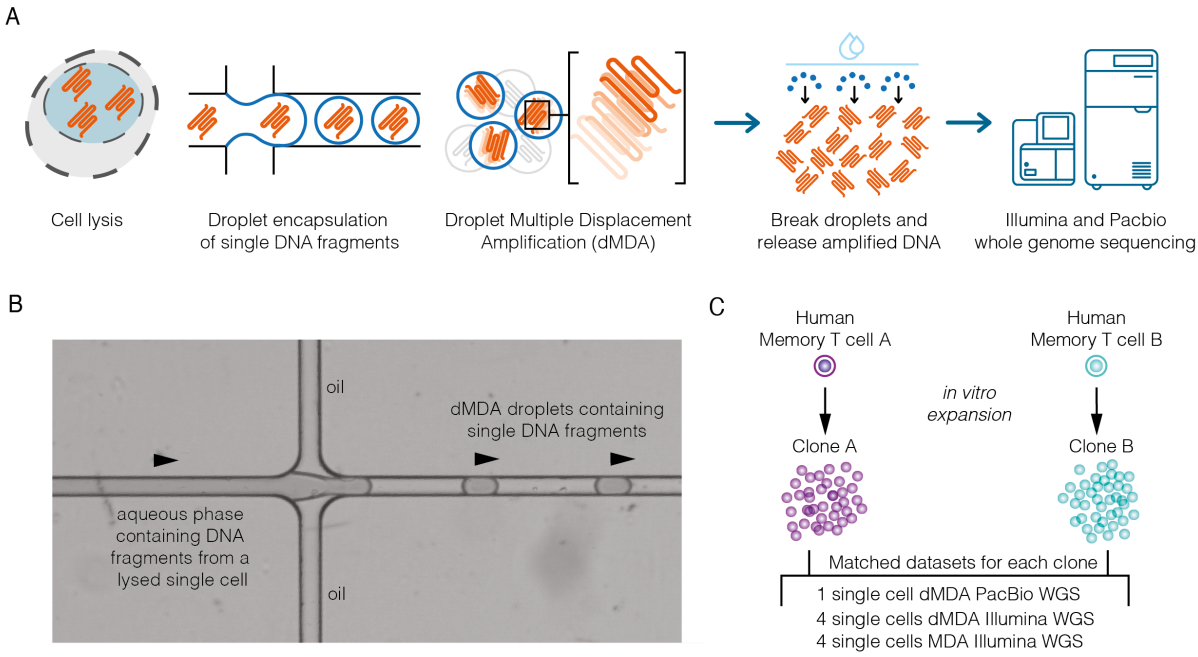

Figure 1. Overview of the single-cell DNA amplification and sequencing experiment. A) An individual cell is isolated by fluorescence activated cell sorting (FACS) and placed into a well containing lysis buffer. DNA molecules from the lysed single cell are then encapsulated in picoliter droplets using the Xdrop microfluidic system, after which dMDA whole genome amplification takes place inside each droplet. After amplification, the droplets are broken and DNA is released, followed by library preparation and whole genome sequencing using short- (Illumina) and long-read (PacBio) technologies. **B)** Image showing how droplets are formed in the Xdrop microfluidic system. An aqueous phase containing lysed DNA and dMDA reagents encounters an oil layer, resulting in <100 µm diameter droplets where single DNA fragments are captured. The Xdrop system has capacity to produce around 50,000 droplets in 45 seconds. **C)** Two human memory T-cells from the same individual (A and B) were used as starting point for the experiments. Collections of daughter cells were obtained by *in vitro* expansion, and individual cells from clones A and B were analyzed using Illumina and PacBio whole genome sequencing.

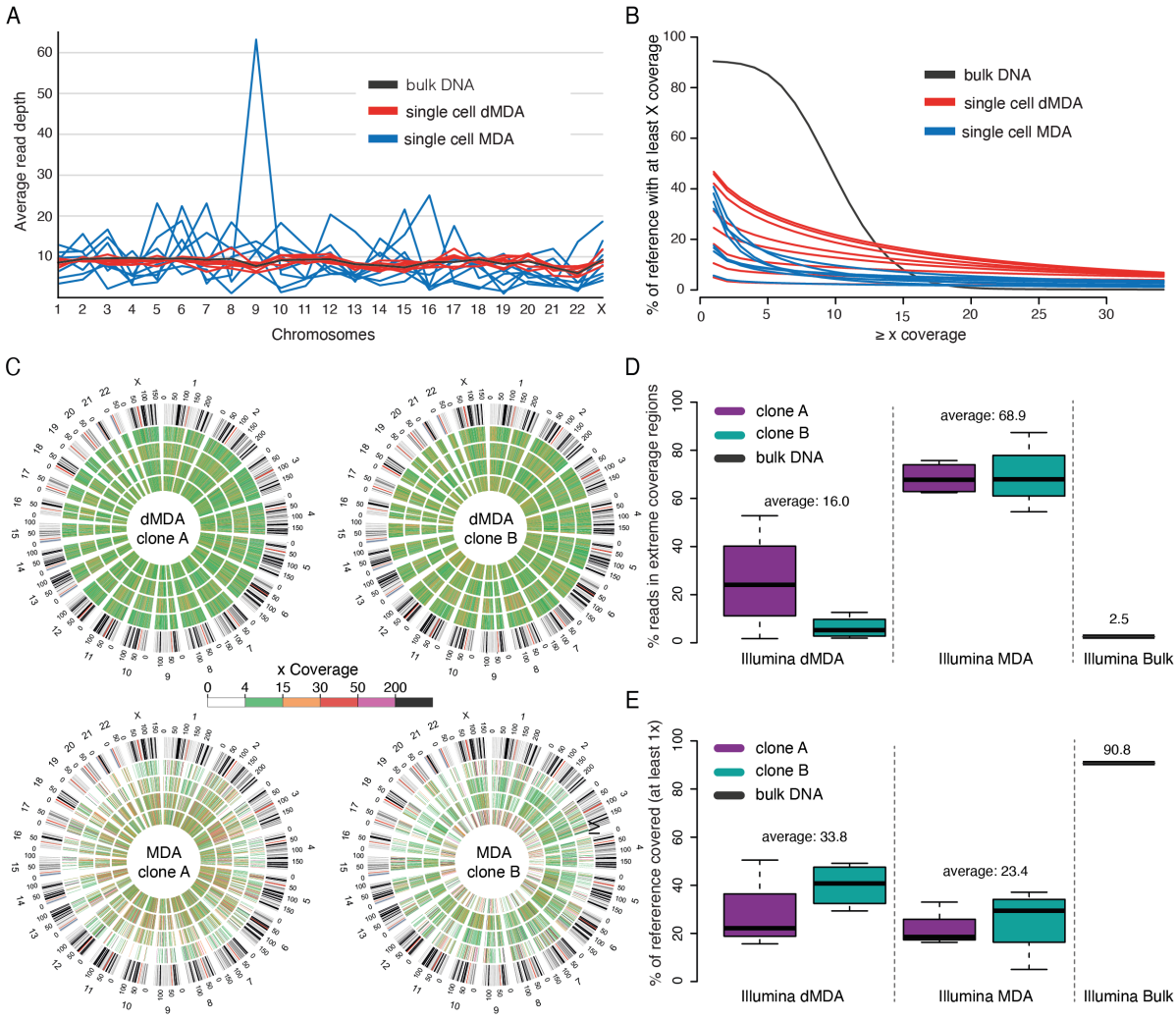

Figure 2. Comparison of MDA and dMDA for whole genome amplification. These results are based on Illumina MDA, dMDA and bulk sequencing where the datasets that have been randomly downsampled to contain the same number of reads. **A)** The figure displays the average sequencing depth across the human chromosomes. The dMDA single-cell samples display good uniformity of coverage, whereas the MDA data show high spikes due to amplification bias. **B)** Plot showing the percentage of bases in the reference genome (y-axis) having a minimal coverage (x-axis). On average the dMDA samples have more bases covered at a range 10x-30x, as compared to the single-cell samples subjected to regular MDA. **C)** Circle plots showing sequencing coverage in 500kb bins for all of the Illumina single-cell samples, color coded from 0x coverage (white) to over 200x coverage (black). Four replicate samples are included in each of the circle plots, and the chromosomal coordinates are displayed in the outermost circle. The dMDA samples at the top row display more even coverage than the MDA samples below, with more of the bins having average coverage in 4x-15x coverage range (green). **D)** Box plot showing the percentage of reads aligning to regions of extreme (≥200x) coverage. **E)** Box plot showing the percentage of reference bases that are covered by at least one read.

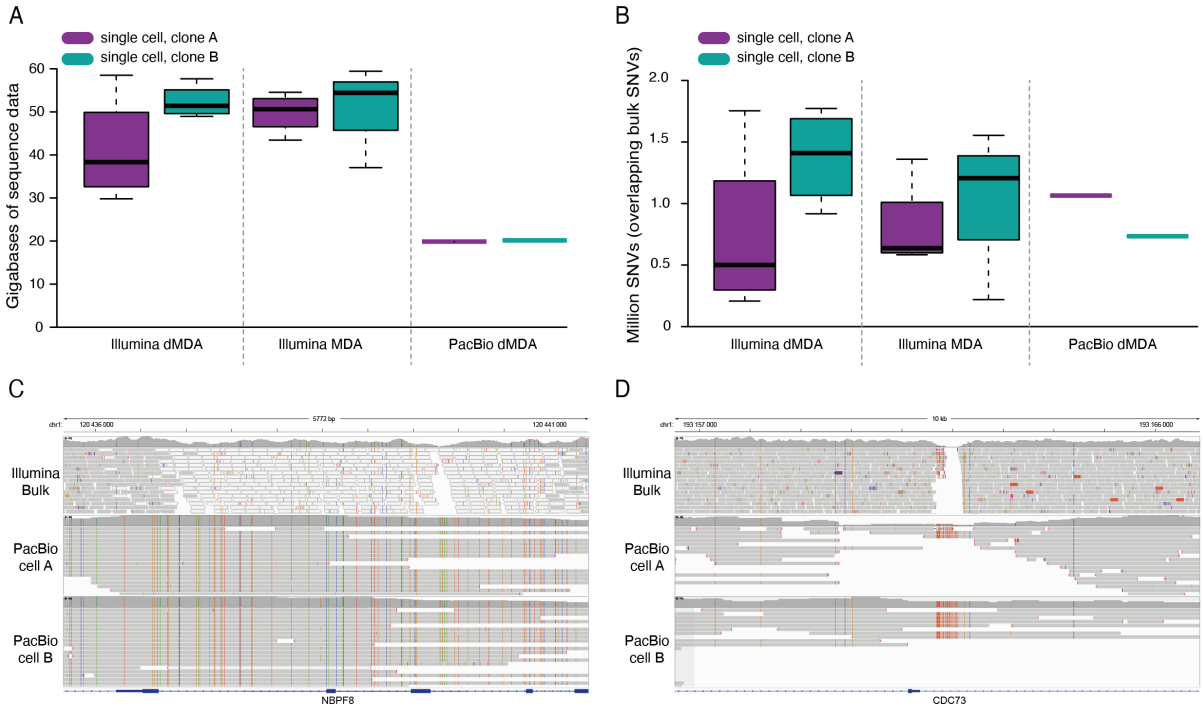

Figure 3. Results of SNV analyses in single-cell samples. A) Total amount of data generated for the single cell samples. On average, 48.7 Gb was generated for the sixteen Illumina samples, and 20.0 Gb for the two PacBio samples. **B)** Number of SNV overlapping in the single cells that were found to be overlapping with SNVs identified in the Illumina bulk sample. 992,338 such SNVs/sample were found in the sixteen Illumina samples, 899,827 SNVs/sample for the two PacBio samples. **C)** Example of a "dark" genic region (*NBPF8*) where Illumina data fails to align uniquely, while SNVs can be identified and phased in the PacBio single cell data. **D)** Another example of a "dark" genic region (*CDC73*), where PacBio reads from the two single cells span across a repetitive region that lacks coverage in the Illumina bulk sequencing data.

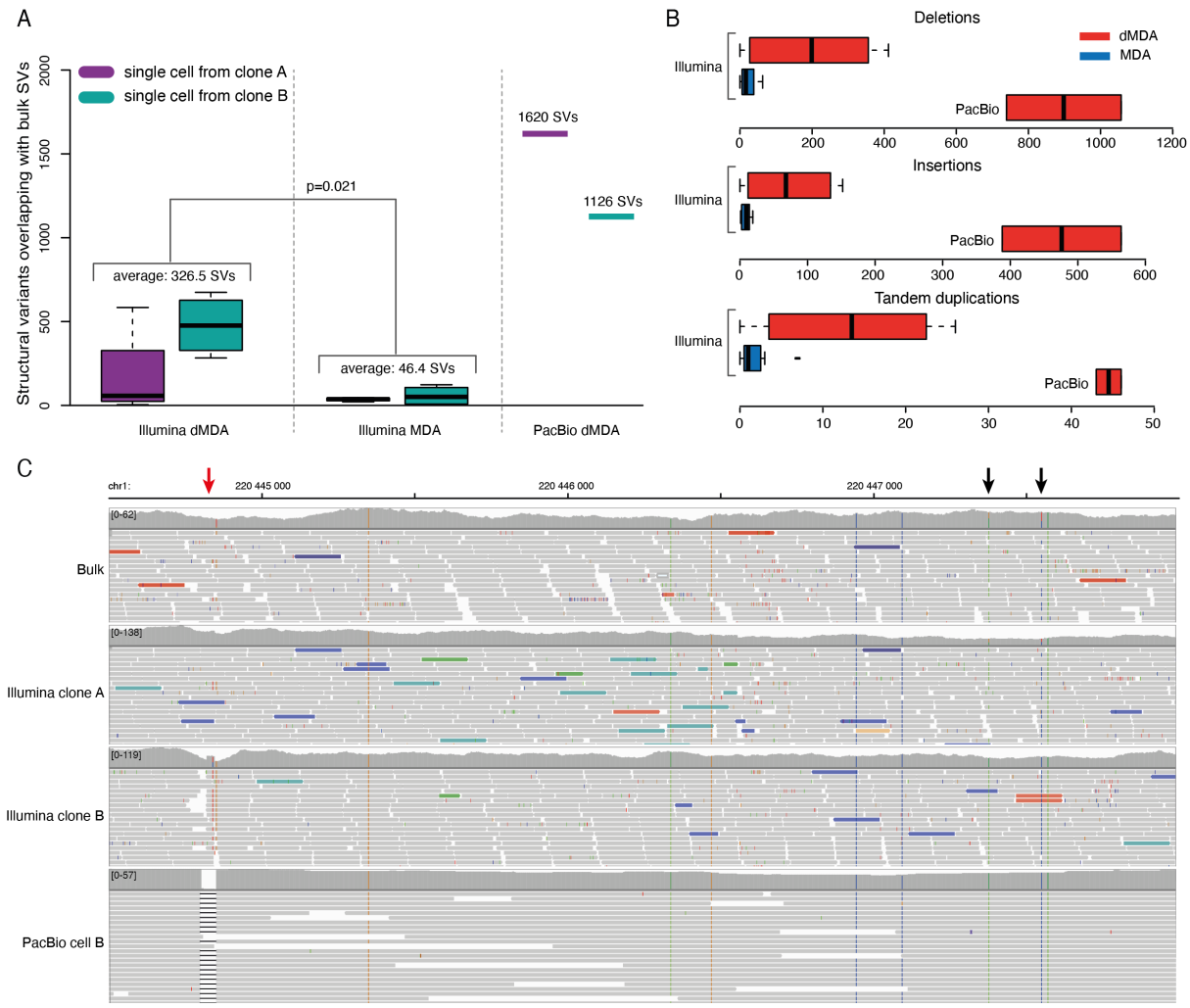

Figure 4. Structural variants detected in single-cell whole genome sequencing data. A) The box plots show the number of SVs that were found in a single cell, while also being detected by SV analysis of the Illumina bulk sample. A higher number of bulk-supported SVs are found in the dMDA samples (average 326.5 SVs) as compared to the regular MDA samples (average 46.4 SVs), and a Welsh t-test resulted in rejection of the null hypothesis that there is no difference between the two distributions (p-value 0.021). However, by far the highest numbers of bulk-supported SVs was found in the PacBio single-cell data; 1620 for T-cell A and 1126 for T-cell B. **B)** The boxplots show the same SVs as in panel A), divided into deletion (top panel), insertion (middle panel) and tandem duplication events (bottom panel). **C)** IGV plot showing one example of a candidate somatic 50 bp deletion, indicated by a red arrow at the top. This event was detected both in the PacBio and Illumina single cell data for T-cell B. However, it is not visible in the bulk sequencing or in the single-cell data from T-cell A. The two black arrows to the right indicate positions with heterozygous SNVs that can be used for phasing of the deletion in the PacBio data for T-cell B.

References

1. Lewin, H.A. et al. Earth BioGenome Project: Sequencing life for the future of life. *Proc Natl Acad Sci U S A* **115**, 4325-4333 (2018).
2. Rhie, A. et al. Towards complete and error-free genome assemblies of all vertebrate species. *bioRxiv*, 2020.2005.2022.110833 (2020).
3. Logsdon, G.A. et al. The structure, function, and evolution of a complete human chromosome 8. *bioRxiv*, 2020.2009.2008.285395 (2020).
4. Miga, K.H. et al. Telomere-to-telomere assembly of a complete human X chromosome. *Nature* (2020).
5. Audano, P.A. et al. Characterizing the Major Structural Variant Alleles of the Human Genome. *Cell* **176**, 663-675 e619 (2019).
6. Chaisson, M.J.P. et al. Multi-platform discovery of haplotype-resolved structural variation in human genomes. *Nat Commun* **10**, 1784 (2019).
7. Ebert, P. et al. Haplotype-resolved diverse human genomes and integrated analysis of structural variation. *Science* (2021).
8. Mantere, T., Kersten, S. & Hoischen, A. Long-Read Sequencing Emerging in Medical Genetics. *Front Genet* **10**, 426 (2019).
9. Ameur, A., Kloosterman, W.P. & Hestand, M.S. Single-Molecule Sequencing: Towards Clinical Applications. *Trends Biotechnol* **37**, 72-85 (2019).
10. Ebbert, M.T.W. et al. Systematic analysis of dark and camouflaged genes reveals disease-relevant genes hiding in plain sight. *Genome Biol* **20**, 97 (2019).
11. Navin, N. et al. Tumour evolution inferred by single-cell sequencing. *Nature* **472**, 90-94 (2011).
12. Evrony, G.D. et al. Single-neuron sequencing analysis of L1 retrotransposition and somatic mutation in the human brain. *Cell* **151**, 483-496 (2012).
13. Lu, S. et al. Probing meiotic recombination and aneuploidy of single sperm cells by whole-genome sequencing. *Science* **338**, 1627-1630 (2012).
14. Wang, J., Fan, H.C., Behr, B. & Quake, S.R. Genome-wide single-cell analysis of recombination activity and de novo mutation rates in human sperm. *Cell* **150**, 402-412 (2012).
15. Zong, C., Lu, S., Chapman, A.R. & Xie, X.S. Genome-wide detection of single-nucleotide and copy-number variations of a single human cell. *Science* **338**, 1622-1626 (2012).
16. Brazhnik, K. et al. Single-cell analysis reveals different age-related somatic mutation profiles between stem and differentiated cells in human liver. *Sci Adv* **6**, eaax2659 (2020).
17. Kirkness, E.F. et al. Sequencing of isolated sperm cells for direct haplotyping of a human genome. *Genome Res* **23**, 826-832 (2013).
18. Evrony, G.D. et al. Cell lineage analysis in human brain using endogenous retroelements. *Neuron* **85**, 49-59 (2015).
19. Lodato, M.A. et al. Aging and neurodegeneration are associated with increased mutations in single human neurons. *Science* **359**, 555-559 (2018).
20. Lodato, M.A. et al. Somatic mutation in single human neurons tracks developmental and transcriptional history. *Science* **350**, 94-98 (2015).
21. Dong, X. et al. Accurate identification of single-nucleotide variants in whole-genome-amplified single cells. *Nat Methods* **14**, 491-493 (2017).
22. Lan, F., Demaree, B., Ahmed, N. & Abate, A.R. Single-cell genome sequencing at ultra-high-throughput with microfluidic droplet barcoding. *Nat Biotechnol* **35**, 640-646 (2017).
23. Vitak, S.A. et al. Sequencing thousands of single-cell genomes with combinatorial indexing. *Nat Methods* **14**, 302-308 (2017).

24. Zahn, H. et al. Scalable whole-genome single-cell library preparation without preamplification. *Nat Methods* **14**, 167-173 (2017).
25. Zhang, L. et al. Single-cell whole-genome sequencing reveals the functional landscape of somatic mutations in B lymphocytes across the human lifespan. *Proc Natl Acad Sci U S A* **116**, 9014-9019 (2019).
26. Bohrson, C.L. et al. Linked-read analysis identifies mutations in single-cell DNA-sequencing data. *Nat Genet* **51**, 749-754 (2019).
27. Hard, J. et al. Conbase: a software for unsupervised discovery of clonal somatic mutations in single cells through read phasing. *Genome Biol* **20**, 68 (2019).
28. Hazen, J.L. et al. The Complete Genome Sequences, Unique Mutational Spectra, and Developmental Potency of Adult Neurons Revealed by Cloning. *Neuron* **89**, 1223-1236 (2016).
29. Lee-Six, H. et al. Population dynamics of normal human blood inferred from somatic mutations. *Nature* **561**, 473-478 (2018).
30. McConnell, M.J. et al. Mosaic copy number variation in human neurons. *Science* **342**, 632-637 (2013).
31. Satas, G. & Raphael, B.J. Haplotype phasing in single-cell DNA-sequencing data. *Bioinformatics* **34**, i211-i217 (2018).
32. Cai, X. et al. Single-cell, genome-wide sequencing identifies clonal somatic copy-number variation in the human brain. *Cell Rep* **8**, 1280-1289 (2014).
33. Baslan, T. et al. Optimizing sparse sequencing of single cells for highly multiplex copy number profiling. *Genome Res* **25**, 714-724 (2015).
34. Knouse, K.A., Wu, J. & Amon, A. Assessment of megabase-scale somatic copy number variation using single-cell sequencing. *Genome Res* **26**, 376-384 (2016).
35. Upton, K.R. et al. Ubiquitous L1 mosaicism in hippocampal neurons. *Cell* **161**, 228-239 (2015).
36. Evrony, G.D., Lee, E., Park, P.J. & Walsh, C.A. Resolving rates of mutation in the brain using single-neuron genomics. *Elife* **5** (2016).
37. Gawad, C., Koh, W. & Quake, S.R. Single-cell genome sequencing: current state of the science. *Nat Rev Genet* **17**, 175-188 (2016).
38. Chen, C. et al. Single-cell whole-genome analyses by Linear Amplification via Transposon Insertion (LIANTI). *Science* **356**, 189-194 (2017).
39. Dean, F.B. et al. Comprehensive human genome amplification using multiple displacement amplification. *Proc Natl Acad Sci U S A* **99**, 5261-5266 (2002).
40. Fu, Y. et al. Uniform and accurate single-cell sequencing based on emulsion whole-genome amplification. *Proc Natl Acad Sci U S A* **112**, 11923-11928 (2015).
41. Leung, K. et al. Robust high-performance nanoliter-volume single-cell multiple displacement amplification on planar substrates. *Proc Natl Acad Sci U S A* **113**, 8484-8489 (2016).
42. Marcy, Y. et al. Nanoliter reactors improve multiple displacement amplification of genomes from single cells. *PLoS Genet* **3**, 1702-1708 (2007).
43. Wenger, A.M. et al. Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nat Biotechnol* **37**, 1155-1162 (2019).
44. Vollger, M.R. et al. Improved assembly and variant detection of a haploid human genome using single-molecule, high-fidelity long reads. *Ann Hum Genet* **84**, 125-140 (2020).
45. Porubsky, D. et al. Fully phased human genome assembly without parental data using single-cell strand sequencing and long reads. *Nat Biotechnol* (2020).

46. Garg, S. et al. Chromosome-scale, haplotype-resolved assembly of human genomes. *Nat Biotechnol* (2020).
47. Madsen, E.B., Hoijer, I., Kvist, T., Ameur, A. & Mikkelsen, M.J. Xdrop: Targeted sequencing of long DNA molecules from low input samples using droplet sorting. *Hum Mutat* **41**, 1671-1679 (2020).
48. Poplin, R. et al. A universal SNP and small-indel variant caller using deep neural networks. *Nat Biotechnol* **36**, 983-987 (2018).
49. Chen, X. et al. Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. *Bioinformatics* **32**, 1220-1222 (2016).
50. Eisfeldt, J., Vezzi, F., Olason, P., Nilsson, D. & Lindstrand, A. TIDDIT, an efficient and comprehensive structural variant caller for massive parallel sequencing data. *F1000Res* **6**, 664 (2017).
51. https://github.com/PacificBiosciences/pbsv.
52. Cheng, H., Concepcion, G.T., Feng, X., Zhang, H. & Li, H. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nat Methods* **18**, 170-175 (2021).
53. Simao, F.A., Waterhouse, R.M., Ioannidis, P., Kriventseva, E.V. & Zdobnov, E.M. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**, 3210-3212 (2015).
54. https://www.pacb.com/blog/introducing-the-ultra-low-input-protocol-for-smrt-sequencing/. (2020).
55. Jain, M. et al. Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nat Biotechnol* **36**, 338-345 (2018).
56. Logsdon, G.A., Vollger, M.R. & Eichler, E.E. Long-read human genome sequencing and its applications. *Nat Rev Genet* **21**, 597-614 (2020).
57. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv:1303.3997 (2013).
58. Li, H. et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-2079 (2009).
59. Tan, A., Abecasis, G.R. & Kang, H.M. Unified representation of genetic variants. *Bioinformatics* **31**, 2202-2204 (2015).
60. Krzywinski, M. et al. Circos: an information aesthetic for comparative genomics. *Genome Res* **19**, 1639-1645 (2009).
61. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094-3100 (2018).
62. Thorvaldsdottir, H., Robinson, J.T. & Mesirov, J.P. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief Bioinform* **14**, 178-192 (2013).
63. Altschul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J. Basic local alignment search tool. *J Mol Biol* **215**, 403-410 (1990).