Method for comprehensive detection of somatic mosaicism using single cell sequencing data.

A THESIS

SUBMITTED TO THE FACULTY OF THE

UNIVERSITY OF MINNESOTA

BY

Vivekananda Sarangi

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY IN

BIOINFORMATICS AND COMPUTATIONAL BIOLOGY

Adviser: Alexej Abyzov

OCTOBER 2023

Acknowledgement:

I am profoundly grateful to all those who have contributed to my academic journey and the completion of this PhD thesis. Your unwavering support, encouragement, and guidance have been instrumental in this endeavor.

First and foremost, I would like to extend my heartfelt gratitude to the Mayo Clinic PDAP program for their financial assistance, which made my doctoral studies possible.

I owe a debt of gratitude to the individuals who believed in me and encouraged me to pursue a PhD. Yan and Edward Asmann, John Copland and Jaime Davila, your constant support and faith in my abilities have been a driving force throughout this journey.

My time at Mayo Clinic was enriched by the camaraderie and stimulating scientific discussions with my talented colleagues and friends. Aditya Bhagwate, Daniel O'Brien, Stephen Johnson, Jagadeshwar Balan, Aaron Norman, Nicholas Boddicker, Steve Hart, Gavin Oliver and many others, your insights and conversations during meetings, coffee breaks, lunches, and happy hours have been invaluable in shaping my research.

I am deeply grateful for the mentorship and guidance provided by James Cerhan, Susan Slager, Yan Asmann, Jean-Pierre Kocher, Anne Novak, and Flora Vaccarino and her team at Yale University. Your expertise and dedication have not only influenced my research but also my personal growth.

To my wife, Valentina Zanfagnin, I extend my heartfelt thanks for pushing me to join the program when I had doubts and for continually challenging me to step out of my comfort zone. Your unwavering support has been my cornerstone. I would also like to express my

gratitude to my family back in India for their enduring love, support, and for instilling in me a sense of scientific curiosity from an early age.

I am indebted to my committee members, Chad Myers, Steven Hart, and Tim Starr, for their guidance and insightful questions that exposed gaps in my knowledge and helping me to bridge them.

A special note of appreciation goes out to the members of the Abyzov lab—Taejong Bae, Milovan Suvakov, Arijit Panda, Yifan Wang, and Yeongjun Jang—for their collaboration and contributions to my research.

Finally, I would like to acknowledge my advisor, Alexej Abyzov, for being an exceptional mentor. Your dedication to fostering my intellectual growth and research skills has been unwavering. Your patience in the face of challenges, your firm guidance during moments of uncertainty, and your gentle steering towards research excellence have been instrumental in keeping me on the right path throughout my doctoral journey. Your mentorship has not only shaped my academic and research capabilities but has also inspired me to strive for excellence and embrace the ever-evolving landscape of science.

Last but certainly not least, I want to express my gratitude to my children, Luca and Enea, whose joy and presence enriched my life in the past few years, even if it occasionally slowed my progress. Your love and support mean the world to me. This work would not have been the same without you. Thank you all for being an integral part of my academic and personal growth.

Abstract:

Mutations acquired in each cell during and after embryogenesis are passed to the descendant cells such that, within the same individual, different populations of somatic cells have slightly different DNA, resulting in genomic mosaicism. These mosaic mutations might give the cells proliferative advantage, and ultimately cause cancer, or can affect the cellular functions without a proliferative effect as in case of diverse neurological diseases. This makes the detection of mosaic mutations important for understanding the mechanism of various diseases. Although whole genome sequencing of bulk tissue has been used for detecting mosaic mutations, it is not sensitive enough to detect mosaic mutations present in less than 2% of the cells. This hurdle has been overcome by single-cell DNA sequencing (scDNA-seq) which in recent times has emerged as an efficient tool for discovering and analyzing mosaic mutations. However, there are pitfalls and drawbacks of scDNA-seq that need to be addressed.

First, the amount of DNA in a single cell is not sufficient for sequencing and needs to be amplified. The most used amplification method MDA (Multiple Displacement Amplification) has drawbacks such as uniformity of amplification and high error rate. Some cells go through the uniform amplification process with less errors than others and it is important to identify the quality of the data prior to mosaic mutation discovery. In Chapter 2 I discuss a method we have developed which can rank amplification quality using low coverage (1X) sequencing data to give an estimate of uniformity of amplification which can be used to select single cells for high coverage (>30X) sequencing.

Second, for detection of mosaic mutations in single cells, the most frequently used approach is to compare single cell genomes to that of a matched reference bulk. While this approach works well to find private mutations in a cell, it misses mutations that are present at higher frequency, and consequently present in multiple cells in the reference bulk. To address this, I have developed, described in Chapter 3, a bioinformatic tool to detect somatic mosaicism including SNVs and INDELs using pair-wise comparison of single cell data and provide data demonstrating that the method outperforms existing methods.

TABLE OF CONTENTS

List of Figures:

List of Tables:

Abbreviations:

MDA: Multiple displacement amplification

PTA: Primary Template Amplification

VAF: Variant allele frequency

scDNA-seq: Single cell DNA sequencing

scRNA-seq: Single cell mRNA sequencing

WGA: Whole genome amplification

DOP-PCR: Degenerate oligonucleotide-primed polymerase chain reaction

MALBAC: Multiple annealing and looping-based amplification cycles

MAPD: Median absolute pairwise difference

WGS: Whole genome sequencing

PCR: Polymerase chain reaction

HETs: Heterozygous SNPs

SNPs: Single nucleotide polymorphism

SNV: Single nucleotide variation

INDELS: Insertions and deletions

VCF: Variant call format (file format)

CNV: Copy number variation

BED: Browser extensible data (file format)

SV: Structural variation

ADA: Allele dropout analysis

iPSC: Induced pluripotent stem cells

CHAPTER 1 – Introductions

1.1 SOMATIC MOSAICISM

Somatic mosaicism refers to the presence of a distinct population of cells in the body harboring slightly different DNA code from each other. One of the most common mechanisms for somatic mosaicism is the incorporation of postzygotic mutations during cell division. Once the cell is fertilized, it develops into a blastocyst, which further divides and forms the inner cell mass and trophoblast. Further divisions lead to the germ layers (ectoderm, endoderm, and mesoderm) and the pluripotent cells in the germ layers keep dividing, producing more specialized cells, which forms different tissues, organs and eventually to a fully developed fetus. Now, as cells start dividing from the fertilized egg, at each division, due to imperfect DNA replication (~0.27 to 0.99 errors in 10^9 nucleotides) (1), de-novo mutations arise and are transferred to the daughter cells. These early somatic mutations are passed on to all subsequent future cells and are accumulated over time. As the cells divide into different lineages, each lineage will tend to have its lineage specific private mutations. It gets more complex as different lineages converge together to form tissues and organs (2). It is estimated that approximately 1.2 SNVs are introduced per cell division during development leading to an accumulation of about 50 to 100 de-novo SNVs in each newborn human (3-8). The accumulation of somatic mutation increases with age (9-11), and is roughly estimated to be about 16-50 SNVs per cell per year when looking at multiple tissues (6, 12, 13). Environmental factors like tobacco smoking and alcohol consumption can also increase the rate of somatic mutation accumulation (14). These mutations present at a low frequency in the body, by themselves, or with help of future mutations can lead to a variety of health conditions.

1.2 CONDITIONS ASSOCIATED WITH SOMATIC MOSAICISM

Mosaic mutations arising early in embryogenesis pose a cancer predisposition risk if they affect cancer susceptibility genes (CSG) and their detection may affect therapeutic and prophylactic measures (15). Almost all cancers expand from a single cell, such that the genome of the founder cell is replicated in all cells of the cancer. As the cancer progresses and proliferates uncontrollably, more mutations are incorporated and genomic heterogeneity of the mass increases, making the detection of driver mutations challenging. However, multiple evidence suggests that most variants in cancer originate before the malignant transformation and thus can inform about mutagenesis in normal cells and its relationship to age (11, 16, 17). A benign somatic mutation early in life can aggravate the pathogenesis of a cancer when a second de-novo mutation is introduced later in life (11).

Neuronal **(**non-dividing cells**)** mosaic mutations rate also increases with age at the rate of 16-21 SNVs per year due to oxidation and age specific defective DNA damage repair (18, 19). In case of an Alzheimer's brain, this number increases compared to neurotypical individuals and is probably due to the increase of ROS (reactive oxygen species) caused by accumulation of amyloid and tau proteins (18). In conditions such as autism spectrum disorder (ASD), mosaic mutations occur more frequently in affected individuals compared to healthy siblings and contribute to 5.1 % of autism diagnosis (20). Mosaic mutation in the gene SLC34A2 in a pediatric epileptic patient, present between 4%-20% allele frequency in 12 different tissue samples showed correlation between the proportion of the mutation and the severity of the disease (21).

This makes study of somatic mosaicism extremely vital to understanding their occurrence in unaffected tissues and their potential to lead to more aggressive conditions.

1.3 SINGLE CELL SEQUENCING FOR DETECTION OF MOSAIC MUTATION

Although whole genome sequencing of bulk tissue has been used for detecting somatic mutations, it is not sensitive enough to detect mosaic mutations present below 1% variant allele frequency (VAF), i.e., a heterozygous mutation present in less than 2% of the cells. This is because bulk tissue is made up of heterogeneous cell populations and the genomic material of all the cells are mixed during DNA extraction. After sequencing , it is easy to identify mutations which are present in the majority of the cells (germline) but the mutations coming from a minority cell population becomes indistinguishable from noise (22). This hurdle has been overcome by single-cell DNA sequencing (scDNA-seq) which in recent times has emerged as an efficient tool for studying mosaic mutations (23-25). Since we are looking at the DNA of a single cell, the private somatic mutations have the same profile as the germline mutation and are easier to differentiate from sequencing and/or library preparation artifacts. Additionally, somatic mutations detected from single cells can be used for lineage tree construction which can give insight on the origin of the condition such as cancer's founder cell and for the study of developmental biology (14, 26-28). While in bulk sequencing, mosaic mutation can be differentiated from germline based on their variant allele frequencies, in single cell sequencing the same method is not effective. Since the germline and mosaic mutations have the same allele frequency in a single cell, a

supplemental bulk sequencing data is required to identify and filter the germline mutation from the single cell data leaving behind the cell's private mutations.

1.4 CHALLENGES

There are challenges in single cell sequencing that need to be addressed for effective identification of somatic mosaicism. Since the starting DNA amount in a single cell is very low, an additional step of DNA amplification is required. There are two types of broad methods for DNA amplification: cell cloning and enzymatic whole genome amplification (WGA). Depending on the experimental design one of the two methods can be used. WGA methods, unlike cell cloning, directly isolate extracted DNA from single cells and then amplify it, making it possible to sequence the DNA of cells which cannot be cultured, such as neurons. There are five types of WGA methods: DOP-PCR (Degenerate Oligonucleotide–Primed Polymerase Chain Reaction) (29), MDA (Multiple Displacement Amplification) (30), PTA (Primary Template Amplification) (31), MALBAC (Multiple Annealing and Looping–Based Amplification Cycles) (32) and LIANTI (33), each having its advantages and drawbacks. MDA is the most widely used method for WGA owing to its longer fragment length (up to 70 kbps), low error rate during amplification and higher fraction of the genome being amplified as compared to the other WGA methods (34). However, one of the major drawbacks of MDA is introduction of allelic imbalance where one allele is overrepresented than the other which can suppress the real signal and make an artifact indistinguishable from real mutation. Additionally, the phi29 polymerase used in MDA is sensitive to DNA breaks and doesn't amplify the whole genome uniformly leading to difficulties in identifying somatic mutations in less amplified regions. In Chapter 2, I

discuss this issue in more detail and introduce a method that can identify the amplification quality using a low coverage data prior to sending the samples for high coverage sequencing, hence reducing the cost of sequencing and bioinformatics efforts.

Another challenge using single cell sequencing is detecting high frequency mutations that are present at a higher proportion in the bulk sample. As discussed earlier, to call somatic mutations in a single cell we need the corresponding bulk tissue data to be able to differentiate between somatic and germline mutations. In order to do so, a common practice is to use somatic mutations caller such as Mutect (35) or Strelka (36) which takes the bulk data and the single cell data as input, and for each mutation in the single cell, check if that specific mutation is present in the bulk. If it's present in the bulk, the mutation is not considered a somatic mutation call. However, if there are cells (carrying somatic mutations) present at a higher percentage in the bulk and the single cell being compared is from that specific clone, the somatic mutations will never be called by this method. All the somatic mutations will be deemed as germline since the callers will see their presence in the bulk. The problem can get exacerbated if the single cells are amplified by MDA. There are single cell callers that correct amplification bias within local genomic regions (37) or use heterozygous bulk SNPs to create an allelic balance model to correct for allele imbalance in MDA amplified cells (38). Even though these callers reduce the noise introduced by MDA, the noise persists making detection of high frequency mosaic mutations challenging. In Chapter 3, I discuss a method which uses a comprehensive all to all comparison of single cells without the need of a bulk sample to identify high frequency somatic mutations both from clonally amplified and MDA amplified single cells.

CHAPTER 2 - SCELLECTOR: ranking amplification bias in single cells using shallow sequencing.

Vivekananda Sarangi[1], Alexandre Jourdon[2], Taejeong Bae[1], Arijit Panda[1], Flora Vaccarino[2,3] and Alexej Abyzov[1*]

Affiliations:

1. Department of Health Sciences Research, Center for Individualized Medicine, Mayo Clinic, Rochester, MN 55905, USA
2. Child Study Center, Yale University, New Haven, CT 06520, USA
3. Department of Neuroscience, Yale University, New Haven, CT 06520, USA

Corresponding Author: Alexej Abyzov

Contribution:

VS: Writing—Original Draft Preparation, Validation, Methodology, Software, Visualization. AJ: Writing—Review and Editing, Resources (Wet lab experiments). TB: Conceptualization, Methodology, AP: Validation. FV: Supervision, Resources. AA: Supervision, Resources, Conceptualization, Methodology.

Adapted from Sarangi, V. et al. SCELLECTOR: ranking amplification bias in single cells using shallow sequencing. BMC Bioinformatics 21, 521 (2020).

2.1 RATIONALE

MDA is an exponential amplification method where the DNA is amplified using a high fidelity phi29 polymerase with proofreading activity under isothermal conditions (30). However, phi29 polymerase is sensitive to template fragmentation happening during cell lysis as well as presence of blocking sites where DNA damage prevents amplification. This may lead to uneven coverage, over-fragmented or completely damaged DNA, which may further lead to allelic imbalance when one of the alleles is under-amplified and the other allele is over-amplified. Even though MDA results in high yield of DNA material, introduction of biases such as allelic imbalance and over representation of C to T mutation introduced during lysis can affect the variant detection downstream.

Before moving forward with high coverage Whole Genome Sequencing (WGS), it is important to select cells with successful amplification, exhibiting little or no biases. Uneven amplification, with the ultimate manifestation of allelic dropouts (i.e., random, and drastic overrepresenting of one allele over the other), challenges separating false positives from real somatic variants. For example, deamination of cytosine happening during cell lysis on one strand of one allele are expected to have 25% allele frequency in a balanced amplification and based on that, can be marked as artifact. However, if the other non-deaminated allele is not amplified, the allele frequency for the artifact will become 50%, making it indistinguishable from a heterozygous variant. So, using a cell with high allele drop-out rate will result in more false positives and reduce sensitivity, as variants in dropped out regions cannot be discovered.

PCR can be used as a first quality control to test the presence of several random genomic loci, usually chosen on different chromosomes, in the amplified DNA. Multiplex-PCR of 4 loci in one PCR reaction can for instance be used as a rapid quality control where cells are considered to have good quality amplification if at least 3 loci are detected(39). However, this test is quite limited as there might be regions outside of the 4 loci with un-uniform amplification. Similarly, failing the test doesn't imply low amplification quality outside of the 4 loci. It is therefore essential to look at the genome as a whole. A few methods for checking amplification quality in silico from WGS data were proposed. Statistical models have been used to detect amplification bias using depth of sequence(40). Amplification quality prior to sequencing has also been determined by using power spectral density to estimate uniformity of amplification which can be otherwise masked by non-unique read mapping, assembly gaps and locus dropouts (both alleles are not amplified)(41), and median absolute difference (MAPD) (42). However, these methods either rely on at least 20X-30X coverage or do not evaluate allelic imbalance, which is important to access to have full coverage of all haplotypes in a cell.

Here, we describe a method to determine the extent of allelic imbalance introduced by MDA into the amplified DNA using shallow (<1X) sequencing coverage. The method is based on considering allele frequency distribution of the heterozygous SNPs, which, for diploid genome, should have a Gaussian distribution centered around 50%. In case of a non-uniform amplification, the distribution of most of the SNPs will support homozygosity, suggesting high rate of allelic dropouts during amplification.

2.2 TESTING AND VALIDATION

Each single cell sequencing experiment can involve hundreds of single cells. After WGA, not all cells are amplified uniformly owing to the allelic imbalance described earlier. Allelic imbalance can be checked from the VAF of heterozygous SNPs (HETs) in the cell. When sequencing in bulk, the VAF distribution of HETs should be centered at 50% and be bell-shaped (Fig. 1-1A). For a balanced single cell amplification, the distribution should follow the same shape, but can have wide dispersion. For an unbalanced amplification the distribution will not be bell-shaped, and one allele will be drastically overrepresented over the other one.

At shallow coverage, most SNPs will either have just a few or no reads supporting them, making assessment of amplification quality impossible (Fig. 1-2). Therefore, the underlying idea of the method is to judge the quality of amplification based on VAF of multiple consecutive HETs from the same haplotype, rather than on individuals HETs. This however requires that HETs are phased to haplotypes. When HETs from the same haplotype are combined, it allows reaching per unit read counts that are comparable to those for individuals HET at high sequencing coverage (Fig. 1-1B). Furthermore, it is important to note that the implicit assumption is that multiple consecutive SNPs are amplified together. For MDA, which is known to have around 50 - 70 kb amplified fragments (30), it is a valid assumption.

Our QC workflow proceeds as follows (Fig. 1-1B). First, we determine HETs from a bulk sample sequenced at high coverage. These SNPs are then phased into maternal and paternal haplotypes using the SHAPEIT2 method (43, 44), which has been shown to be the most accurate method for phasing sets of known genotypes (45). Multiple consecutive HETS are

merged to form a SNP unit. The number of SNPs in the SNP unit is determined by the coverage of the cell. For a high coverage data (~30X) with 100bp reads, we use each heterozygous SNP for calculating VAF across the genome. Proportionally, for coverage of 0.3X with 100 bp reads, the number of SNPs to be used in a SNP unit is 100 (30X divided by 0.30X). The number of SNPs in a SNP unit is inversely proportional to the coverage. The reads supporting SNPs within the SNP unit from only one haplotype are used to calculate the allele frequency over that SNP unit. An allele frequency plot is then generated using all the SNP units similar to how it is done for VAF distribution of individual HETs at high coverage.

The described approach was implemented in a modular pipeline written in python. The pipeline consists of three scripts and each script can be run independent of each other if the user has the required input file (Fig. 1-3). Script-1 takes a VCF file from the bulk sample (the germline SNPs can be called either using sequencing or any other genotyping methods), subsets it into SNPs present in the catalogues of germline variants provided by the 1000 Genomes Project (46), followed by phasing the SNPs using SHAPEIT2, and provides a phased VCF file. Script-2 uses the phased VCF and the low coverage bam file from the single cell to generate allele frequency over all SNPs. Script-2 can be used independently of Script-1, which allows users to use phasing tools other than SHAPEIT2 as long as the input is in VCF format. Script-3 takes the allele frequency of the SNPs from Script-2 and the phased VCF from Script-1 (or user specified phased VCF) to generate the allele frequency plot and ranks cells using only one of the parental haplotypes. The SNP unit is automatically calculated and applied by default using the equation mentioned earlier.

It must be noted that, for coverage lower than 0.3 X, the number of SNP in a SNP unit increases beyond a single MDA amplified fragment and can lead to averaging of multiple amplified fragments. In case of very low coverage, this may lead to a poorly amplified cell being represented as a good cell (Fig. 1-4). For this reason, we also provide an option where the user can override this with their own SNP unit. The result of the final script is a plot showing the distribution of the SNP units allele frequency (Fig. 1-3).

To test our method, we did shallow sequencing on human iPSC-derived single neuronal cell which were amplified using MDA. Cellular DNA was sequenced at various read coverages (0.11 to 0.38) and data were then processed through Scellector. SNP unit size was determined for each cell based on read coverage. Based on the obtained VAF distribution and allelic dropout rate, we ranked single cells as having good, moderate, and bad amplifications. Cells with standard deviation less than 0.26 were considered as uniformly amplified (good) cells and cells with standard deviation between 0.26 and 0.35 were considered moderate cells (Fig. 1-5). Bad cell with standard deviation higher than 0.35 were used as negative control. Out of 14 single cells with shallow sequencing we picked 2 good cells, 5 moderately good cells, 1 bad cell as a negative control and 1 cell (i.e., B01) for which amplification quality could not be determined due to too shallow (0.11X). The selected 8 cells were then re-sequenced at high coverage (at least 30X) using DNBseq platform and their amplification quality was assessed through VAF distribution for individual HETs.

We saw a good concordance between shallow and deep coverage indicating that our method can accurately estimate the effects of non-uniform amplification from shallow

sequencing data (Fig. 1-3A). We noticed that the standard deviation was slightly higher in the deep coverage data. We reasoned that this is because SNP units can span more than one MDA amplified fragments (of typical size of 50-70 kbp), which averages the amplification bias making it seem less to that of high coverage data. Using Spearman correlation, we estimated the concordance between high and low coverage data for the same cells to be 0.92 (Fig. 1-3B). We also found similar high correlations using allelic dropout rate only and additive effects of standard deviation and allelic drop out. Above mentioned cell B01, which was excluded due to low coverage also turned out to be well amplified (Fig. 1-6).

2.3 USAGE GUIDELINES

Bias in amplification may result not only in allelic imbalance but also in non-uniform coverage across genome. We found that the quality of amplification measured using our method correlates with coverage uniformity (Fig. 1-7) and more balanced amplification likely to result in more reliable CNV calls (Fig. 1-8). Furthermore, there is an increase in percent of not covered bases as the standard deviation and allelic dropout rate increases (Fig. 1-9). Additionally, our analysis suggests that our method is more sensitive than pairwise bin comparison approach like MAPD (Fig. 1-10). Finally, allelic imbalance is independent of combination of nucleotide substitution in SNPs (Fig. 1-11). Therefore, we suggest haplotype imbalance as a universal indicator of biased amplification.

Currently, VAF distribution of HETs from bulk is the target that none of single cell amplification methods can achieve. We also note that there exists no clear standard about what is good and what is bad amplification. To address this issue, we take an empirical approach by considering amplification quality of single cell from different independent

studies, including our own, Lodato et al (23) and Sanchez-Luque et al (47) data. From these studies the consensus emerges that standard deviation of ~0.27 with allelic dropouts of less than 10-15 % indicate the best currently achievable amplification (Fig. 1-12). As discussed above, using SNP units large than typical length of amplified fragments leads to averaging amplification bias and we therefore recommend using for QC coverage of ~0.3X or higher. Using these guidelines, we estimated that study of single cell genomes can save a significant amount of funds on sequencing (Fig. 1-13)

2.4 METHODS

Cell samples origin and genome amplification

Single cell DNA used here for validation of Scellector originated from induced pluripotent stem cell -derived human neurons cultivated *in-vitro* and harvested by FACS before conservation at -80°C. Amplification using MDA were obtained through Accusomatic service (SingulOmics), which consisted of a custom cold lysis preliminary step followed by amplification with REPLI-g kit (Qiagen) and DNA purification with AMPure XP-beads kit (Beckman Coulter). To be selected for sequencing, amplification samples were selected based on total yield (above 5µg) and 4-loci PCR test(39).

Bulk DNA sample of induced pluripotent stem cell was used as a reference genome. DNA was purified through DNeasy Blood & Tissue kit (Qiagen) before sequencing at high coverage.

Sequencing

The low coverage sequencing was conducted at Yale Stem Cell Center Genomics Core facility. The library preparation was done using Nextera XT (DNA library kit, Illumina)

and the samples were pooled together to be sequenced on Hiseq4000 (2x100bp) at low coverage per sample (0.1X to 0.4X). For the high coverage sequencing (requested coverage above 30X) of bulk and validated amplified DNA, the library preparation and sequencing (DNBseq) were conducted by the BGI sequencing company (China).

Data analysis

The bulk sample, shallow and high coverage samples were analyzed using the same pipeline. We started with raw fastq files which were aligned to the GRCh37 human reference genome using BWA mem version 0.7.10(48), the bam files were then realigned and recalibrated using GATK 3.6. The germline variant calling for the bulk sample was performed using GATK haplotype caller version 3.6(49). The resulting bam files and vcf file were analyzed using Scellector.

FIGURES AND FIGURE LEGENDS

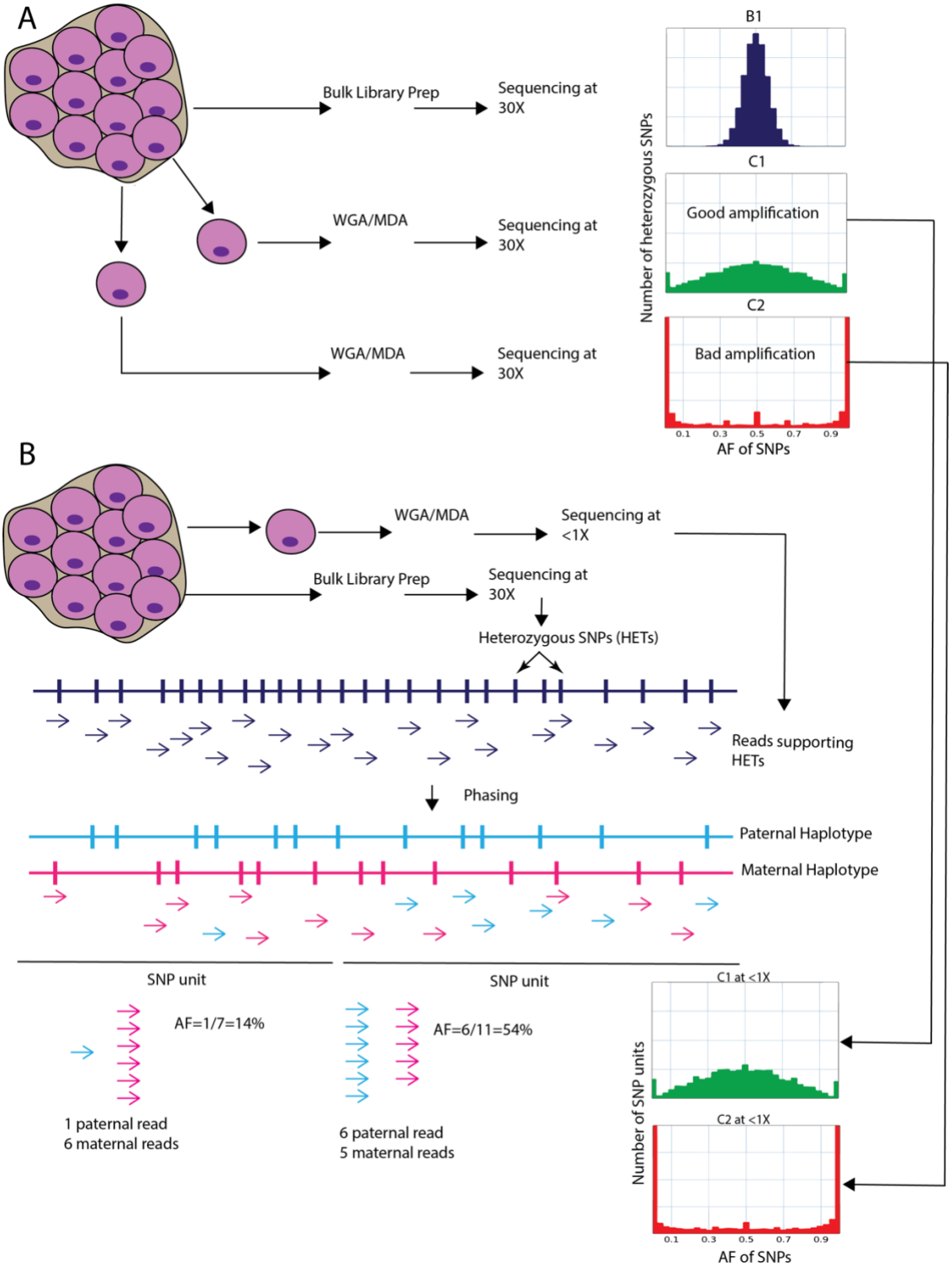


Figure 1-1. Concept and workflow of the approach. A) VAF distribution of HETs at 30X sequencing coverage in three cases: Bulk sample, uniformly amplified cell, and un-uniformly amplified cell. The distribution from bulk shows a peak around 50%, which is expected. Then we have a single cell sequenced at 30X with good amplification. The allele frequency plot still has a peak around 50%, but not as sharp as the bulk sample. The last example is a single cell also sequenced at 30X but with non-uniform amplification. B) Conceptual description of the approach. First, SNPs are phased. The reads supporting the SNPs are divided into two haplotypes, named maternal and paternal, although the exact origin of each haplotype is unknowns. With less than 1 read supporting each SNP (coverage <1X), multiple SNPs are merged to form a SNP unit. Reads supporting SNPs within the SNP unit from only one haplotype are used to calculate the allele frequency over that SNP unit. The allele frequency plot for high coverage data closely resembles the one from shallow coverage data.

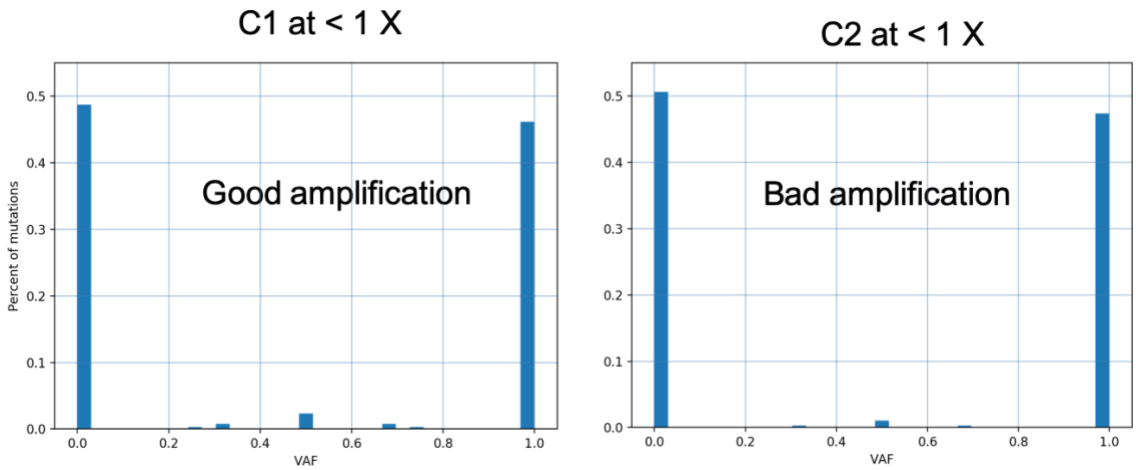


Figure 1-2. Low coverage VAF distribution for cells with good and poor amplification. C1 and C2 are the well amplified and poorly amplified cell respectively from Figure 1. Without using binning and phasing of SNPs, it is impossible to distinguish amplification quality of these cells.

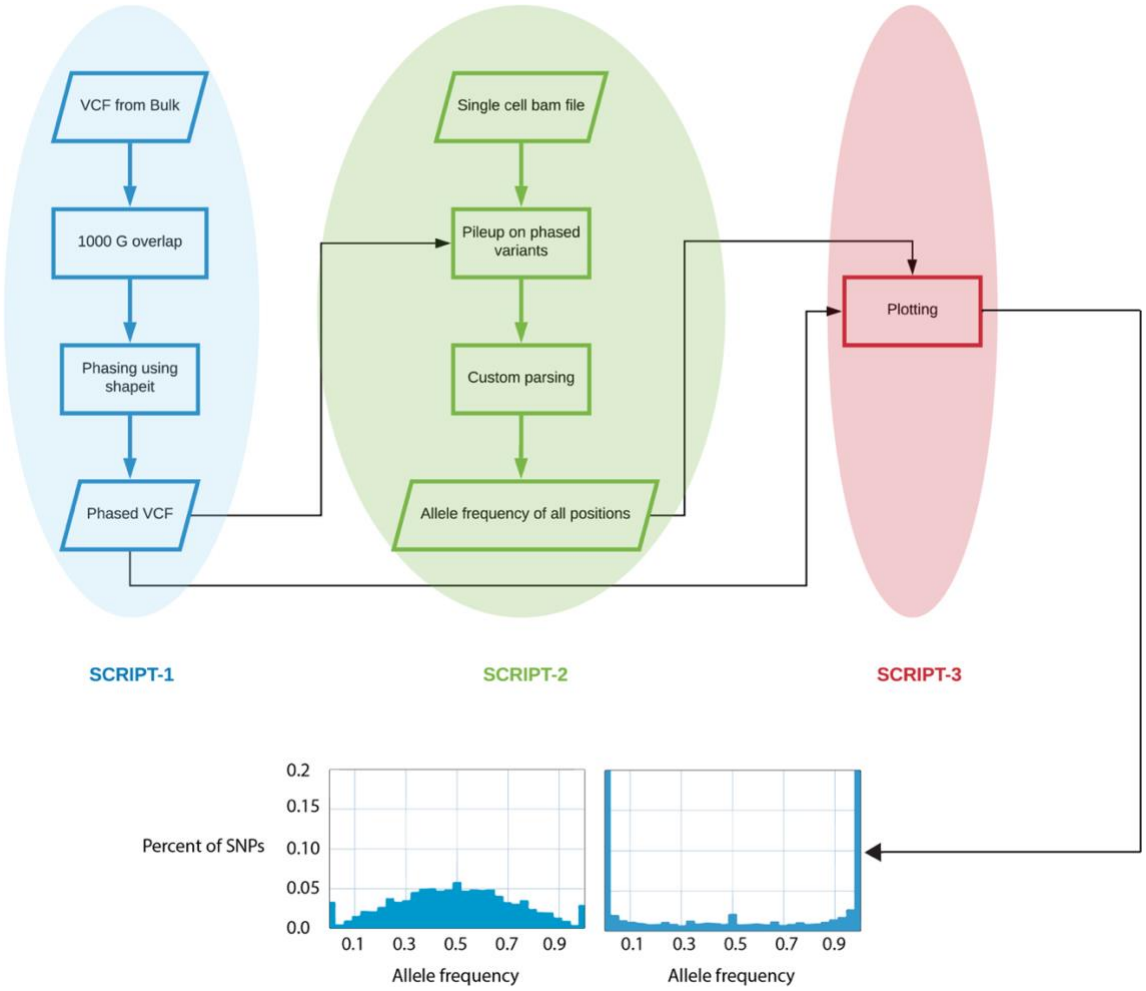


Figure 1-3. Flowchart of method implementation. Script 1 through 3 should be executed in sequence, however, they are independent of each other and as long as the input are correct, user can start with any script. The final Script-3 produces a VAF plot for each sample. Two examples of uniformly (on the left) and un-uniform (on the right) amplified cells are shown.

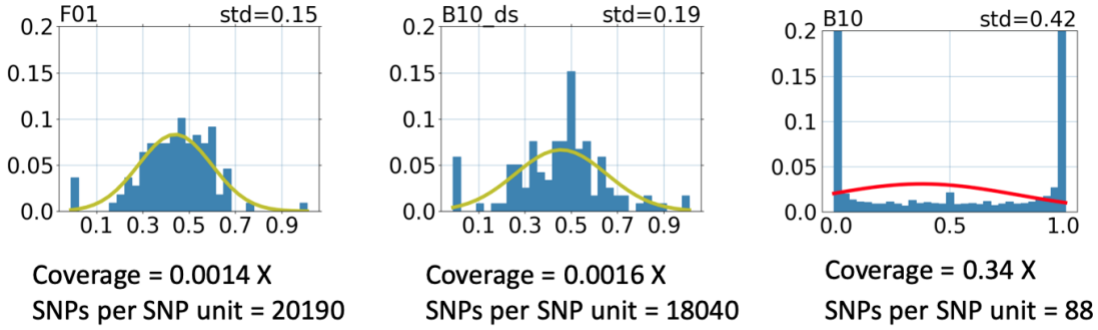


Figure 1-4. Impact of large SNP unit on VAF plots. When the coverage gets lower, the size of the SNP unit gets higher. Cells with very low coverage such as F01, can exhibit properties of uniform amplified cell even though they might be poorly amplified. We took B10 (with bad amplification) and down-sampled it (B10_ds) to match the coverage of F01, to demonstrate that, using very low coverage data is not ideal to determine amplification quality.

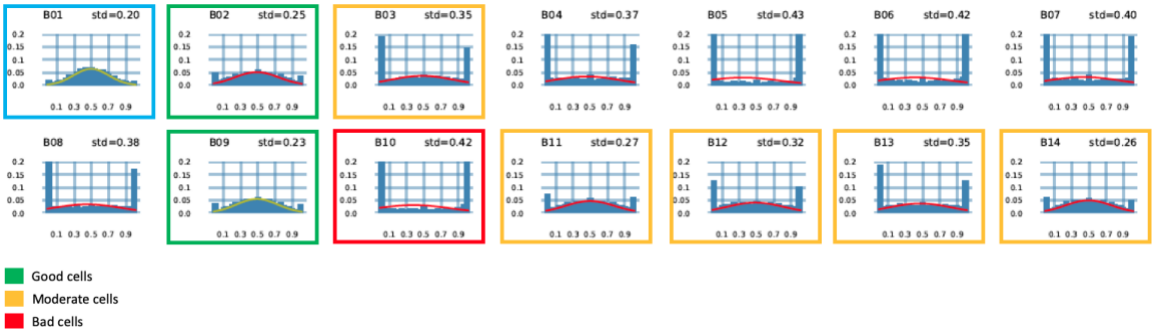


Figure 1-5. Selection of cells for high coverage sequencing. The figure shows VAF distribution of phased SNP units from shallow sequencing for amplified cells. All cells with good (B02 and B09; green squares) and moderate (yellow squares) amplification were selected. One cell with bad amplification (B10; red square) was selected as a negative control. Additionally, we select cell B01 (blue square) with too shallow coverage of 0.11X.

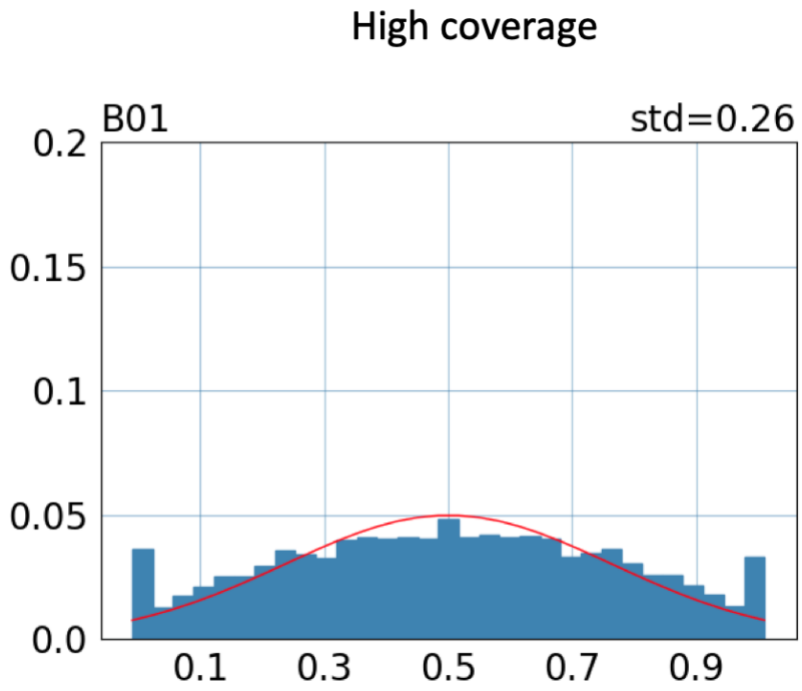


Figure 1-6. Cell B01 has good amplification quality from high coverage.

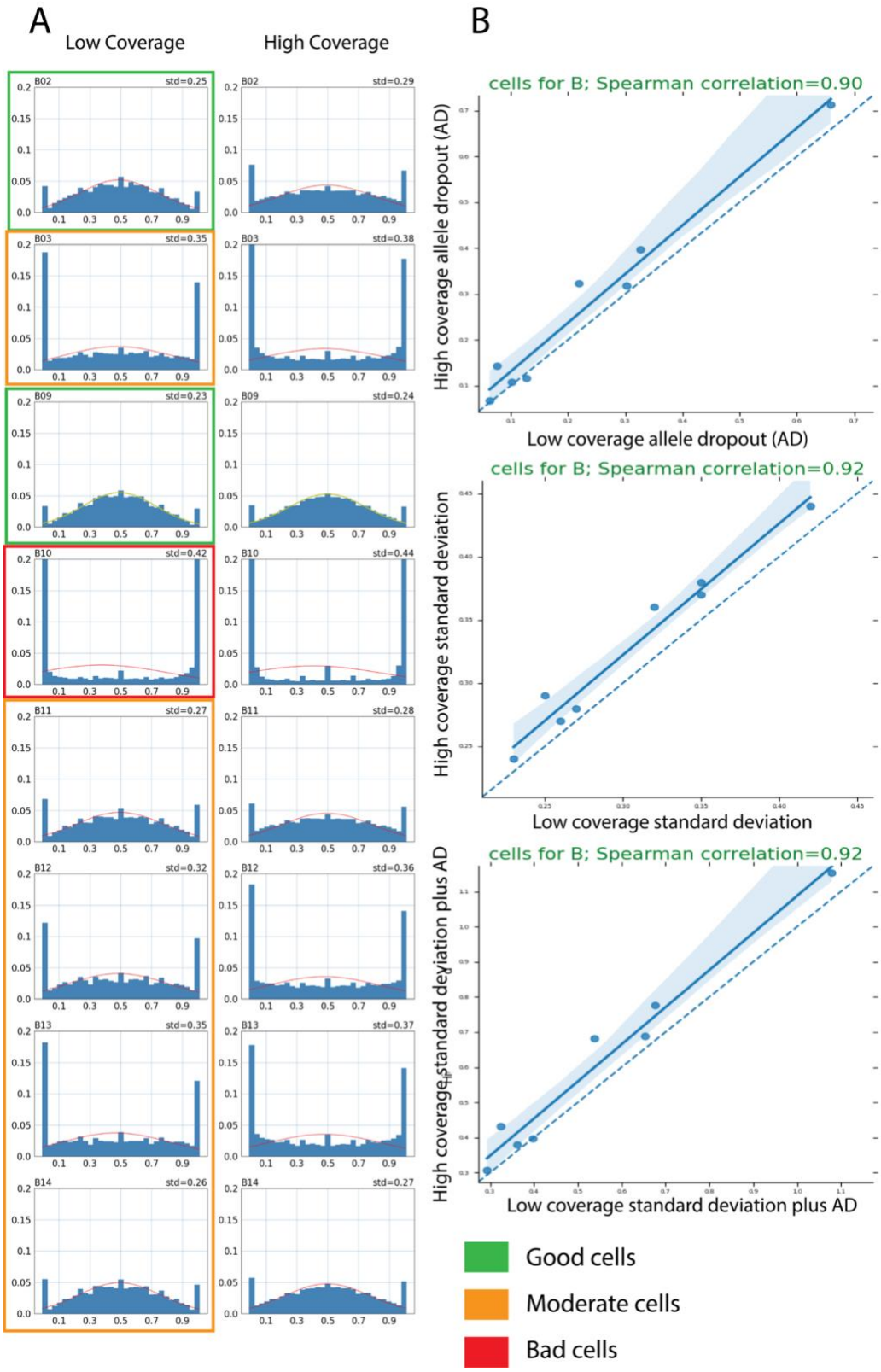

A
Low Coverage
High Coverage
B
B02 std=0.25
B02 std=0.29
B03 std=0.35
B03 std=0.38
B09 std=0.23
B09 std=0.24
B10 std=0.42
B10 std=0.44
B11 std=0.27
B11 std=0.28
B12 std=0.32
B12 std=0.36
B13 std=0.35
B13 std=0.37
B14 std=0.26
B14 std=0.27
cells for B; Spearman correlation=0.90
High coverage allele dropout (AD)
Low coverage allele dropout (AD)
cells for B; Spearman correlation=0.92
High coverage standard deviation
Low coverage standard deviation
cells for B; Spearman correlation=0.92
High coverage standard deviation plus AD
Low coverage standard deviation plus AD
Good cells
Moderate cells
Bad cells

Figure 1-7. Validation of Scellector using 9 cells subjected to shallow sequencing by high coverage sequencing. A) Side by side comparison of the allele frequency plots from shallow coverage and high coverage. B) Scatter plot showing high correlation between the shallow and high coverage. Three comparison using allele dropout, standard deviation and a combination of standard deviation and allele dropout (AD) show similar results.

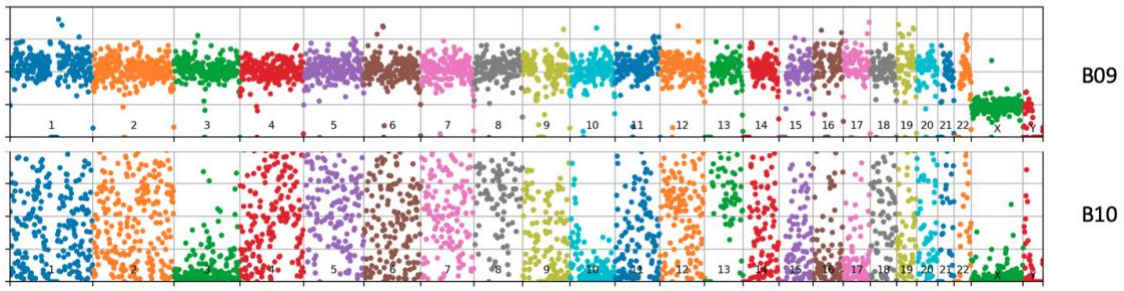
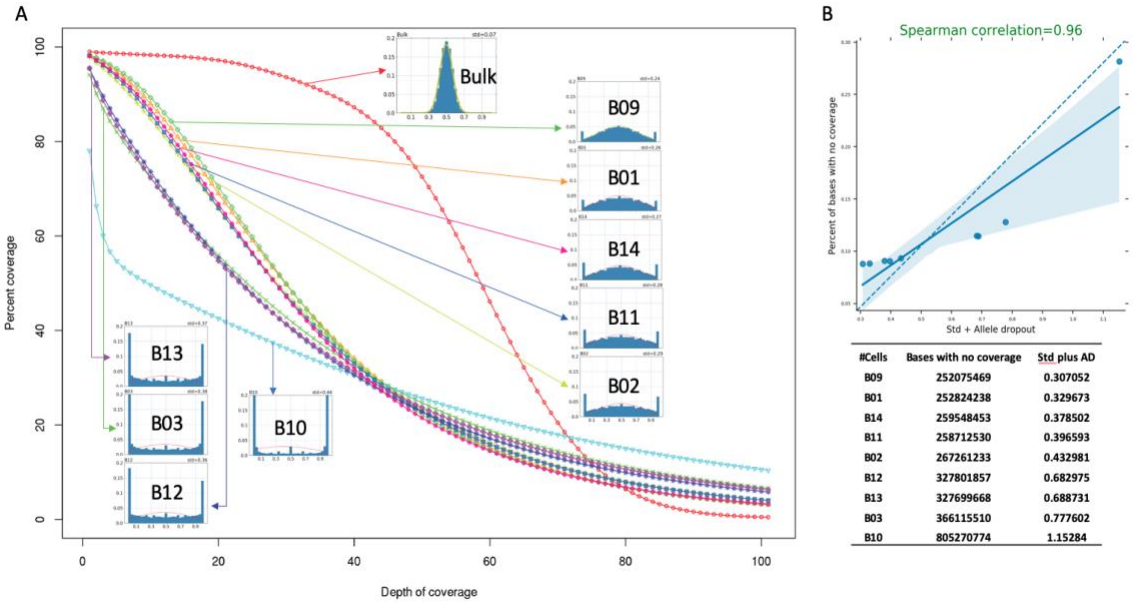


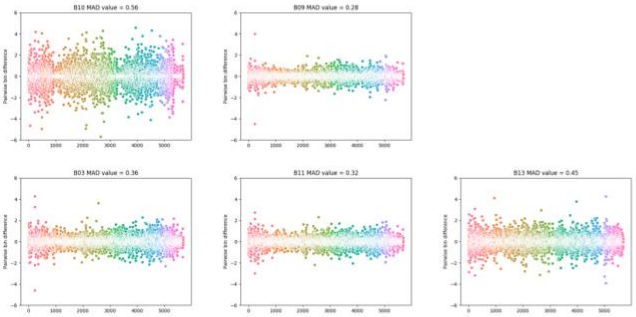
Figure 1-8. Read depth distribution across genome for cells with good (B09) and bad (B10) amplification. B09 has a tight read depth distribution as compared to unevenly amplified B10. Uneven coverage in B10 will likely affect CNV calling by reducing sensitivity and leading to false positive calls.



#Cells	Bases with no coverage	Std plus AD
B09	252075469	0.307052
B01	252824238	0.329673
B14	259548453	0.378502
B11	258712530	0.396593
B02	267261233	0.432981
B12	327801857	0.682975
B13	327699668	0.688731
B03	366115510	0.777602
B10	805270774	1.15284

Figure 1-9. Relation between rate of coverage distribution and variant allele frequency plots for each cell with high coverage data. A) The shape of coverage distribution corelates with amplification quality (standard deviation and allelic dropout rate). B) The combined value of standard deviation and allele dropout rate correlates with number of bases not covered in a cell.

A



B

Cells	Std plus AD	MAPD	Coverage
B09	0.293027	0.28	0.28 X
B02	0.32503	0.31	0.25 X
B14	0.361049	0.31	0.34 X
B11	0.396791	0.32	0.38 X
B12	0.538496	0.42	0.28 X
B13	0.652225	0.45	0.33 X
B03	0.676985	0.36	0.29 X
B10	1.07943	0.56	0.34 X

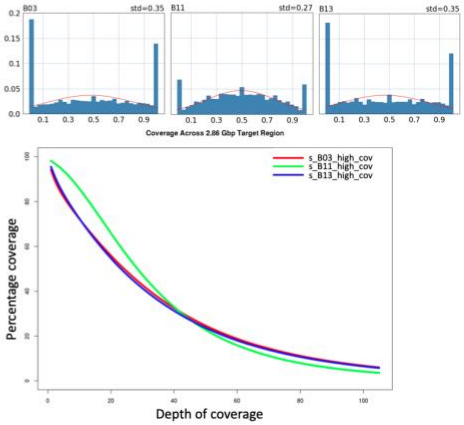


Figure 1-10. Comparing QC approach by Scellector and by MAPD for shallow coverage data. A) MAPD score was able to differentiate between the best amplified cell and the worst amplified cell as we see a wider dispersion of the pairwise bin difference for B10 compared to B09. However, the dispersion of cell B03 and B11 are similar and their MAPD scores are also very similar. B) Scellector unambiguously indicates that amplification quality of cell B11 is much better than that for cell B03 and B13 (highlighted scores in the table). The VAF plot also shows an apparent difference between the two cells. Lastly, Scellector reports B03 to be more similar to B13 (MAPD shows B03 to be similar to B11 than B13 in (A)) and the coverage plot clearly shows the coverage profile of B13 is more similar to B03, indicating our method is more sensitive than MAPD.

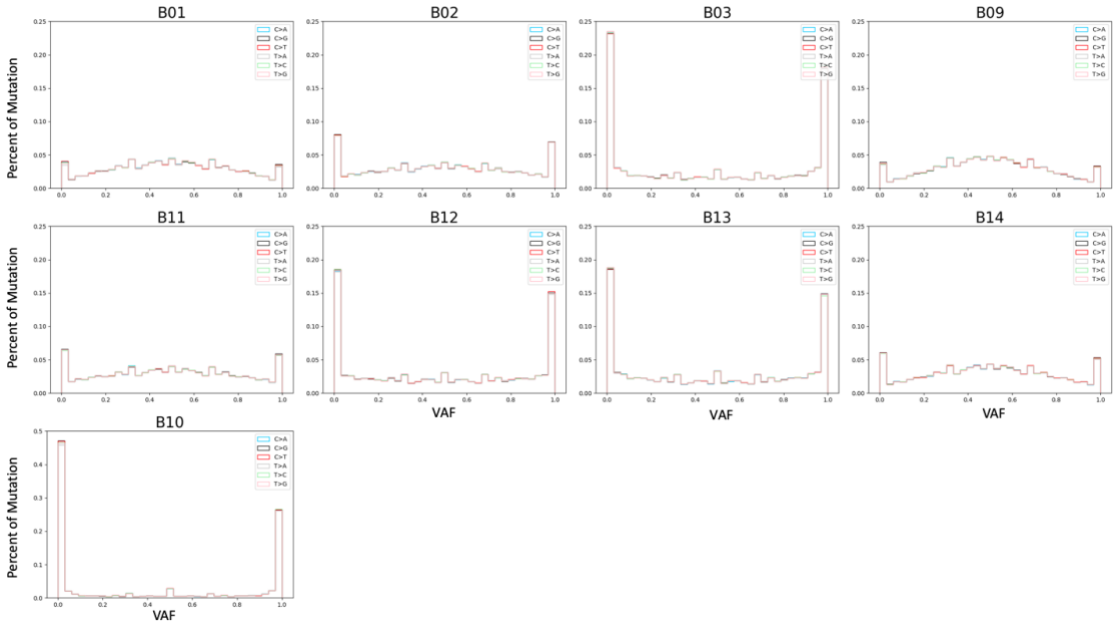


Figure 1-11. Imbalance of amplification is not related to base type. Histogram of variant allele frequency broken down by six mutation types. All mutation types are evenly distributed in the respective variant allele frequency bins.

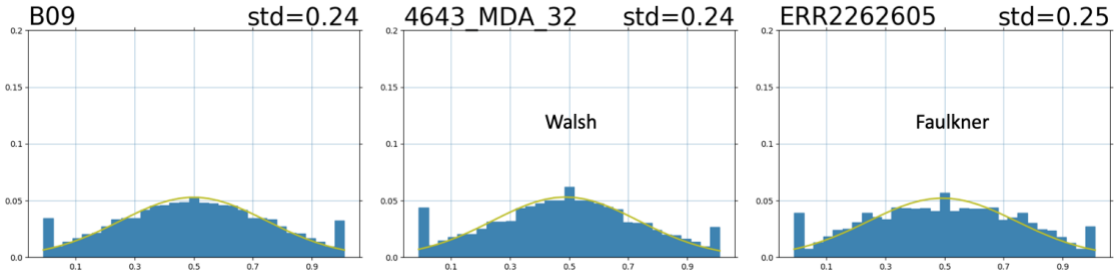


Figure 1-12. VAF plots for cells with best amplification quality across different MDA methods. Comparing our data with data from “Lodata et al, 2015 PMID:2643012” and “Sanchez-Luque et al,2019 PMID:31230816 ” we see the standard deviation and allele drop out for well amplified cells being similar.

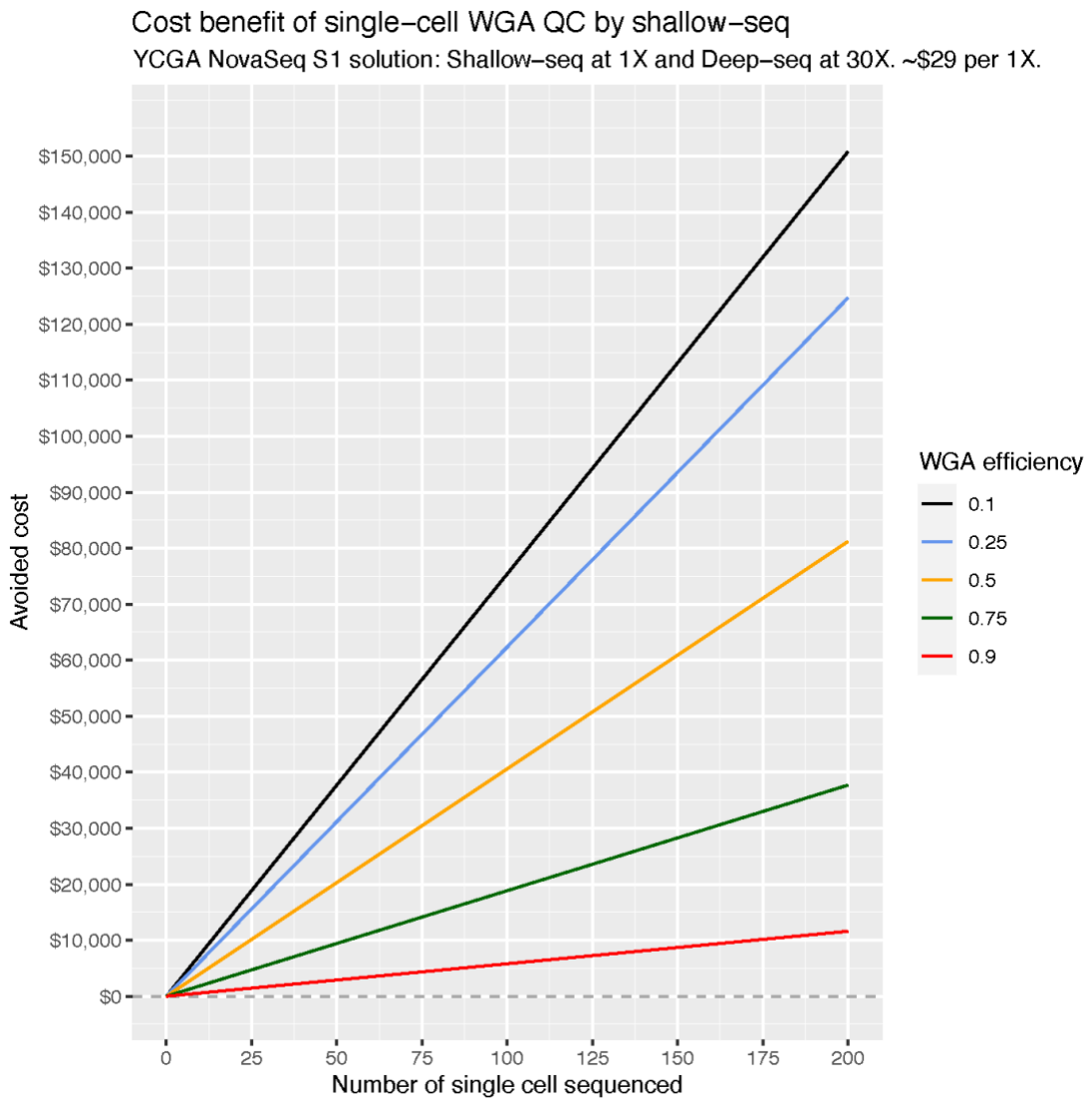


Figure 1-13. Estimated cost saving for amplification QC with shallow sequencing. Here we show the cost savings, if low coverage is being performed at 1X and high coverage at 30X. With WGA efficiency being fraction of well amplified cells, the cost saving increase with decrease in WGA efficiency.

CHAPTER 2 - All^2: A tool for selecting mosaic mutations from comprehensive multi-cell comparisons.

Vivekananda Sarangi[1], Yeongjun Jang[1], Milovan Suvakov[1], Taejeong Bae[1], Liana Fasching[2], Shobana Sekar[1], Livia Tomasini[2], Jessica Mariani[2], Flora M. Vaccarino[2,3], Alexej Abyzov[1]

Affiliations:

1. Department of Quantitative Health Sciences, Center for Individualized Medicine, Mayo Clinic, Rochester, MN 55905, USA
2. Child Study Center, Yale University, New Haven, CT 06520, USA
3. Department of Neuroscience, Yale University, New Haven, CT 06520, USA

Corresponding Author: Alexej Abyzov

Contribution:

VS: Writing—Original Draft Preparation, Validation, Methodology, Software, Visualization. YJ: Validation, Resources. MS: Resources, Visualization. TB: Conceptualization, Methodology. LF: Resources (Wet lab experiments). SS: Methodology. LT: Resources (Wet lab experiments). JM: Resources (Wet lab experiments). FV: Supervision, Resources. AA: Supervision, Resources, Conceptualization, Methodology.

Adapted from Sarangi V. et al. (2022) All^2: A tool for selecting mosaic mutations from comprehensive multi-cell comparisons. PLOS Computational Biology 18(11): e1010703

2.1 RATIONALE:

For detection of mosaic mutations in single cells, the most frequently used approach is to compare single cell genomes to that of a matched reference bulk. While this approach works well to find private mutations in a cell, it misses mutations that are present at higher frequency, and consequently present in multiple cells in the reference bulk. It also requires one to have bulk data which might not be always available. Here we present a tool called All^2 (pronounced 'all square') which detects mosaic mutations without the need for a reference bulk by relying on comprehensive cell-to-cell comparisons. By consolidating information from all comparisons, every call is categorized as either a germline variant, mosaic mutation or noise/false positive.

2.2 CONCEPT

All^2 is an easy-to-use tool which extends and implements an algorithm initially proposed in Bae et al. 2018 (50). All^2 takes mutation calls from all pair-wise comparisons of N cells in the study and, for every non-redundant call, creates a NxN pairwise binary matrix corresponding to comparing different pairs of cells, where 1 corresponds to a call and 0 to no call. Patterns of values in the matrix are used to determine whether a call is a mosaic mutation, germline variant or false positive (Fig. 2-1A-D). In theory, the presence of these patterns should be sufficient to make the determination, however, real data has noise, smearing the patterns (Fig. 2-2).

For effective categorization, we developed a scoring system which reflects how likely it is for a call to be a mosaic mutation or a germline variant. The tool calculates two scores: a germline score and a mosaic score, each within a range between 0 and 1. A real mosaic

variant could only be discovered when comparing a cell carrying the variant and a cell not carrying the same variant. The number of times a call for a variant shows up in the matrix is determined by

$n = f(1 - f)N^2$,

where n is the number of times a variant is seen in all comparison, f is fraction of cells carrying the variant, N is the total number of cells. By solving the above quadratic equation, we get two solutions:

$f_m \approx 0.5 - \sqrt{0.25 - n/N^2}$ and $f_g \approx 0.5 + \sqrt{0.25 - n/N^2}$

Since the mosaic mutations are typically present in a small fraction of cells in the bulk, and germline variants are present in (almost) all the cells, we conditionally call f_m as frequency of a mosaic mutation and f_g as frequency of a germline variant. Note, a germline variant can be lost or undetected in some cells, and that is why its cell frequency in a bulk may be below 1.

Since $f_m = 1 - f_g$, we can just use one equation, such as $f = f_m$, where f is the fraction of cells with mosaic mutation or the fraction of cells without germline variant. Now, we can calculate the number of cells N' carrying the mosaic mutation or the number of cells not carrying germline variant as $N' \approx fN$. In case of a true mosaic mutation, the corresponding calls are arranged in rows in the matrix (Fig. 2-1B), and would sum up to

$$n_m = \sum_{i=1}^{N'} nr_i,$$

where nr_i is the number of calls for the variant in a row corresponding to the ith cell. From the data, the best estimate of n_m is the maximum from all possible subset of N' cells from N (Fig. 2-3 and 2-4). Similarly, for a germline variant, corresponding calls are arranged in columns (Fig. 2-1C), and would sum up to

$$n_g = \sum_{i=1}^{N'} nc_i,$$

where nc_i is the number of calls for the variant in a column corresponding to ith cell. And best estimate of n_g is the maximum from all possible subset of N' cells from N (Fig. 2-3 and 2-4). The mosaic and germline scores are then defined as

$$Mosaic\ Score = \max\,(n_m)/n$$

$$Germline\ Score = \max\,(n_{\mathrm{g}})/n$$

A call having a high mosaic score and low germline score is defined as a mosaic mutation. Similarly, a call with high germline score and low mosaic score is defined as a germline score. When a call has both high germline and mosaic scores, we define it as a high

frequency mosaic mutation. Such mutations are likely present at a higher fraction of cells in a tissue. For example, such mutations could occur during early development and be present in a high fraction of cells across tissues in the human body (51). The distribution of mutations (as dots) on a plane with axes corresponding to the two scores can be used to divide the calls into mosaic mutations, germline variants, noise or false positive (low mosaic and low germline score) and high frequency mosaic mutations (high mosaic and high germline score) (Fig. 2-1E).

2.3 IMPLEMENTATION AND USAGE

Genomes of all pairs of cells need to be compared prior to using All^2. Variant calls can be made by a caller of choice (see Methods). All^2 is written in python and has three commands: 'score', 'call', and 'matrix'. The first command takes a manifest file with names of single cells, along with the VCF file containing calls (SNVs and INDELs) as rows. Case and control fields in the manifest file are used to define the directions of pairwise comparisons, where the case is compared to control. A user can optionally provide a BED file with the inclusion list of genomic regions where to apply the filtering. The output of this command is a file with mosaic and germline scores for each of the calls as well as density scatter plot of the scores showing distribution of calls based on their scores (Fig. 2-5A). The second command relies on the output of the 'score' command and annotates the calls as mosaic mutation, germline variant, noise, or high frequency mosaic mutations based on default or user specified score cut-offs. This command also annotates the density scatter plot (Fig. 2-1E), provides a file with annotated calls for each cell, per category plots of call counts, per sample plots of call counts, VAF (variant allele frequency)

plot, and mutation spectrum plots (Fig. 2-5B-D). The third command plots a matrix of pairwise comparison for one or multiple calls. The plots also display calculated scores along with VAF for the call(s) in each cell (Fig. 2-6). Analogous to SNVs and indels, All^2 is capable of filtering structural variant (SV) calls using commands 'score_sv', 'call_sv' and 'matrix_sv'. Two SV calls are considered the same if they have at least 50% reciprocal overlap. For this purpose, the tool supports VCF file as input, e.g., VCFs generated by the SV caller MANTA (52, 53).

One implicit underlying assumption of the approach is that in each compared cell, the genome is covered/sequenced uniformly. This is true in case of the single cell cloning approach, however, single cell genome amplification may result in non-uniform coverage which, at the extreme, manifests in allelic dropouts (54). To handle this, we have implemented a dedicated allele dropout analysis (ADA) mode, which considers allele dropout regions when calculating the scores, thereby reducing the noise. The ADA mode can also be used for running All^2 on exome data where the exome capture region can be specified per cell in the manifest file. More details of the command and description of the results can be found on the dedicated GitHub page https://github.com/abyzovlab/All2.

2.4 APPLICATION TO RECONSTRUCTING CELL LINEAGE TREE

To demonstrate the uniqueness of All^2 approach, we applied it to reconstruct post-zygotic cell divisions in a living individual. Analysis of developmental cell lineages is one of the central questions in developmental biology, resolving which can shed light on the etiology of developmental diseases. Unlike model organisms, lineage tracing in humans can only be done retrospectively using naturally occurring somatic variants that serve as permanent

marks of the lineages. Mutations that occur during early development are present in a high fraction of cells across tissues in the human body, and their discovery is challenging for existing methods.

In the analyzed individual, we compared mosaic variant discovery using three approaches: 1) by analysis of bulk blood and saliva; 2) by pairwise comparison of 25 clonal iPSC lines (representing 25 fibroblast single cells) with the bulk blood; and 3) by comparing the clonal lines followed by application of All^2. To reconstruct the lineage tree, we selected mosaic variants shared by clones or by multiple bulk tissues (51) (Fig. 2-7). Analysis of bulks alone allowed discovering only high frequency mutations but not all. For example, mutations a, b, and c defining branches of the first zygotic cleavage (Fig. 2-7B) could not be discovered because of resembling germline variants by frequency of occurrence in the bulks (i.e., in 80% to 90% of cells). Pairwise comparisons between clones and bulk tissues are powered to find mutations present in the analyzed cell and at low frequency (typically <1% VAF) in bulks but missed high frequency mutations. Remarkably, the All^2 approach was able to call both high and low frequency mutations resulting in the most complete lineage tree – a tree that cannot be reconstructed even if we combine comparisons of clones relative to bulk tissues and analysis of bulks.

In this comparison we utilized data from bulk blood and saliva as these samples are easier and cost-effective, as compared to bulks of fibroblasts, to collect for an individual. Also, blood and saliva are made up of multiple early developmental cell lineages (51), while fibroblasts from a biopsy can be dominated by just a few lineages (55). So, samples from fibroblasts can only have a fraction of early mosaic mutations. Our additional analysis of

65 SNVs in the tree using high coverage data suggests that using any of the bulk samples (saliva, blood, or fibroblasts) does not outperform ALL^2 for lineage tree construction (Table 2-1).

The advantage of calling mosaic mutations in bulk is that it allows discovering mutations with intermediate VAF (between 1% to 10%), which were not sampled by the 25 analyzed clones and consequently, not discovered by All^2. Increasing the number of analyzed clones will likely increase the overlap in discovered mutations between those two approaches but would also increase experimental cost. Thus, this observation suggests complementarity in analyzing clones/single cells and bulks for lineage reconstruction.

2.5 ALLELE DROPOUT MODE FOR WHOLE GENOME AMPLIFIED SINGLE CELLS

Using clones as gold standard, we applied All^2 in Allele Dropout Analysis (ADA) mode (see Methods) to MDA amplified single cells, to demonstrate the effectiveness of this mode to filter out spurious calls originating from biases in the amplification process. MDA uses Phi29 polymerase under isothermal condition, which results in an exponential DNA amplification. The exponential amplification leads to uneven coverage and over representation of one allele over the other (allele imbalance). In extreme, a locus can have only one allele amplified and germline variants on the unamplified locus will be lost. ADA mode is designed to address this issue. In ADA mode, All^2 takes a list of genomic regions (in BED format) where no allele dropout is observed (see Methods). Using this, for each call, All^2 excludes from the score calculation those cells where a call is not made, and the surrounding region has allele dropout. This exclusion may change the number of

considered cells and pairwise comparisons, which eventually affects the mosaic and germline scores.

We called mosaic mutations in 11 clones (representing 11 brain progenitor cells) derived from a human fetal brain (specimen 316) (50), as well as in an MDA amplified single cell taken from one of the clones. Just adding the single MDA-amplified cell into the analysis, more than doubled the count of high frequency mosaic mutation (Fig. 2-8A, B). Next, we applied a single cell specialized caller SCOUT (56). We observed that it reduced the effect of MDA amplification artifacts; 47% reduction in high frequency calls and 33% reduction in call in the single cell (Fig. 2-8C). Further application of the ADA mode resulted in calling all but one (28 vs 29) high frequency mosaic calls and further 11% reduction in mosaic calls in the single cell (Fig. 2-8D). Additionally, there was a 69% reduction in the germline variants after applying ADA. These variants, falsely called as homozygous reference due to allele dropout in the single cell, are effectively filtered by the ADA mode. Mutation counts per clone (Fig. 2-8D) were also similar to those found when analyzing only clones (Fig. 2-8A). This comparison yields evidence that even though the number of mutations called in the single cell is high, by applying ADA mode, we were able to reduce the number of likely false calls introduced by single cell amplification by half, without compromising the mutation calls from clones not affected by allele dropout. ADA mode is also effective in reducing false positive calls when using data for cells with low (i.e., highly non-uniform) amplification quality, however, that can come with the tradeoff of filtering likely true mosaic mutations (Fig. 2-9).

2.6 RUNTIME

Runtime depends on the number of cells in the study and the variant caller used (since some variant callers will output higher number of calls than others). For the first example with 25 iPSC lines (Fig. 2-7B), application of All^2 using 8 GB memory on a 2.4 GHz dual-core processor took less than 15 minutes for the 'score' module and less than 10 minutes for the 'call' module to compute mutation annotation and plot the mutation count and VAF plots. For the second example with 11 clones and one single cell (Fig. 2-8), application of All^2 in ADA mode took less than 90 minutes to complete the 'score' module and less than 20 minutes for the 'call' module. In this case, the runtime is longer because of the longer list of variant candidates from the single cell.

2.7 METHODS

iPSC line generation

The iPSC lines were derived from fibroblasts using the Epi5 Episomal iPSC Reprogramming Kit (Invitrogen catalog A15960) delivering the five reprogramming factors Oct4, Sox2, Klf4, L-Myc, and Lin28. The iPSC lines were propagated using mTeSR1 media (Stem Cell Technologies) on 1X Matrigel-coated dishes (Matrigel). Genomic DNA was extracted at passage six, using QIamp DNA Minikit (Qiagen) following the manufacturer instructions.

Saliva collection and DNA extraction

Saliva DNA was collected and purified using the Oragene-Discover kit (DNA Genotek) following the manufacturer instructions. Saliva DNA was extracted using DNeasy Blood and Tissue kit (Qiagen) with the following modifications: 5 ml AL-buffer and 200 µl Proteinase K were added to saliva and incubated at 56°C for 30 minutes. RNA was digested

using 20µl RNAse A (Qiagen) for 5 minutes and DNA was extracted using 4 extraction columns in parallel to optimize the yield.

Blood collection and DNA extraction

10-15 ml of blood was collected using BD Vacutainer ACD tubes. DNA was extracted using the Gentra Puregene Blood Kit (Qiagen) following standard manufacturer protocols.

Whole genome sequencing (WGS)

DNA extracted from iPSC lines were sequenced at 30X, while DNA extracted from saliva and blood was sequenced at 200X. All sequencing was conducted at BGI using with 2x100 bp paired reads. The sequencing library preparation was PCR-free.

Fetal brain tissue and MDA

Collection of fetal brain tissues for subject 316, derivation of clonal neurosphere lines and sequencing has been previously described (50). Single cells from a clonal neurosphere line were manually picked using a micropipette under an inverted microscope. Whole genome amplification was performed by multiple displacement amplification (MDA) using the REPLI-g Single Cell Kit (QIAGEN) following the manufacturer recommendations. Genomic DNA was extracted using the DNeasy Blood & Tissue Kit (QIAGEN). Multiplex PCR for four arbitrary loci from different chromosomes was used to exclude single cells if less than four loci were amplified (39). Five out of eight single cells (62.5%) passed the 4-locus multiplex PCR quality control and were selected for sequencing. Illumina Truseq DNA PCR-free libraries were prepared for the five cells and sequenced on a HiSeq X (2X150 bp) at 30X coverage.

Allele dropout analysis mode

We started with raw FASTQ files which were aligned to the GRCh37 human reference genome using BWA mem version 0.7.10 (48) , the BAM files were then realigned and recalibrated using GATK 3.6 (49). The clones and the single cells were compared to each other using Mutect2 (35), Strelka2 (36) and SCOUT (56). For the clones, mutations called by both Mutect2 (35) and Strelka2 (36) with depth of 10 or more reads as well as a PASS value by both callers were used as input to All^2. For the single cells, mutations called by Mutect2 (35), Strelka2 (36) and SCOUT (56) with a depth of greater than 10 reads and a PASS value from all callers were used. All^2 was run four times with four different settings as depicted in Fig. 2-7. Post All^2, only mutations which had an allele frequency of 35% or more were considered, to further filter noise introduced during clone amplification, library preparation and sequencing. The allele dropout regions for single cells were calculated using CNVpytor(57), where the entire genome was divided into 5000 base pair bins. For each bin, a likelihood score was calculated using allele frequency of SNPs within the bin. Bins were marked as allele dropout if they satisfied both of the following conditions: i) at least one heterozygous SNP in the bin had VAF smaller than 0.01 or larger than 0.99 ii) the maximum likelihood VAF within the bin deviated from 0.5 by more than 0.1. Additionally, we marked as dropout neighboring bins of the bin satisfying the above conditions. For each mutation call, only cells with no call (or calls with VAF < 50%) and marked as having a drop out at the corresponding locus were excluded from calculation of score by All^2 (Fig. 2-10). Single cell QC on the MDA amplified cells was performed using Scellector (54) and only one cell (cell5) passed QC (Fig. 2-1).

Mutation calling for lineage analyses

The FASTQ files were processed the same way as the clones above. Calls were made using an allele frequency cut-off of 35% to remove mutations introduced during culturing clones. Additionally, only INDELs shorter than 10bp (most confident calls) were used. Pairwise comparison between bulk data and the clones were done using consensus calls between Mutect2 and Strelka2. Mutations with a depth greater than 10 reads, at least 2 alternate supporting reads, and PASS values from both callers were used. For the allele frequency plots (Fig. 2-7A&C), all mutations from All[2], bulk, and pairwise comparison were used. For details, including calling mosaic mutation from bulk tissue and lineage tree construction, please refer to the method section of Fasching et al (51).

FIGURES AND FIGURE LEGENDS

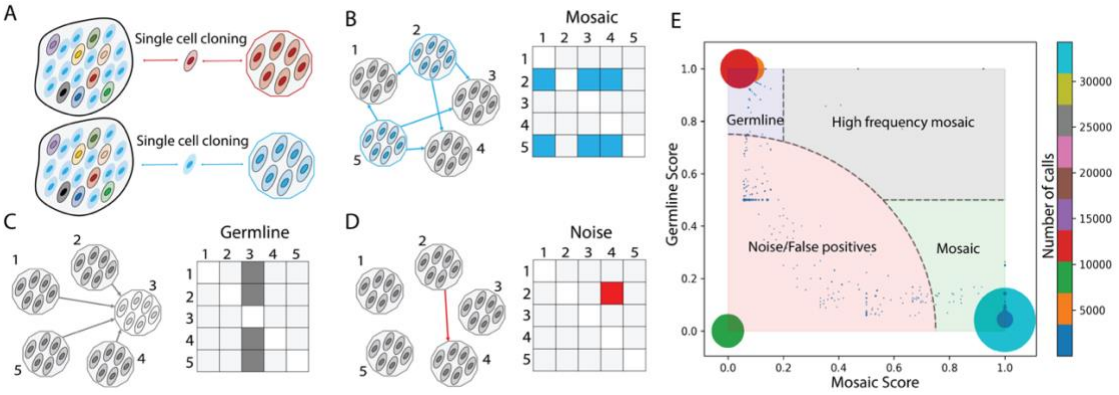


Fig 2-1. Conceptual overview of All[2] approach and scoring. (A) A tissue/sample is made up of different cells (ovals) carrying various mosaic mutations (reflected by different colors). Post single cell clonal expansion, rare mosaic mutations (in red) can be easily detected by comparing the clone to the bulk tissue. However, frequent mutations (in blue) will be missed by this approach. (B-D) Each mutation in clone-to-clone (which is cell-to-cell) comparison can be represented by a NxN matrix of pairwise clone comparisons, where each box represents the call between a clone in the row versus a clone in the column. (B) In case of a true mosaic mutation, the calls are arranged as rows in the matrix. The pattern in the matrix shows that the mutation is called in clone 2 and clone 5 when comparing them to other clones. (C) In case of a germline variant, the calls are arranged in a column(s) in the matrix. The displayed pattern suggests that the mutation is present in all clones except clone 3. (D) The pattern has a sporadic distribution of calls in the pairwise matrix and does not suggest either mosaic mutations or germline variants. Such call is deemed as a false positive or noise. (E) Distribution of mosaic and germline scores for calls (the size of the

dot/circle corresponds to the number of calls with the same scores; the color represents number of calls depicted in the colorbar. The plot can be divided into four areas: mosaic mutations (light green area, where the mutations have high mosaic scores and low germline scores), germline variants (light blue area, where the mutations have high germline and low mosiac scores), high frequency mosaic mutations (light gray area, where calls have both high mosaic and high germline scores) and, lastly, noise or false positive calls (light red area).

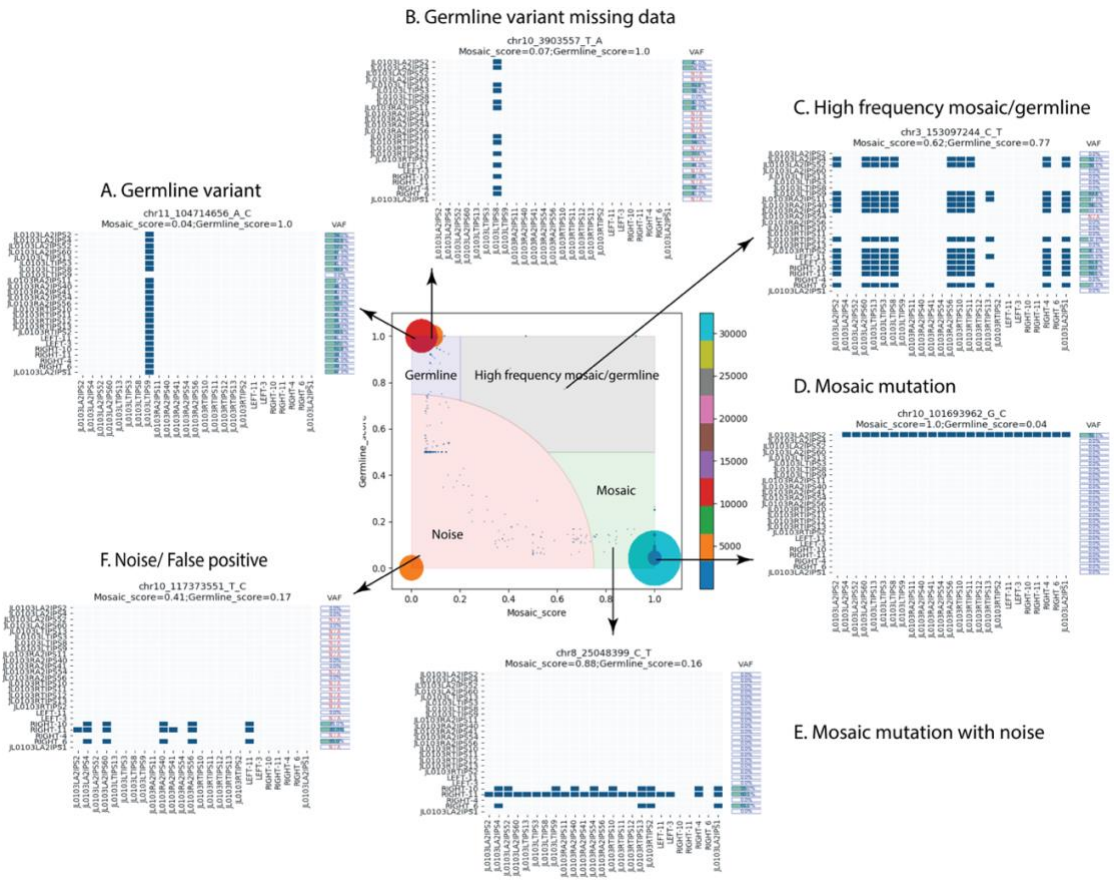


Fig 2-2. Real data introduces noise/missing data that masks mutation type pattern. At the center is an example distribution of mosaic and germline scores for generated calls. Arrows show NxN pairwise binary matrices for select calls. A) An ideal matrix for a germline variant. Calls for the variant are made in all but one cell. B) A matrix for a germline variant that is similar to example in A, however there are multiple cells for which call was not made. C) This call exhibits pattern that is a "combination" of patterns for germline variant and mosaic mutations. This call is thus a mosaic mutation that is present in high fraction cells and is missed when comparing cells to bulk. D) An ideal matrix for a mosaic mutation showing that the mutation is present in only one cell. E) An example of a matrix for a

mosaic mutation present in 3 cells. One cell exhibits an ideal pattern (i.e., a call is made when comparing to each other cell), but other two cells have missing calls. F) A noisy call with pattern not matching to either germline variant or mosaic mutation.

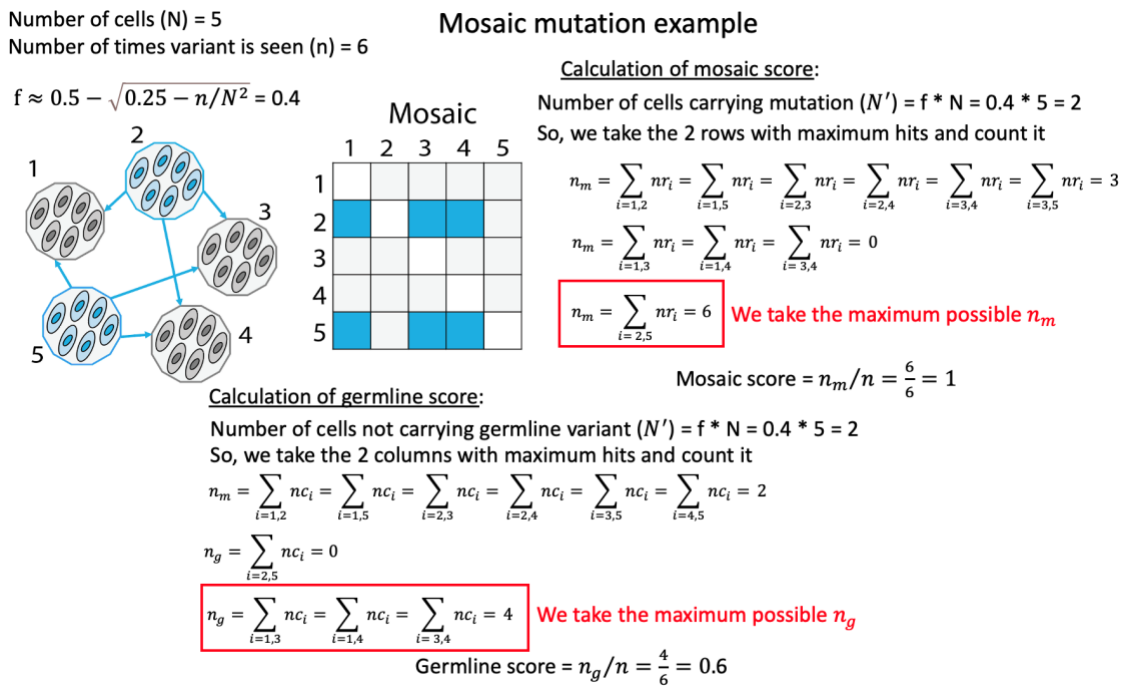


Fig 2-3. Example of calculating mosaic and germline scores for a mosaic variant.

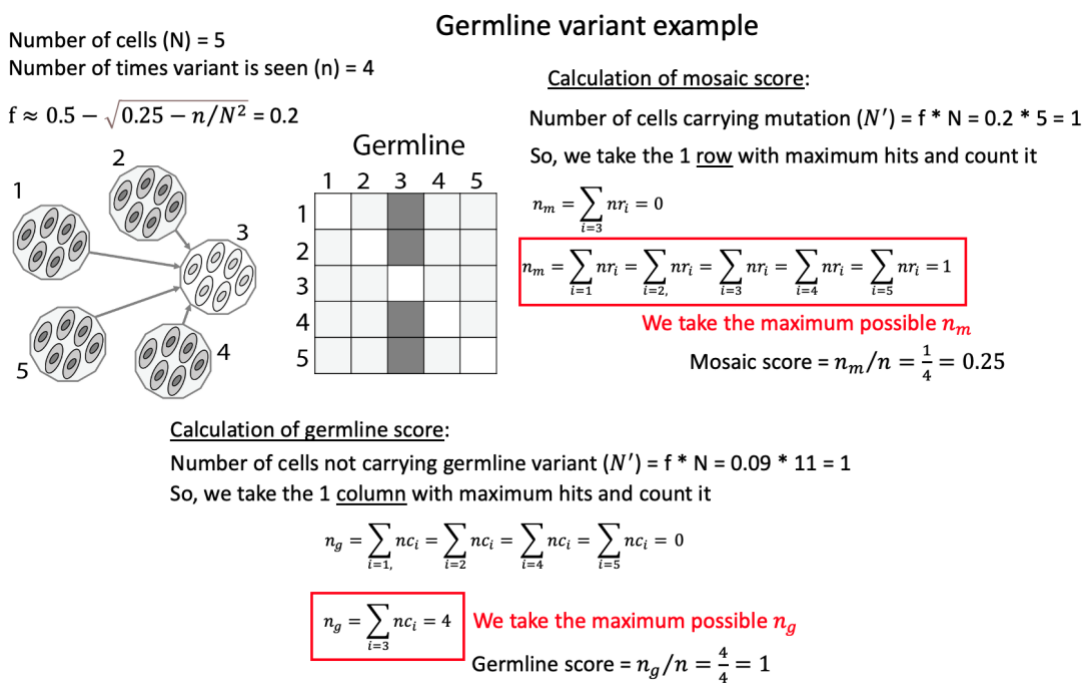


Fig 2-4. Example of calculating mosaic and germline scores for a germline variant.

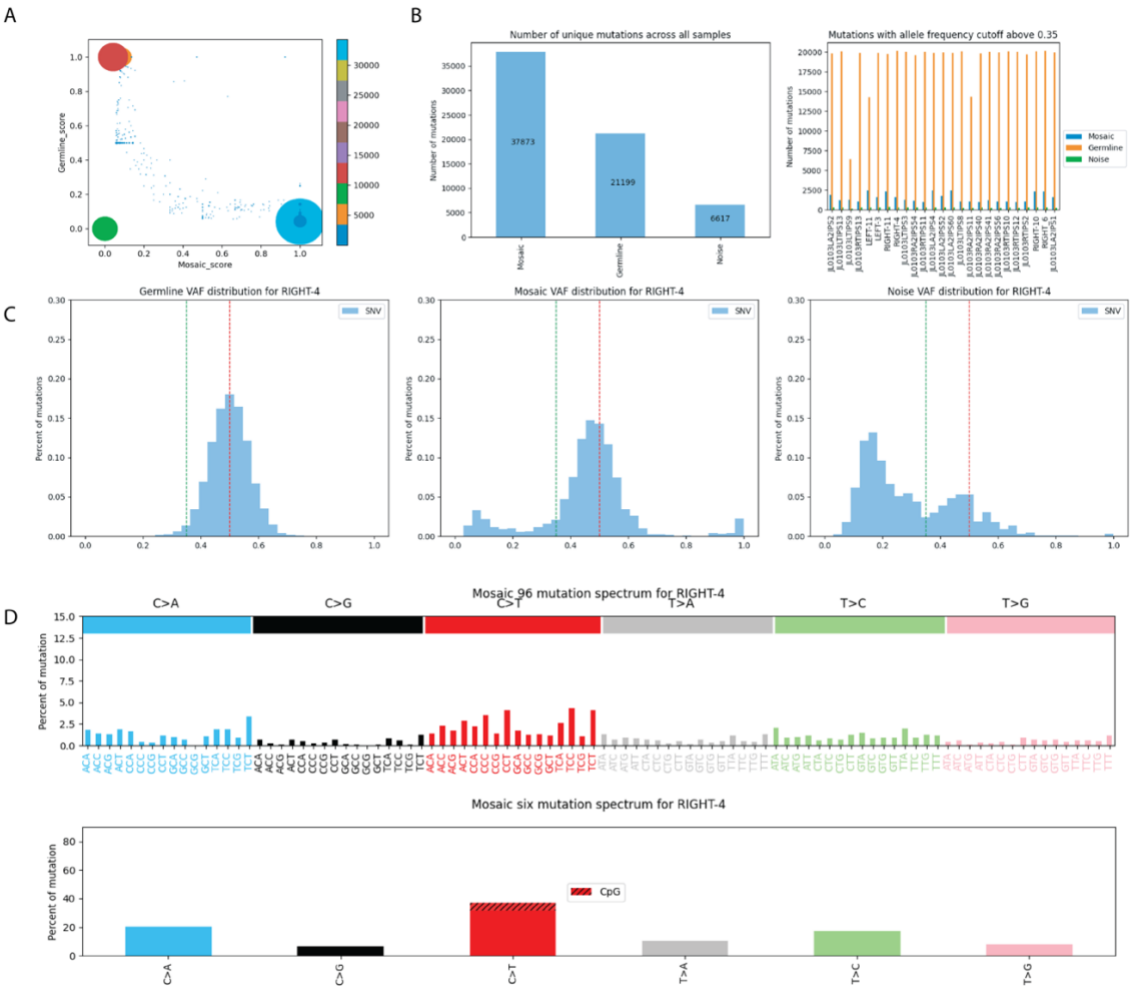


Fig 2-5. Plots generated by All[2] 'call' command. A) Density scatter plot showing the distribution of calls by germline and mosaic scores. This gives the user an idea on what cut-off they may want to use when running the 'call' command. B) Counts of call by classified type (i.e., germline variant, mosaic mutation, or false positive/noise). Calls in the summary plot (left) are redundant, e.g., the same germline variant is counted as many times as it was called in various comparisons. Call in the per cell plot (right) are by definition are non-redundant. C) Example of variant allele frequency (VAFs) for calls classified by types:

germline variants, mosaic mutations and false positives/noise. D) Mutation spectra plots for mosaic mutations.

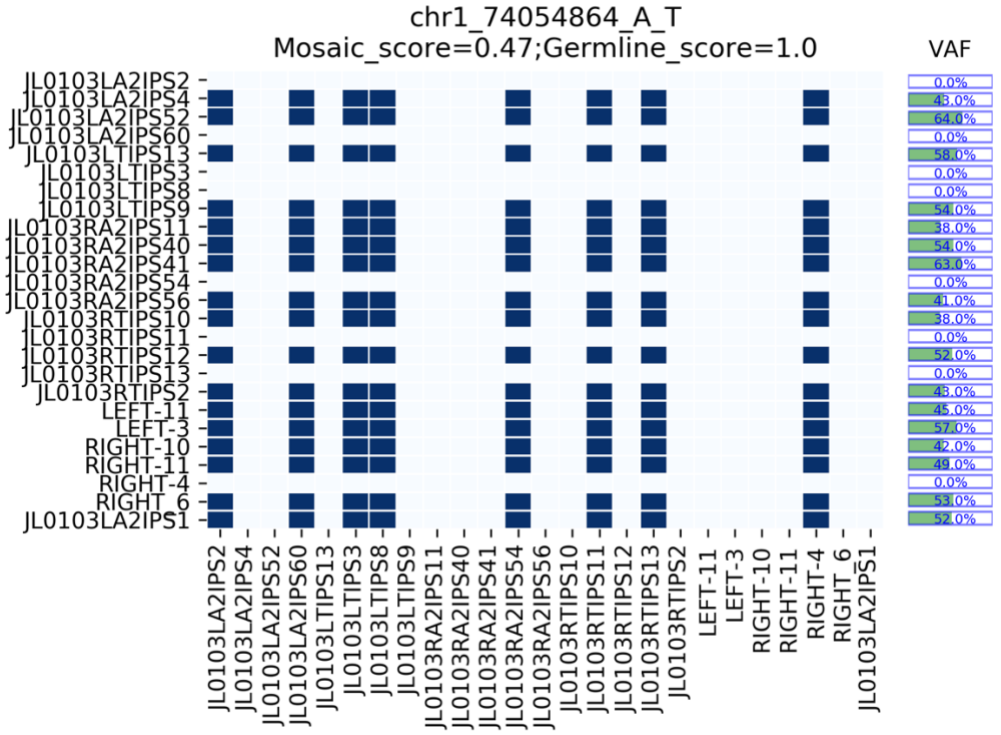


Fig 2-6. NxN pairwise binary matrices for an exemplar call. The plot is annotated with germline and mosaic scores and with variant allele frequencies of the call in each cell.

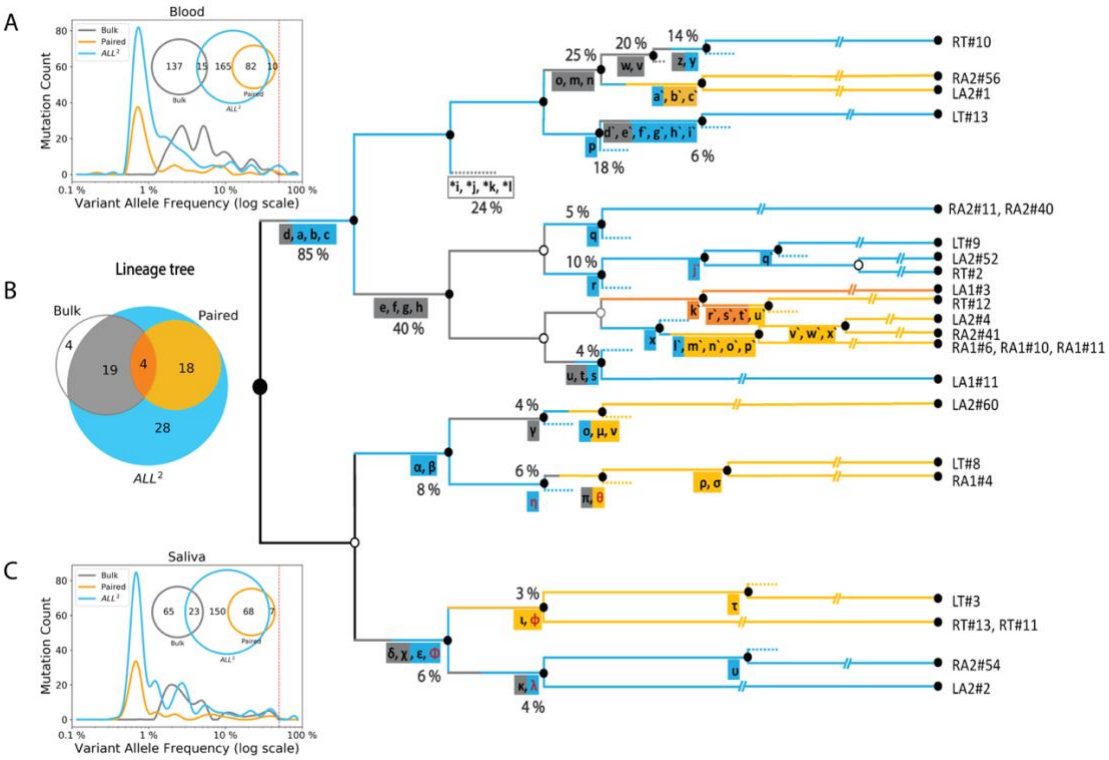


Fig 2-7. Calls from All^2 enable reconstruction of high-resolution lineage tree. (A, C) Application of All^2 to iPSC clones discovers more variants (cyan) than analysis of deeply sequenced bulk tissues (gray) or pairwise comparison of clonal lines and the bulk (orange). The approach also calls variants across entire VAF spectrum. Analysis of bulk may discover variants with intermediate VAF (1%-10%) which are not sampled in clones. For the displayed comparison, variants with at least two supporting reads in the bulks are considered for each discovery approach. (B) Lineage tree reconstructed from the analysis of 25 clones from an adult individual. Variants discovered from either bulks (gray) or from pairwise (orange) comparisons provide limited information as compared to All^2, which is the most comprehensive. Multiple branches in the lineage tree can be traced when using additional variants (cyan) discovered by applying only the All^2 approach, which is also

reflected in the Venn diagram. SNVs found only in the bulk tissues are marked with asterisks and define putative branch not sampled by clones. INDELs are colored in red and SNVs are colored in black. The percentage values next to branches denote the average fraction of the cells in bulks carrying the mutations. Clone names are shown on the right.

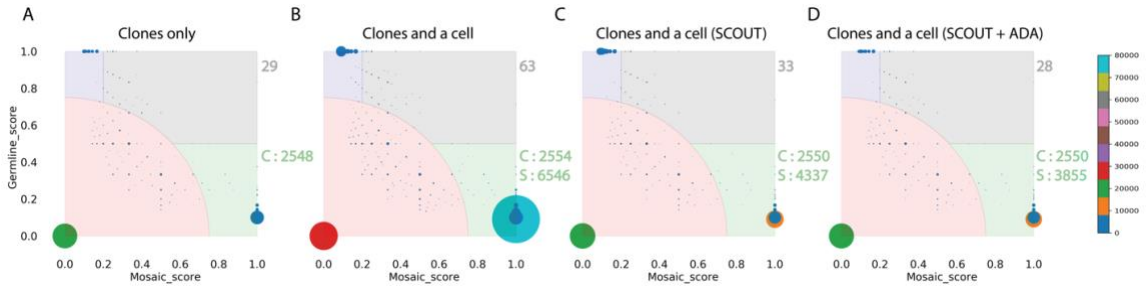


Fig 2-8. All^2 in ADA mode reduces false positive calls from allele dropout in MDA. (A) Score distribution when applying All^2 to 11 clones derived from single brain progenitor cells. There are 29 calls for high frequency (gray area) and 2548 calls for low frequency (green area) mosaic mutations. The 'C' points to mosaic calls in the clones). (B) Adding one MDA amplified cell to the analysis results in double the number of calls for high frequency mosaic mutations. Noise also increases. The 'S' points to the calls coming from the single cell. (C) Application of a specialized single cell caller SCOUT on the single cell partially mitigates issues with calling, i.e., reduces the noise and the number of mosaic calls. (D) Applying the ADA mode results in almost the same set of high frequency mosaic mutations. The mode also reduced calls for mosaic mutations in single cell without affecting calls in the clones. The color (and size) of the circles corresponds to the number of mutations sharing the same scores as depicted in the colorbar. The mosaic mutations are represented by the light green area, germline muations are represented in light blue area, high frequency mosaic mutations are represented in the light gray area and noise is represented by the light red area.

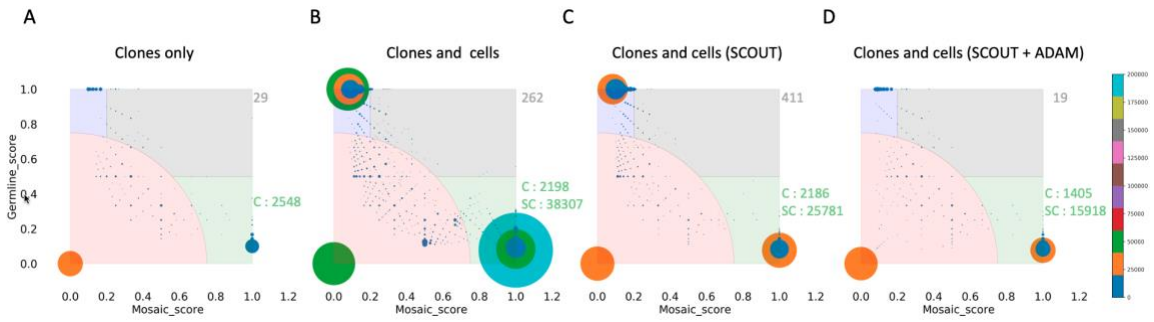


Fig 2-9: All[2] in ADA mode including 3 single cells (cell1, cell3, and cell5). Plots shown here are like those in Figure 3 and display results of mutations calling with addition of two single cells. Due to the low quality of the added single cells, the amount of noise increased 6-fold. Additionally, likely real mutations (both from clones and single cell5) were filtered out (see panel D).

Fig 2-10. Regions with allelic dropout for single cell. Upper panel: cross genome percentage of 5 Mbp bins marked as allele drop-outs. Bottom panel: the distribution of allele frequency for heterozygous variants in all and non alle drop-outs bins.

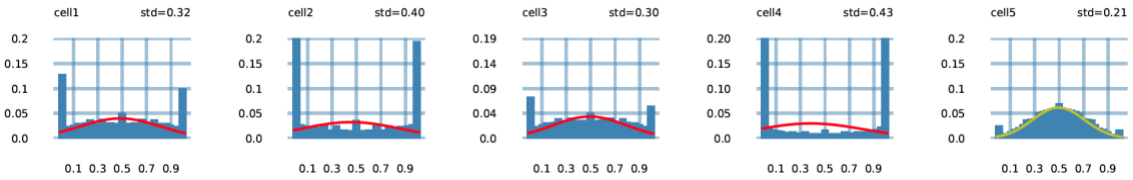


Fig 2-11. Allele frequency distribution of heterozygous germline variants in 5 MDA-amplified cells. Cell5 was selected for further analysis owing to its most uniform amplification and lower allele drop out as compared to the other cells.

Samples	Number of SNV in tree	SNVs with coverage > 200X	# of SNVs with VAF >= 4% < 35%	# of SNVs with VAF <= 1%	Missed	Missed(As fraction)
LA1	65	50	14	24	12	24.00%
LA2	65	49	10	15	24	48.98%
RA1	65	51	0	19	32	62.75%
RA2	65	51	10	15	26	50.98%
LT	65	51	22	12	17	33.33%
RT	65	50	14	21	15	30.00%
Blood	65	50	16	19	15	30.00%
Saliva	65	50	17	12	21	42.00%

Table 2-1. Fraction of mosaic mutations (SNVs) missed using different tissue types. From capture sequencing of 65 SNVs from the lineage tree (Figure 2B), we only considered the one with coverage of 200X or higher. SNVs with VAF greater than 4% and less than 35% represents mutations that can be derived from bulk whereas SNVs which VAF less than 1% can be derived from bulk to single cell paired comparison. By subtracting the number of SNVs that can be found by bulk analyses and paired analyses from the number of "SNVs with coverage > 200X" we can calculate the fraction of missed SNVs. The column "Missed (As Fraction)" shows that LA1, LT and RT have comparable missed SNVs with blood and saliva, whereas in LA2, RA1 and RA2, half of the SNVs are being missed. These results suggest that using any of the bulk samples does not outperform ALL[2] for lineage tree construction.

CHAPTER 4 – Conclusion

4.1 SUMMARY

Single cell omics experiments are becoming increasingly crucial for mapping cell heterogeneity in tissues and organs from many different perspectives, from transcriptomics and DNA variations to epigenomic such as chromatin accessibility (i.e., scATAC-seq). Single cell sequencing experiments can be very costly, and it is important to optimize the sequencing cost by choosing cells which have been amplified uniformly over the whole genome. We have developed a tool Scellector (discussed in Chapter-2) which implements a method to detect amplification quality from shallow coverage data (<1X) and prioritizes well amplified cells for high coverage sequencing. With the advent of single cell DNA sequencing from companies like Chromium Single Cell CNV profiling solutions (10X Genomics), which uses an isothermal amplification protocol similar to MDA, we believe that our tool can be extended to estimate uniformity of amplification from these platforms. This platform can profile hundred to thousand cells in a single sample to detect copy number variation and provide information on genomic heterogeneity as well as clonal evolution. Not all cells will have uniform amplification and Scellector can be used to detect and remove low quality cells, which will make the downstream analyses of CNV detection more robust. Scellector is an open-source tool and source code can be found at https://github.com/abyzovlab/Scellector.

Another challenge is the accurate discovery of somatic mutations in a cell, and it partially lays in immaturity of dedicated analytical approaches. Approaches comparing a cell's genome to a control bulk sample miss common mutations, while approaches to find such mutations from bulk suffer from low sensitivity. We have developed and implemented All^2

(discussed in Chapter-3), which can discover mosaic SNVs, indels and SVs from exhaustive cell-to-cell comparison of WGS data from single cells or clones. Our method is superior to using deep sequencing of bulk tissues and/or paired comparison of single cells versus bulk for detection of both low and high frequency mosaic mutations. A limitation relative to bulk method is that the mutations that are not sampled by the analyzed single cells cannot be discovered. This can be addressed by increasing the number of analyzed single cells. We have also applied All^2 for comprehensive reconstruction of a developmental lineage tree, showing that All^2 allows a vastly more comprehensive lineage discovery. Furthermore, the method is general and can be applied to any problem of lineage tracing that relies on the analysis of multiple cells, such as tracing cancer evolution.

We further demonstrate that All^2 facilitates removal of false positive calls (in ADA mode) from amplified single cells. Additionally, since ADA mode takes a BED file with inclusive regions as input, All^2 can be applied to the analyses of exome sequencing where a user can provide a file with target regions. The same mode can also be applied to exclude copy number altered regions when analyzing cancer cells. All^2 provides visualizations such as allele frequency distribution, mutation spectrum, mutation counts and score distribution plots to help guide the user to better understand their data as well as change parameter setting for calling mosaic mutations. The tool is open source and is freely available on GitHub: https://github.com/abyzovlab/All2.

4.2 FUTURE DIRECTIONS

Other than the issues we addressed in this thesis, there are other factors that need to improve to make scDNA-seq scalable and cost efficient for somatic mutation detection. scRNA-seq

has higher abundance of mRNA molecules compared to DNA making it easier to obtain enough material for sequencing and making it possible to analyze hundreds of cells for relatively lower cost using platforms like 10X genomics. However, with scDNA-seq, similar platforms have not been successful due many factors such as faulty amplification and lower coverage for calling mutations, requiring the need for sequencing each single cell as a single sample. Since each single cell must be sequenced independently at 30X coverage for mutation detection, this makes the experiment design expensive and limits the number of single cells that can be analyzed. With future improvement in the amplification methods such as those shown by PTA (primary template amplification)(31), it will be possible to use scalable platforms for scDNA-seq analyses such as 10X genomics for the analyses of hundreds of cells at a time. Additionally, with sequencing costs going down, such as shown by the Ultima genomics(58) which promises $300 genome, the possibility to analyze high volumes of scDNA-seq data is a possibility in not so far future. Another advantage of scRNA-seq over scDNA-seq is the fact that the former can identify the cell type from the sequencing data which is used to identify different cell types in each sample. This is particularly helpful to study cancer samples to understand the immune infiltration within the tissues as well as different clones representing the mass. Platforms such as Pacbio have shown that it is possible to identify methylation from DNA sequencing data(59). The methylation pattern thus obtained can be used to identify different cell types(60, 61). Preliminary data from our lab on PTA amplified single cells shows promise to do the same. Equipped with improved amplification methods, scalable single cell tagging platforms, low-cost sequencing, and the ability to identify cell type from DNA, the

future looks very promising for robust, accurate and high scale detection of somatic mosaicism which will give us a deeper biological understand of normal development and disease and accelerate the field further.

Bibliography

1. Lynch M. Rate, molecular spectrum, and consequences of human mutation. Proc Natl Acad Sci U S A. 2010;107(3):961-8.
2. Galvão V, Miranda JG, Andrade RF, Andrade JS, Jr., Gallos LK, Makse HA. Modularity map of the network of human cell differentiation. Proc Natl Acad Sci U S A. 2010;107(13):5750-5.
3. Conrad DF, Pinto D, Redon R, Feuk L, Gokcumen O, Zhang Y, et al. Origins and functional impact of copy number variation in the human genome. Nature. 2010;464(7289):704-12.
4. Maretty L, Jensen JM, Petersen B, Sibbesen JA, Liu S, Villesen P, et al. Sequencing and de novo assembly of 150 genomes from Denmark as a population reference. Nature. 2017;548(7665):87-91.
5. Michaelson JJ, Shi Y, Gujral M, Zheng H, Malhotra D, Jin X, et al. Whole-genome sequencing in autism identifies hot spots for de novo germline mutation. Cell. 2012;151(7):1431-42.
6. Milholland B, Dong X, Zhang L, Hao X, Suh Y, Vijg J. Differences between germline and somatic mutation rates in humans and mice. Nature communications. 2017;8:15183.
7. Rahbari R, Wuster A, Lindsay SJ, Hardwick RJ, Alexandrov LB, Turki SA, et al. Timing, rates and spectra of human germline mutation. Nat Genet. 2016;48(2):126-33.
8. RK CY, Merico D, Bookman M, J LH, Thiruvahindrapuram B, Patel RV, et al. Whole genome sequencing resource identifies 18 new candidate genes for autism spectrum disorder. Nat Neurosci. 2017;20(4):602-11.
9. Blokzijl F, de Ligt J, Jager M, Sasselli V, Roerink S, Sasaki N, et al. Tissue-specific mutation accumulation in human adult stem cells during life. Nature. 2016;538(7624):260-4.
10. Lee-Six H, Obro NF, Shepherd MS, Grossmann S, Dawson K, Belmonte M, et al. Population dynamics of normal human blood inferred from somatic mutations. Nature. 2018;561(7724):473-8.
11. Martincorena I, Fowler JC, Wabik A, Lawson ARJ, Abascal F, Hall MWJ, et al. Somatic mutant clones colonize the human esophagus with age. Science. 2018;362(6417):911-7.
12. Podolskiy DI, Lobanov AV, Kryukov GV, Gladyshev VN. Analysis of cancer genomes reveals basic features of human aging and its role in cancer development. Nature communications. 2016;7:12157.

13. Cagan A, Baez-Ortega A, Brzozowska N, Abascal F, Coorens THH, Sanders MA, et al. Somatic mutation rates scale with lifespan across mammals. Nature. 2022;604(7906):517-24.
14. Yokoyama A, Kakiuchi N, Yoshizato T, Nannya Y, Suzuki H, Takeuchi Y, et al. Age-related remodelling of oesophageal epithelia by mutated cancer drivers. Nature. 2019;565(7739):312-7.
15. Pareja F, Ptashkin RN, Brown DN, Derakhshan F, Selenica P, da Silva EM, et al. Cancer-Causative Mutations Occurring in Early Embryogenesis. Cancer Discov. 2022;12(4):949-57.
16. Tomasetti C, Li L, Vogelstein B. Stem cell divisions, somatic mutations, cancer etiology, and cancer prevention. Science. 2017;355(6331):1330-4.
17. Milholland B, Auton A, Suh Y, Vijg J. Age-related somatic mutations in the cancer genome. Oncotarget. 2015;6(28):24627-35.
18. Miller MB, Huang AY, Kim J, Zhou Z, Kirkham SL, Maury EA, et al. Somatic genomic changes in single Alzheimer's disease neurons. Nature. 2022;604(7907):714-22.
19. Lodato MA, Rodin RE, Bohrson CL, Coulter ME, Barton AR, Kwon M, et al. Aging and neurodegeneration are associated with increased mutations in single human neurons. Science. 2018;359(6375):555-9.
20. Freed D, Pevsner J. The Contribution of Mosaic Variants to Autism Spectrum Disorder. PLoS Genet. 2016;12(9):e1006245.
21. Miller KE, Koboldt DC, Schieffer KM, Bedrosian TA, Crist E, Sheline A, et al. Somatic SLC35A2 mosaicism correlates with clinical findings in epilepsy brain tissue. Neurol Genet. 2020;6(4):e460.
22. Evrony GD, Hinch AG, Luo C. Applications of Single-Cell DNA Sequencing. Annual review of genomics and human genetics. 2021;22:171-97.
23. Lodato MA, Woodworth MB, Lee S, Evrony GD, Mehta BK, Karger A, et al. Somatic mutation in single human neurons tracks developmental and transcriptional history. Science. 2015;350(6256):94-8.
24. Wang Y, Waters J, Leung ML, Unruh A, Roh W, Shi X, et al. Clonal evolution in breast cancer revealed by single nucleus genome sequencing. Nature. 2014;512(7513):155-60.
25. Zhang CZ, Spektor A, Cornils H, Francis JM, Jackson EK, Liu S, et al. Chromothripsis from DNA damage in micronuclei. Nature. 2015;522(7555):179-84.
26. Martincorena I, Fowler JC, Wabik A, Lawson ARJ, Abascal F, Hall MWJ, et al. Somatic mutant clones colonize the human esophagus with age. Science. 2018;362(6417):911-7.
27. Lee-Six H, Øbro NF, Shepherd MS, Grossmann S, Dawson K, Belmonte M, et al. Population dynamics of normal human blood inferred from somatic mutations. Nature. 2018;561(7724):473-8.
28. Zhang L, Dong X, Lee M, Maslov AY, Wang T, Vijg J. 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. 2019;116(18):9014-9.

29. Cheung VG, Nelson SF. Whole genome amplification using a degenerate oligonucleotide primer allows hundreds of genotypes to be performed on less than one nanogram of genomic DNA. Proc Natl Acad Sci U S A. 1996;93(25):14676-9.
30. Dean FB, Nelson JR, Giesler TL, Lasken RS. Rapid amplification of plasmid and phage DNA using Phi 29 DNA polymerase and multiply-primed rolling circle amplification. Genome Res. 2001;11(6):1095-9.
31. Gonzalez-Pena V, Natarajan S, Xia Y, Klein D, Carter R, Pang Y, et al. Accurate genomic variant detection in single cells with primary template-directed amplification. Proc Natl Acad Sci U S A. 2021;118(24).
32. Zong C, Lu S, Chapman AR, Xie XS. Genome-wide detection of single-nucleotide and copy-number variations of a single human cell. Science. 2012;338(6114):1622-6.
33. Chen C, Xing D, Tan L, Li H, Zhou G, Huang L, et al. Single-cell whole-genome analyses by Linear Amplification via Transposon Insertion (LIANTI). Science. 2017;356(6334):189-94.
34. Huang L, Ma F, Chapman A, Lu S, Xie XS. Single-Cell Whole-Genome Amplification and Sequencing: Methodology and Applications. Annual review of genomics and human genetics. 2015;16:79-102.
35. Cibulskis K, Lawrence MS, Carter SL, Sivachenko A, Jaffe D, Sougnez C, et al. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. Nat Biotechnol. 2013;31(3):213-9.
36. Saunders CT, Wong WS, Swamy S, Becq J, Murray LJ, Cheetham RK. Strelka: accurate somatic small-variant calling from sequenced tumor-normal sample pairs. Bioinformatics. 2012;28(14):1811-7.
37. Tu K, Lu K, Zhang Q, Huang W, Xie D. Accurate single-cell genotyping utilizing information from the local genome territory. Nucleic acids research. 2021;49(10):e57.
38. Luquette LJ, Bohrson CL, Sherman MA, Park PJ. Identification of somatic mutations in single cell DNA-seq using a spatial model of allelic imbalance. Nature communications. 2019;10(1):3908.
39. Evrony GD, Cai X, Lee E, Hills LB, Elhosary PC, Lehmann HS, et al. Single-neuron sequencing analysis of L1 retrotransposition and somatic mutation in the human brain. Cell. 2012;151(3):483-96.
40. Zhang CZ, Adalsteinsson VA, Francis J, Cornils H, Jung J, Maire C, et al. Calibrating genomic and allelic coverage bias in single-cell sequencing. Nature communications. 2015;6:6822.
41. Sherman MA, Barton AR, Lodato MA, Vitzthum C, Coulter ME, Walsh CA, et al. PaSD-qc: quality control for single cell whole-genome sequencing data using power spectral density estimation. Nucleic acids research. 2018;46(4):e20.
42. Cai X, Evrony GD, Lehmann HS, Elhosary PC, Mehta BK, Poduri A, et al. Single-cell, genome-wide sequencing identifies clonal somatic copy-number variation in the human brain. Cell Rep. 2014;8(5):1280-9.
43. Delaneau O, Zagury JF, Marchini J. Improved whole-chromosome phasing for disease and population genetic studies. Nature methods. 2013;10(1):5-6.

44. Delaneau O, Marchini J, Zagury JF. A linear complexity phasing method for thousands of genomes. Nature methods. 2011;9(2):179-81.
45. Delaneau O, Marchini J. Integrating sequence and array data to create an improved 1000 Genomes Project haplotype reference panel. Nature communications. 2014;5:3934.
46. Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, Korbel JO, et al. A global reference for human genetic variation. Nature. 2015;526(7571):68-74.
47. Sanchez-Luque FJ, Kempen MHC, Gerdes P, Vargas-Landin DB, Richardson SR, Troskie RL, et al. LINE-1 Evasion of Epigenetic Repression in Humans. Mol Cell. 2019;75(3):590-604.e12.
48. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009;25(14):1754-60.
49. Poplin R, Ruano-Rubio V, DePristo MA, Fennell TJ, Carneiro MO, Van der Auwera GA, et al. Scaling accurate genetic variant discovery to tens of thousands of samples. bioRxiv. 2018:201178.
50. Bae T, Tomasini L, Mariani J, Zhou B, Roychowdhury T, Franjic D, et al. Different mutational rates and mechanisms in human cells at pregastrulation and neurogenesis. Science. 2018;359(6375):550-5.
51. Fasching L, Jang Y, Tomasi S, Schreiner J, Tomasini L, Brady MV, et al. Early developmental asymmetries in cell lineage trees in living individuals. Science. 2021;371(6535):1245-8.
52. Chen X, Schulz-Trieglaff O, Shaw R, Barnes B, Schlesinger F, Källberg M, et al. Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. Bioinformatics. 2016;32(8):1220-2.
53. Sekar S, Tomasini L, Proukakis C, Bae T, Manlove L, Jang Y, et al. Complex mosaic structural variations in human fetal brains. Genome Res. 2020;30(12):1695-704.
54. Sarangi V, Jourdon A, Bae T, Panda A, Vaccarino F, Abyzov A. SCELLECTOR: ranking amplification bias in single cells using shallow sequencing. BMC Bioinformatics. 2020;21(1):521.
55. Abyzov A, Tomasini L, Zhou B, Vasmatzis N, Coppola G, Amenduni M, et al. One thousand somatic SNVs per skin fibroblast cell set baseline of mosaic mutational load with patterns that suggest proliferative origin. Genome Res. 2017;27(4):512-23.
56. Wei J, Zhou T, Zhang X, Tian T. SCOUT: A new algorithm for the inference of pseudo-time trajectory using single-cell data. Comput Biol Chem. 2019;80:111-20.
57. Suvakov M, Panda A, Diesh C, Holmes I, Abyzov A. CNVpytor: a tool for copy number variation detection and analysis from read depth and allele imbalance in whole-genome sequencing. GigaScience. 2021;10(11).
58. Almogy G, Pratt M, Oberstrass F, Lee L, Mazur D, Beckett N, et al. Cost-efficient whole genome-sequencing using novel mostly natural sequencing-by-synthesis chemistry and open fluidics platform. bioRxiv. 2022:2022.05.29.493900.
59. Ni P, Nie F, Zhong Z, Xu J, Huang N, Zhang J, et al. DNA 5-methylcytosine detection and methylation phasing using PacBio circular consensus sequencing. Nature communications. 2023;14(1):4054.

60. Loyfer N, Magenheim J, Peretz A, Cann G, Bredno J, Klochendler A, et al. A DNA methylation atlas of normal human cell types. Nature. 2023;613(7943):355-64.
61. Zhu T, Liu J, Beck S, Pan S, Capper D, Lechner M, et al. A pan-tissue DNA methylation atlas enables in silico decomposition of human tissue methylomes at cell-type resolution. Nature methods. 2022;19(3):296-306.