Fall 12-14-2018

Red Panda: A Novel Method for Detecting Variation in Single-Cell RNA Sequencing

Adam Cornish
University of Nebraska Medical Center

Recommended Citation

Cornish, Adam, "Red Panda: A Novel Method for Detecting Variation in Single-Cell RNA Sequencing" (2018). *Theses & Dissertations*. 315.
https://digitalcommons.unmc.edu/etd/315

RED PANDA: A NOVEL METHOD FOR DETECTING VARIATION IN SINGLE-CELL RNA SEQUENCING

by

Adam Cornish

A DISSERTATION

Presented to the Faculty of
the University of Nebraska Graduate College
in Partial Fulfillment of the Requirements
for the Degree of Doctor of Philosophy

Genetics, Cell Biology & Anatomy
Graduate Program

Under the Supervision of Professor Chittibabu Guda

University of Nebraska Medical Center
Omaha, Nebraska

November 2018

Supervisory Committee:
Jyothi Arikkath, Ph.D.
Andrew Dudley, Ph.D.
James Eudy, Ph.D.
Fang Yu, Ph.D.

DEDICATION

To my immensely wonderful support network. You saved me.

ACKNOWLEDGMENTS

First, I would like to thank my mentor Dr. Guda for his unfailing guidance during the course of my graduate career. His direction and encouragement allowed me to learn, make my own mistakes, and grow into the scientist that I am today. I would also like to thank my committee members, Dr. Jyothi Arikkath, Dr. Andrew Dudley, Dr. James Eudy, and Dr. Fang Yu. They provided astounding support and critical feedback that greatly influenced this project in a positive way.

I would also like to thank the Genetics, Cell Biology and Anatomy department, especially those who provide such passionate support of their students. Dr. Gould and Dr. Dudley provide constant aid for those who seek it and they have helped foster a thriving student culture we can exceed. Speaking of those other students, thank you to all of you. You have made this difficult process at times, dare I say it, fun.

Thank you to all of the lab members in the Guda lab, both past and present, who have aided me. A special thank you to Kristin for always being there to help me through the complexities and frustrating simplicities of this project.

Lastly, I would like to thank my incredible support network. All of you have made it possible for me to get through this with a surprisingly large amount of sanity left over. I am truly grateful that I have such incredible people at my side. Cecily, I will never be able to repay you for the massive investment you have made in helping me with this monolith. And to Molly. You have been unfailingly selfless and have picked me up every time I stumbled or floundered. I could not have done this without you.

RED PANDA: A NOVEL METHOD FOR DETECTING VARIATION IN SINGLE-CELL RNA SEQUENCING

Adam Cornish, Ph.D.

University of Nebraska, 2018

Supervisor: Chittibabu Guda, Ph.D.

Single-cell sequencing enables the rapid acquisition of genomic and transcriptomic data from individual cells to better understand genetic diseases, such as cancer or autoimmune disorders, which are often affected by changes in rare cells. Currently, no existing software is aimed at identifying single nucleotide variations or micro (1-50bp) insertions and deletions in single-cell RNA sequencing (scRNA-seq) data. However, generating high quality data is vital to the study of the aforementioned diseases, among others. Our goal is to create such a tool and use in-house sequencing to validate its effectiveness. Our software, Red Panda, employs the unique information found in scRNA-seq data to more accurately identify variants in ways not possible with software designed for bulk sequencing. We intentionally isolate variants based on three different classes: homozygous-looking, heterozygous, and bimodally-distributed heterozygous, the last of which can only be identified in scRNA-seq. To properly validate the results from this method, variants were called on: scRNA-seq and exome sequencing jointly performed on human articular chondrocytes, scRNA-seq from mouse embryonic fibroblasts (MEFs), and simulated data stemming from the MEF alignments. The chondrocyte exome sequencing was used to validate the chondrocyte scRNA-seq results. For Red Panda, on average, 913 variants were shared with the exome and had a Positive

Predictive Value (PPV) of 45.0%. Other tools—FreeBayes, GATK HaplotypeCaller, GATK UnifiedGenotyper, and Platypus—ranged from 65-705 variants and 5.8%-31.7% PPV. Sanger sequencing was performed on a subset of the variants identified in the MEFs, and simulated data was generated to assess the sensitivity of each tools. From the latter, Red Panda had the highest sensitivity at 72.44%. The other tools ranged from 18.22% to 39.09%. We show that our method provides a novel and improved mechanism to identify variants in scRNA-seq as compared to currently-existing software.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS AND INITIALISMS

AP	Apurinic/Apyrimidinic
AF	Allele Frequency
CI	Confidence Interval
DP	Depth of sequencing
FDR	False Discovery Rate
FN	False Negative
FP	False Positive
GATK	Genome Analysis ToolKit
GATK-HC	GATK HaplotypeCaller
GATK-UG	GATK UnifiedGenotyper
indel	Insertion and Deletion
MEF	Mouse Embryonic Fibroblast
MNV	Multi-Nucleotide Variant
NGS	Next Generation Sequencing
PPV	Positive Predictive Value
scRNA-seq	Single-Cell RNA Sequencing
SCS	Single-Cell Sequencing
-seq	Sequencing
SNV	Single Nucleotide Variation
TN	True Negative
TP	True Positive
TPM	Transcripts Per Million
TPR	True Positive Rate

UTR	Untranslated Region
v.	Version
VCF	Variant Call Format
WGS	Whole Genome Sequencing

INTRODUCTION

Single-cell sequencing

Single-cell sequencing (SCS) is a relatively new technique that saw its first use in 2011[1]. Since its introduction, it has been used to investigate the heterogeneity of different cancers[2–4], determine copy number variation in enhanced detail[5,6], and better characterize circulating tumor cells using differential expression analysis[7–10]. Multiple recent studies using SCS have also shown that tumors are genetically diverse and produce subclones that contribute to the pathogenicity of the disease by conferring chemotherapy resistance and metastatic capabilities to the tumor[11–13]. Application of this new technology is not limited to cancer research; it has proven very useful in areas such as characterizing somatic mutations in neurons[14], identifying rare intestinal cell types[15], and discriminating cell types in healthy tissues[15–17].

Regarding the exact methodology, SCS itself is composed of three steps: 1.) cell capture, 2.) RNA or DNA library preparation, and 3.) sequencing of those libraries. The different technologies used at each step can greatly influence the type and quality of data generated. Our study captured cells with the C1 Fluidigm System, which uses a microfluidic circuit design (Integrated Fluidics Circuit, or IFC) to accurately capture single cells into 96 or 800 different reaction chambers[18]. These cells were then lysed and had their mRNA isolated, converted into cDNA, amplified, and prepared for sequencing using the Smart-seq2 protocol[19]. This library preparation method is unique in that it amplifies the entire transcript, as opposed to just the 3′ end. Following this, the cDNA is sequenced using standard Illumina protocol on the NextSeq500 or a similar instrument to generate short reads that can be used in downstream analyses.

Variant detection in SCS

Single Nucleotide Variants (SNVs) and micro (1-50bp) insertions and deletions (indels) can have a large impact on human disease[20–22] and are typically identified using exome sequencing or whole genome sequencing (WGS)[23]. In these datasets where of millions of cells are sequenced—otherwise referred to as bulk sequencing—, reads are aligned to a reference genome, and variations are identified by one of a number of different tools, such as FreeBayes[24], GATK-HaplotypeCaller[25], GATK-UnifiedGenotyper[26], or Platypus[27].

The ability to accurately assess the presence of SNVs and indels in an individual without having to perform both exome sequencing and RNA sequencing would be a great boon to researchers. Further, as some studies rely on obtaining data from rare cell types, it is important that it be possible to identify mutations and characterize gene expression in the same cell. While it is possible to sequence both DNA and RNA from the same cell as described by Macaulay *et al*.[28], this technique is one that the typical lab or core facility will not be able to perform due to the custom procedures and chemistry involved. Further, the method developed by Macaulay *et al*. makes it impossible to accurately identify variants[28–30]. Instead investigators are limited to using standard reagent kits, such as the extant SMARTer Ultra Low RNA Kit by Clontech for Illumina Sequencing, to obtain material of high enough quality to identify variants.

In addition to the aforementioned types of studies performed in SCS (copy number variation detection, differential expression analysis, rare cell identification, etc.), effort has been made to apply standard bulk sequencing bioinformatic methods to identify variants in SCS datasets[14,31], and while this is feasible, it does not take advantage of the unique nature of the data produced by the SCS platform. Further, it has been necessary to rely on tools that were designed for Next Generation Sequencing (NGS) data derived from genomic DNA instead of data generated from mRNA.

This study introduces a novel method, Red Panda, that is designed to identify variants in single-cell RNA sequencing (scRNA-seq) and tests how it compares to currently-available variant callers. These tools are those that have previously been determined by our group to be the most accurate[32]. They are FreeBayes[24], HaplotypeCaller found in the Genome Analysis Toolkit (GATK) package[25], Platypus[27], and UnifiedGenotyper as found in the GATK package[26]. FreeBayes is a Bayesian statistical framework capable of modeling multiallelic loci regardless of copy number. GATK HaplotypeCaller (GATK-HC) uses a De Bruijn-like graph to reassemble regions of the genome that show evidence of significant variation. GATK UnifiedGenotyper uses a Bayesian genotype likelihood model to estimate the most likely genotypes and allele frequency. Platypus uses local realignment of reads and local reassembly to accurately identify variants in a genome. All four tools infer information about changes in sequencing data when compared to a reference genome.

These tools were originally developed for calling variants using bulk DNA sequencing data, but can also identify variants in bulk mRNA-seq data (e.g., The Broad

Institute has put together a Best Practices guide using GATK HaplotypeCaller[33]); however, ultimately this is software that was not designed for the unique case of scRNA-seq data . Ideally variant calling would be performed on data derived from genomic DNA from single cells as these would suit such software better as well as provide more comprehensive results, but there exist too many problems with these data to make this approach viable, namely: amplifying genomic DNA by any method leads to allelic dropout; coverage nonuniformity reduces the probability of identifying variants in useful areas of the genome, such as exons; and False Positive (FP) amplification errors are very common[34]. While it is certainly possible to utilize these programs on data derived from single-cell genomic sequencing, the aforementioned problems ultimately make variant calling untenable. For this reason developing a method that can utilize the higher quality scRNA-seq data to perform variant calling is necessary, as it does not suffer the same shortcomings. Once such a strategy has been implemented, it will also have the benefit of allowing us to investigate allele-specific expression, which play a role in understanding different cell processes and how it correlates with diseases[35,36].

Currently all bioinformatic tools used for variant identification in mRNA-seq data treat variants within a transcript as independent events because they assume the sample is composed of source material from millions of cells. This approach works well for bulk sequencing but loses its power when used in scRNA-seq. To address this, our method, Red Panda, employs the unique information found in scRNA-seq data to more accurately identify variants in ways not possible with software designed for bulk sequencing. The fact that transcripts represented by scRNA-seq reads necessarily only originate from the chromosomes present in a single cell is important. Where applicable, this fact is used to decide what is and is not a heterozygous variant. For example, if 20% of the transcripts in a cell originate from the maternal chromosome and 80% originate from the paternal, then every heterozygous variant in the expressed transcript will be represented by reads in the scRNA-seq data at a reference allele to alternate allele ratio of either 1:4 or 4:1. This is because the expressed transcripts must have been derived from either the maternal or maternal chromosomes in one cell. These types of heterozygous variants are termed bimodally-distributed heterozygous. As part of the process of identifying this class of variant, Red Panda creates three different classes: homozygous-looking, bimodally-distributed heterozygous, and non-bimodally-distributed heterozygous. This partitioning strategy, as well as treating bimodally-distributed variants differently, grants an advantage compared to currently available tools.

Datasets

In order to properly test Red Panda, a number of datasets were used and/or generated. To be useful in a testing environment, test data needed to satisfy the following criteria:

1. Bulk genomic sequencing data must pair with scRNA-seq data generated from Smart-seq libraries.
2. The tissue used to generate these libraries must be isogenic.
3. The sequencing data must be of high quality.

4. The data must be from a well-annotated genome, specifically either human or mouse.
5. The data must be from healthy tissue to ensure as few variables as possible be introduced to the software testing environment.

The first criterion was especially important because the bulk sequencing data will be used to corroborate the findings from the scRNA-seq data.

To ensure high quality data was available for algorithm development, the Genomics Core Facility at UNMC performed sequencing using the Smart-seq2 protocol for single cells on human articular chondrocytes. In addition to RNA from 30 cells being sequenced, exome sequencing data were also generated from these chondrocytes using traditional exome sequencing. Ultimately, data from 22 cells were used after eliminating poor quality data from eight of the cells. Additionally, 55 normal mouse embryonic fibroblasts (MEFs) have been sequenced for additional validation using the Smart-seq2 protocol.

Validation

To confirm the existence or nonexistence of variants identified by Red Panda as well as the four bulk sequencing variant callers (FreeBayes, GATK HaplotypeCaller, GATK UnifiedGenotyper, and Platypus), Sanger sequencing was used attempt to verify the existence of 40 randomly identified variants: 20 unique to Red Panda and 20 identified by all five variant callers. Simulated data was also generated from the MEF sequencing to accurately predict the sensitivity of each variant caller.

Statistics

Sensitivity and specificity are the metrics most often used to determine how accurate data collection is or how accurate bioinformatic tools are. Sensitivity measures the proportion of True Positives present in the measured dataset from the total that should exist, and specificity measures the proportion of True Negatives that are correctly measured as such. In this dataset:

- A True Positive (TP) is a variant that truly exists as compared to the reference genome and is identified as such.
- A True Negative (TN) is a position on the genome that is correctly identified as not differing from the reference genome.
- A False Positive (FP) is a position on the genome that is incorrectly identified as a variant when compared to the reference.
- A False Negative (FN) is a variant that truly exists but is incorrectly identified as not differing from the reference genome.

For our purposes, specificity (TN/(TN+FP)) and its derivatives such as False Positive Rate (1-specificity), were not calculated for the exome and single-cell RNA sequencing results due to how many True Negatives exist in the dataset as compared to the number of False Positives: often the number of True Negatives in these datasets are in the millions as opposed to the hundreds or thousands of False Positives. Instead, True Positive Rate (TPR) and Positive Predictive Value (PPV) are focused on in these analyses, as those numbers are more meaningful when using the results from the scRNA-seq and exome sequencing. However, sensitivity is calculated using the simulated results from the MEF sequencing.

Software distribution

To ensure ease of access to and adoption of this tool, it shall be published via GitHub as a stand-alone package where the source code will also be made available. The stand-alone package will be a binary that any Linux-based system can run in conjunction with the Genome Analysis Toolkit.

Hypothesis

Modern variant calling software, designed for bulk sequencing, cannot take advantage of the information found only in scRNA-seq. Our method, Red Panda, efficiently utilizes this unique information resulting in greater accuracy, opening up more ways to analyze scRNA-seq data as it was previously not possible to use SNVs and micro indels to investigate diseases at the cellular level.

CHAPTER 1: ALGORITHM DEVELOPMENT OF RED PANDA

Introduction

Proper development of the software required a dataset that satisfied a number of criteria for testing purposes. The sequencing data needed to:

1. Be isogenic for both a bulk genomic sequencing sample and also a scRNA-seq sample.
2. Be high quality
3. Belong to an organism with a well-annotated genome
4. Come from a healthy sample individual

The data obtained from Borel *et al.* 2015 was the testing dataset employed to assess the algorithm because, upon first inspection, it fit all four criteria. There are 163 cells from the UCF1014 cell line and 40 cells from the TN2A cell line on which variant calling was performed, and there is WGS from both cell lines.

The largest challenge when identifying SNVs in any NGS data is reducing the number of False Positives while maximizing the number of True Positives. The novel method outlined here utilizes the uniqueness of the scRNA-seq data to classify variants: all heterozygous variants in a given isoform can be pooled together into an expected bimodally-distributed pattern to determine which variants in the isoform are real. This strategy results in sorting variants into three different classes: homozygous-looking, heterozygous and not bimodally-distributed, and bimodally-distributed heterozygous.

This partitioning strategy and treatment of bimodally-distributed variants grants an advantage compared to currently available tools.

Normally, by necessity, variant callers look at each SNV or indel as independent events: they scan the alignment files for discrepancies between the reference genome and the reads aligned to the reference genome. If there is a difference between the two, then multiple statistical measures are calculated for that location. The particular method depends on the tool, e.g., GATK HaplotypeCaller performs a *de novo* assembly of the reads at the location in question and then uses Hidden Markov Models to determine the haplotype of the variant at that position. If the results of those statistical tests meet certain cutoff criteria, then that location, whether it is an SNV or an indel, is reported as a potential variant and the variant caller moves on to the next position. While this works very well for bulk sequencing datasets, improvements are possible through utilizing the extra information specifically obtained from SCS data.

Red Panda utilizes the unique information found in scRNA-seq data to more accurately identify variants. We capitalize on the fact that, because data come from a single cell, transcripts represented by the scRNA-seq reads necessarily only come from the two chromosomes present (or more, if there is aneuploidy), and we factor that into our decision-making process when establishing what is and is not a variant. In a diploid cell, one would expect transcripts to originate from two chromosomes, and thus, any heterozygous variant present in a transcript as seen in the sequencing data will be represented in a fraction consistent with the fraction of transcripts coming from a specific chromosome. For instance **Figure 1** shows that if 30% of the transcripts in a cell originate from the maternal chromosome and 70% from the paternal chromosome, then reads in the scRNA-seq data will represent every heterozygous variant present in that transcript at either a 7:3 ratio (reference:alternate allele) or a 3:7 ratio. This type of variant is considered to be bimodally-distributed heterozygous. Using this concept of read distributions, Red Panda can accurately remove False Positives—often artefacts from the library preparation, sequencing, or alignment—that modern variant callers would not remove, as well as pick up variants supported by a low fraction of reads.

With this concept, we were able to provide a novel and improved method for identifying variants in datasets generated in a rapidly-evolving technology. However, to accurately develop this algorithm, an appropriate dataset is needed for testing.

Primary fibroblast data analysis

To properly evaluate the development of our algorithm, an appropriate dataset is needed for testing. Ideally, this is NGS data generated from a single sample but by two different methods: bulk sequencing of genomic DNA and scRNA-seq of many individual cells. Such a dataset has been acquired through the European Genome Archive and is described below.

In 2015, Borel *et al.* performed a study to test the differential allelic distribution in human primary fibroblasts[31]. Their interest lay in whether alleles were expressed uniformly or preferentially. To determine the pattern of allelic expression, the researchers sequenced 203 cells using scRNA-seq from two human primary fibroblast cell lines, TN2A and UCF1014. **Figure 2** shows a schematic detailing the full sequencing

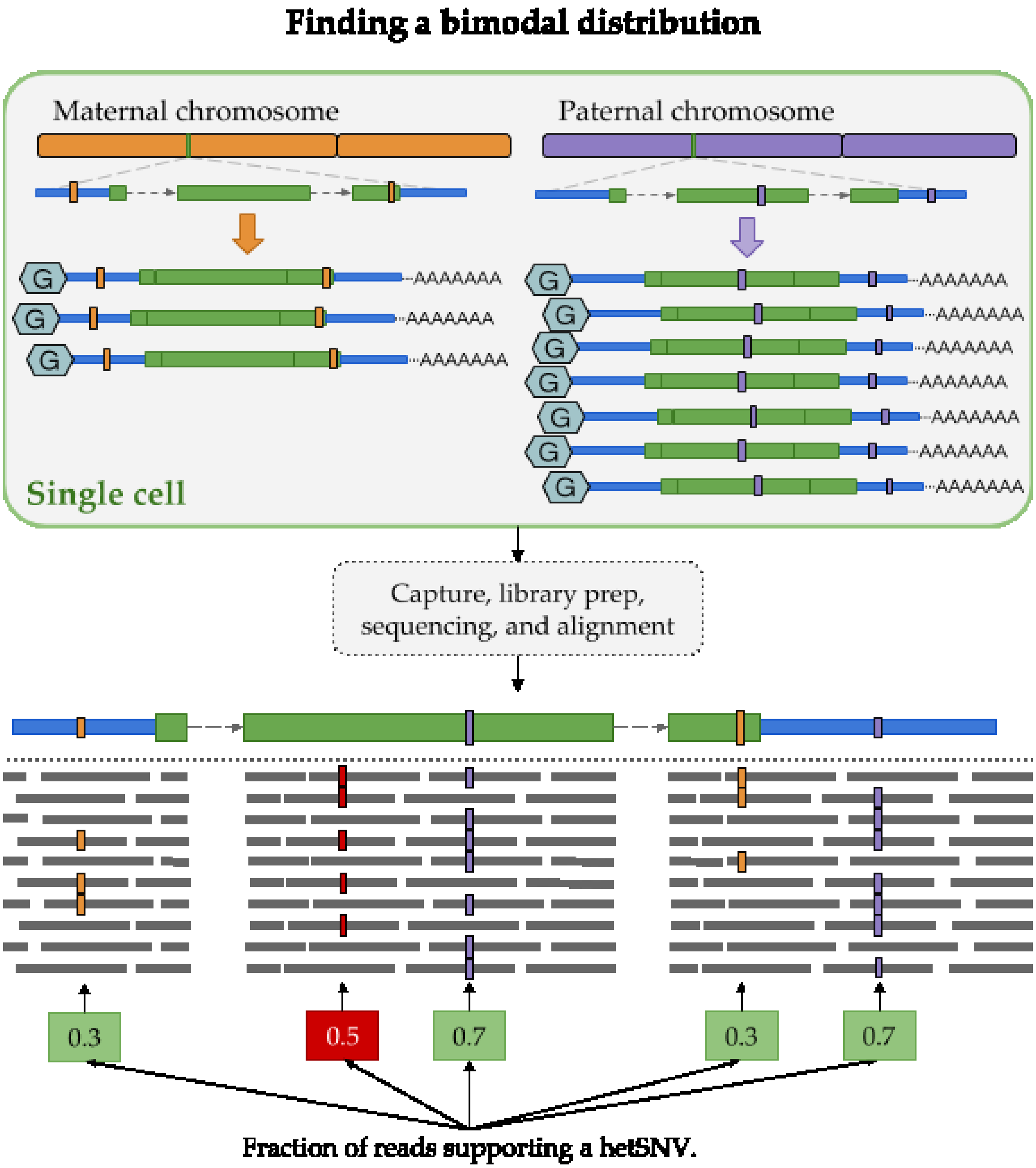
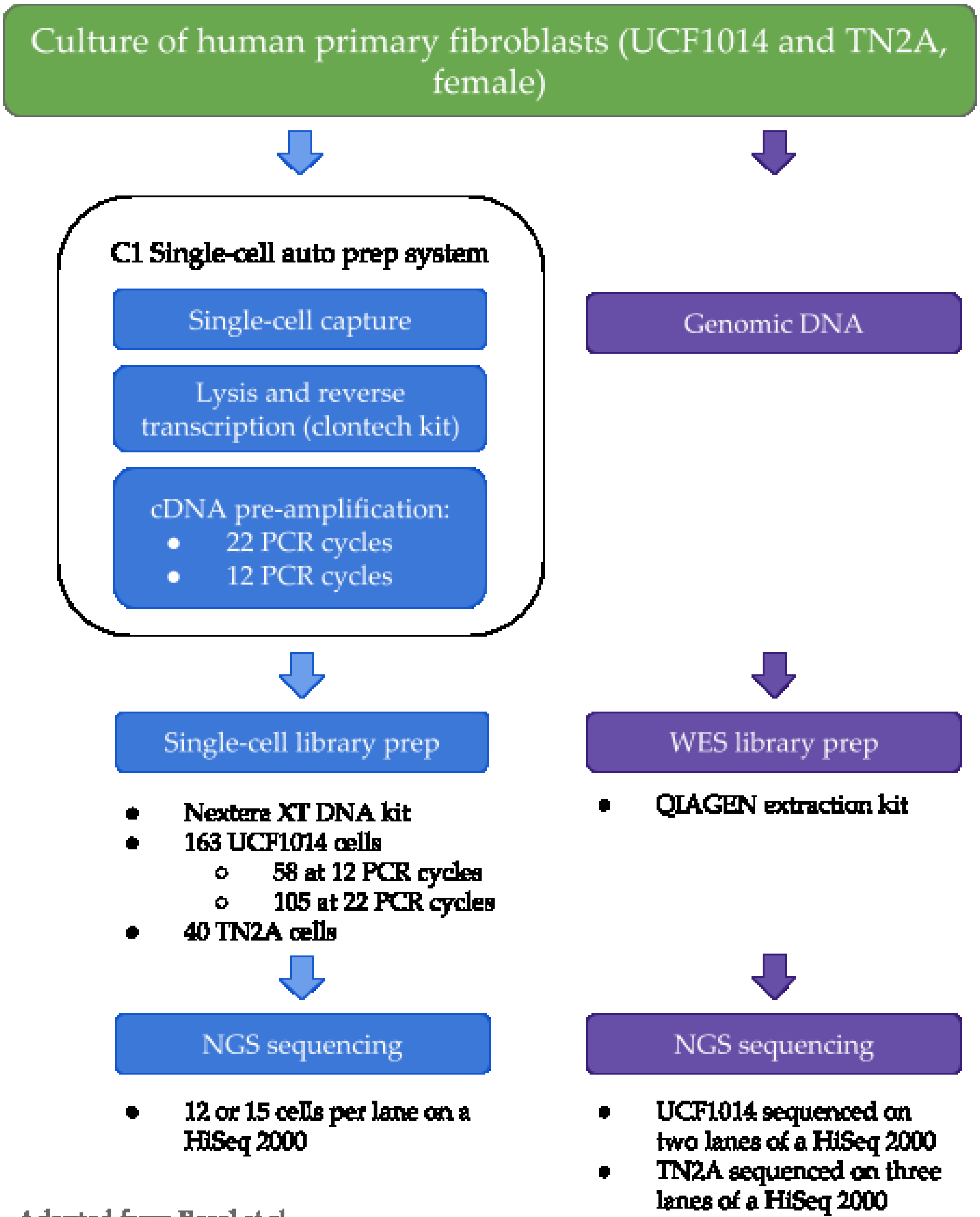

Finding a bimodal distribution
Maternal chromosome
Paternal chromosome
G
AAAAAAA
Single cell
Capture, library prep, sequencing, and alignment
0.3
0.5
0.7
0.3
0.7
Fraction of reads supporting a hetSNV.

Figure 1. Finding a bimodal distribution. Any variant (green box) that fits into the expected distribution of reads stays. Any that do not are removed: here the variant existing at a fraction of 0.5 (red box) would be removed.


Culture of human primary fibroblasts (UCF1014 and TN2A, female)
C1 Single-cell auto prep system
Single-cell capture
Lysis and reverse transcription (clontech kit)
cDNA pre-amplification:
• 22 PCR cycles
• 12 PCR cycles
Genomic DNA
Single-cell library prep
• Nextera XT DNA kit
• 163 UCF1014 cells
○ 58 at 12 PCR cycles
○ 105 at 22 PCR cycles
• 40 TN2A cells
WES library prep
• QIAGEN extraction kit
NGS sequencing
• 12 or 15 cells per lane on a HiSeq 2000
NGS sequencing
• UCF1014 sequenced on two lanes of a HiSeq 2000
• TN2A sequenced on three lanes of a HiSeq 2000

Adapted from Borel *et al.*

Figure 2. UCF1014 and TN2A sequencing strategy. Bulk WGS is paired with scRNA-seq for the two primary fibroblast cell lines. The library prep for the single cells was performed using the original Smart-seq protocol.

strategy for sample preparation. The amplification kit used to generate the cDNAs of the transcripts was the SMARTer Ultra Low RNA Kit for Illumina Sequencing (Clontech). This is crucial because, while many other kits create transcripts of only the 3′ ends, this kit generates cDNAs of the full transcripts, and for our study the ability to look at the entire transcript for variants—not just the 3′ end—is necessary. This amplified cDNA was turned into a proper mRNA-seq library using the Nextera XT DNA Kit, following which the cells were multiplexed with 12 or 16 samples per lane and sequenced on a HiSeq 2000.

In addition to the above, bulk WGS was performed for both the TN2A and UCF1014 cell lines. Genomic DNA was harvested the same day as the single cells using the QIAGEN kit, followed by library prep using the Illumina TruSeq DNA Kit. Two lanes on the HiSeq 2000 were allotted for genomic DNA of the UCF1014 cell line and three lanes were used for the TN2A cell line.

Validation requires both genomic DNA and scRNA-seq data from the same source. Variant calling of genomic NGS data, especially whole genome (as opposed to whole exome), is a very well-established practice and can determine True Positives, False Positives, True Negatives, and False Negatives when identifying putative SNVs in the scRNA-seq dataset. Also, while Borel *et al*. did perform variant calling on this data, it was determined that the methods used to perform these analyses were outdated and we opted to generate our own list using modern methods. Specifically, the Best Practices outlined by the Broad Institute to identify variants in RNA-seq data[37] was followed, and the bcbio-nextgen pipeline, which implements the Best Practices outlined by the Broad

Institute, was used to identify genomic variants[38]. We are uniquely qualified to assess the quality of the type of bioinformatic analysis used due to previous work we have published on assessing the quality of various variant calling pipelines[32]. It is important to perform variant identification correctly, since these are the data used during software development. The goal of this experiment was to compare any variants identified in the scRNA-seq data by Red Panda and the other four variant callers to those found in the dataset derived from the WGS results. Any SNVs that are present in both datasets are True Positives; SNVs found in the genomic dataset and not the scRNA-seq dataset (assuming appropriate read coverage in the scRNA-seq dataset) will be considered False Negatives; SNVs found in the scRNA-seq dataset and not the genomic dataset will be considered False Positives; and SNVs not found in either dataset will be classified as True Negatives.

The bcbio-nextgen version (v.) 1.0.3 pipeline was used for variant calling to align reads and identify variants in the bulk genome data. Reads were aligned to the human genome v. 38 (hg38) using BWA MEM v. 0.7.15. Following this, three variant callers were used to identify SNVs and indels: FreeBayes (v. 1.1.0)[24], GATK HaplotypeCaller (v. 3.7.0)[25], and Platypus (v. 0.8.1)[27]. Only variants identified by at least two out of the three algorithms were kept as this has been shown to work well[39]. Following this, MultiQC v. 1.0.dev0[40] was run to aggregate QC statistics from bcbio-nextgen, samtools v. 1.4[41], bcftools v. 1.4, and FastQC v. 0.11.5[42]. **Table 1** displays the number of reads, coverage, mapping rates, number of SNVs identified, and the average number of SNVs per gene for the genomic DNA. Of particular note are the mean number of heterozygous variants

Sample	Reads	Coverage	Map Rate	hetVariants found in genes	Mean hetVariants found per gene
UCF1014	8,460,80,874	27.9x	98.9%	105,779	5.32
TN2A	14,40,893,964	47.5x	99.0%	110,412	5.56

Table 1. Genomic sequencing statistics for UCF1014 and TN2A. Sequencing and analysis statistics of the WGS from the two cell lines, UCF1014 and TN2A, to show that the quality is appropriate as well as that there are enough hetVariants per gene to evaluate the concept unique to Red Panda.

(hetVariants) found per gene. For Red Panda to work, it was necessary that there be enough heterozygous variants available in a gene to establish a bimodal distribution in the first place. With ~5-6 hetVariants per gene—this number naturally increases or decreases with the length of the gene—it was determined that there were enough to continue with development of our software.

Once the genomic data had been analyzed, testing began on the scRNA-seq data, specifically the TN2A cells, for quality. Testing the number of reads aligning outside of exons found that the number was unexpectedly high: 41.70%. To determine if this was a normal number, ten samples from each of six other datasets were investigated[43–48]. Three datasets using cancer samples and three datasets using non-cancer samples (i.e., normal) were chosen. Additionally, because the human fibroblast samples used version 1 of the Smart-seq protocol, we checked the alignment rates of reads outside exons in Smart-seq2[49] to check its performance. As can be seen in **Figure 3**, Smart-seq2 did perform better. This matches expectations, as Picelli *et al.* proved that the newer protocol produces longer and more complete transcripts[49].

Due to the degree to which Smart-seq2 outperforms Smart-seq, it was determined that it would not be appropriate to develop software using the fibroblast data since it was generated using the inferior capture kit. Instead, data were generated in-house using the Smart-seq2 protocol.

Articular chondrocyte sequencing

Since it was determined that Smart-seq2 produced higher quality data, in-house sequencing was performed in collaboration with Dr. Andrew Dudley and his graduate

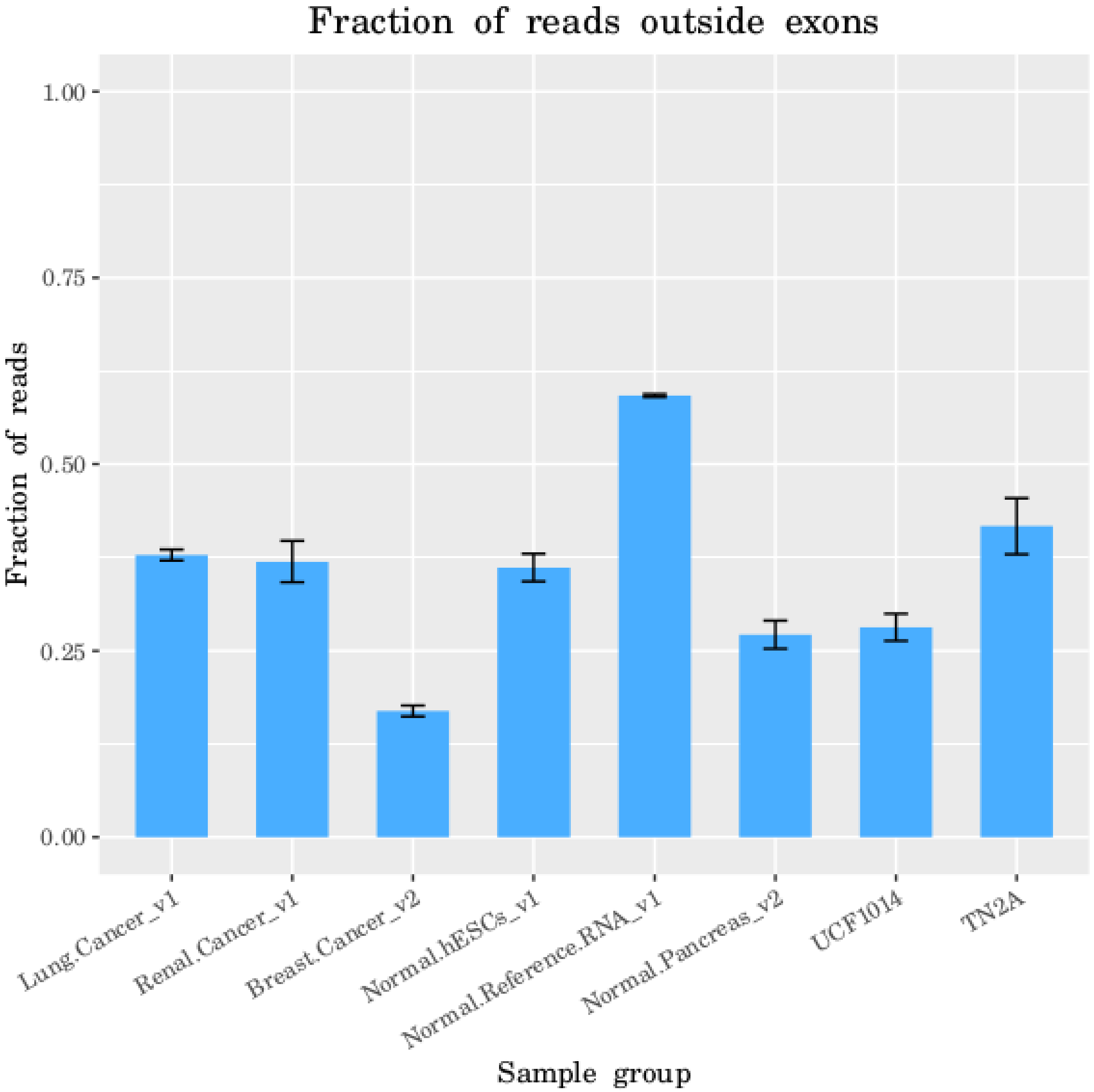


Figure 3. Average fraction of reads aligned outside exons using two versions of Smart-seq protocols. Here, v1 refers to version 1 of the Smart-seq protocol and v2 refers to Smart-seq2. Both samples using the Smart-seq2 protocol contain a lower fraction of reads outside exons than the other samples.

student Krishna Sarma. This dataset met all of the four previously-mentioned criteria for testing the software. The sequencing strategy employed for this sample can be seen in **Figure 4.**

Krishna Sarma processed the articular chondrocytes harvested from a female patient of Caucasian origin undergoing total knee replacement, who provided informed consent prior to the study. Beau S. Konigsberg (orthopedic surgeon), Dillon R. Ellis (Clinical research associate), and Dana M. Schwarz (Research nurse coordinator) (IRB #691-13-EP) at UNMC approved this tissue for use. Human articular chondrocytes were specifically chosen because they are the only cell type present in the cartilage of the human knee and because they are locked in G_0[50]. This means that during software development aberrations in isoform expression would be minimized due to the absence of different cell types and cells in different phases of their cell cycle.

Cells were extracted from shavings of articular cartilage through sequential digestion in .2% Pronase (Roche) for 2 hours followed by overnight digestion in .2% collagenase (Gibco), all while shaking at 37°C. Cell suspensions were passed through 70μM cell strainers (BD Falcon) and centrifuged at 500xG for 10 minutes to recover chondrocytes. The cells were subsequently embedded in three-dimensional alginate bead cultures at a final concentration of about 75 million cells per mL. The cultures were maintained in DMEM/F12 (1:1) media supplemented with 1% penicillin-streptomycin-glutamine (Invitrogen, 10378-016), Amphotericin B (Gibco, 15290026), insulin-transferrin-sodium selenite (Sigma, I2771), 50μg/mL Vitamin C, 10ng/mL FGF2, and 10ng/mL TGF-bb3 (PeproTech®, 100-36E) and maintained at 37°C in a 5% CO2

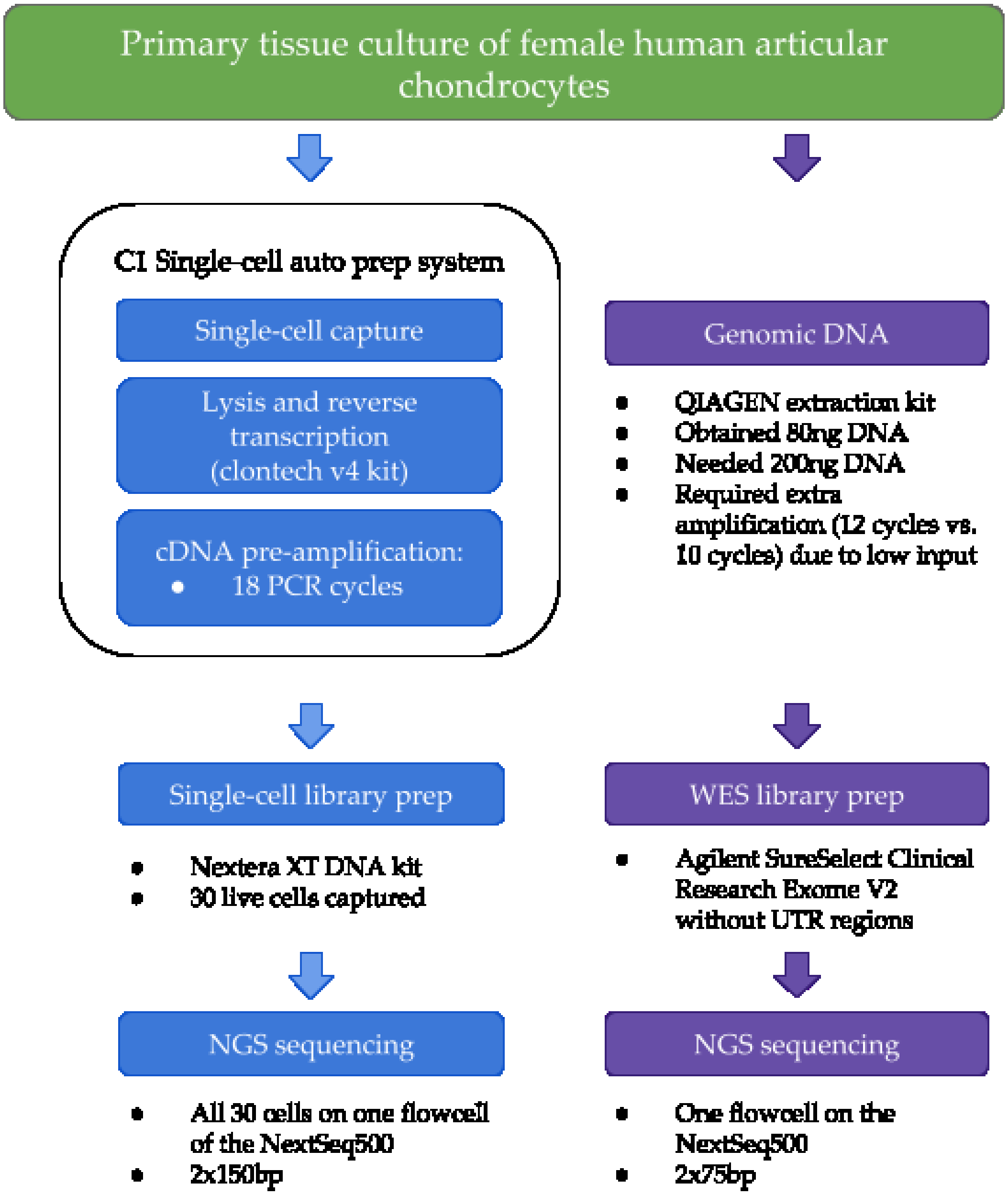

Primary tissue culture of female human articular chondrocytes
C1 Single-cell auto prep system
Single-cell capture
Lysis and reverse transcription (clontech v4 kit)
cDNA pre-amplification:
• 18 PCR cycles
Genomic DNA
• QIAGEN extraction kit
• Obtained 80ng DNA
• Needed 200ng DNA
• Required extra amplification (12 cycles vs. 10 cycles) due to low input
Single-cell library prep
• Nextera XT DNA kit
• 30 live cells captured
WES library prep
• Agilent SureSelect Clinical Research Exome V2 without UTR regions
NGS sequencing
• All 30 cells on one flowcell of the NextSeq500
• 2x150bp
NGS sequencing
• One flowcell on the NextSeq500
• 2x75bp

Figure 4. Human articular chondrocyte sequencing strategy. Exome sequencing was paired with scRNA-seq from the primary tissue culture of human articular chondrocytes. The library prep for the single cells was performed using the updated Smart-seq2 protocol.

atmosphere for 14 days. The day before single-cell capture, cells were lysed using Trizol® reagent (Life Technologies) according to the manufacturer’s protocol. These cells were split into two groups for DNA and RNA extraction.

Exome sequencing

Krishna Sarma performed the cell prep and DNA extraction on cells harvested the same day as the single-cell capture. Genomic DNA was extracted using the QIAGEN DNA extraction kit using the manufacturer's instructions. The UNMC Sequencing Core Facility performed DNA prep. Due to the low amount of DNA (80ng) captured by the QIAGEN kit, instead of the normal 10 PCR amplification cycles, 12 were performed prior to library preparation to obtain enough DNA. Therefter, the Agilent SureSelect Clinical Research Exome V2 kit was used to capture coding regions on the genome and generate a library. Notably, the Clinical Research Exome V2 kit used does not include 5’-UTR and 3’-UTR regions which limits what can be compared between the exome and the scRNA-seq data, the latter of which will naturally have coverage in those regions. The exome library was sequenced on two lanes of the NextSeq500 using 75 base pair paired-end sequencing.

The bcbio-nextgen v. 1.0.3 pipeline was used for germline variant calling to align reads and identify variants in the bulk exome data. For this analysis, the pipeline was run on human genome v. 38 (hg38). The aligner BWA MEM v. 0.7.15 was used to align reads to the human genome. Results of this can be found in **Table 2**. Following this, three variant callers were used to identify SNVs and indels: FreeBayes (v. 1.1.0), GATK

Total Reads	Paired Reads	PCR Duplicate	Alignment	Coverage	On-target rate
182M	111M (61%)	38.3%	98.7%	74.67x	55%

Table 2. Human articular chondrocyte exome sequencing statistics. Sequencing and analysis statistics of the exome data from the human articular chondrocytes.

HaplotypeCaller (v. 3.7.0), and Platypus (v. 0.8.1). After variant calling, only those identified by at least two out of the three algorithms were kept. MultiQC v. 1.0.dev0 was run to aggregate QC statistics from bcbio-nextgen, samtools v. 1.4, bcftools v. 1.4, and FastQC v. 0.11.5.

The output of the exome sequencing resulted in good coverage of the coding exons, however the percentage of reads that were PCR duplicates was particularly high, likely due to the low amount of starting DNA. Normally 200ng of DNA is used during library preparation, but only 80ng could be extracted which required two extra cycles of PCR amplification. The ratio of homozygous to heterozygous variants is in line with what one would expect from this type of sequencing[32,51,52], however the numbers were high since, originally, variants within 100bp of the coding exon boundary were included. This means that variants were identified in the intronic regions as well as the exonic region, but since this dataset is being used to compare against scRNA-seq, only variants found within exons were included. This led to numbers more in line with what is expected from exome sequencing as seen in the second row in **Table 3**[32].

Single-cell RNA sequencing

The UNMC Sequencing Core Facility performed capture and sequencing. For single-cell capture, 735 cells were loaded on to a 10-17 μm Fluidigm C1 Single-Cell Auto Prep IFC (with 96 wells), and the cell-loading script was performed using the manufacturer's instructions. Each of the 96 capture sites were inspected under a confocal microscope to remove sites containing dead cells as identified by the LIVE/DEAD Cell

Bed file contains 100bp ± coding exons	Total Variants	Homozygous Variants	Heterozygous Variants	Ratio	SNVs	indels
Yes	85,128	33,856	45,769	0.74	79,627	5,504
No	20,315	7,777	12,538	0.62	20,057	258

Table 3. Human articular chondrocyte exome variant calling statistics. Variant analysis statistics of the exome data from the human articular chondrocytes using the ensemble approach where 2/3 variant caller tools had to agree to call a variant.

Viability Assay and to remove capture sites containing more than one cell. Cells that were not identified as either alive or dead by the LIVE/DEAD assay were retained for RNA sequencing.

Following capture, reverse transcription and cDNA amplification were performed in the C1 system using the Clontech SMARTer Ultra Low Input RNA Kit for Sequencing v3 which was done according to the manufacturer's instructions. Only 27 Single-cell cDNA libraries were obtained at a concentration of 0.09 to 0.55 ng/µl. Three libraries were below a concentration of 0.08 ng/µl which may have been dying cells and did not have a LIVE/DEAD staining. They have "NC" attached to their sample name signifying "No Color". The majority of failed cells on the capture plate were either a single dead cell (37) or a combination of live and dead cells (17) as seen in **Table 4** and **Figure 5**. Amplification was performed using the Nextera XT DNA Sample Preparation Kit and the Nextera XT DNA Sample Preparation Index Kit (Illumina) was used for indexing. After quantification using an Agilent Bioanalyzer, sequencing was performed on two lanes of the NextSeq500 using 150 base pair paired-end sequencing.

The bcbio-nextgen v. 1.0.3 pipeline was used for RNA-seq to align reads and perform transcript quantification for each of the cells. For this analysis, the pipeline was run twice, once on human genome v. 19 (hg19) and once on human genome v. 38 (hg38). For hg19 STAR v. 2.5.3a was used to align reads to the human genome; however, hisat2 v. 2.0.5 was used to align reads to hg38 due to its ability to correctly handle the alt alleles present in that version of the human genome. Following this, MultiQC v. 1.0.dev0 was run to aggregate QC statistics from bcbio-nextgen, samtools v. 1.4, QualiMap v. 2.2.2a[53],

Live Cells	Dead Cells	No Color (NC)	Live and Dead	>1 Live	>1 Dead	Empty
27	37	3	17	2	3	7

Table 4. Summary of the cells captured on the C1.

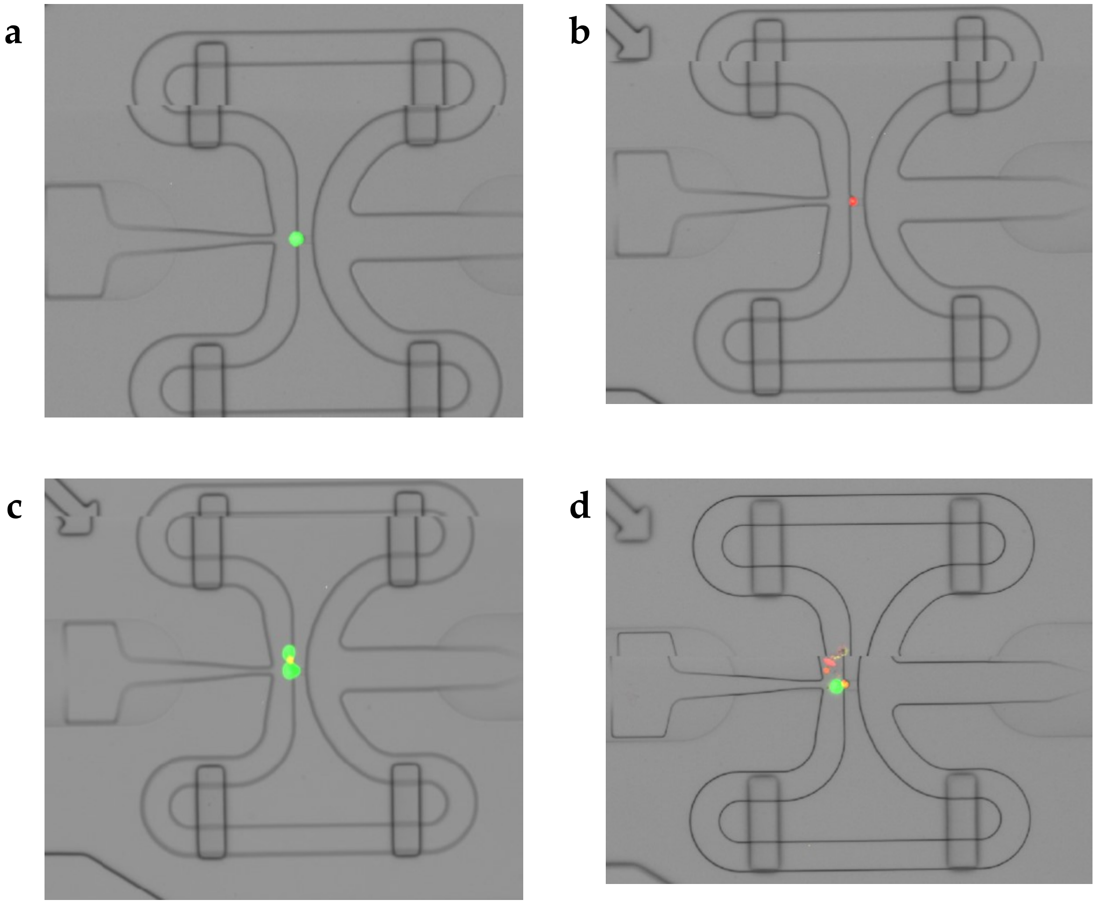


Figure 5. Four capture sites on the C1 chip. Here we see **(a)** one cell stained as LIVE, **(b)** one cell stained as DEAD, **(c)** three cells: two stained as LIVE, and one as DEAD, **(d)** cell debris, one cell stained as LIVE and one stained as DEAD. The latter three illustrate the type of difficulties present when trying to capture the articular chondrocytes.

and FastQCv. 0.11.5. In addition to performing this type of analysis on each cell individually, it was also performed on two bulk samples: one in which all 30 cells were pooled together, and also a smaller pool of 26 cells where four (A3-C1NC, C10-C64, D12-C72, and H7-C46) were removed for quality reasons. Full alignment statistics can be found in **Table 5**. Once alignments had been performed, the genomic origin of the reads in each cell was assessed. Any cells that had more than 30% of their reads originating from outside exons are considered poor quality and will not be considered for further analysis. This results in excluding the cells A3-C1NC, C10-C64, D12-C72, and H7-C46 as seen in **Figure 6**. Since this is paired-end sequencing, the insert size of each fragment was also calculated to see if there were any outliers. As seen in **Figure 7**, the cells A3-C1NC, C10-C64, D12-C72, and G2-C38NC all have significantly smaller insert sizes than the rest of the cells. Interestingly, three of these are the same as those cells that have reads whose origin is largely intronic or intergenic.

To further assess the quality of the cells that were sequenced, the expression of the transcripts needed to be assessed. As these cells are all of the same type and in the same stage of the cell cycle, their expression profiles should be highly correlative. To generate raw counts of genes expressed in each cell, htseq v. 0.6.1[54] was run using default parameters. Normalized quantification was then performed using sailfish v. 0.10.1[55]. Following this, custom scripts were used to generate a matrix containing the expression counts generated by htseq for each gene in each cell and the two pooled sample groups. In total there were 33 columns, one for the gene name, 30 for each cell, one for the high quality data of cells (pooled26), and one for the total batch

Sample Name	Reads	% Dup	rRNA pct	5'-3' bias	% GC
A3-C1NC	9.80 M	24.8%	3.8%	4.24	49%
A7-C6	11.04 M	25.3%	0.6%	1.18	47%
A8-C5	9.69 M	16.9%	0.7%	1.12	46%
B1-C9	9.50 M	16.9%	0.6%	1.15	47%
B12-C60	7.24 M	22.5%	0.9%	1.32	47%
B3-C7	14.80 M	15.4%	0.4%	1.21	46%
B6-C57	11.47 M	24.0%	0.6%	1.20	46%
B7-C12	11.62 M	19.4%	0.5%	1.14	45%
B8-C11	8.10 M	24.3%	1.7%	1.28	47%
C10-C64	11.76 M	17.2%	2.8%	2.66	46%
C11-C65	7.80 M	19.8%	0.4%	1.20	47%
C12-C66	0.08 M	1.5%	0.7%	1.32	53%
C5-C62	10.55 M	19.4%	0.3%	1.22	46%
C8-C17	16.34 M	25.7%	0.5%	1.14	46%
D11-C71	10.37 M	22.3%	0.4%	1.14	47%
D12-C72	6.79 M	20.6%	5.4%	3.24	47%
E1-C25	9.89 M	23.1%	0.5%	1.13	46%
E11-C77	9.59 M	26.7%	0.6%	1.14	46%
E2-C26	7.84 M	19.7%	0.6%	1.13	45%
E4-C75	7.06 M	18.3%	3.0%	1.38	48%
E5-C74	5.86 M	17.1%	0.4%	1.13	47%
F2-C32	9.33 M	18.9%	0.3%	1.18	46%
F3-C33	9.96 M	16.7%	0.5%	1.10	46%
G1-C37	10.75 M	17.7%	0.3%	1.19	45%
G2-C38NC	7.65 M	27.2%	0.7%	7.60	50%
G5-C86	7.65 M	17.2%	0.4%	1.20	46%
G8-C41	13.05 M	21.0%	0.6%	1.10	46%
H4-C93NC	8.57 M	19.2%	0.3%	1.12	47%
H6-C91	7.12 M	17.1%	0.5%	1.14	46%
H7-C46	5.43 M	17.8%	15.1%	1.76	52%

Table 5. Alignment statistics for the 30 cells captured on the C1.

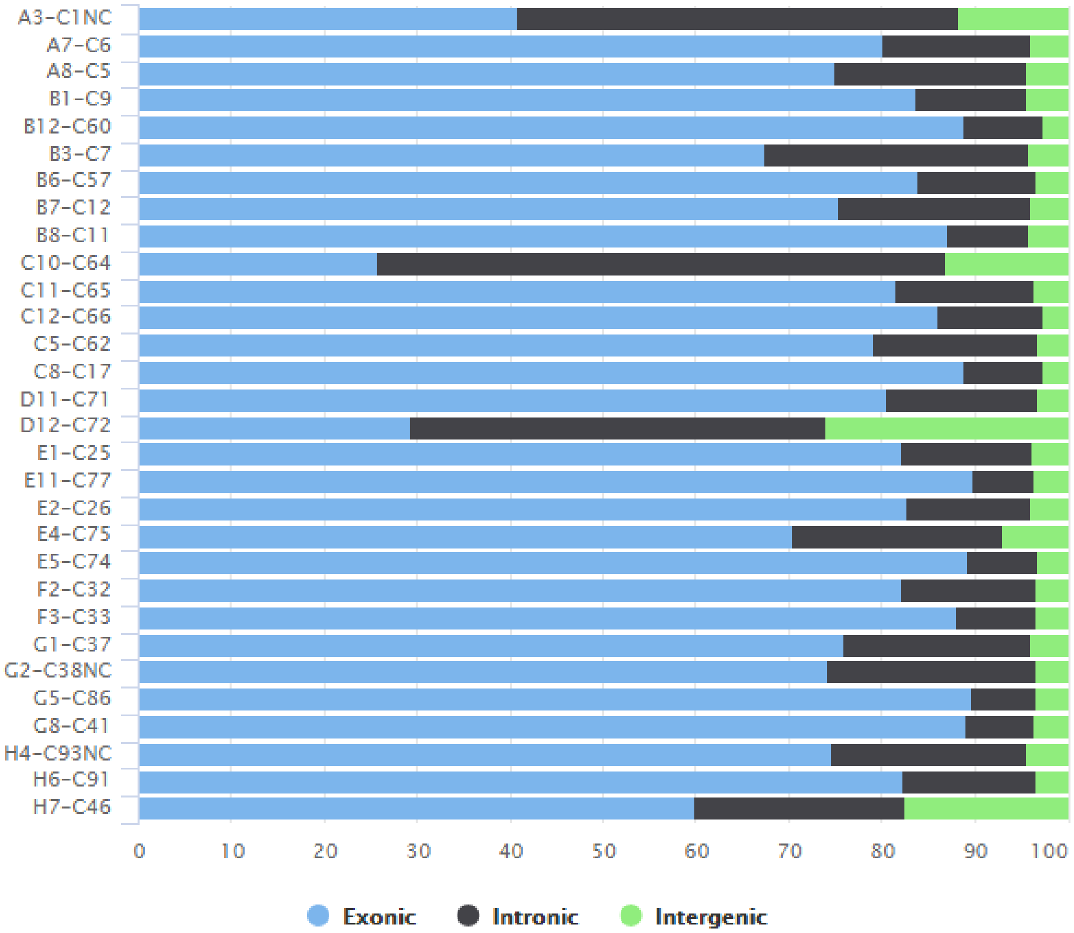


Figure 6. The genomic origin of reads found in each cell. Here one can see what percentage of reads originate from exons (blue), introns (black) or intergenic space (green). The cells A3-C1NC, C10-C64, D12-C72, and H7-C46 have significantly more reads originating outside the exonic region than other samples.

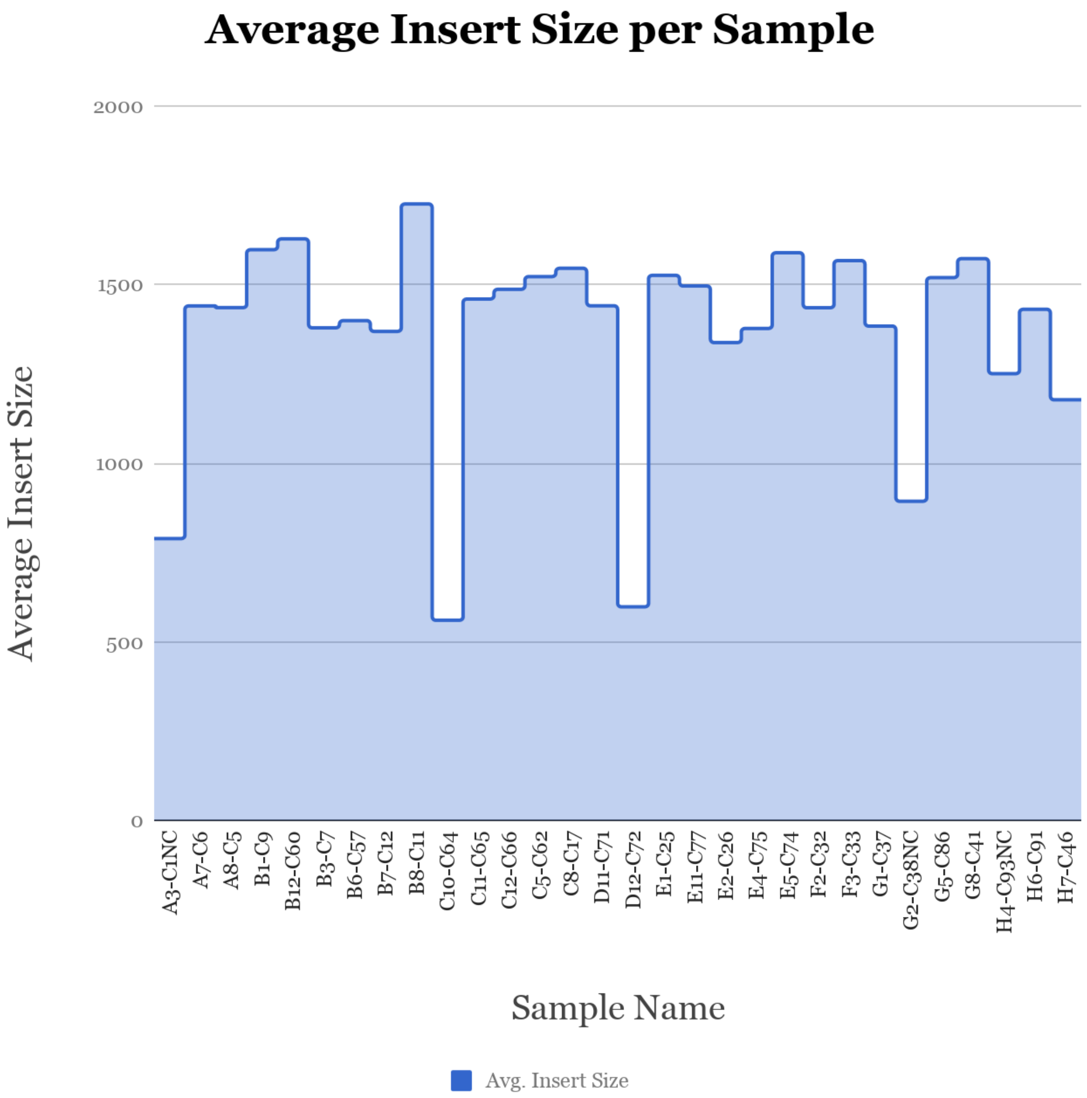


Figure 7. The average insert size for each paired read for each cell. The average insert size of the pair-end fragment is calculated from the alignment. The cells A3-C1NC, C10-C64, D12-C72, and G2-C38NC have significantly smaller average insert sizes than other samples.

of cells (pooled30) as seen in **Figure 8**. Here, pooled26 contains all cells excluding the four that have been previously identified as low quality due to genomic origin of the reads present: A3-C1NC, C10-C64, D12-C72, and H7-C46**.** Using this matrix as input, cormat[56] was used in R to generate a Pearson's correlation coefficient for all pairwise comparisons between the cells as well the two pooled samples. Following this, ggplot2[57] was used generate a heat map of these comparisons.

From the correlation data, one can clearly see the four poor quality cells not correlating to the rest of the batch as well as identify three other cells that do not correlate well based on their expression patterns: H4-C93NC, G2-C38NC, and E4-C75. All three (0.31, 0.39, and 0.35) are well below the median of 0.785 for all 30 cells. Because these three do not correlate well with the rest of the samples, they will not be used for further analysis. Unsurprisingly, all three cells that did not have a LIVE/DEAD stain, labeled as NC, have now been removed for quality reasons.

Scatter plots were generated for all five of the samples that did not correlate with the rest of the cells and one sample that did correlate well with the rest of the cells, and they can be seen in **Figure 9**. From these it is clear that the expression profile of the cell that has a high correlation coefficient, E2-C26, is much more tightly clustered along the diagonal than the other samples.

The last method for removing low quality samples was to check the total read count. Confidence Intervals (CI) were originally used to determine if there were enough reads in a sample for it to be kept, and those intervals can be seen in **Table 6**. Instead of arbitrarily removing sample C12-C66 because it had too few reads (80k vs. millions for

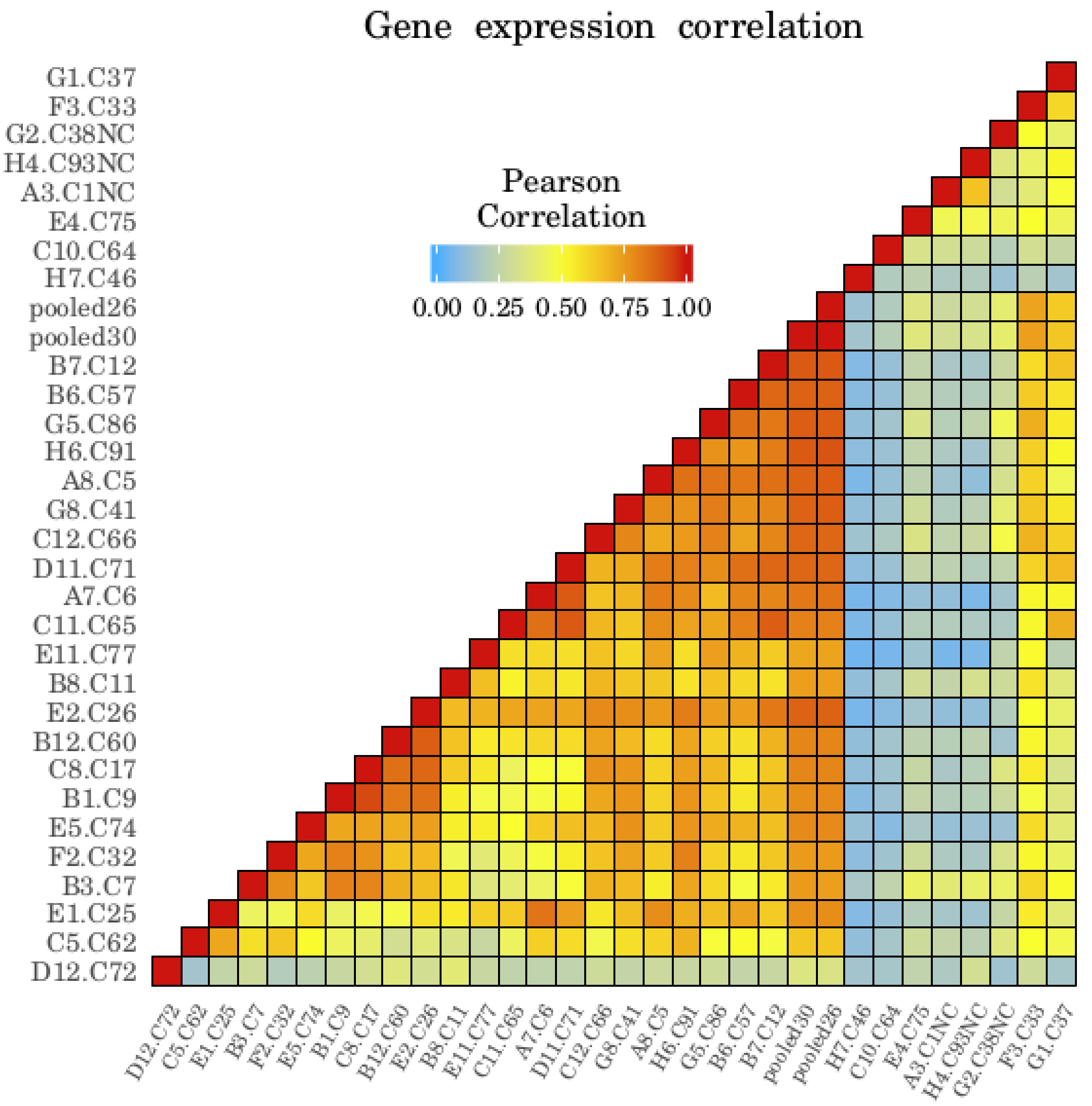


Figure 8. Expression correlation between articular chondrocytes. Pearson Correlation Coefficient calculated for every possible comparison of cells to each other and the two batches of cells. The darker the color red, the higher the correlation between each cell. One can clearly see the four poor quality cells not correlating to the rest of the batch as well as identify three other cells that do not correlate well based on their expression patterns: H4-C93NC, G2-C38NC, and E4-C75.

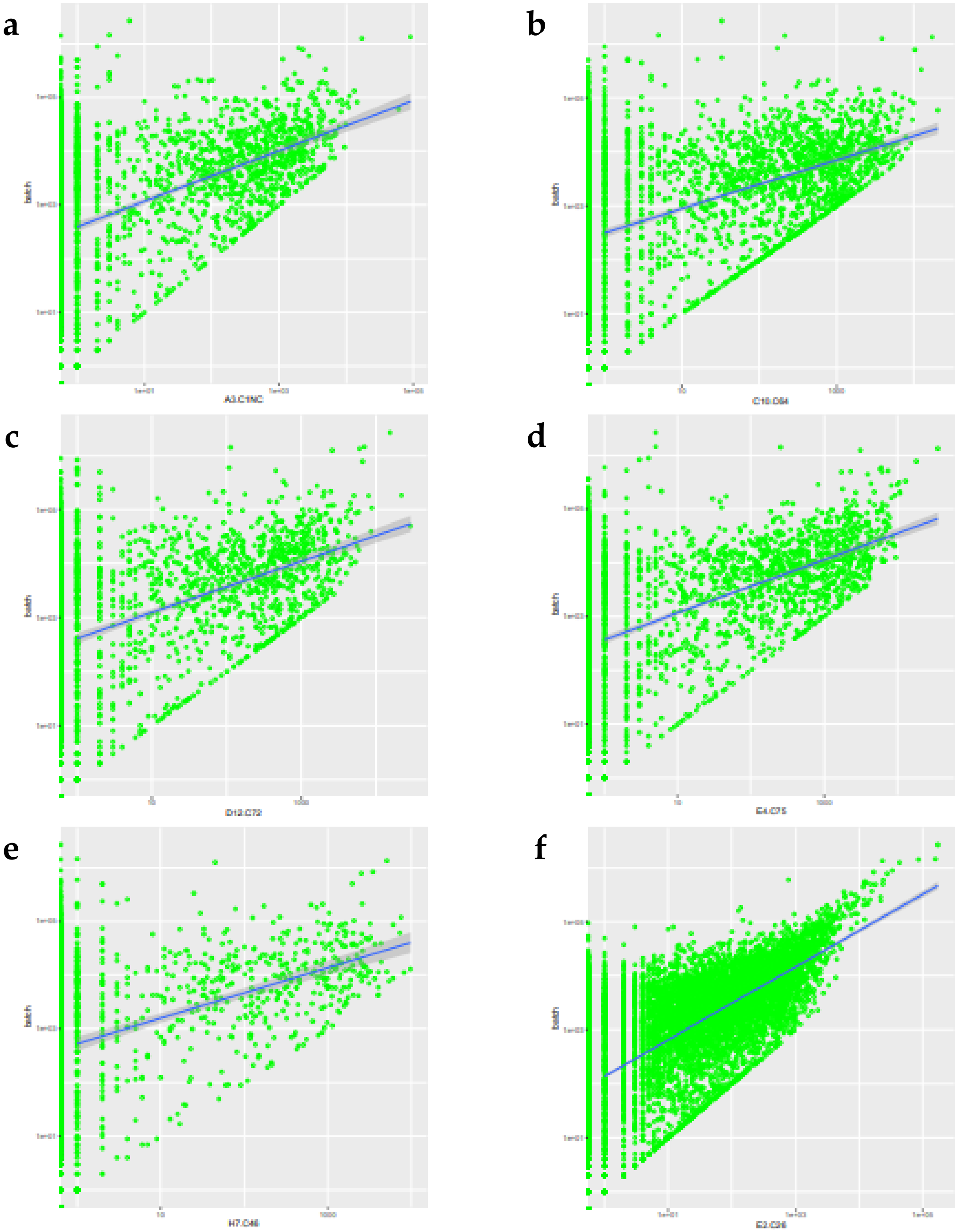


Figure 9. Scatter plots generated based on expression. Five cells with low correlation and one cell with high correlation as compared to the total batch. **(a)** A3-C1NC, R = 0.31 **(b)** C10-C64, R=0.19 **(c)** D12-C72, R=0.34 **(d)** E4-C75, R=0.35 **(e)** H7.C46, R=0.13 **(f)** E2-C26, R=0.88

	n	Mean	Std Dev.	Confidence	CI	Samples outside CI	# above upper-bound
Confidence intervals for reads generated per sample.							
95% CI	30	9.22	3.01	2.41	9.22 (7.26–11.19)	12	5
99% CI	30	9.22	3.01	3.17	9.22 (6.64–11.81)	5	3
Confidence intervals for reads **mapped** per sample.							
95% CI	30	7.37	3.25	3.17	7.37 (4.96–9.78)	11	5
99% CI	30	7.37	3.25	2.41	7.37 (4.20–10.54)	8	3

Table 6. Confidence intervals for number of reads. Confidence intervals for reads generated per sample and reads mapped per sample. This data was used to determine if a statistical cutoff based on confidence intervals could be calculated.

the other samples), confidence intervals were used for the total number of reads generated per sample and total number of reads mapped per sample.

Unfortunately it does not make a lot of sense to use confidence intervals here since they imply that a sample falling outside the CI is an aberration in the population of data and should be discarded; however that is inappropriate logic since more sequencing data per sample is good thing even though those samples fall outside the CI. Given that, it makes sense to use 500k reads as a lower-bound cutoff since that is what one aims for in experiments using the Smart-Seq2 protocol[58]. This cutoff only removes sample C12-C66 as it is the only one that has fewer than 500k reads. All of the above filtering and QC steps have removed eight cells as seen in **Table 7**, leaving us with 22 cells of high quality data.

After establishing which samples to remove due to quality reasons, it became necessary to prove that there exist transcripts that support the model that spawned our algorithm. Such a proof of concept exists in **Figure 10** where four variants were found in the scRNA-seq data, but only three in the exome data. In cell A7-C6, gene CWC22, there are two heterozygous variants and one homozygous variant. In the scRNA-seq two of the heterozygous variants support each other when you add their fraction of reads together (0.2 and 0.8 together add up to 1 as one would expect). However, there is another putative heterozygous variant where 53% of the reads contribute to its existence, but since this doesn’t fit what is expected (i.e., either 20% or 80%), it would be discarded. Corroborating this is the fact that no such variant exists in the exome data. Also seen is the homozygous variant in both the scRNA-seq data and the exome.

Sample Name	Reason for removal
A3-C1NC	Too many reads outside exon; Poor correlation coefficient
C10-C64	Too many reads outside exon; Poor correlation coefficient
C12-C66	Number of reads < 500k
D12-C72	Too many reads outside exon; Poor correlation coefficient
E4-C75	Poor correlation coefficient
G2-C38NC	Poor correlation coefficient
H4-C93NC	Poor correlation coefficient
H7-C46	Too many reads outside exon; Poor correlation coefficient

Table 7. Reasons for removing eight samples from further analysis.

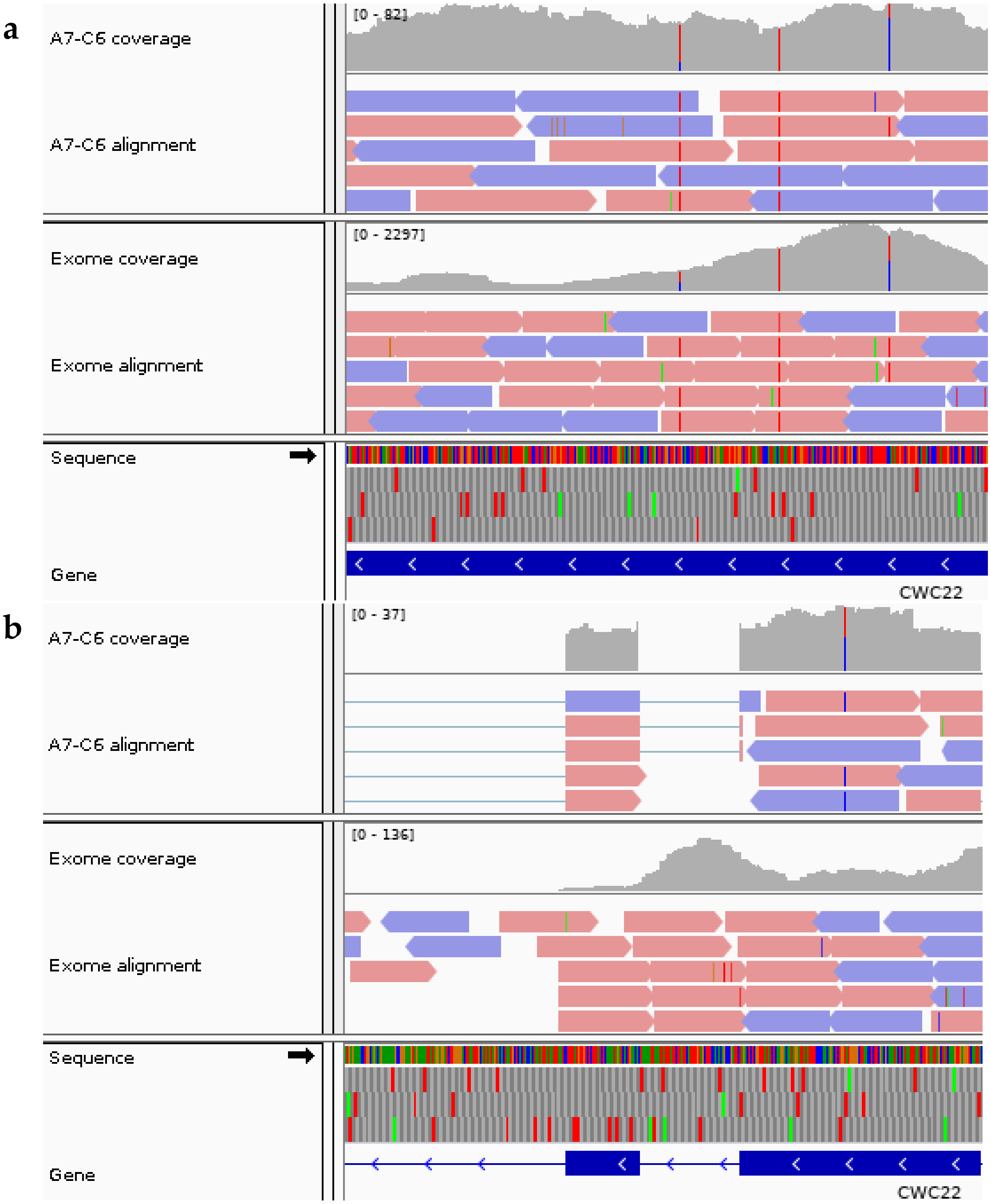


Figure 10. Proof of concept data in articular chondrocytes. An example of the variations, from gene CWC22, that we find in the scRNA-seq data as compared to the exome. The main area of interest is the coverage track (the gray histograms). Red corresponds to T and blue corresponds to a C. When there are two colors, the top color corresponds to the alternate allele. **(a)** Two hetSNVs found in the cell A7-C6 have reads supporting them at percentages of 80% (left) and 20% (right). The same hetSNVs are found in the exome data at 50%. There is also a homozygous variant (middle) seen in both. **(b)** One hetSNV found in the same gene at 53% in the cell A7-C6 is absent in the exome sequencing. This is expected as it does not fit the existing biomodal distribution at 80% or 20%.

Algorithm Development

Logic

With an established dataset in hand, algorithm development commenced. The basic workflow can be found in **Figure 11**. For each cell, it is first determined which transcripts are expressed. This is done using the quantification numbers provided by running sailfish v. 0.10.1 to filter out transcripts with a Transcripts Per Million (TPM) value < 1. Once a list of expressed transcripts is established, samtools mpileup v1.4 is used to generate a list of every possible variant contained within the provided alignment file for this cell.

This list is then broken down into two lists containing variants that are likely heterozygous or likely homozygous. Heterozygous variants are first filtered to exclude those where, proportionally, few reads support their existence:

1. Remove variants where the fraction of reads supporting them is <20% and read depth is <20x.
2. Remove variants where the fraction of reads supporting them is <10% and read depth is <40x.
3. Remove variants where the fraction of reads supporting them is <5%.

Then, the remaining are checked to see if a potential bimodal distribution exists where the fraction assigned to each mode adds up to 1, and if such a distribution does exist, remove all heterozygous variants that do not fit, assuming a tolerance of 5%. For example, if heterozygous variants are expected to be either 30% or 70%, anything falling

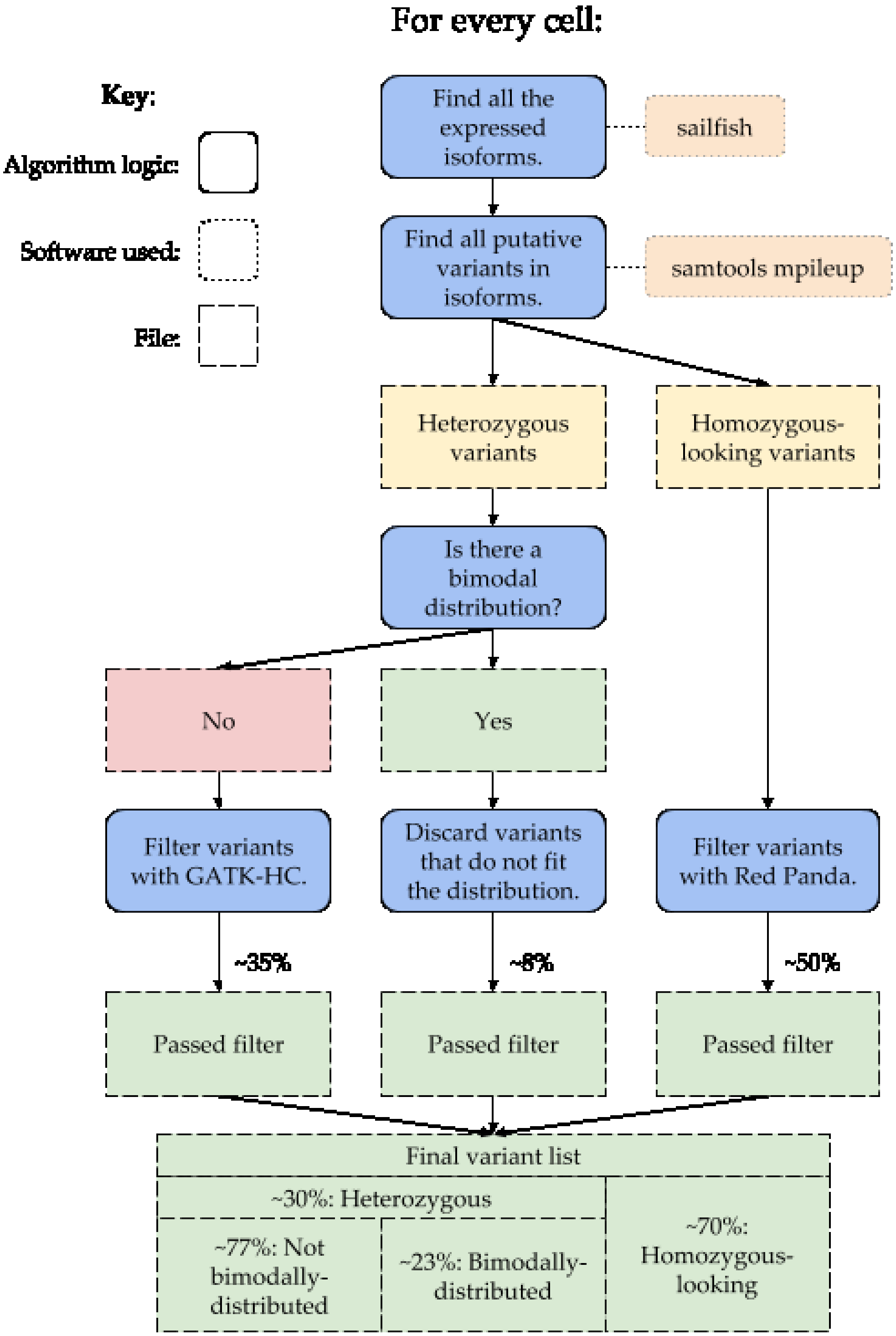

For every cell:
Key:
Algorithm logic:
Software used:
File:
Find all the expressed isoforms.
sailfish
Find all putative variants in isoforms.
samtools mpileup
Heterozygous variants
Homozygous-looking variants
Is there a bimodal distribution?
No
Yes
Filter variants with GATK-HC.
Discard variants that do not fit the distribution.
Filter variants with Red Panda.
~35%
~8%
~50%
Passed filter
Passed filter
Passed filter
Final variant list
~30%: Heterozygous
~77%: Not bimodally-distributed
~23%: Bimodally-distributed
~70%: Homozygous-looking

Figure 11. A simple schematic of the logic used in Red Panda. For every cell, every expressed isoform is identified with sailfish. All putative variants are then identified in each isoform and split into a homozygous-looking VCF file and a heterozygous VCF file. The latter is then filtered using Red Panda if the variants are bimodally-distributed or GATK-HC if they are not. Homozygous-looking variants are filtered by Red Panda using quality cutoffs. These three sets of variants are then combined into a single VCF file.

in the range of 25%-35% and 65%-75% would be allowed. If a bimodal distribution does not exist, then all heterozygous variants that have sufficient read support are written to a file that will later be evaluated by GATK HaplotypeCaller. Similarly, all variants that look to be homozygous are added to this sample's file that will later be assessed. This software is used because, as was established in our previous work, it is the best variant caller available[32] and it makes sense to use it in scenarios where the unique information afforded to us by single-cell sequencing was no longer available. Specifically, in scRNA-seq those scenarios are heterozygous variants that do not have enough supporting variants to determine an appropriate bimodal distribution.

The final list of variants that is presented to the user contains those that: are heterozygous and fit a bimodal distribution, are heterozygous and did not fit a bimodal distribution but were supported by GATK HaplotypeCaller, and those that appeared to be homozygous and had a read depth of at least 10x.

It should be noted that this method of taking advantage of the fraction of reads supporting a heterozygous allele is also used to identify insertions and deletions. This is especially important because indels are frequently disruptive while also being notoriously difficult to accurately identify in sequencing data due to the length of the typical next generation sequencing read as well as problems with short read aligners being consistent in their alignment of reads to the reference genome[59–64].

Results

All putative variants were generated for each of the 22 samples using samtools mpileup and the number of variants per sample ranged from 415,888 for B8-C11 to 1,398,345 for B3-C7 as seen in **Figure 12a**. To get more realistic numbers in line with what is used in other methods, a minimum depth cutoff of 10x was used as can be seen in **Figure 12b**. The number of variants per sample now ranges from 6,239 for E5-C74 to 15,737 for C8-C17. As expected, the largest bin is the 10x-14x bin due to the relatively low amount of sequencing performed for each sample.

After the total putative variants are identified and filtered by depth, in the instance of a bimodal distribution, variants are removed or kept based on whether they fit said distribution. **Figure 13** shows that when a bimodal distribution existed for a transcript, on average 7.96%, or ~215, of the total putative variants heterozygous variants in that transcript were kept. However, in the instance where there was not a bimodal distribution present for a transcript and there was sufficient read support, variants were filtered using HaplotypeCaller. On average, 34.89% of the variants checked at this stage were kept after being evaluated as seen in **Figure 14**.

After all heterozygous variants were identified, both those identified by Red Panda and those filtered by GATK HaplotypeCaller, they were combined with the homozygous-looking variants into a final VCF file. **Table 8** shows that, on average, there were 1,369 variants per cell, of which 69.78% were homozygous-looking and

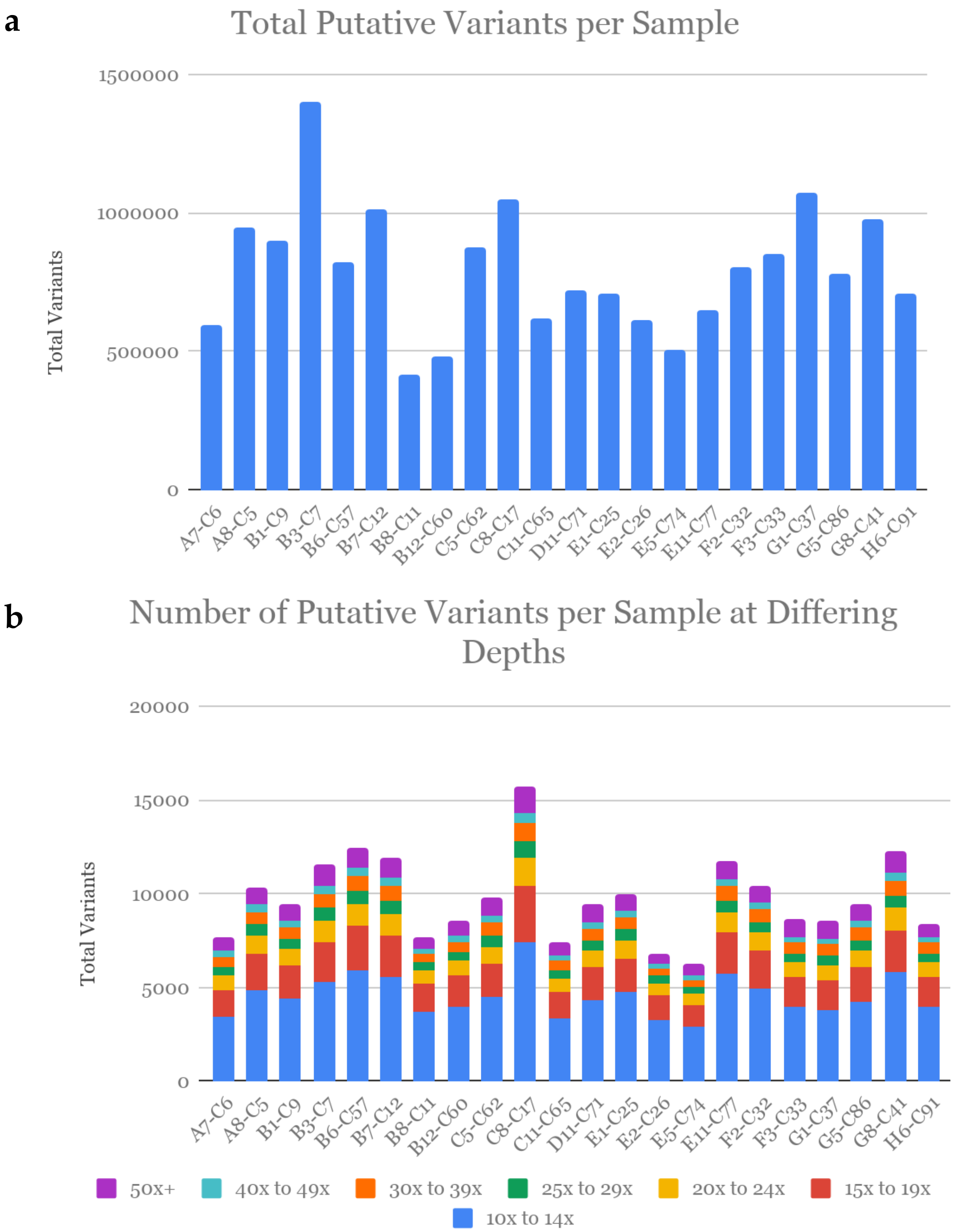


Figure 12. Putative variants per Sample. (a) Total Putative Variants per Sample. There is a very large number of putative variants per sample generated by samtools mpileup. These numbers are before any filtering has taken place. **(b)** Number of Putative Variants per Sample at Differing Depths. This is to see what proportion of putative variants exist at different depths.

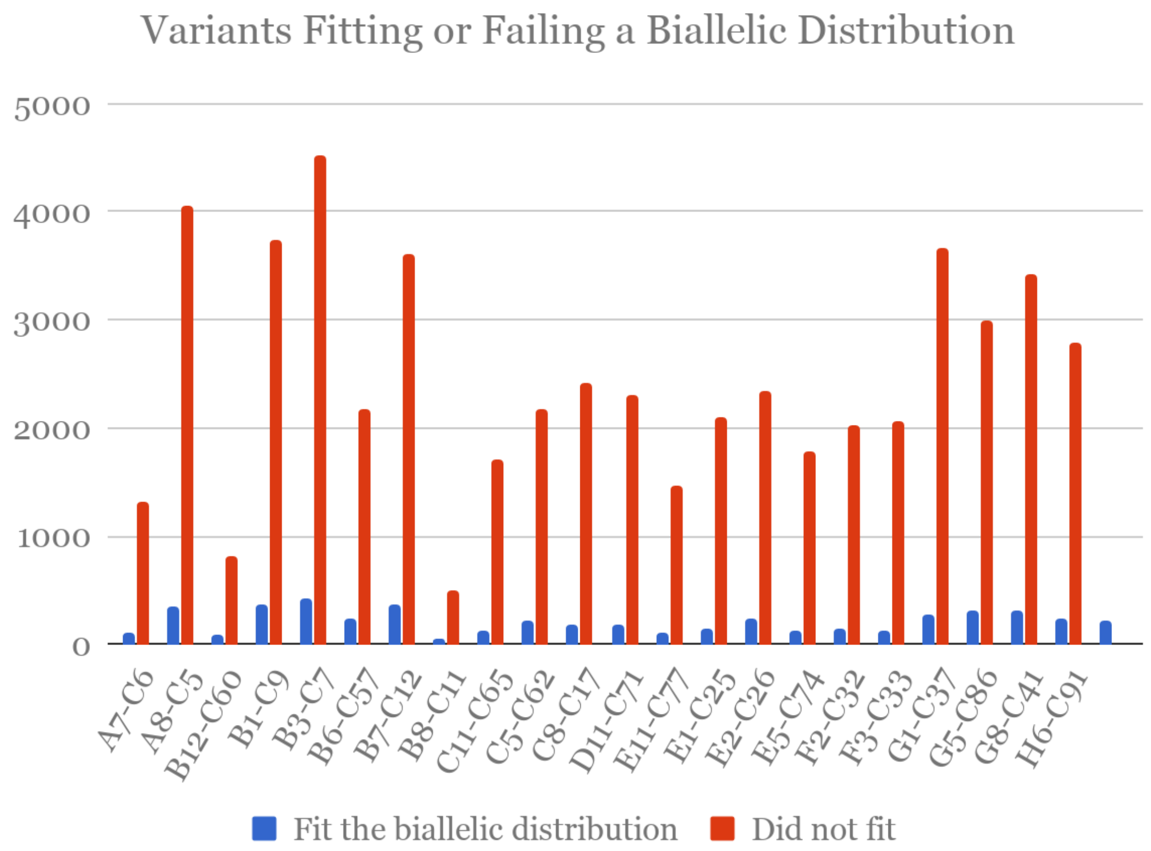


Figure 13. Variants fitting or failing a bimodal distribution. On average 7.96% of variants were kept per cell. That's an average of 215 variants. It is unclear why B8-C11 has so few variants compared to other cells despite having a high read count.

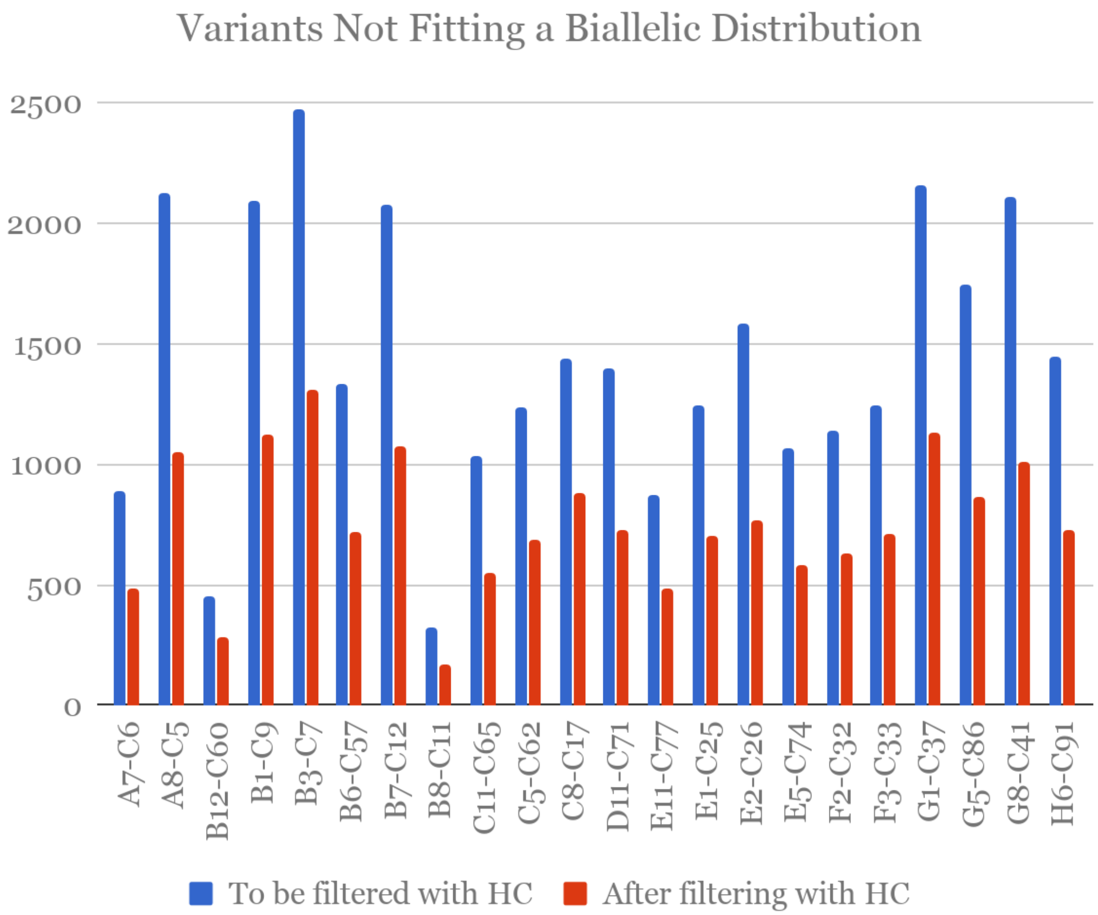


Figure 14. Variants not fitting a bimodal distribution. These variants were found in isoforms that did not contain an obvious bimodal distribution and need to be filtered with GATK HaplotypeCaller (HC). On average 34.89% of variants were kept after being evaluated. It is unclear why B8-C11 has so few variants compared to other cells despite having a high read count.

	Total
Percent of total variants that are homozygous	69.78%
Percent of total variants that are heterozygous	30.22%
Percent of heterozygous variants that are not bimodally-distributed	77.17%
Percent of heterozygous variants that are bimodally-distributed	22.83%
Total Variants	1369.5

Table 8. Summary table of variants identified by Red Panda. Percent of variants that are homozygous-looking, heterozygous, heterozygous and not bimodally-distributed, heterozygous and bimodally-distributed are calculated. Average total number of variants in the final VCF file is also shown.

30.22% were heterozygous. Of the latter group, 22.83% on average were bimodally-distributed variants.

After initial algorithm development, it was determined that a more deliberate approach should be taken to determine whether a variant is being correctly classified as heterozygous or homozygous. In the first iteration, variants were considered heterozygous if one of the following criteria held true where AF = allele frequency, DP = depth of the sequencing at this location, and C = the cutoff fraction at which this variant is no longer considered heterozygous:

1. DP < 20 and AF < C where C = 0.85
2. 21 <= DP < 40 and AF < C where C = 0.90
3. DP >= 40 and AF < C where C = 0.95

To ensure that this Method, termed Method A, was the most accurate, ten total strategies—Methods A-J—were created and compared: for each cell, all variants that were identified by a certain method as heterozygous in the cell were cross-referenced with that location in the exome. Assuming there was proper coverage in the exome to come to an accurate conclusion, that method was then scored for that location.

Methods A-J largely follow the same logic where different bins, composed of two cutoffs are used. The cutoffs are for depth and allele frequency at that location. The general idea is that, as the depth at a location increases, it is more certain that the variant at this spot is either homozygous or heterozygous. The methods are as follows:

Method A: three bins

1. DP < 20 and AF < C where C = 0.85

2. 21 <= DP < 40 and AF < C where C = 0.90
3. DP >= 40 and AF < C where C = 0.95

Method B: three bins, with the cutoff fractions reversed

1. DP < 20 and AF < C where C = 0.95
2. 21 <= DP < 40 and AF < C where C = 0.90
3. DP >= 40 and AF < C where C = 0.85

Method C: three bins with laxer cutoffs for C

1. DP < 20 and AF < C where C = 0.90
2. 21 <= DP < 40 and AF < C where C = 0.933
3. DP >= 40 and AF < C where C = 0.967

Method D: three bins with laxer cutoffs for C, and the fractions reversed

1. DP < 20 and AF < C where C = 0.967
2. 21 <= DP < 40 and AF < C where C = 0.933
3. DP >= 40 and AF < C where C = 0.90

Method E: six bins

1. DP < 20 and AF < C where C = 0.8
2. 21 <= DP < 40 and AF < C where C = 0.84
3. 41 <= DP < 60 and AF < C where C = 0.88
4. 61 <= DP < 80 and AF < C where C = 0.92
5. 81 <= DP < 100 and AF < C where C = 0.94
6. DP > 100 and AF < C where C = 0.96

Method F: six bins; reversed fractions

1. DP < 20 and AF < C where C = 0.96
2. 21 <= DP < 40 and AF < C where C = 0.94
3. 41 <= DP < 60 and AF < C where C = 0.92
4. 61 <= DP < 80 and AF < C where C = 0.88
5. 81 <= DP < 100 and AF < C where C = 0.84
6. DP > 100 and AF < C where C = 0.80

Method G: six bins with laxer cutoffs

1. DP < 20 and AF < C where C = 0.900
2. 21 <= DP < 40 and AF < C where C = 0.915
3. 41 <= DP < 60 and AF < C where C = 0.930
4. 61 <= DP < 80 and AF < C where C = 0.945
5. 81 <= DP < 100 and AF < C where C = 0.960
6. DP > 100 and AF < C where C = 0.975

Method H: six bins with laxer cutoffs; reversed fractions

1. DP < 20 and AF < C where C = 0.975
2. 21 <= DP < 40 and AF < C where C = 0.960
3. 41 <= DP < 60 and AF < C where C = 0.945
4. 61 <= DP < 80 and AF < C where C = 0.930
5. 81 <= DP < 100 and AF < C where C = 0.915
6. DP > 100 and AF < C where C = 0.900

Method I: 10 bins

1. DP < 10 and AF < C where C = 0.80

2. 11 <= DP < 20 and AF < C where C = 0.82
3. 21 <= DP < 30 and AF < C where C = 0.84
4. 31 <= DP < 40 and AF < C where C = 0.86
5. 41 <= DP < 50 and AF < C where C = 0.88
6. 51 <= DP < 60 and AF < C where C = 0.90
7. 61 <= DP < 70 and AF < C where C = 0.92
8. 71 <= DP < 80 and AF < C where C = 0.94
9. 81 <= DP < 90 and AF < C where C = 0.96
10. DP > 90 and AF < C where C = 0.98

Method J: 10 bins

1. DP < 10 and AF < C where C = 0.80
2. 11 <= DP < 20 and AF < C where C = 0.82
3. 21 <= DP < 30 and AF < C where C = 0.84
4. 31 <= DP < 40 and AF < C where C = 0.86
5. 41 <= DP < 50 and AF < C where C = 0.88
6. 51 <= DP < 60 and AF < C where C = 0.90
7. 61 <= DP < 70 and AF < C where C = 0.92
8. 71 <= DP < 80 and AF < C where C = 0.94
9. 81 <= DP < 90 and AF < C where C = 0.96
10. DP > 90 and AF < C where C = 0.98

Figure 15 shows that Method C identifies the highest number of variants correctly identified as heterozygous, but it has one of the lowest total percent of variants

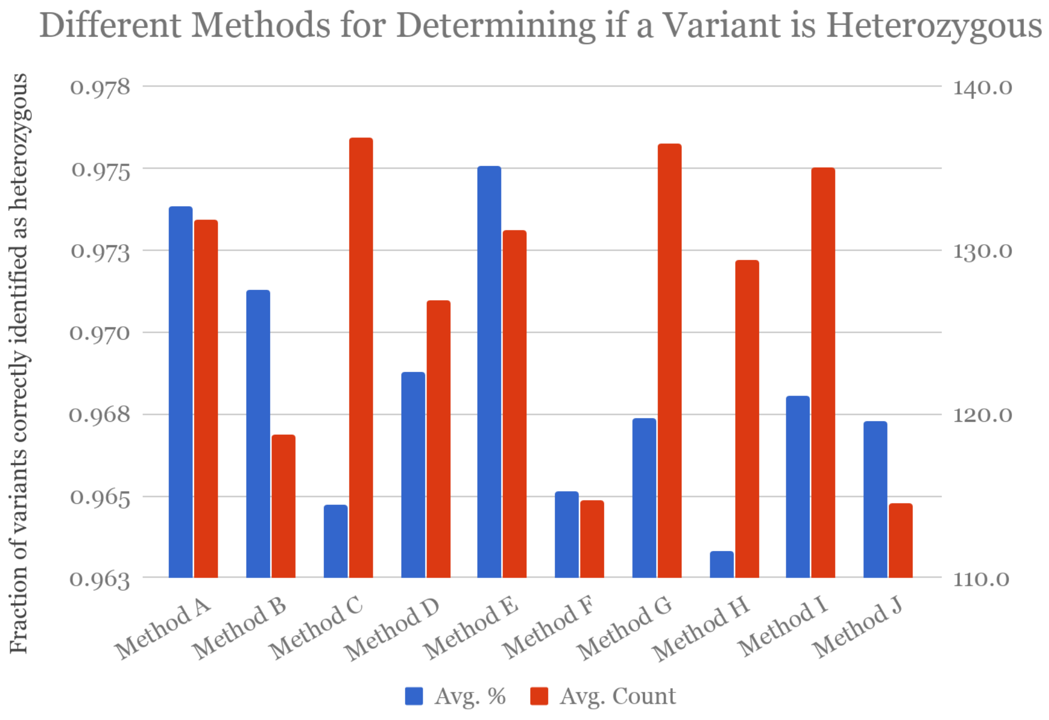


Figure 15. Different methods for determining if a variant is heterozygous. Ten different methods were written to determine if a variant is considered heterozygous by only looking at the scRNA-seq data. This graph is an average across all 22 cells of the percentage and counts of correctly identified heterozygous variants by each method.

identified correctly. Interestingly the original method used, Method A, is among the best performers, but Method E will be used as it slightly outperformed everything else.

CHAPTER 2: VALIDATION USING SIMULATED AND EXPERIMENTAL DATA

Introduction

To determine the effectiveness of Red Panda, we tested the method using both simulated and experimentally-generated datasets. Our results are compared to those from four currently available variant calling tools: FreeBayes, Genome Analysis Toolkit HaplotypeCaller, Genome Analysis Toolkit UnifiedGenotyper, and Platypus. It is important to compare these tools because, in addition to being popular in the bioinformatics community, their performance has been assessed in bulk sequencing settings[32,65,66], but not in single-cell sequencing.

The first dataset used is the human articular chondrocyte single-cell RNA sequencing described previously. To ensure consistency of comparisons, all tools are given identical inputs: alignment files which are uniformly generated using the scRNA-seq data from the articular chondrocytes. All variant calling software is then run using their recommended settings. These variant calling data from each cell are compared to the results from the human articular chondrocyte exome sequencing to determine its veracity. To abrogate any False Negatives, the variant calls are restricted to those locations that have sufficient supporting alignments from the scRNA-seq datasets as well as the exome sequencing data. Once these regions have been identified, all variants are compared to the exome sequencing variant calling results to assess the Positive Predictive Value (PPV) of Red Panda as well as the other variant calling tools.

The second dataset is generated from MEFs. These cells are isolated using the C1 Fluidigm 96-well chip and have libraries generated using version three of the SMARTer Ultra Low RNA Kit for Illumina Sequencing in the same manner as the articular chondrocytes. Alignment and variant calling is performed the same as with the first dataset; however validation is performed differently as this data will not be paired with exomic bulk sequencing. Sanger sequencing was used to verify the existence of 40 randomly identified variants: 20 unique to Red Panda and 20 identified by all five variant callers. Simulated data was also generated from the MEF sequencing to accurately predict the sensitivity of each variant caller.

Comparing variant callers using human articular chondrocyte data

Alignment file preparation

For four of the five variant calling pipelines, alignments are prepared by running hisat2 version 2.1.0 using human genome version hg38 as the reference. The exception to this is Platypus which used BWA MEM v. 07.17[67] due to its requirements for alignment files. Hisat2 is specifically designed for aligning RNA-seq reads and will split reads across exons to ensure the highest-quality alignments. Unfortunately this results in very large fragments that Platypus cannot process. Instead, BWA MEM, another high quality alignment tool, was used in a manner that generated alignments that Platypus could process.

Alignment is followed by running the Genome Analysis Toolkit version 3.8.0 module, SplitNTrim + ReassignMappingQuality. This method of alignment preparation ensures that the highest quality alignment files are generated when trying to identify variants, and thus, that the only deficiencies identified are due to the variant caller itself instead of other processing steps. Specifically, reads spanning exon-intron junctions have overhanging regions clipped to remove any intronic sequence if they have been incorrectly aligned. Lastly, for every tool, a bed file containing the regions of probes pulled down by the SureSelect Clinical Research Kit V2 was used to limit the location of where variants could be identified to ensure proper overlap of the single-cell data and the exome data.

FreeBayes

FreeBayes is Bayesian statistical framework capable of identifying, in a reference genome, small variations including SNVs, indels, multi-nucleotide variants (MNVs), and composite insertion and substitution events so long as these events are shorter than the length of the sequencing read belonging to the alignment. Version 1.1.0.46 was used to generate Variant Call Format (VCF) files containing all variants present in the alignment file for each cell. FreeBayes was run using the following arguments in addition to default parameters:

- --min-alternate-fraction 0.01
- --targets SureSelect_Clinical_Research_Exome_v2.bed
- --no-partial-observations

GATK HaplotypeCaller

GATK HaplotypeCaller uses a De Bruijn-like graph to perform local *de novo* assembly of regions of the genome that show evidence of significant variation, including SNVs, MNVs, and indels. This has the advantage of more accurately identifying more complex variation in a sample such as indels, and, importantly for our study, splice junctions. However these benefits come at the expense of increased computation time[68]. Version 3.8-0-ge9d806836 was used to generate VCF files containing all variants present in the alignment file for each cell. GATK HaplotypeCaller was run using the following arguments in addition to default parameters:

- --filter_reads_with_N_cigar
- --standard_min_confidence_threshold_for_calling 4.0
- --dbsnp dbsnp-150.vcf.gz
- -L SureSelect_Clinical_Research_Exome_v2.bed

GATK UnifiedGenotyper

GATK UnifiedGenotyper uses a Bayesian genotype likelihood model to estimate the most likely genotypes (SNVs, MNVs, and indels) and allele frequency. GATK UnifiedGenotyper also benefits from not assuming ploidy for the organism being analyzed which is not the case for GATK HaplotypeCaller[68]. Version 3.8-0-ge9d806836 was used to generate VCF files containing all variants present in the alignment file for each cell. GATK UnifiedGenotyper was run using the following arguments in addition to default parameters:

- --filter_reads_with_N_cigar
- --standard_min_confidence_threshold_for_calling 4.0
- --dbsnp dbsnp-150.vcf.gz
- -L SureSelect_Clinical_Research_Exome_v2.bed
- --geontype_likelihoods_model BOTH

Platypus

Platypus uses local realignment of reads and local reassembly to accurately identify variants--SNVs, MNVs, indels, and long-range insertions and deletions--in a genome. Version 0.8.1.1 was used to generate VCF files containing all variants present in the alignment file for each cell. Platypus was run using the following arguments in addition to default parameters:

- --regions=SureSelect_Clinical_Research_Exome_v2.bed
- --filterDuplicates=0

Comparison of all tools' results with the exome data

Comparisons are performed by looking at the data produced by each tool for each cell; one VCF file is created per cell per tool, totaling 110 files (22 cells×5 tools). Statistical metrics are then calculated to evaluate each tool using the exome variant analysis results as the reference against which these 110 files are compared. This process involves calculating the intersection of regions in the genome that are common between the SureSelect Clinical Research Exome V2 library preparation kit and only the transcripts that are expressed in the cell being analyzed. This is needed because the scRNA-seq is a subset of the exome data, and not limiting the search space to those regions corresponding to genes expressed in this cell leads to the erroneous identification of thousands of False Negatives. All of these regions of interest are contained in a bed file in a format that can be seen in **Table 9**. The intersection of the exome bed file and the expression bed file for that cell is created using bedtools[69] and results in a file that only contains regions common to both files as seen in **Figure 16**.

As an example, a new bed file that is the intersection between the exome bed file and the transcripts expressed in cell G1-C37 (named here: exome_G1-C37.bed, or the bounded exome bed file) is created. Following this, five new VCF files are generated, one per tool (these are bounded VCF files as they are restricted by the boundaries of the exome + transcripts bed file):

1. Variants identified by FreeBayes (FreeBayes_G1-C37.vcf) that fall into the regions contained in exome_G1-C37.bed produce a file containing variants found only in the intersected regions: FreeBayes_exome_G1-C37.vcf.
2. Variants identified by GATK HaplotypeCaller (GATK_HaplotypeCaller_G1-C37.vcf) that fall into the regions contained in exome_G1-C37.bed produce a file containing variants found only in the intersected regions: GATK_HaplotypeCaller_exome_G1-C37.vcf

Chromosome	Chromosome start location	Chromosome end location
chr1	14211	15031
chr1	61724	62229
...	...	...
chrY	56874614	56876385

Table 9. The bed file format. These files contain the regions that will be analyzed for variants in each cell.

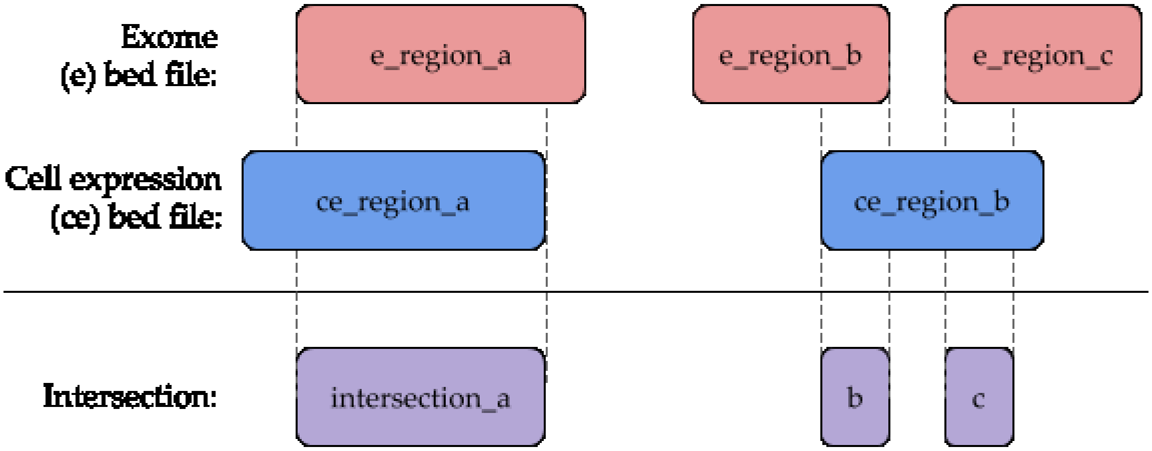


Figure 16. Intersection example of exome and expression bed files. The resulting file containing the purple regions is what is used to determine concordance of variants found exome and those found in the cell.

3. Variants identified by GATK_UnifiedGenotyper (GATK_UnifiedGenotyper_G1-C37.vcf) that fall into the regions contained in exome_G1-C37.bed produce a file containing variants found only in the intersected regions: GATK_UnifiedGenotyper_exome_G1-C37.vcf
4. Variants identified by Platypus (Platypus_G1-C37.vcf) that fall into the regions contained in exome_G1-C37.bed produce a file containing variants found only in the intersected regions > Platypus_exome_G1-C37.vcf
5. Variants identified by Red Panda (Red_Panda_G1-C37.vcf) that fall into the regions contained in exome_G1-C37.bed produce a file containing variants found only in the intersected regions: Red_Panda_exome_G1-C37.vcf

Once these files are created, PPV (specificity and False Positive Rate are not calculated as the number of True Negatives is so large that it results in values so close together that they cannot be meaningfully distinguished) can be calculated for each tool:

- Positive Predictive Value: (TP)/($TP + FP$) where:
    a. TP = number variants in both the tool's bounded VCF file (e.g., Red_Pand_exome_G1-C-37.vcf) AND the exome VCF file that only contains variants found in the intersected bed file (e.g., exome_G1-C37.bed)
    b. FP = number of variants in the tool's bounded VCF file that are not also in the bounded exome VCF file.

In addition to the PPV statistics, the total number of variants—SNVs and indels—produced by each tool in each cell was intersected with the variants found in the exome as can be found in **Figure 17**. Specifically, only those regions that were supported by read data in both the exome and the single-cell data were compared.

It is immediately apparent that Red Panda finds more variants than every other tool. On average, Red Panda identifies 913 variants per cell that are in accordance with the exome whereas FreeBayes identifies 65, GATK HaplotypeCaller identifies 705, GATK UnifiedGenotyper identifies 222, and Platypus identifies 386.

The overlap between the tools was also assessed by creating UpSet[70] plots (a method of showing intersection between datasets) between each tool for each cell as seen in **Figure 18**. There is consistent overlap between the tools, even for FreeBayes and GATK UnifiedGenotyper which typically did not identify as many variants as the other tools. Of note is the fact that while Red Panda shares significant overlap with the other tools, it also identifies a large number of unique variants. This is expected given that Red Panda consistently identified more variants than every other tool.

To assess the effectiveness of these software with regards to heterozygous variant identification, the same analysis was performed using just the heterozygous SNVs and indels in each sample. This posed a new problem as it is difficult to determine what is heterozygous and what is homozygous in scRNA-seq data. This is because transcription does not always occur from both chromosomes at the same time, but rather can happen in a burst fashion resulting in only monoallelic expression being seen[71,72]. Given this, it is possible that what might look like a homozygous variant in the sequencing data, might actually be heterozygous. To handle this, the method created for Red Panda—Method E from the Algorithm section of Chapter 1—was utilized to

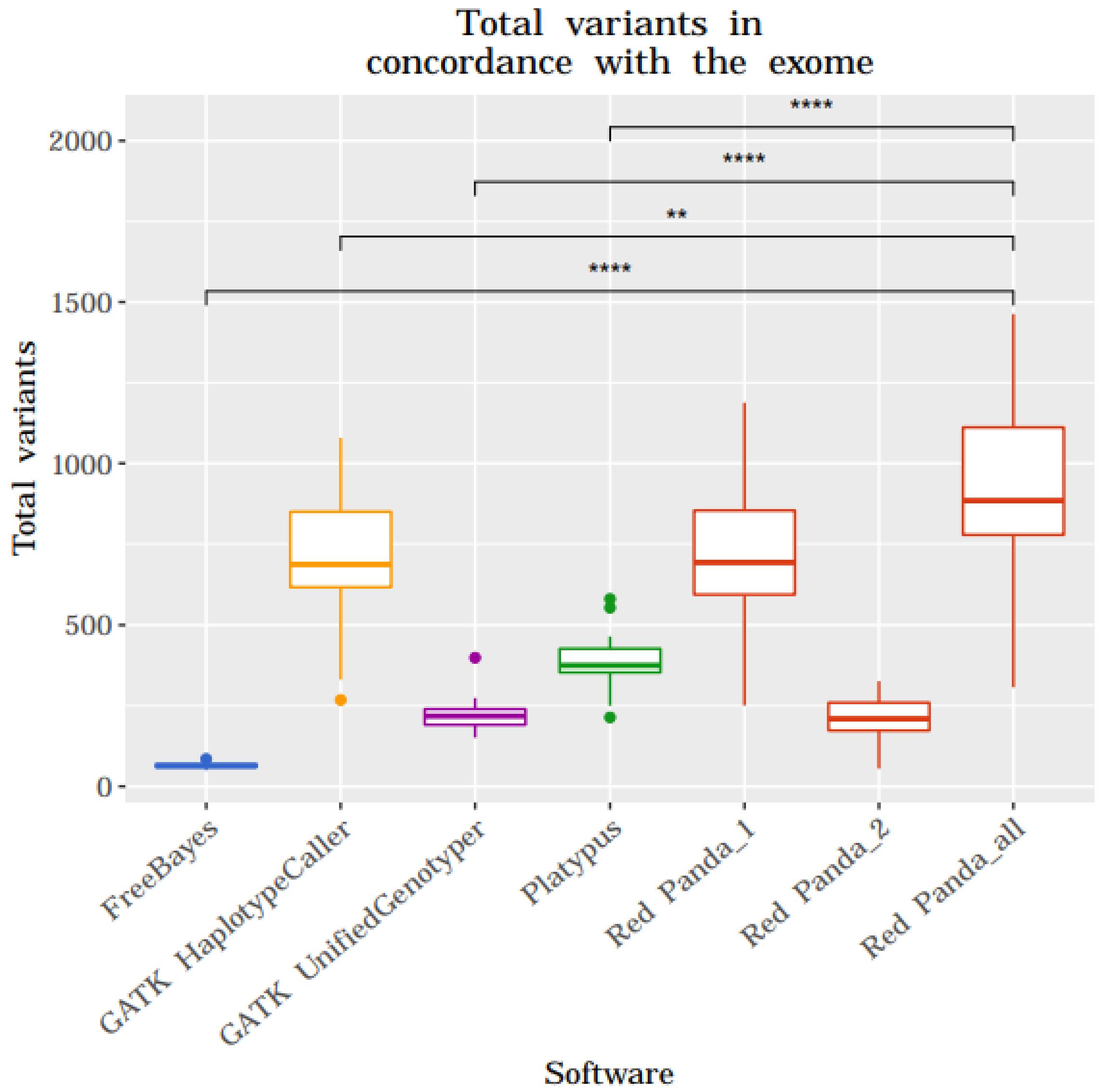


Figure 17. Total variants in concordance with the exome. The total number of variants in concordance with the exome for every cell as identified by each tool. Each cell had variants identified by FreeBayes, Red Panda, GATK HaplotypeCaller, Platypus, and GATK UnifiedGenotyper, after which they were compared to the exome sequencing data to determine their veracity. Red Panda is characterized by three box plots: 1, 2, and all. Red Panda_1 contains variants exclusive to Red Panda logic: homozygous-looking variants and bimodally-distributed heterozygous variants. Red Panda_2 contains non-bimodally-distributed heterozygous variants that are called by GATK-HaplotypeCaller. Red Panda_all is a superset of the two. Comparisons were performed using T-tests: ns = not significant, * is $p < 0.05$, ** is $p < 0.01$, *** is $p < 0.001$, and **** is $p < 0.0001$.

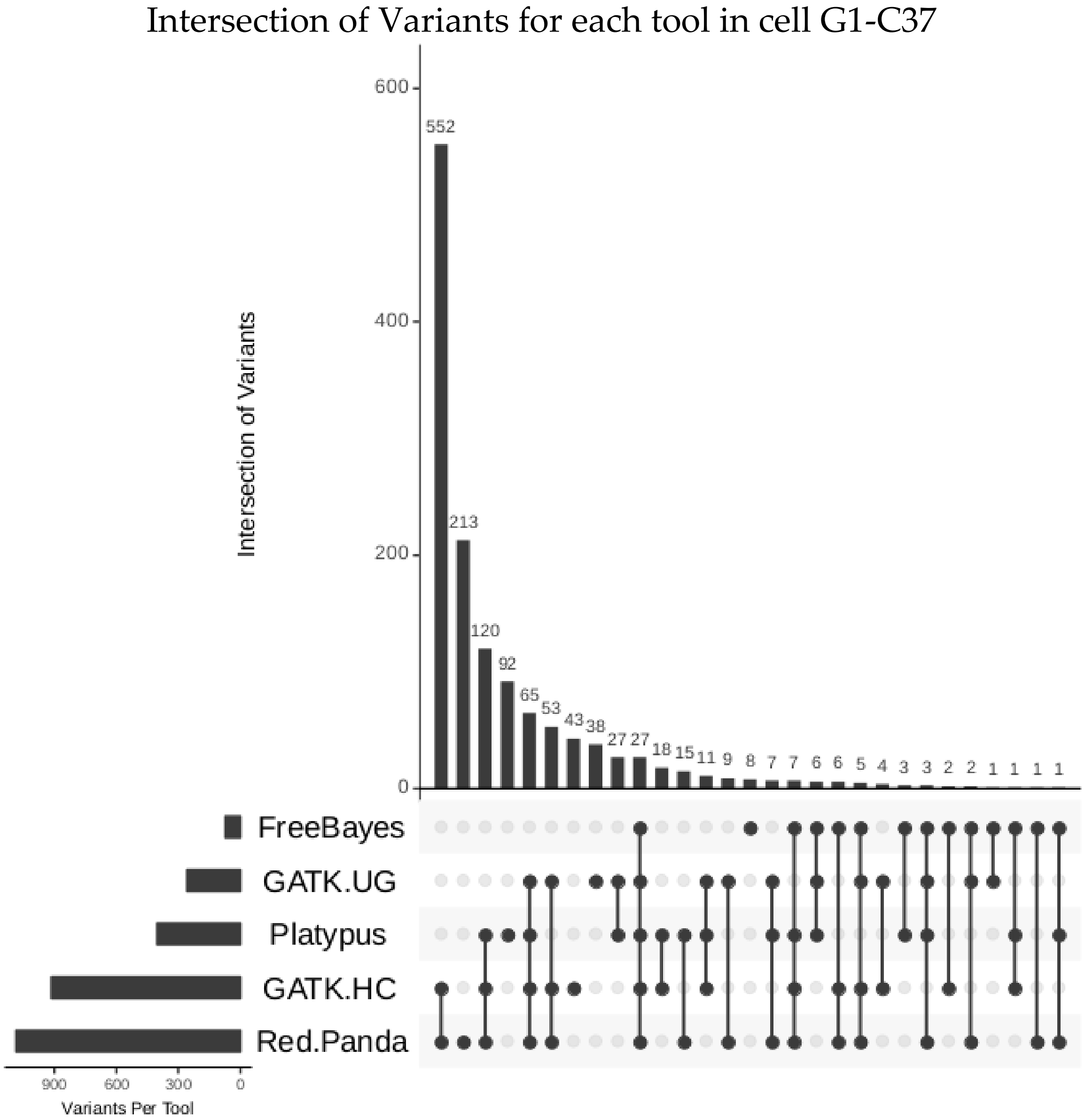


Figure 18. UpSet plots of the overlap between each tool. The overlap of the variants identified by each tool can be seen for the cell G1-C37. Red Panda identifies the most variants as well as the most unique variants in concordance with the exome. Each column of the X-axis shows the overlap between each tool represented by a filled-in dot. For example, the first column indicates that GATK-HC and Red Panda shared 552 variants, the second shows that there were 213 variants unique to Red Panda, the third column indicates that there were 120 variants shared between Platypus, GATK-HC, and Red Panda, and so on.

determine whether something was homozygous-looking or heterozygous. This variant is then cross-referenced with the exome sequencing data to confirm that it is, in fact, heterozygous. Method E utilizes metadata found in the VCF files to identify AF (Allele Frequency = # alt alleles/(# alt alleles + # ref alleles)) and DP (depth of the sequencing at this location). In this method, C = cutoff fraction at which this variant is no longer considered heterozygous (i.e., if AF > C, it is considered homozygous). If a variant fits one of these six criteria then it is considered heterozygous for our comparisons:

1. DP < 20 and AF < C where C = 0.8
2. 21 <= DP < 40 and AF < C where C = 0.84
3. 41 <= DP < 60 and AF < C where C = 0.88
4. 61 <= DP < 80 and AF < C where C = 0.92
5. 81 <= DP < 100 and AF < C where C = 0.94
6. DP > 100 and AF < C where C = 0.96

The five tools had their variants filtered so that heterozygous variants were split out into separate files for further analysis. Method E was used to determine whether something was heterozygous for Platypus and Red Panda as they contained appropriate metadata to calculate DP and AF. For FreeBayes, GATK UnifiedGenotyper, and GATK HaplotypeCaller, there is no metadata available in the VCF file for how many reads supported the alternate allele vs. the reference allele. Instead for FreeBayes and GATK UnifiedGenotyper, the AF field provided in the variant's metadata (not to be confused with the AF value used in Method E) was used to determine the fraction of reads that support the alternate allele. Unfortunately, this is not a perfect comparison since this value is calculated *after* reads are filtered out by the variant caller from the alignment (typically reads that are low quality or have poor mapping scores are removed), whereas the fractions calculated by Method E are performed *before* reads are filtered. For GATK HaplotypeCaller, the MLEAF (Maximum Likelihood Expectation for the Allele Frequency) value was used which attempts to approximate the original proportion of the allele in the context of a diploid organism[73]. For this data, MLEAF is either 0.5 (heterozygous variant) or 1.0 (homozygous); the former is used to filter variants into the file containing heterozygous-looking variants.

Once an accurate method has been developed for identifying how to determine what a heterozygous variant is, the effectiveness of these software with regards to heterozygous variant identification was performed. **Figure 19** shows **t**he total number of heterozygous SNVs and indels in concordance with the exome for each tool and each cell. While the improvement is not as drastic as compared the total number of variants in agreement with the exome data, Red Panda still improves on the other methods. On average Red Panda identifies 154 variants in agreement with the exome, 31 for FreeBayes, 136 for GATK HaplotypeCaller, 118 for GATK UnifiedGenotyper, and 36 for Platypus.

After identifying the total number of variants found per cell per tool that agreed with the exome, PPV and False Discovery Rate (FDR) are calculated. As seen in **Table 10** and **Figure 20**, Red Panda has the highest average PPV of any of the other tools. This shows that, compared to the variant calling software traditionally used for bulk sequencing, Red Panda correctly identifies a variant more consistently. Despite

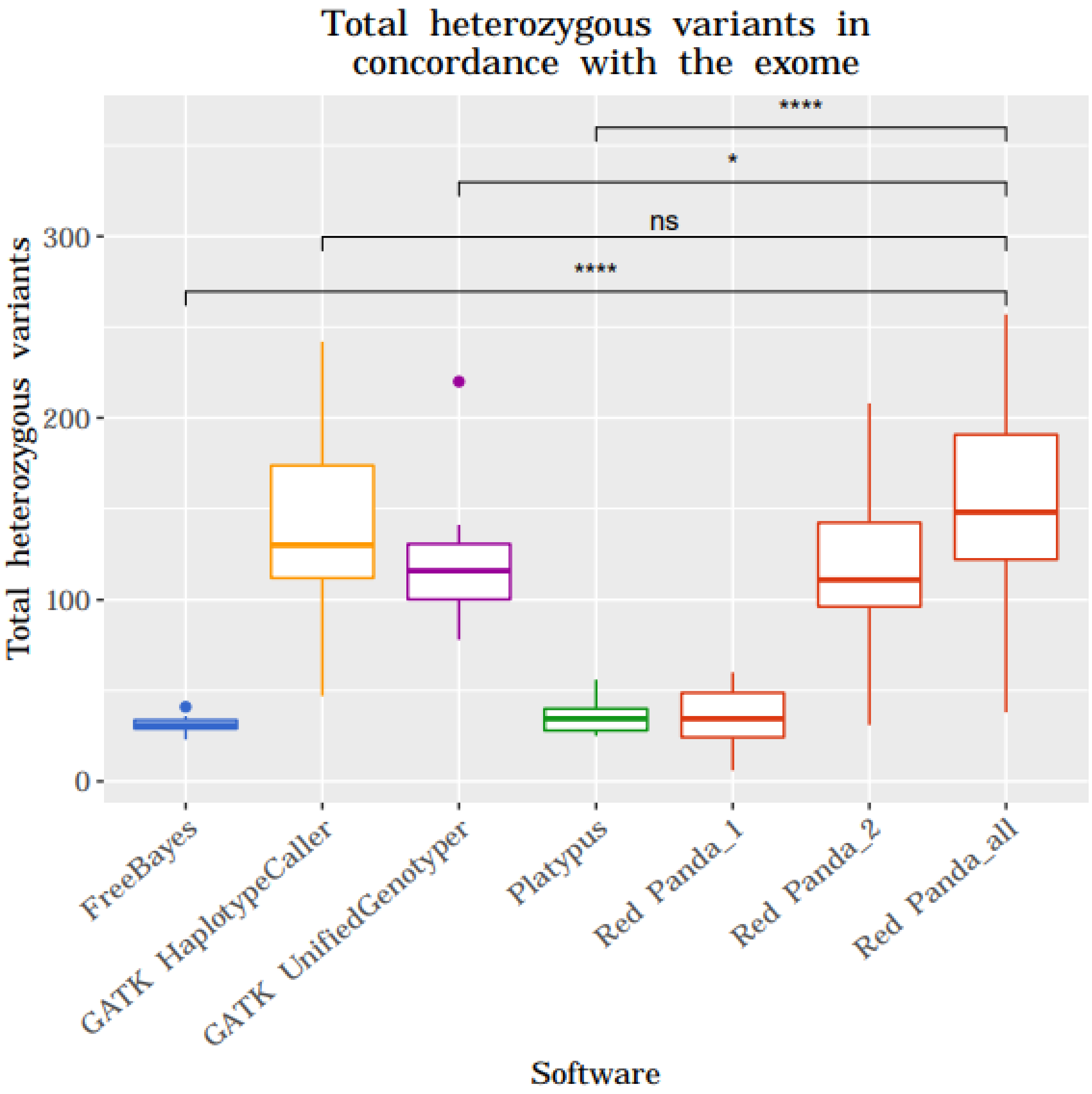


Figure 19. Total heterozygous variants in concordance with the exome. The total number of heterozygous variants in concordance with the exome for every cell as identified by each tool. Each cell had variants identified by FreeBayes, Red Panda, GATK HaplotypeCaller, Platypus, and GATK UnifiedGenotyper, after which they were compared to the exome sequencing data to determine their veracity and then selected for comparison if they were heterozygous-looking. Red Panda consistently identifies more variants in concordance with the exome than every other tool. Red Panda is characterized by three box plots: 1, 2, and all. Red Panda_1 contains bimodally-distributed heterozygous variants. Red Panda_2 contains non-bimodally-distributed heterozygous variants. Red Panda_all is a superset of the two. Comparisons were performed using T-tests: ns = not significant, * is p < 0.05, ** is p < 0.01, *** is p < 0.001, and **** is p < 0.0001.

Algorithm	Average PPV (%)	Average FDR (%)
FreeBayes	8.69% ± 0.35%	91.31% ± 0.35%
GATK- HaplotypeCaller	31.67% ± 2.08%	68.33% ± 2.08%
GATK-UnifiedGenotyper	5.84% ± 0.45%	94.16% ± 0.45%
Platypus	6.95% ± 0.49%	93.05% ± 0.49%
Red Panda	44.96% ± 3.15%	55.04% ± 3.15%

Table 10. PPV and FDR for each tool. The average PPV and FDR with standard deviations for each tool using the exome as reference.

this, PPV is still low compared to traditional datasets[32,74]. For all of the comparisons in **Figures 17, 19** and **20**, with the exception of Red Panda and GATK-HC heterozygous variant calls, the differences between the results for Red Panda and the other software was statistically significant when assessed with T-tests.

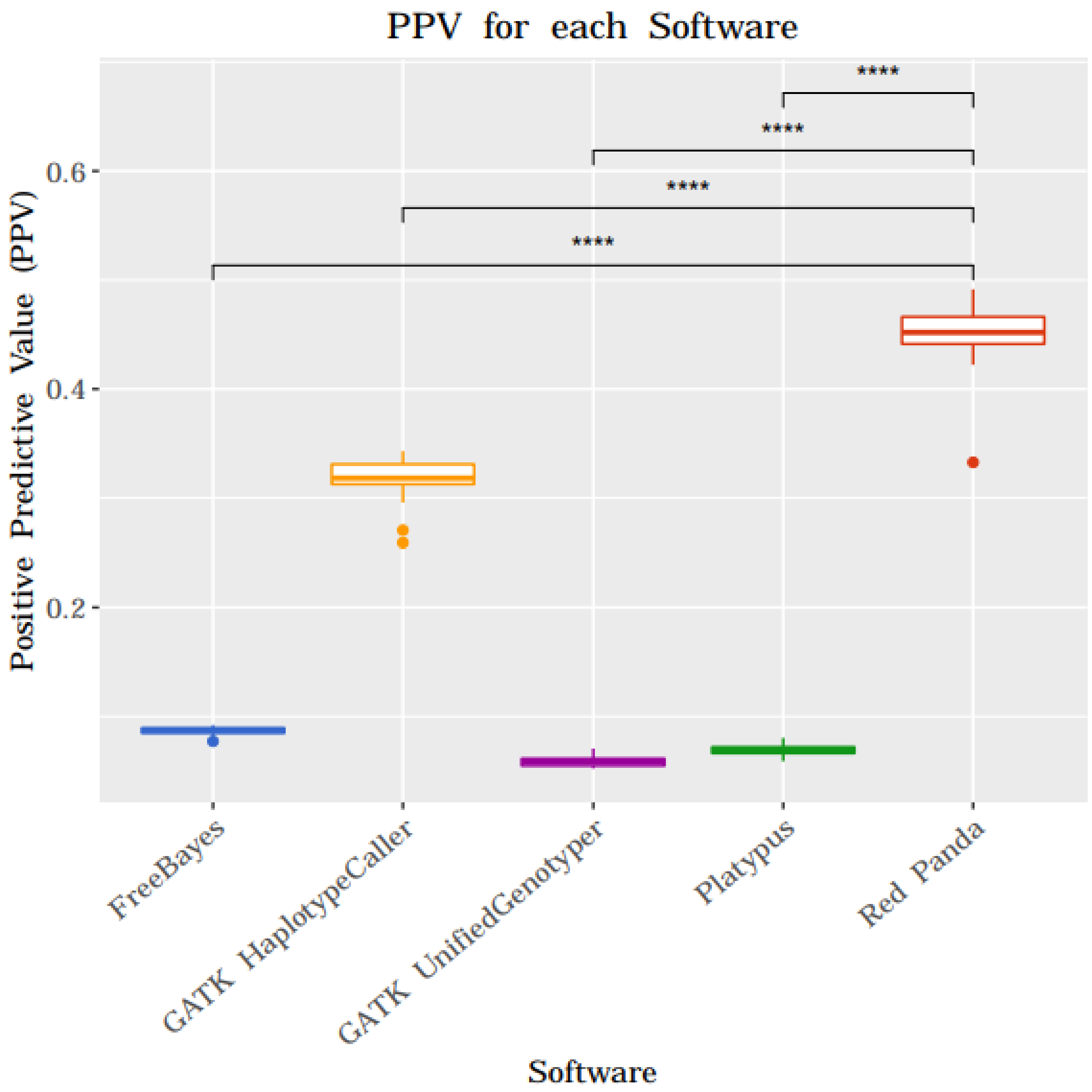


Figure 20. The average PPV calculated for each tool. Red Panda has the highest average PPV than other tools indicating that it has far fewer False Positives. Comparisons were performed using T-tests: ns = not significant, * is p < 0.05, ** is p < 0.01, *** is p < 0.001, and **** is p < 0.0001.

Mouse Embryonic Fibroblasts sequencing

After the development of Red Panda was finished, further testing needed to be performed on other datasets to ensure that this software works as a generalized tool and is not specifically tailored to the conditions found in human articular chondrocytes. The data that is being used to perform additional tests on Red Panda is generated from single-cell RNA sequencing on Mouse Embryonic Fibroblasts (MEFs).

The MEFs used for sequencing are generated in collaboration with Dr. Kishor Bhakat and his graduate student Shrabasti Roychoudhury, the full sequencing strategy of which can be seen in **Figure 21**. Their study involves investigating the effect the gene Ape1—also known as Apex1—has at the cellular level when it is mutated in a specific way. Ape1's function is to repair apurinic/apyrimidinic (AP) sites—DNA lesions—in mammalian cells[75]. Two MEF samples have been generated: one normal/wild-type and one mutated. The mutated sample has had two amino acids altered in the gene Ape1. Lysine 6 and Lysine 7 are both converted to Alanine to remove its ability to be acetylated at those sites[76,77]. The mutated sample—also designated "Ape1$_{K6A,K7A}$" and the normal sample, designated "Ape1$_{K/K}$"—was modified using the Easi-CRISPR[78] system to induce these mutations. Both of these samples, Ape1$_{K,K}$ and Ape1$_{K6A,K7A}$, had single cells isolated on the Fluidigm C1 system and had RNA sequenced on a NextSeq500.

Following sequencing, variant calling is performed on both sets of samples using the same tools used with the articular chondrocytes—FreeBayes, GATK Haplotype Caller, GATK Unified Genotyper, Platypus, and Red Panda—but for the purposes of this study, focus is placed solely on the normal Ape1$_{K,K}$ sample. This is because it is the only

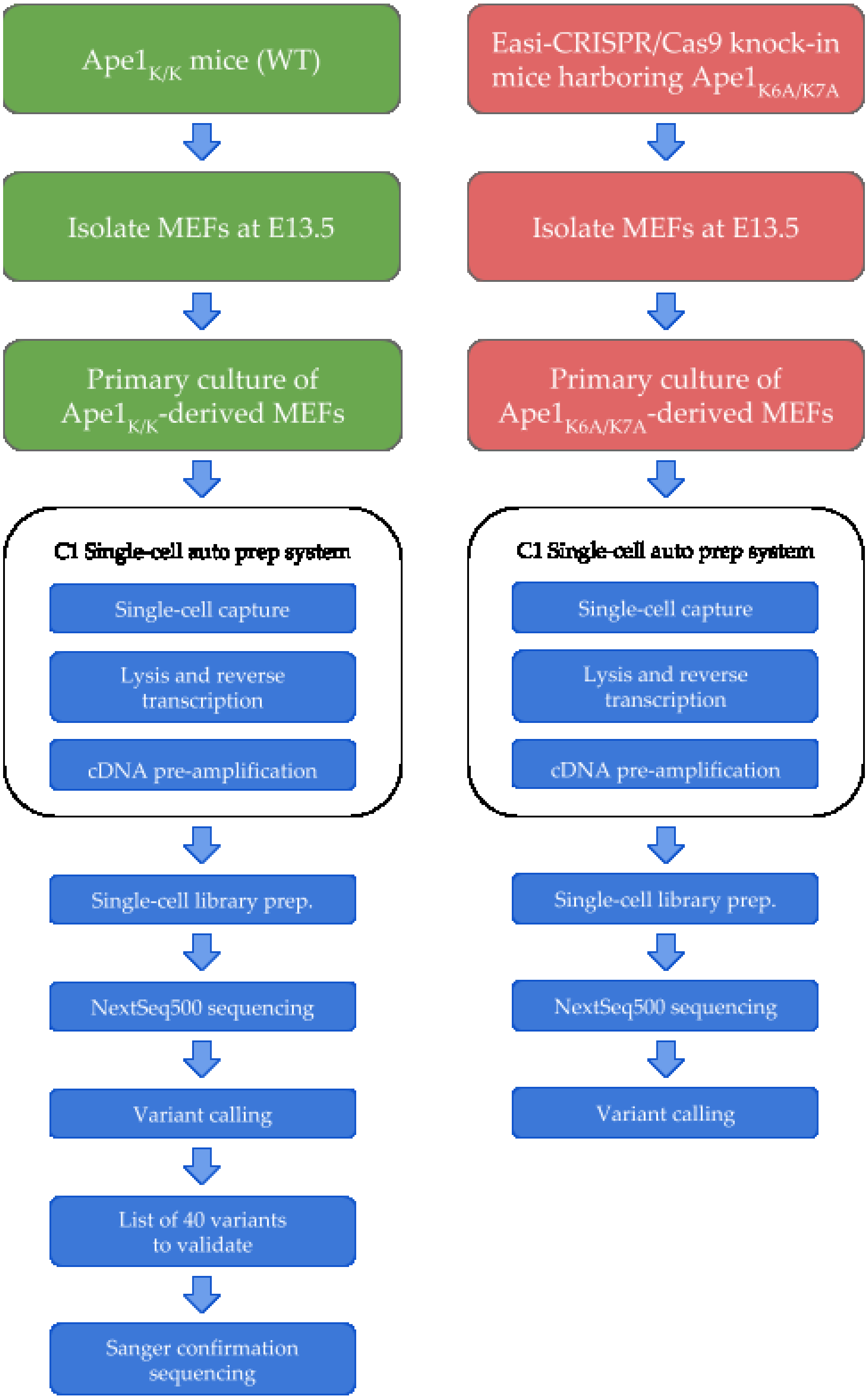

Ape1$_{K/K}$ mice (WT)
Isolate MEFs at E13.5
Primary culture of Ape1$_{K/K}$-derived MEFs
C1 Single-cell auto prep system
Single-cell capture
Lysis and reverse transcription
cDNA pre-amplification
Single-cell library prep.
NextSeq500 sequencing
Variant calling
List of 40 variants to validate
Sanger confirmation sequencing
Easi-CRISPR/Cas9 knock-in mice harboring Ape1$_{K6A/K7A}$
Isolate MEFs at E13.5
Primary culture of Ape1$_{K6A/K7A}$-derived MEFs
C1 Single-cell auto prep system
Single-cell capture
Lysis and reverse transcription
cDNA pre-amplification
Single-cell library prep.
NextSeq500 sequencing
Variant calling

Figure 21. The sequencing strategy for the MEFs. Two groups of MEFs are sequenced, one WT and one mutated. The WT cells have variant calling performed on them with five variant callers as with the articular chondrocytes. Validation is performed by Sanger sequencing on 40 variants.

sample that will have Sanger sequencing[79] and simulation used to confirm the existence of variants identified. The $Ape1_{K6A/K7A}$ mutation will likely lead to an accumulation of spurious mutations throughout the genome but only at the single-cell level, hence these results are not appropriate for validation studies. That is, the increase in mutations would not necessarily lead to an increase in mutations shared across all of the cells, but rather only mutations found in single cells. If this hypothesis is correct, then it would not be possible to identify cell-specific mutations via Sanger sequencing.

Results from variant calling by the five different software packages are compared in a number of ways: total number of variants identified by each tool, total number of homozygous-looking variants identified, and total number of heterozygous variants identified. The distinction between heterozygous variants and homozygous-looking variants is an important one in this analysis as has been mentioned previously. Variants will either have a fraction of reads that supports an unambiguously heterozygous variant, or they will have a fraction of reads that, in a single cell, appears to be a homozygous variant, but could potentially be heterozygous. This is due to the stochastic nature of RNA transcription leading to allele-specific expression[80–82]. This pattern of expressing RNA from a single chromosome—otherwise known as monoallelic expression—can lead to a heterozygous variant looking like a homozygous variant[71,83]. Due to this ambiguity, variation in the genome that has full read coverage supporting an alternate allele is hereafter termed “homozygous-looking” rather than “homozygous”.

In addition to the total number of variants identified, variant overlap between cells is assessed. As these cells are isogenic, each cell should share a large portion of their variants with the other cells sequenced. This overlap is evaluated with the assumption that a high overlap identified by a variant caller is an indicator of that software performed well.

The final way these five tools are compared is using the results of Sanger sequencing validation as well as simulated variants inserted into the normal MEF alignment files. Sanger validation is performed on a set of 20 random variants identified by all variant callers, and on a set of 20 random variants identified exclusively by Red Panda. The first group is meant to assess the accuracy of all the tools taken as a whole. The second is to address whether the Red Panda-specific variants are reliable. One requirement of the variants being checked is that they are identified in at least two cells.

Single-cell RNA sequencing

Shrabasti Roychoudhury performed cell prep and DNA extraction for these samples. This data came from MEFs that were harvested from embryos at E13.5. Cells were extracted using previously standardized methods[84]. After isolation, cells were cultured in Dulbecco's modified Eagle medium (DMEM) medium containing 10% FBS and 1% Penicillin and Streptomycin at 37°C in a 5% CO2 atmosphere for 2 days. On the day of single-cell capture, cells were trypsinized (0.05% Trypsin-EDTA solution), counted and resuspended in media at 10^5 cells/mL concentration.

The UNMC Sequencing Core Facility performed cell capture and sequencing. Cells were loaded on to a 17-25 μm Fluidigm C1 Single-Cell Auto Prep IFC (with 96 wells), and the cell-loading script was performed using the manufacturer's instructions.

Each of the 96 capture sites were inspected under a confocal microscope to remove sites containing dead cells as identified by the LIVE/DEAD Cell Viability Assay and also to remove capture sites containing more than one cell. Only cells that were labeled as LIVE were kept for sequencing. Total DNA was isolated from the remaining MEF cells for Sanger validation purposes.

Following capture, reverse transcription and cDNA amplification were performed in the C1 system using the Clontech SMARTer Ultra Low Input RNA Kit for Sequencing v3 which was done according to the manufacturer's instructions. Only 56 Ape1$_{K6,K7}$ and 55 Ape1$_{K6A,K7A}$ single-cell cDNA libraries were obtained at a concentration of 0.09 to 0.55 ng/μl. The majority of failed cells on the capture plate were dead cells. Amplification was performed using the Nextera XT DNA Sample Preparation Kit, and the Nextera XT DNA Sample Preparation Index Kit (Illumina) was used for indexing. After quantification using an Agilent Bioanalyzer, sequencing was performed on two lanes each of the NextSeq500 for the 56 normal cells and 55 mutated cells. Paired-end reads totaling 150 base pairs were generated.

scRNA-seq processing and QC

As mentioned above, only the normal MEFs are used for software testing purposes. Given this, all following QC metrics and processing steps will be exclusive to that dataset.

The bcbio-nextgen v. 1.0.3 pipeline[85] was used to process the RNA-seq data by way of generating Quality Control (QC) checks, alignments to a reference genome, and expression values for genes in each cell. For this analysis, the pipeline was run on the mouse genome v. 10 (mm10) and its annotation was acquired from Ensembl, release 93. MultiQC v. 1.0.dev0 was run to aggregate QC statistics from bcbio-nextgen, samtools v. 1.4, QualiMap v. 2.2.2a[53], and FastQCv. 0.11.5. The aligner hisat2 v. 2.1.0 was used to align reads to the reference genome, and sailfish v. 0.10.1 was used generate expression values.

The QC metrics aggregated by MultiQC were used to determine if it was immediately apparent that any of the cells had failed sequencing. Quality scores along the length of the read are a good way of determining if there was a systematic problem with the library prep or sequence generation. As seen in **Figure 22**, the quality scores for every sample appeared to be high with the exception of a slight dip at the end of the reads. This dip is expected with Illumina sequencing and it was trimmed off using fqtrim v. 0.9.7—this software was also used to trim adapter sequences common to RNA-seq—using default parameters. This trimming resulted in an average length of 139bp for the normal samples.

Sequencing metrics looked good for all cells with each producing around 5 million reads on average as seen in **Table 11**. One exception to this is Cell 07 which only produced 70,000 reads. This cell was removed from the final analysis due to the low number of reads attributed to it.

To further assess the quality of the cells that were sequenced, the origin of the reads was checked to assess whether the reads are from exonic regions. As this is RNA-seq, the majority of reads should be derived from these regions, and if they are not, then

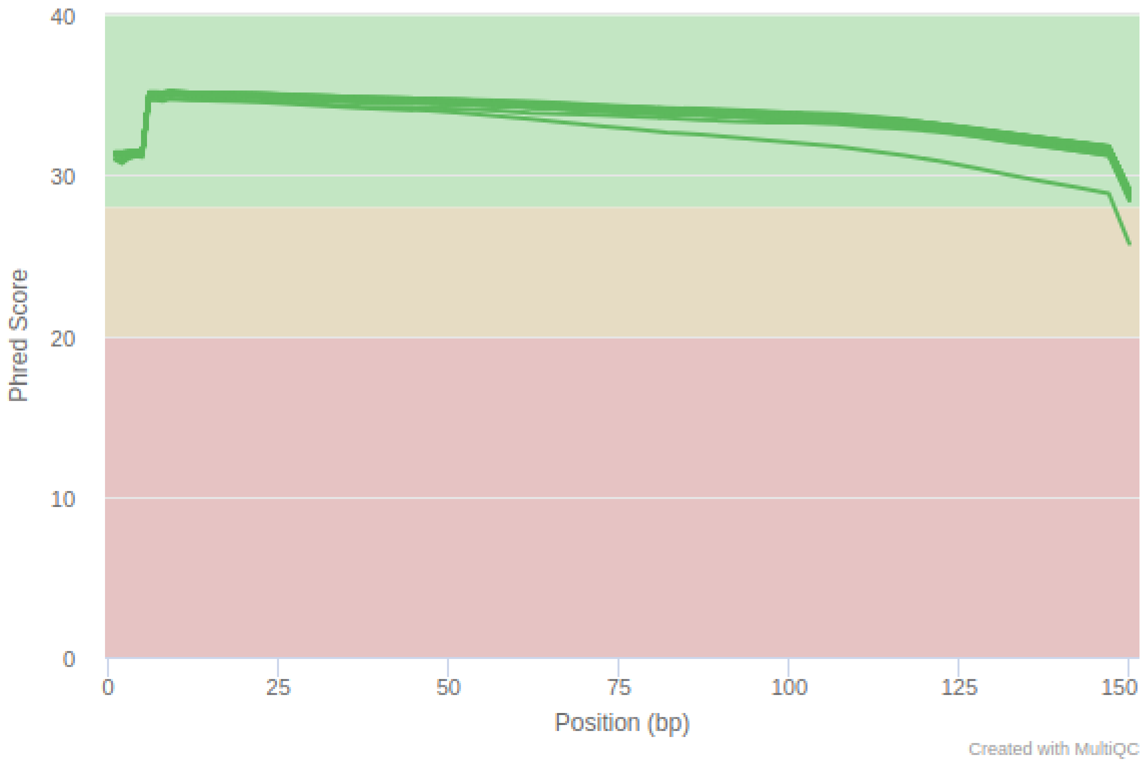


Figure 22. Mean quality scores for reads sequenced from each MEF. The quality score graph generated by FastQC shows the mean quality scores for each normal MEF were generated across the length of the sequenced read.

Sample	Reads	% Dup	rRNA pct	5'-3' bias	% GC
Normal MEFs	5.35 M	40%	1.13%	1.09	46%

Table 11. Average alignment statistics for the MEFs. Average alignment statistics for the 56 normal cells captured on the C1.

there was likely something wrong with the cell upon capture (most likely that the cell was in the process of dying). Any cell that had < 60% of the reads coming from exons were removed. Using this filter, as can be seen in **Figure 23**, only one cell was removed: cell C47.

The last quality control metric employed was checking the Pearson correlation coefficient calculated by comparing the expression profile of all of the cells. To generate these coefficients, a matrix was created where each column to a cell and its expression values for every possible isoform found in the mouse Ensemble 93 database. The value in each cell of the matrix is the expression of the isoform measured by TPM as calculated by sailfish for that particular cell. As these MEFs are from the same population of cells, it is expected that they all contain similar expression patterns. Any cells that are largely deviating from entire the group are likely to be in an altered state that may affect the downstream analyses. As seen in **Figures 24**, only one cell of the 56 sequenced falls into this category and is already being removed due to having a low read count: cell C07. There does exist a block of cells seen at the bottom of the heat map that clusters together, but as they still show significant similarity with a large number of cells, they are retained.

Comparing variant callers using MEF data

Alignment file preparation

For four of the five variant calling pipelines, alignments are prepared by running

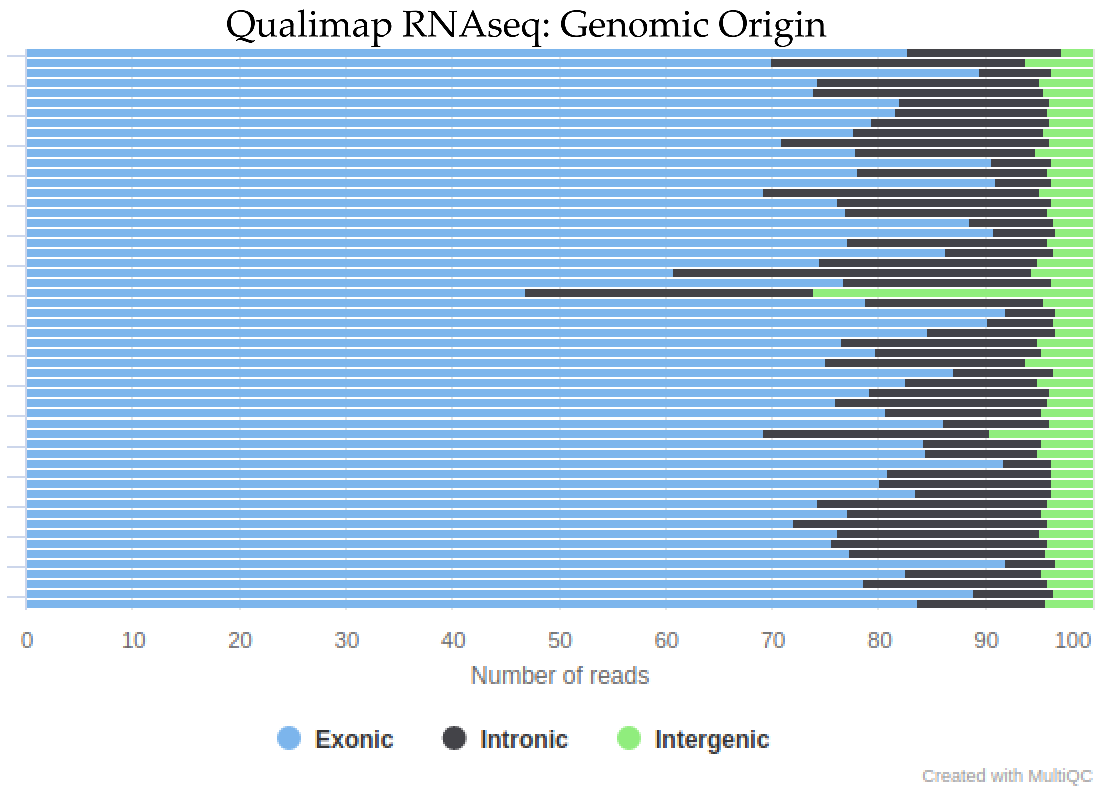


Figure 23. The genomic origin of reads found in each MEF. Here one can see what percentage of reads originate from exons (blue), introns (black) or intergenic space (green). The cell C47 in the normal group is the only cell to have significantly more reads originating outside the exonic region than other samples.

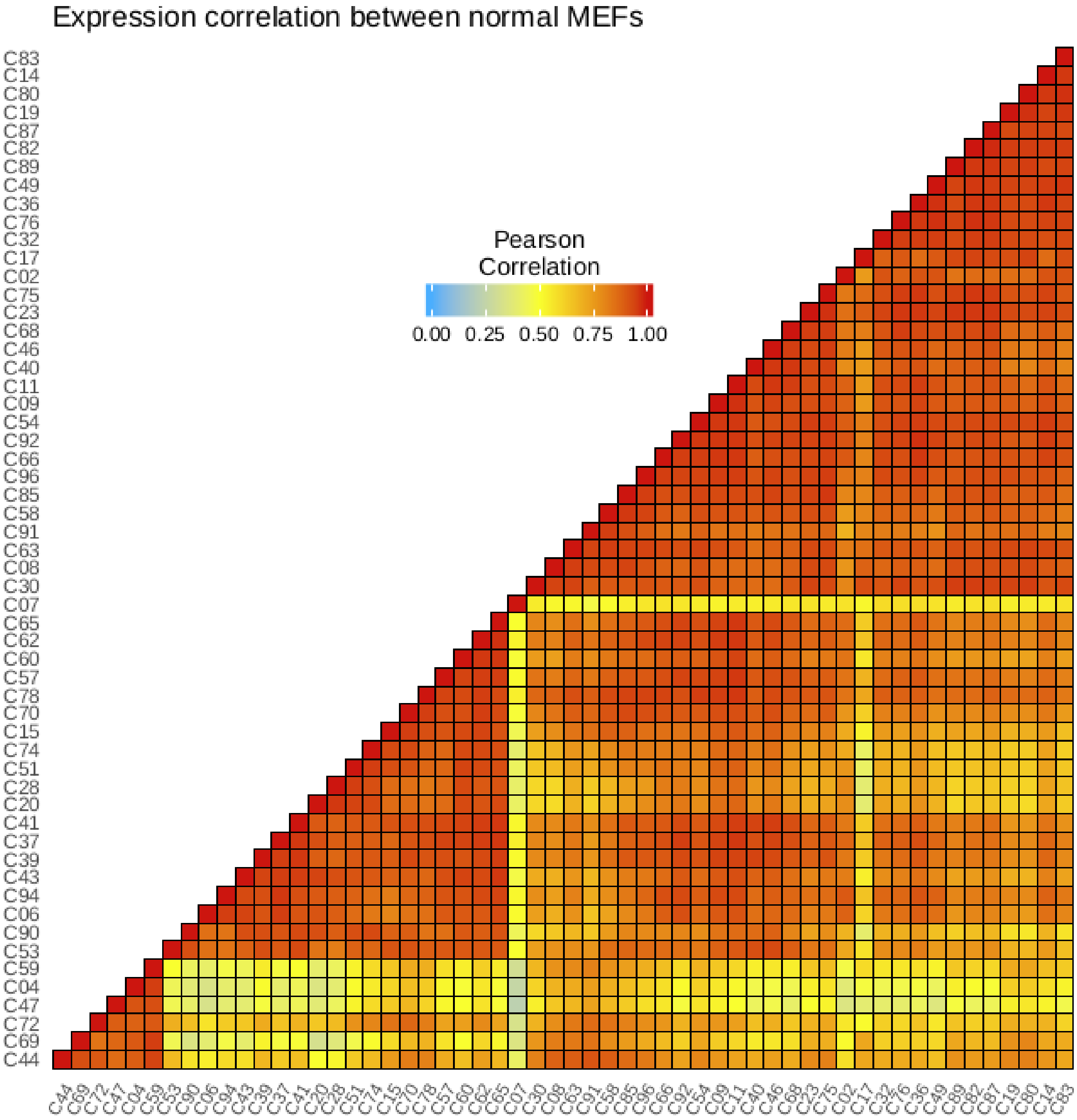


Figure 24. Expression correlation between MEF. Pearson Correlation Coefficient calculated for every possible comparison of cells to each other for the normal MEFs. The darker the color red, the higher the correlation between each cell. Only one cell fails to correlate will with any of the other cells: C07. The bottom block of cells significantly correlates with a high number of cells and they are therefore retained.

hisat2 version 2.1.0 using mouse genome version mm10 as the reference. The exception to this is for the variant caller Platypus which used BWA MEM v. 07.17[67]. Hisat2 is specifically designed for aligning RNA-seq reads and will split reads across exons to ensure the highest-quality alignments. Unfortunately this results in very large fragments that Platypus cannot process. Instead, BWA MEM, another high quality alignment tool, was used in a manner that generated alignments that Platypus could process.

Alignment is followed by running the Genome Analysis Toolkit version 3.8.0 module, SplitNTrim + ReassignMappingQuality. This method of alignment preparation ensures that the highest quality alignment files are generated when trying to identify variants, and thus, that the only deficiencies identified are due to the variant caller itself instead of other processing steps. Specifically, reads aligned around exon junctions have overhanging regions clipped to remove any intronic sequence if that intronic sequence is spurious. Lastly, for every tool, a bed file containing the regions of coding exons in the Ensembl 93 mouse database was used to limit the location of where variants could be identified. The file ref-transcripts.bed contains all of those locations.

FreeBayes

FreeBayes is Bayesian statistical framework capable of identifying, in a reference genome, small variations including SNVs, indels, MNVs, and composite insertion and substitution events so long as these events are shorter than the length of the sequencing read belonging to the alignment. Version 1.1.0.46 was used to generate VCF files containing all variants present in the alignment file for each cell. FreeBayes was run using the following arguments in addition to default parameters:

- --min-alternate-fraction 0.01
- --targets ref-transcripts.bed
- --no-partial-observations

GATK HaplotypeCaller

GATK HaplotypeCaller uses a De Bruijn-like graph to perform local de-novo assembly of regions of the genome that show evidence of significant variation, including SNVs, MNVs, and indels. This has the advantage of more accurately identifying more complex variation in a sample such as indels, and, importantly for our study, splice junctions. However these benefits come at the expense of increased computation time[68]. Version 4.0.7.0 was used to generate VCF files containing all variants present in the alignment file for each cell. GATK HaplotypeCaller was run using the following arguments in addition to default parameters:

- --filter_reads_with_N_cigar
- --standard_min_confidence_threshold_for_calling 4.0
- --dbsnp dbsnp-20130912.vcf.gz
- -L ref-transcripts.bed

GATK UnifiedGenotyper

GATK Unified Genotyper uses a Bayesian genotype likelihood model to estimate the most likely genotypes (SNVs, MNVs, and indels) and allele frequency. GATK UnifiedGenotyper also benefits from not assuming ploidy for the organism being analyzed which is not the case for GATK HaplotypeCaller[68]. Version 3.8-0-ge9d806836 was used to generate VCF files containing all variants present in the alignment file for each cell. This version is different from the GATK HaplotypeCaller version as GATK version 4 does not contain GATK UnifiedGenotyper. GATK UnifiedGenotyper was run using the following arguments in addition to default parameters:

- --filter_reads_with_N_cigar
- --standard_min_confidence_threshold_for_calling 4.0
- --dbsnp dbsnp-20130912.vcf.gz
- -L ref-transcripts.bed
- --geontype_likelihoods_model BOTH

Platypus

Platypus uses local realignment of reads and local reassembly to accurately identify variants--SNVs, MNVs, indels, and long-range insertions and deletions--in a genome. Version 0.8.1.1 was used to generate VCF files containing all variants present in the alignment file for each cell. Platypus was run using the following arguments in addition to default parameters:

- --regions=ref-transcripts.modified.bed
- --filterDuplicates=0

Also, because Platypus does not work on large read fragments and is using the BWA MEM-aligned files, the ref-transcripts.bed file had to be modified to include entire gene regions as well as modifying other functional elements such as lncRNA annotations.

Comparison of all tools using raw counts and cell-to-cell comparisons

After VCF files were generated for each cell by all five tools, they were then compared by looking at the total number of variants identified as well as the percentage of variants identified by each tool that are shared between each of the cells.

The average number of variants identified can be seen in **Table 12**. FreeBayes identifies the highest average number of variants per cell followed Red Panda, then GATK HaplotypeCaller, GATK UnifiedGenotyper, and Platypus. This is unexpected based on the results from the human articular chondrocyte data where FreeBayes had the fewest number of variants shared between the scRNA-seq results and the exome. One explanation from this is that FreeBayes may identify a high number of variants, but the majority of those are False Positives. This idea is supported by the PPV numbers as seen in **Figure 20**.

After the total variant numbers were calculated, comparisons between cells were performed. To test the initial hypothesis that any two cells should share a high number of variants, only data produced by Red Panda from cells C02 (353 total variants) and C06

	FreeBayes	GATK HaplotypeCaller	GATK Unified Genotyper	Platypus	Red Panda
Average # of variants	567.22	423.93	387.82	315.42	510.02
Standard deviation	161.16	124.44	106.80	107.16	143.26

Table 12. Average variant count and standard deviation for each tool. For this analysis, the total number of variants identified by each tool after filtering steps is reported.

(593 total variants) were compared. This measurement was performed by taking the list of variants in cell C02 and checking to see if there was sufficient sequencing coverage (20x coverage) and variant match at the corresponding location for each variant in cell C06. Variants satisfying both criteria were then added to the list of common variants between the two cells; however, if it is not found in C06's VCF file it is added to a list containing variants that are present in C02 but not C06. Using these two lists it is possible to calculate the percentage of overlap between the two cells.

The resulting percentage of overlap between C02 and C06 was 20.1%. This makes sense as it is difficult to do a one-to-one comparison of locations between the two cells, even after we've guaranteed there are sufficient reads covering that location. As mentioned previously, monoallelic expression makes it impossible to say for sure that a variant is not shared based on scRNA-seq data. One way to help combat this hurdle is by only looking at homozygous-looking variants. Heterozygous variants have the disadvantage of having increased ambiguity, but homozygous-looking ones are more likely to show up in both cells even when factoring in the idea that the list of homozygous-looking variants will contain heterozygous variants that are the result of monoallelic expression.

To test this, variants that are homozygous-looking in both C02 and C06 are compared. This results in a much higher percentage of overlap than identified previously: 80.3% were found to be shared which is what is expected from cells that should be isogenic. Additionally, the fact that it's not 100% is also expected: variants that look homozygous may actually be heterozygous.

After doing this initial analysis, the comparison was expanded to analyzing the percentage of pairwise overlap between six cells for all five tools: C02 vs. [C04, C06, C08, C14, and C40]. This was done for all variants in those cells as well as those that are only homozygous-looking. As seen in **Figure 25**, Platypus and Red Panda perform well in the latter category, and Red Panda fared well over Platypus in four out of five comparisons. But barring one comparison for Platypus (C02 vs C04), all of the tools performed poorly (under 25%) when looking at all variants.

This minimal comparison of variant overlap resulted in more questions than answers which prompted a full reassessment of the variants identified by each tool. To ensure as close of a one-to-one comparison between the human articular chondrocyte data and the MEF data as possible, the MEF variant calling was originally restricted to coding exons, which resulted in poor overlap of variants across cells. However, this restriction precludes a large portion of the genome that contains functionally important elements such as the untranslated regions (UTR) of genes. With this in mind, variant reporting and pairwise comparisons were re-run without the restriction of only keeping those in coding exons.

This led to new results for the average number of variants reported by each tool as seen in **Table 13**. Red Panda identifies the highest number of variants of all the tools. Assuming we can extrapolate from the PPV results from the articular chondrocyte data, this also means that Red Panda is identifying the highest number of True Positives as well. One thing to note is that the numbers for Platypus did not change as its pipeline had already been modified include these regions. This was necessary because Platypus

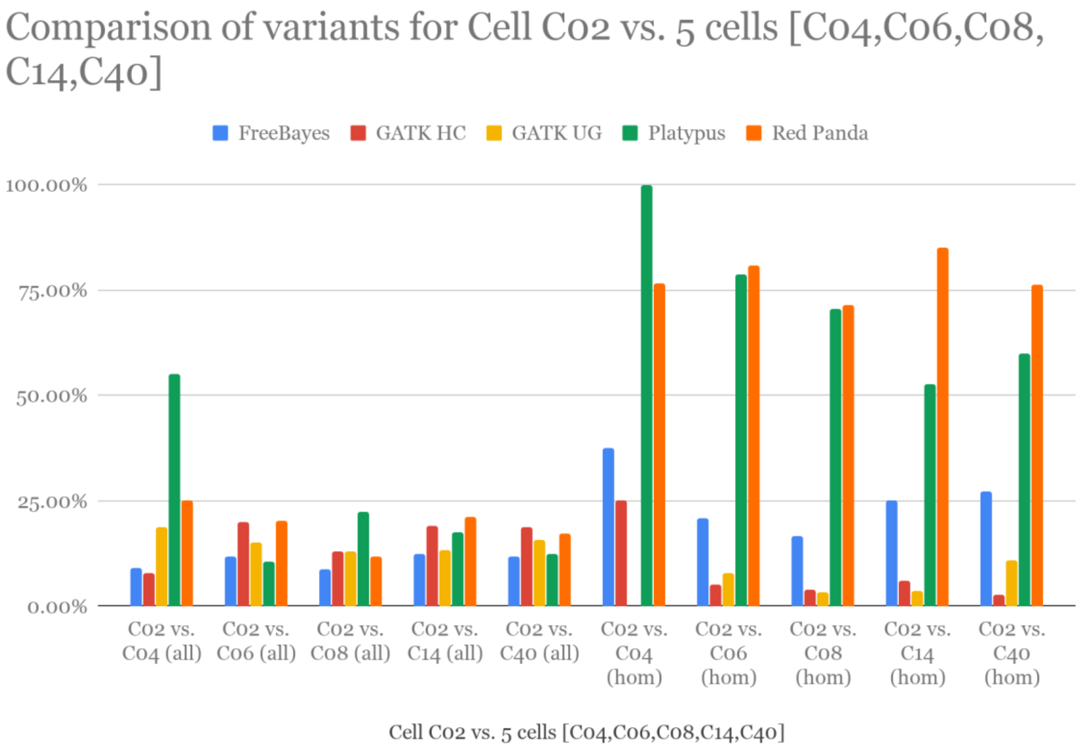


Figure 25. Comparison of variants between Cell C02 and five other cells [C04, C06, C08, C14, and C40]. Pairwise comparisons between C02 and five other cells was performed for every variant caller after filtering was performed on the variants. Here we can see Platypus and Red Panda performing well for homozygous-looking variants and every other tool performing poorly when looking at all variants.

	FreeBayes	GATK-HC	GATK-UG	Platypus	Red Panda
Average # of variants	865.87	611.15	574.73	315.42	1071.83
Standard deviation	235.36	195.17	170.59	107.16	372.67

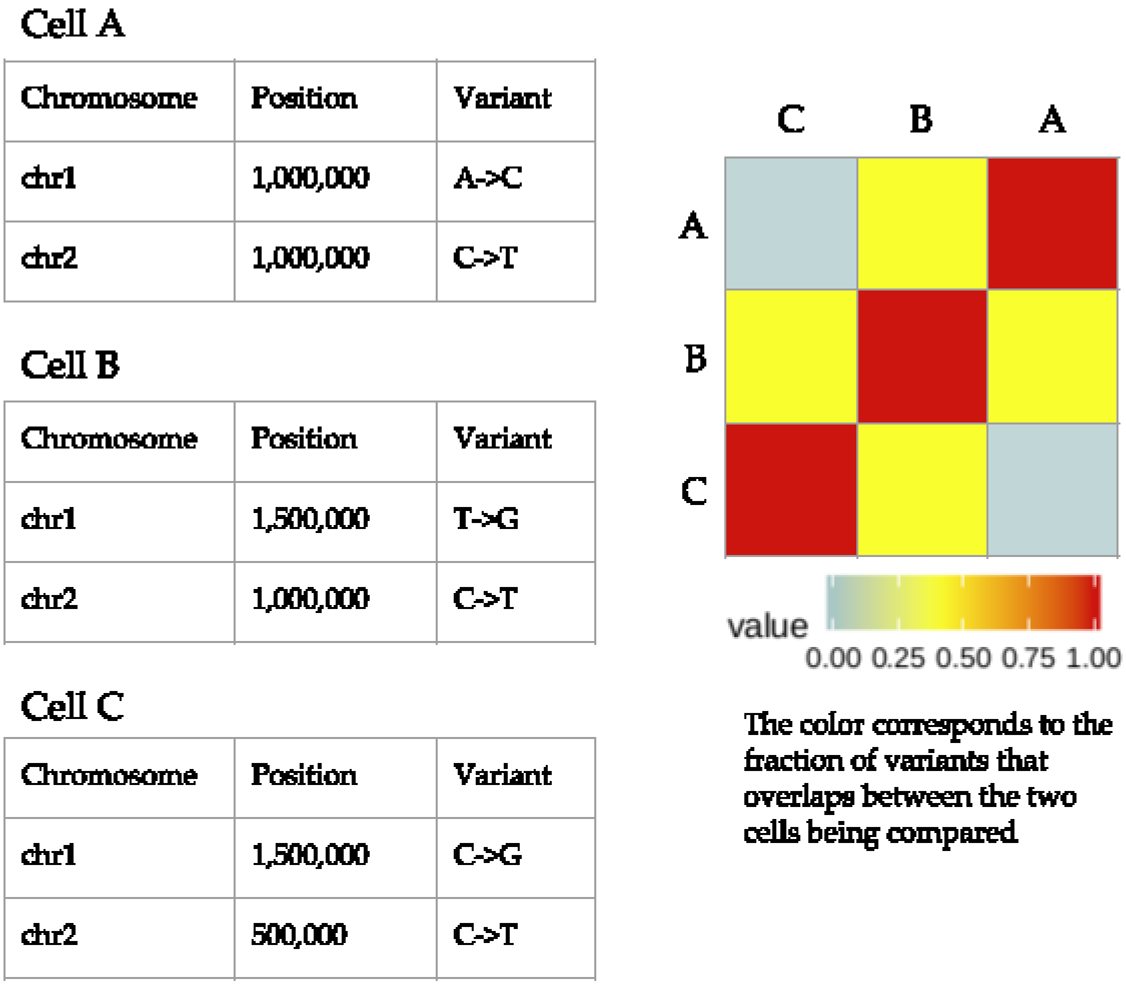
Table 13. Average variant count and standard deviation for each tool after filtering steps were removed. For this analysis, the total number of variants identified by each tool now includes all areas of the genome.

will not run if it has to process reads spanning large regions such as exon-intron junctions. This restriction meant that the original analysis could not be limited to just coding exons for Platypus.

After removing the restriction on where in the genome variant calling can occur, the same pairwise comparison as described previously was performed for all 3,025 (55 cells * 55 cells) possible comparisons. Additionally, three groups of variants were assessed instead of two: percentage of overlap for heterozygous variants was also added to assess how each software handles these types of variants. Custom scripts and the R package ggplot2 were used to generate these comparisons.

To visualize this analysis, a heat map was generated showing the fraction of overlap between each cell. A hypothetical scenario is illustrated in **Figure 26**. Higher overlap leads to a redder color and low overlap leads to a bluish-green color. As we can see in **Figure 27** and **Figure 28**, each tool, especially Red Panda and FreeBayes, performs well when only the homozygous-looking variants are being compared, but they all suffer when assessing purely heterozygous variants.

Interesting to note is the fact that while GATK HaplotypeCaller and GATK UnifiedGenotyper both have a good fraction of overlap among the homozygous-looking variants, it does not appear to affect the overall fraction of variants identified by these tools because the heat maps look almost identical for both the “Heterozygous” and “All Variants” maps. This is because homozygous-looking variants do not contribute much to the total list as seen in **Table 14**. Also important is that Red Panda performs extremely well for homozygous-looking variants, but is average for heterozygous variants. This is



Cell A

Chromosome	Position	Variant
chr1	1,000,000	A->C
chr2	1,000,000	C->T

Cell B

Chromosome	Position	Variant
chr1	1,500,000	T->G
chr2	1,000,000	C->T

Cell C

Chromosome	Position	Variant
chr1	1,500,000	C->G
chr2	500,000	C->T

The color corresponds to the fraction of variants that overlaps between the two cells being compared

Figure 26. Example of comparisons made between cells. In this example heat map, three cells are being compared. There is one variant in common between cells A→B and B→C but none in common between A→C.

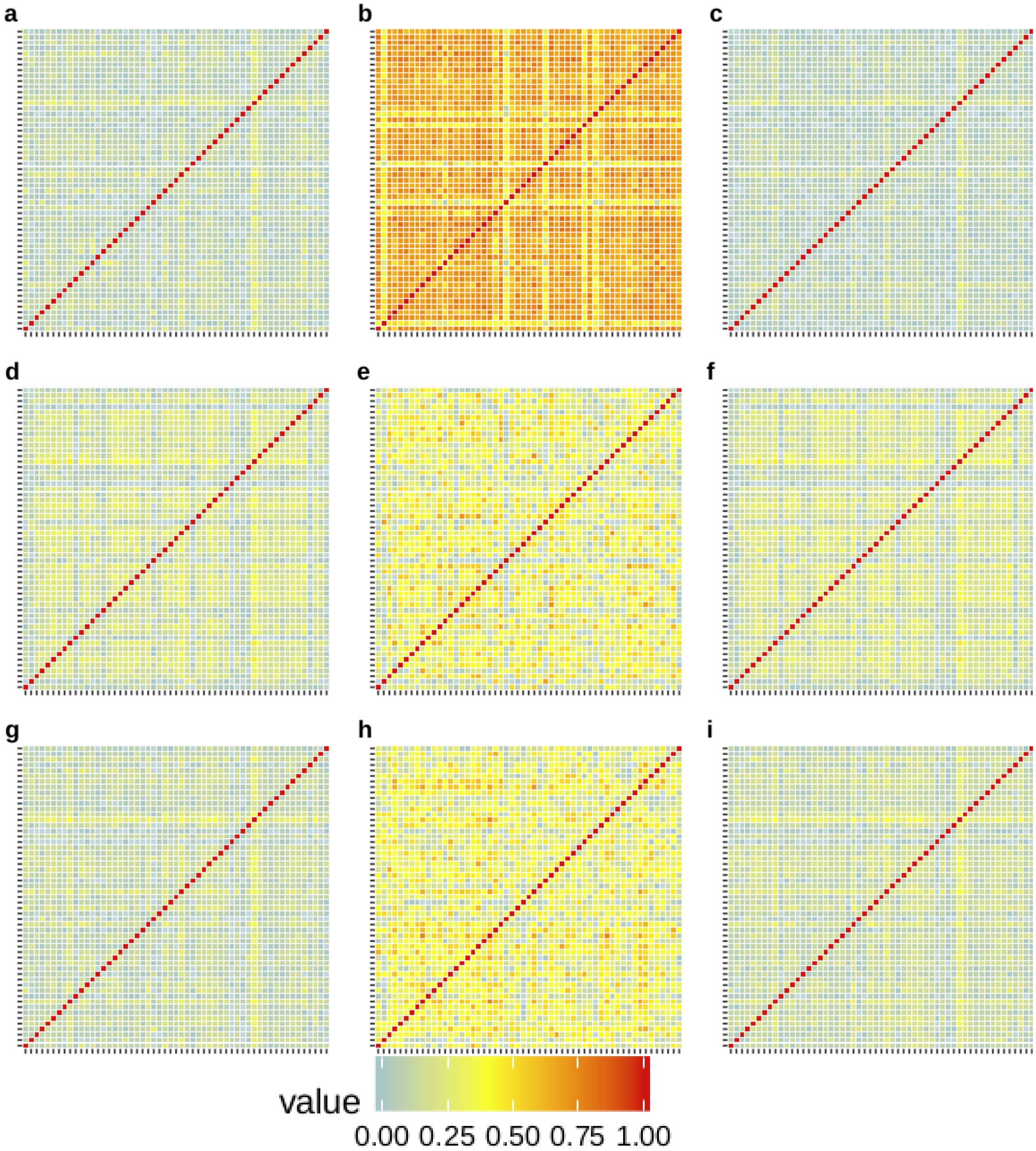


Figure 27. The fraction of overlap in variants for every cell using FreeBayes, GATK HC, and GATK UG. The fraction of overlap for **(a-c)** FreeBayes, **(d-f)** GATK-HaplotypeCaller, and **(g-i)** GATK-UnifiedGenotyper when comparing **(a, d, g)** all variants, **(b, e, h)** homozygous-looking variants, and **(c, f, i)** heterozygous variants. Each box in the matrix is a comparison between two cells.

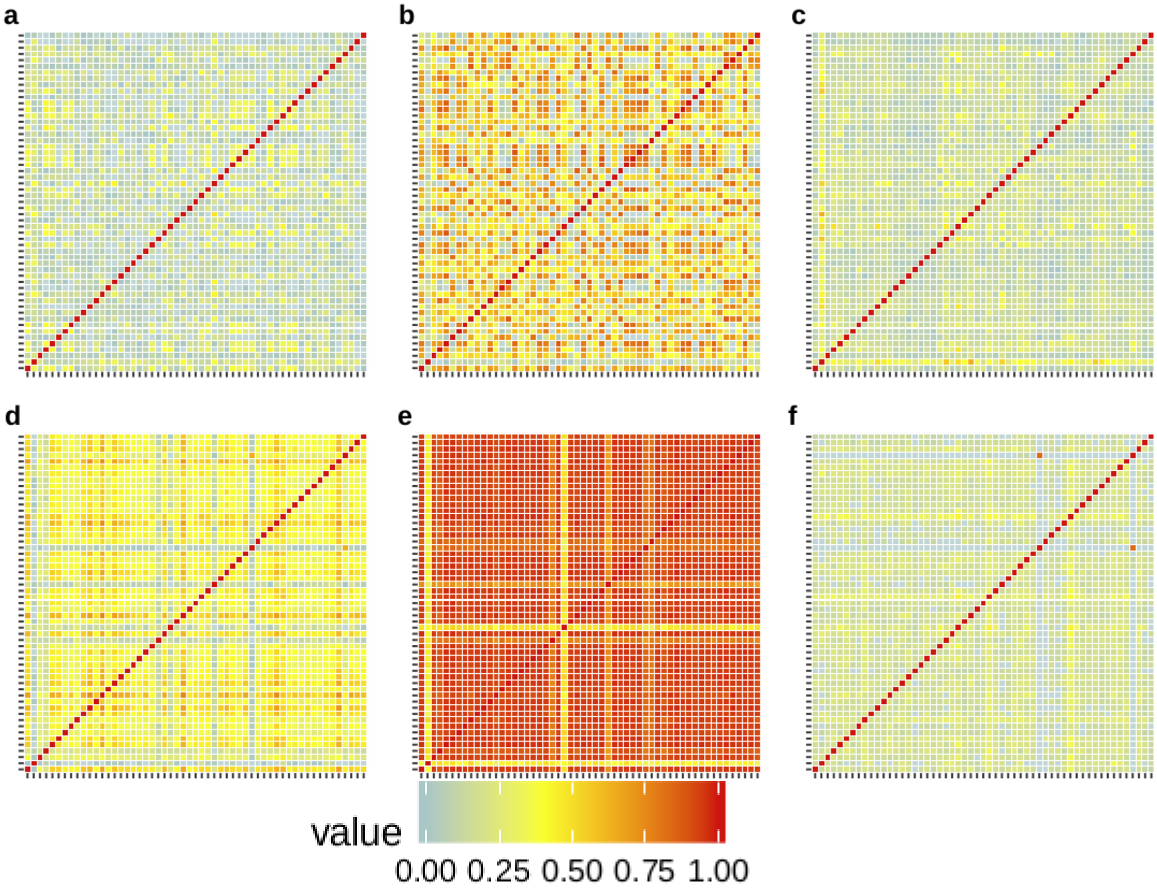


Figure 28. The fraction of overlap in variants for every cell using Platypus and Red Panda. The fraction of overlap for **(a-c)** Platypus and **(d-f)** Red Panda when comparing **(a, d)** all variants, **(b, e)** homozygous-looking variants, and **(c, f)** heterozygous variants. Each box in the matrix is a comparison between two cells.

Tool	All variants	Homozygous variants	Heterozygous variants
FreeBayes	85.70	35.22	50.48
GATK-HC	72.06	6.66	65.40
GATK-UG	58.72	11.35	47.37
Platypus	28.83	19.90	8.93
Red Panda	171.34	115.39	55.95

Table 14. Average number of variants overlapping for pairwise comparisons between cells.

due to the fact that, while Red Panda confers an advantage to identifying heterozygous variants, the majority of those identified are not bimodally-distributed and are thus actually picked up by GATK HaplotypeCaller as seen in **Figure 19**.

To better visualize the distribution of the fraction of overlap of variants as well as the total variants overlapping in the pairwise comparison, violin plots were created for all three classes of variants (**Figure 29)**, and all comparisons between Red Panda and the other tools were statistically significant. These show that there tends to be a tight distribution of values for every tool in each class with the exception of Red Panda and its ability to identify homozygous-looking variants. The total number of homozygous-looking variants has a wide distribution, but the distribution of the fraction of variants in this class is tight and very close to 1. It follows that the same trend can be seen in total variants shared in these comparisons since homozygous-looking variants make up the largest proportion of those identified by Red Panda to be overlapping between two cells as can be seen in **Table 14.**

In **Figure 29** it is clear that Red Panda performs the best both at identifying raw numbers of variants shared between cells as well as fractions of variants shared. This is an important distinction because, as can be seen with FreeBayes, it's possible to have a large number of variants shared while also having a small fraction of the total variants *possible* be shared. As a hypothetical example, there may be 200 variants out of 1000 possible shared between two cells. This results in a large number of variants shared (200), but a small fraction (0.2) of total possible variants shared. This pattern indicates that there are a lot of potential False Positives in the FreeBayes results which fits with

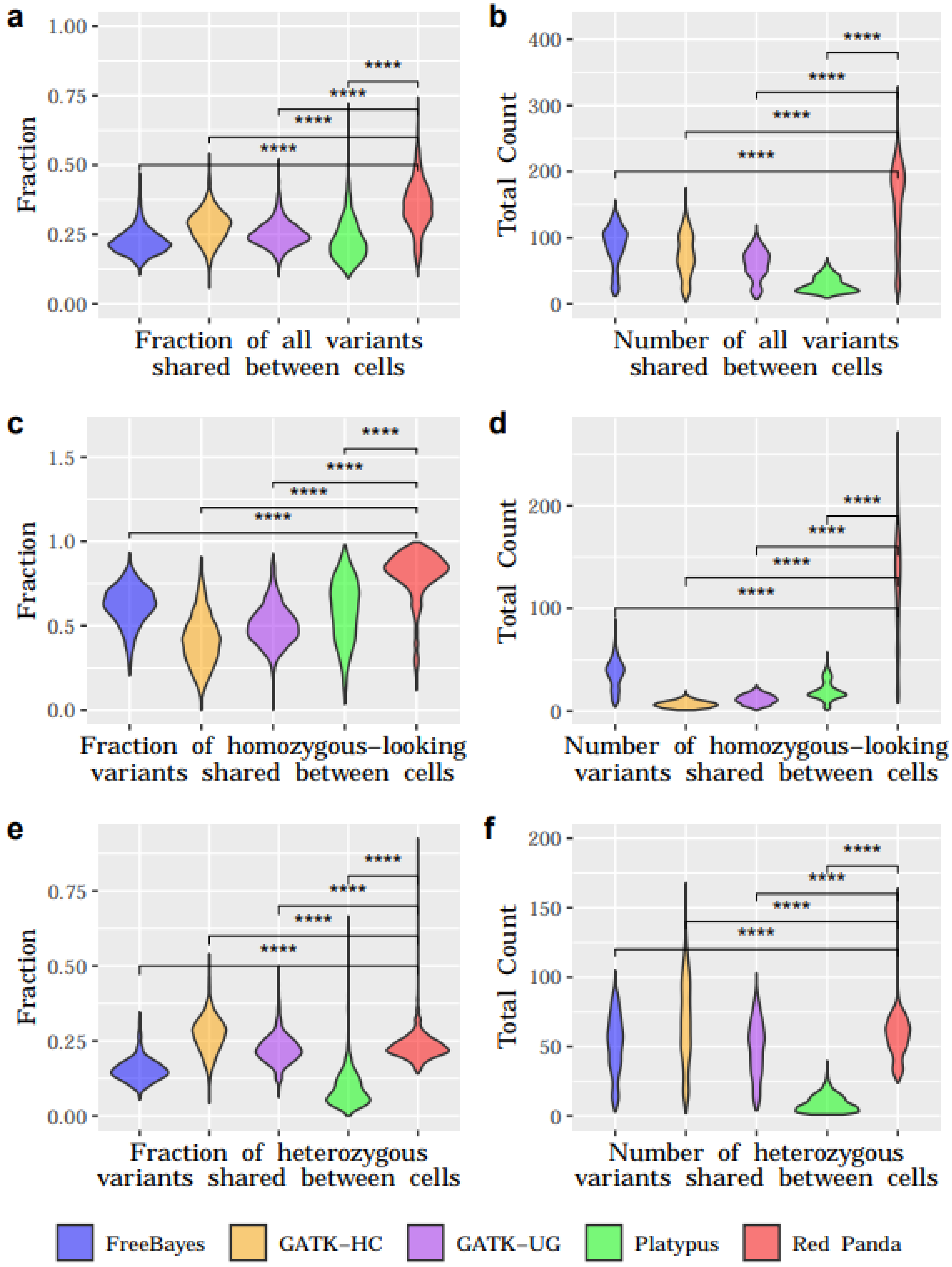


Figure 29. Violin plots for variants shared between cells. Violin plots showing the fraction (left) and quantitative (right) overlap for **(a, b)** all variants, **(c, d)** homozygous-looking variants, and **(e, f)** heterozygous variants shared in every pairwise cell comparison. Comparisons were performed using T-tests: ns = not significant, * is p < 0.05, ** is p < 0.01, *** is p < 0.001, and **** is p < 0.0001.

what was seen in the articular chondrocyte sequencing.

As we saw in the heat maps, it is clear that the highest fraction of variants shared pairwise between two cells comes from the homozygous-looking class. This is made very evident where, again, Red Panda has the highest number of variants and the highest fraction of variants shared. However, it is clear that it is the homozygous-looking class that is contributing most to the total shared between cells for Red Panda. This is significant as it was originally thought that Red Panda would perform the best among the heterozygous variants, rather than the homozygous-looking variants.

As for the other tools, it appears that despite performing the worst at identifying shared homozygous-looking variants GATK HaplotypeCaller performs the best with heterozygous variants. This is useful as it proves that using GATK HaplotypeCaller was a good choice as the supplemental variant caller for Red Panda when it cannot use its unique bimodal distribution model. Lastly, it's interesting to note that it is rare for any tool to have more than 100 heterozygous variants shared between cells. This is likely due to the stochastic nature of allele-specific expression.

Comparing variant callers using simulated data

To accurately assess the sensitivity of each variant calling tool in a controlled environment, simulated data was generated against which we can compare the results from the five different variant callers. This simulation consists of ~1,000 variants generated per cell. They are created by programmatically inserting variants into the alignments generated from the normal MEFs. The list of simulated variants consists of

650 homozygous variants and 350 heterozygous variants, roughly 70 of which are bimodally-distributed. These numbers are used because they are close to the proportions seen in the variants corroborated by the exome sequencing in the articular chondrocyte data, and while these proportions do not match those expected based on bulk sequencing experiments[86–88], they do match what is expected from scRNA-seq data[31].

Once this simulated list is created, the variants are inserted into the original alignment files for each cell. These new alignment files are used as input for the variant calling pipelines described above for the MEF sequencing data. Once a new list of variants has been generated by each of the five tools, the results will exclusively be compared against the list of simulated variants. For this analysis, all other locations identified in the standard variant calling process are not considered as they are due to normal variation in that cell and not part of the simulated set.

To generate the new alignment files containing the simulated variants, a number of important steps have to performed, as illustrated in **Figure 30**. To ensure variants could be identified by each tool, a read depth cutoff of 20 was used for locations where variants could be inserted. This was achieved using the coverage module from the bamtools package[89]. This list of potential insertion locations was then used as the source of the locations of the ~1,000 simulated random variants. For the 650 homozygous and 280 heterozygous variants, the locations were not restricted except that they must originate from this list.

The bimodally-distributed heterozygous variants had a number of extra parameters that determined their placement. It was required that they originate from

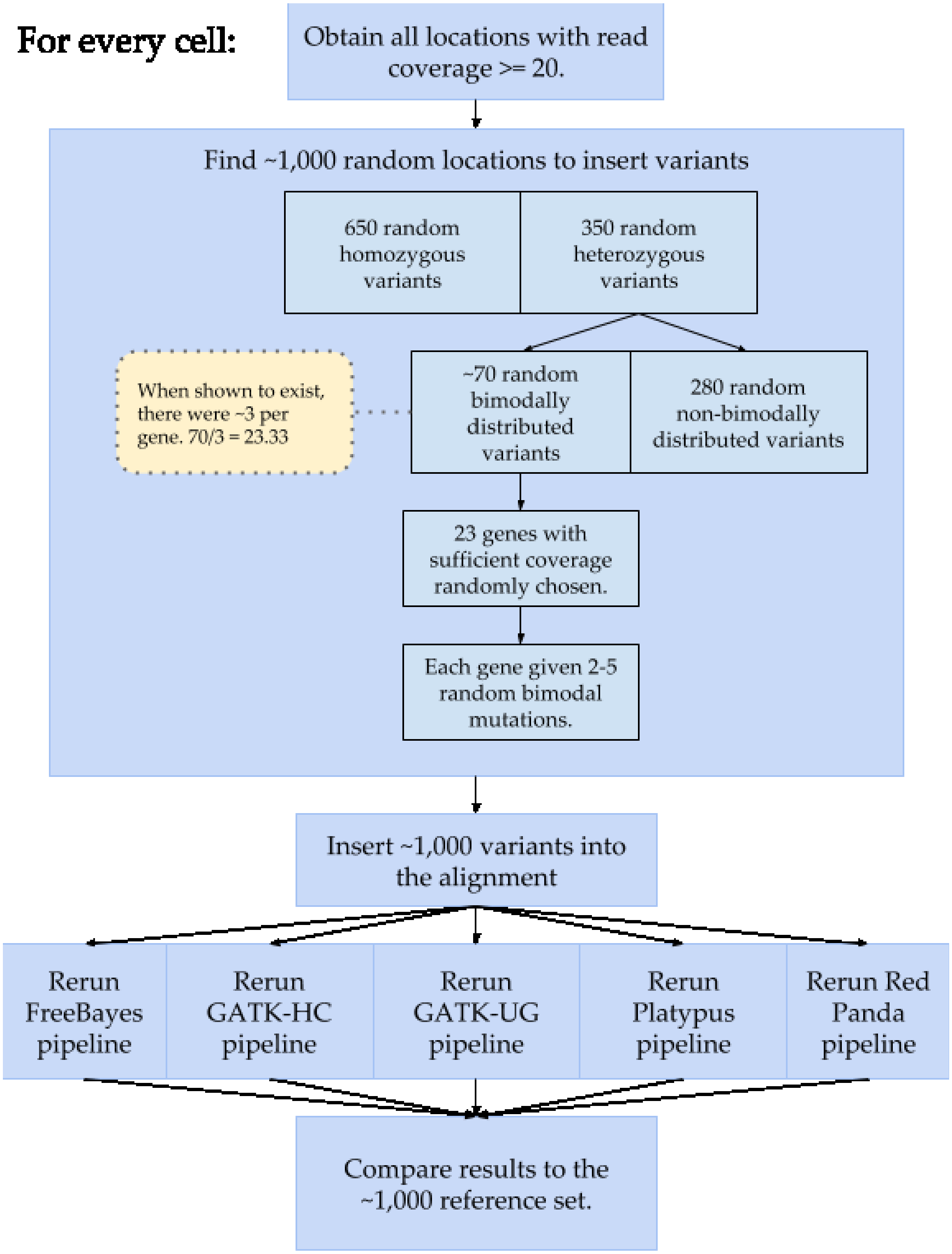

For every cell:
Obtain all locations with read coverage >= 20.
Find ~1,000 random locations to insert variants
650 random homozygous variants
350 random heterozygous variants
When shown to exist, there were ~3 per gene. 70/3 = 23.33
~70 random bimodally distributed variants
280 random non-bimodally distributed variants
23 genes with sufficient coverage randomly chosen.
Each gene given 2-5 random bimodal mutations.
Insert ~1,000 variants into the alignment
Rerun FreeBayes pipeline
Rerun GATK-HC pipeline
Rerun GATK-UG pipeline
Rerun Platypus pipeline
Rerun Red Panda pipeline
Compare results to the ~1,000 reference set.

Figure 30. **Workflow for inserting simulated variants**. To assess each tool, ~1,000 simulated variants (650 homozygous, 280 heterozygous, and ~70 bimodally-distributed heterozygous) were inserted into the alignments for each cell. Standard variant calling was then performed using each tool, and these results were compared to the list of known variants to assess their performance.

genes that were considered to be expressed (TPM > 1) and that they have a minimum of two of variants placed in the expressed gene. From the MEF sequence data, an average of ~3 (a range of about 2–5) variants per gene was observed; this results in roughly 23 genes containing bimodally-distributed variants: 70/3 = 23.33. Thus, to simulate this class of variation, 23 expressed genes were randomly chosen wherein 2, 3, 4, or 5 variants were randomly inserted into the gene, but only if there were more than 250bp of viable locations where a variant could be inserted. That is, a gene was only in consideration to have bimodally-distributed variants added to it if more than 250bp of its sequence had sequencing depth of at least 20 reads.

After ~1,000 locations had been chosen, a new alignment file was created containing the simulated variants. Additionally, a VCF file containing all of the variants was generated for ease of comparison in the downstream analysis.

This type of custom simulation was necessary because, while there are a number of methods available to imitate read counts and expression profiles[90–96], there currently exist no tools to generate scRNA-seq reads *in silico*. Were such a tool available, raw reads with built-in variation would have been generated, from which accuracy metrics for the variant calling tools could have been calculated. However, since this was not possible, random variants were inserted into the alignments already generated in the MEF analysis. This has the benefit of recreating a more realistic simulation environment because all of the artefacts and flaws inherent to scRNA-seq are maintained. One downside however, is that it disallows us from calculating any accuracy statistics requiring False Positive numbers. Since real scRNA-seq data is being used, it already contains variation inherent to the MEFs. It follows that the variant callers will pick up this possibly-real variation that would then be classified as a False Positive because it is not contained in our list of ~1,000 simulated variants for that cell.

Due to this limitation, sensitivity is the main metric by which each tool is measured, and it was calculated for each tool across every cell. These numbers were then plotted using a violin plot to assess the distribution of True Positives identified by each variant caller. **Figure 31** (boxplots for the raw number of True Positives) and **Figure 32** (the aforementioned violin plots for sensitivity) show that for homozygous variants and bimodally-distributed heterozygous variants, Red Panda consistently performs better than the other four tools, and its results are statistically significantly different than the results from every other tool when compared using T-tests. For heterozygous variants taken as a whole, FreeBayes performs the best of the tools. It is unsurprising then that Red Panda does not perform as well in this category because it uses GATK HaplotypeCaller (shown to accurately identify few heterozygous variants in this simulation) to validate heterozygous variants that do not follow a bimodal distribution. In this instance, GATK HaplotypeCaller and GATK UnifiedGenotyper perform poorly because they both utilize a feature where all samples are considered simultaneously. This results in poorer performance on samples that are more genetically diverse, or put another way, single cells that have private mutations. And for our simulation, the mutations being tested are unique to each cell. Red Panda does not suffer from this limitation as it explicitly directs GATK-HC to call variants at specific locations one at a

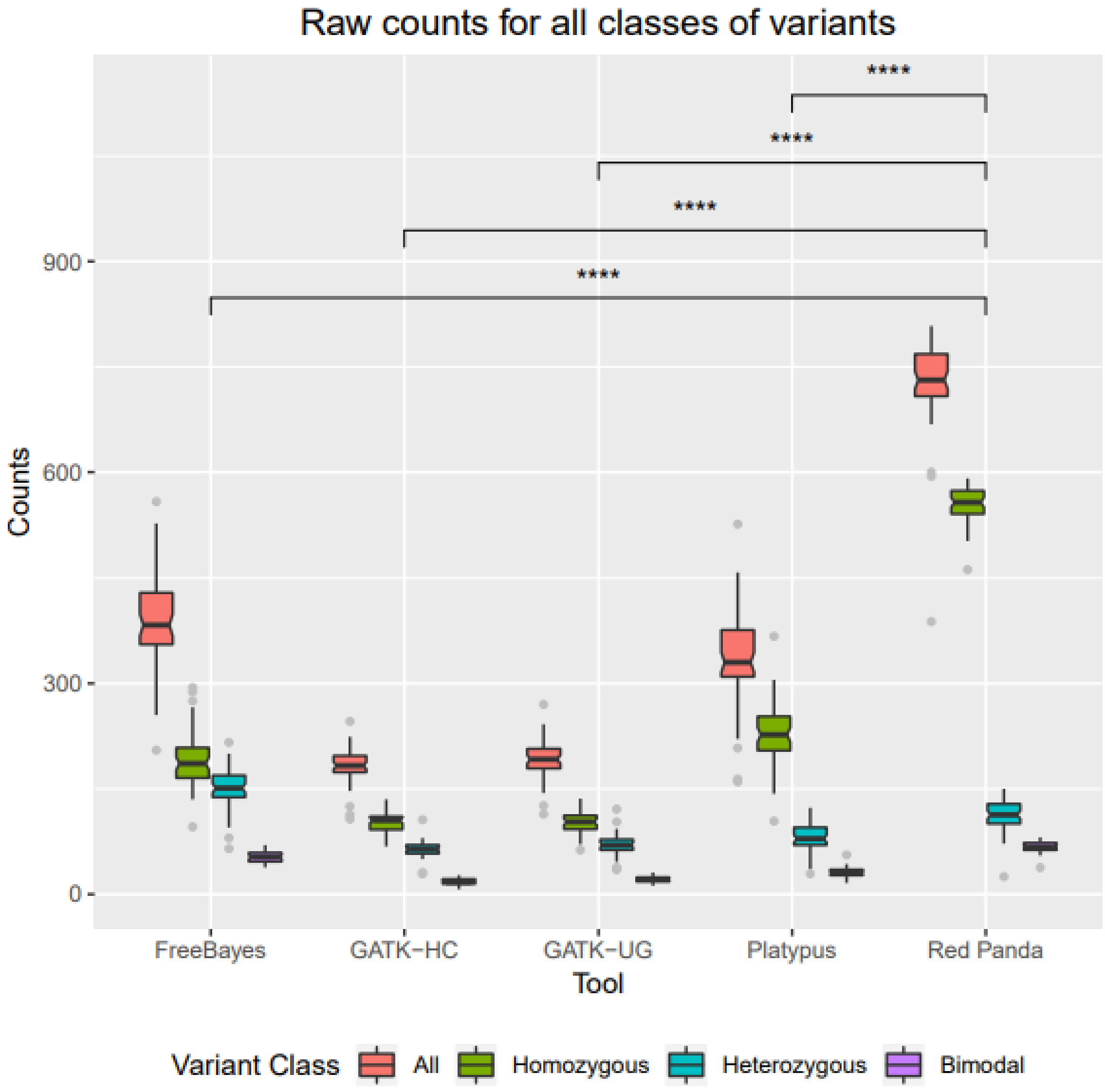


Figure 31. Raw counts of True Positives for each tool. The box plots of the raw number of True Positives show how well each tool is at identifying variants in the simulation for: all variants, all homozygous-looking variants, all heterozygous variants, and all bimodally-distributed heterozygous variants. Due to advantages gained in identifying homozygous and bimodally-distributed variants, Red Panda identifies the highest number of True Positives. Comparisons were performed using T-tests: ns = not significant, * is p < 0.05, ** is p < 0.01, *** is p < 0.001, and **** is p < 0.0001.

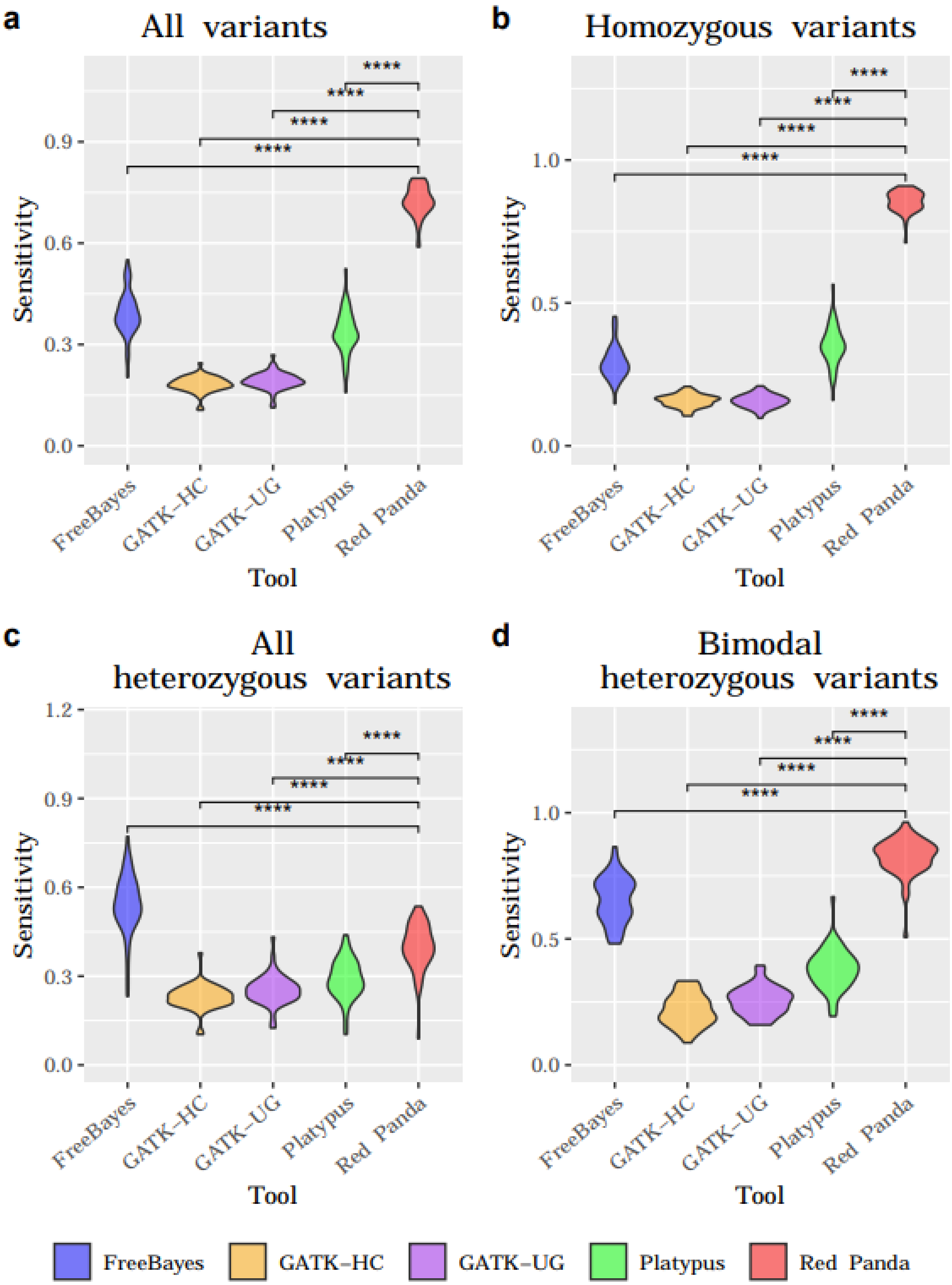


Figure 32. Sensitivity for identifying simulated variants for each tool. The violin plots of the sensitivity, calculated for each cell using each class of simulated variants are shown: (a) all variants, (b) homozygous variants, (c) all heterozygous variants, and (d) bimodally-distributed variants. Comparisons were performed using T-tests: ns = not significant, * is $p < 0.05$, ** is $p < 0.01$, *** is $p < 0.001$, and **** is $p < 0.0001$.

time rather than jointly. However, as seen in the results in **Figure 29f**, this can result in lowered sensitivity for samples that are genetically similar.

Evaluation of Red Panda by Sanger confirmation on MEFs

To complement the simulated data, Sanger sequencing was performed on 40 variants: 20 that were identified by all five software and 20 that were only identified by Red Panda. These variants were pulled from the VCF files generated for cell C14 as seen in **Figure 33**, and cross-referenced with the other 54 cells to make sure that at least one other cell contained the variant. This low number—a minimum of two cells sharing the same variant—was necessary because it was rare for variants to be present in more than two cells when pulling from the list that were identified by all five variant callers as seen in **Table 15**. This table shows a breakdown of the number of variants identified by each tool that were common to: at least 2, 5, 10, 22 (50%), and 41 (75%) of the cells. Ideally, a minimum of 20 variants unique to each tool would have been chosen for Sanger sequencing, but it was outside the scope of this project to do so.

Primer Design

Primers (see Appendix A) were designed to amplify specific target regions containing 39 SNVs and one indel. To generate sequence fragments for primer design, 900bp sequences upstream and downstream of the variant were obtained from the mouse genome mm10 using samtools faidx. This resulted in an 1800bp fragment that was searched for primers using Primer3Plus. The parameters in **Table 16** were used to

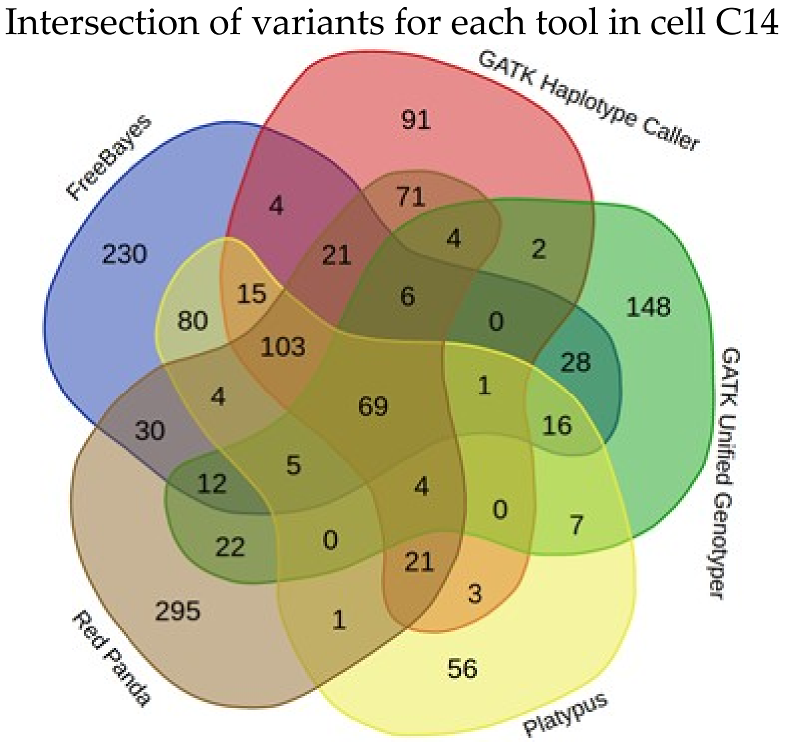


Figure 33. The overlap in the number of variants identified in the cell C14. Here we can see that there is a sizeable (69 variants) overlap between all the tools that identified variants in this cell. Red Panda identifies the most variants unique to a specific variant caller (295). It is from these two populations that variants are chosen for confirmation via Sanger sequencing.

Present in:	FreeBayes	GATK-HC	GATK-UG	Platypus	Red Panda	Intersection of all tools
>= 2/55 cells	2922	2463	1991	2947	3159†	96*
>= 5/55 cells	970	894	693	324	1051	18
>= 10/55 cells	416	398	309	161	565	0
>= 23/55 of cells	129	122	84	66	257	0
>= 42/55 of cells	38	24	27	22	98	0

Table 15. Breakdown by tool of variants present in more than one cell. The number of cells in which a variant was found was broken down into five groups: presence in at least 2, 5, 10, 23, or 42 of cells. Additionally, the variants identified by all tools were checked for their presence in the five groups listed above. The variants submitted for Sanger sequencing were drawn from the two groups labeled with a cross (†) and an asterisk (*).

Parameters used to design primers on Primer3Plus

Product Size Range	401-700*
Min primer	18
Opt primer*	20
Max primer	27
Primer Tm Min	57
Primer Tm Opt	60*
Primer Tm Max	63
Max Tm difference	100
Primer GC% Min	20
Primer GC% Opt	50*
Primer GC% Max	80
Concentration of monovalent cations	50
Concentration of divalent cations	0
Annealing Oligo Concentration	50
Concentration of dNTPs	0
Max Self Complementarity	4*
Max #Ns	0
Max Poly-X	5
CG Clamp	1*
Max 3' Self Complementarity	3
Max 3' Stability	9
Pair Max Repeat Mispriming	24
Pair Max Template Mispriming	24

Table 16. Parameters used to design the primers used for PCR and Sanger. Parameters with a * were changed from their defaults to ensure good sequencing.

design the primers. In total, 38 primer pairs were created as seen in **Appendix A** and were named based on the range of sequence that was searched by Primer3Plus. Only 38 pairs were needed as two variants could be validated by the primer pair for "chr8:85260471-85262071" and two variants can be validated by the primer pair for "chr19:60770223-60771823".

PCR Amplification

Shrabasti Roychoudhury and Suravi Pramanik performed the PCR amplification. For the first round of sequencing, the PCR reaction was performed using GoTaq Hot start polymerase following the manufacturer's protocol with an annealing temperature (Tm) of 50°C. After amplification, PCR products were run on 1.5% agarose gel and visualized in Kodak gel doc and specific DNA bands were recovered using QIAquick Gel Extraction Kit. Purified DNA products paired with their forward primer in 0.2 mL PCR 8 tube-strips were then submitted to Genewiz for Sanger sequencing.

Additional amplification and second sequencing pass was performed on the 18 fragments containing variants specific to Red Panda due to the poor quality as seen in **Table 17**. To attempt to increase the quality of the PCR reactions, a Tm 55°C was used, followed by running the PCR products on 2% agarose gel. Each fragment was added to two 0.2 mL PCR 8 tube-strips wherein one tube contained the forward primer and one tube contained the reverse primer resulting in 36 total products to be submitted to Genewiz for Sanger sequencing. This increased the likelihood of obtaining usable sequence.

Number	Hom/Het	Variant Location	Variant	Validated by Sanger	Cells supported by
1	Het	chr1 43954701	T→G	🟥 N	2
2	Hom	chr1 181176175	C→T	🟦 N	2
3	Het	chr2 3328501	G→T	🟦 N	2
4	Het	chr2 22940605	G→C	⬛	2
5	Het	chr2 33246775	A→G	🟦 N	2
6	Het	chr2 39195366	A→G	🟥 N	2
7	Het	chr3 19133919	A→G	🟥 N	2
8	Het	chr4 43977653	A→G	🟥 N	2
9	Het	chr8 36567823	T→C	🟦 N	2
10	Het	chr8 71359979	A→G	🟥 N	2
11	Het	chr6 83802489	G→T	🟦 N	2
12	Het	chr9 44742670	C→A	🟨 N	2
13	Het	chr10 112926193	T→C	🟥 N	2
14	Hom	chr11 73175960	A→G	🟦 N	2
15	Het	chr13 75771943	G→T	⬛	2
16	Het	chr13 90105223	T→C	🟥 N	2
17	Het	chr16 49868008	C→T	🟦 N	2
18	Hom	chr16 58466497	G→A	🟦 N	2
19	Het	chr17 12683939	A→G	🟦 N	2
20	Het	chr18 43321798	T→C	🟦 N	2

Table 17. Validation of variants identified by all five variant callers. Blue indicates that the sequence was of good quality at the position of the variant. Yellow indicates mediocre quality at the position of the variant. Red indicates bad quality at the position of the variant. Dark grey indicates that there was no sequence available at the location of the variant.

Number	Hom/Het	Variant Location	Variant	Validated by Sanger	Cells supported by
1	Het	chr2 120515974	T→C	N	2
2	Het	chr3 19133919	A→G	N	2
3	Hom	chr3 95734876	T→C	■	38
4	Het	chr4 130165817	T→C	N	2
5	Hom	chr4 132833055	C→G	■	9
6	Hom	chr5 104435120	C→G	■	2
7	Hom	chr7 27205154	TA→T	■	2
8	Hom	chr7 27205568	A→G	■	4
9	Het	chr8 85261271	A→C	■	2
10	Het	chr8 85261288	G→A	■	2
11	Hom	chr10 40251185	G→A	N	2
12	Hom	chr11 72777865	C→A	■	2
13	Hom	chr12 54783425	T→C	■	2
14	Het	chr13 31630905	A→G	N	2
15	Het	chr14 54542219	T→C	■	2
16	Het	chr16 52270742	C→A	N	2
17	Hom	chr16 94468834	C→T	Y	34
18	Het	chr19 60771023	C→A	■	2
19	Het	chr19 60771042	G→T	■	2
20	Hom	chrX 101404519	C→A	N	2

Table 18. First sequencing pass: Validation of variants only identified by Red Panda. Blue indicates that the sequence was of good quality at the position of the variant. Yellow indicates mediocre quality at the position of the variant. Red indicates bad quality at the position of the variant. Dark grey indicates that there was no sequence available at the location of the variant.

Results for first sequencing pass

As seen in **Table 17** and **Table 18**, good quality sequence was only produced in 12 out of 38 samples, but enough valid sequence was generated by 26 (all those except dark grey color) to validate the presence of their corresponding variant. Out of 26, only one variant, which was exclusively identified by Red Panda, was validated by Sanger sequencing. The most likely reason for this is that the variants being validated are part of a small fraction of the total number of cells (frequently the variant was supported by only 2/55 cells). The only variant confirmed by Sanger was supported by 34 cells.

Results for second sequencing pass

Sequence was obtained for 15 of the 20 variants being validated. **Table 19** shows that three out of the 15 variants identified by Red Panda were confirmed to exist including the one confirmed in the first pass. In all three instances, the variants were found in nine or more cells and were homozygous. It is possible that these confirmed variants were identified because they were homozygous, but a likelier explanation is variants identified in more cells are confirmed to be representative of the entire cellular population and thus are able to be seen in the Sanger sequencing. It is unclear then whether the variants identified by each tool in a low number of cells are in fact False Positives because they may be private mutations to a very small subset of cells.

Number	Hom/Het	Variant Location	Variant	Validated by Sanger	Cells supported by
1	Het	chr2 120515974	T→C	N	2
2	Het	chr3 19133919	A→G	N	2
3	Hom	chr3 95734876	T→C	Y	38
4	Het	chr4 130165817	T→C	N	2
5	Hom	chr4 132833055	C→G	Y	9
6	Hom	chr5 104435120	C→G	N	2
7	Hom	chr7 27205154	TA→T	■	2
8	Hom	chr7 27205568	A→G	N	4
9	Het	chr8 85261271	A→C	N	2
10	Het	chr8 85261288	G→A	■	2
11	Hom	chr10 40251185	G→A	N	2
12	Hom	chr11 72777865	C→A	■	2
13	Hom	chr12 54783425	T→C	■	2
14	Het	chr13 31630905	A→G	N	2
15	Het	chr14 54542219	T→C	■	2
16	Het	chr16 52270742	C→A	N	2
17	Hom	chr16 94468834	C→T	Y	34
18	Het	chr19 60771023	C→A	N	2
19	Het	chr19 60771042	G→T	N	2
20	Hom	chrX 101404519	C→A	N	2

Table 19. Second sequencing pass: Validation of variants only identified by Red Panda. Blue indicates that the sequence was of good quality at the position of the variant. Yellow indicates mediocre quality at the position of the variant. Red indicates bad quality at the position of the variant. Dark grey indicates that there was no sequence available at the location of the variant.

CHAPTER 3: DISTRIBUTION OF RED PANDA

Introduction

Adoption, distribution, and ease-of-use is necessary for any bioinformatic application. Given that, the popular source code repository GitHub[97] is used to distribute Red Panda. This is done under the MIT License[98], one of the most permissive Free Use licenses available, to ensure easy adoption by any end-user. It states that: “Permission is hereby granted, free of charge, to any person obtaining a copy of this software and associated documentation files (the "Software"), to deal in the Software without restriction, including without limitation the rights to use, copy, modify, merge, publish, distribute, sublicense, and/or sell copies of the Software, and to permit persons to whom the Software is furnished to do so”. Additionally, Read the Docs[99] is used to create documentation on how to run Red Panda.

Distribution and function

Red Panda, written almost entirely in the Perl programming language, relies on a number of different tools to function. The ‘Statistics::Basic’ Perl package is needed to perform basic statistics within the main Red Panda Perl script. The tool mpileup found in the samtools package[41] is required to generate a list of every variant in a sample. GATK HaplotypeCaller[25] is necessary to check variants that do not fit the expected bimodal distribution; however it is possible to use any standard variant caller for this step. Based on the above results, we recommend GATK HaplotypeCaller or FreeBayes.

Bedtools[100], vcf-sort found in the vcftools package[101], and Picard Tools[102] are all necessary to manipulate the different types of files used during the variant calling process.

As these tools are all supported by different institutions under different licenses, Red Panda does not come prepackaged with them. Instead a script is provided that assists the user in acquiring each tool, with the only exception being the Statistics::Basic package. As this is a Perl package the user will need to install this independently via CPAN[103].

DISCUSSION

Variant calling on human articular chondrocytes

Red Panda performs better than software designed for bulk NGS data and proves that scRNA-seq offers unique information on a small number of variants which Red Panda takes advantage of.

Using the exome data as a reference for comparisons, Red Panda's superior performance is well evidenced. There is consistent overlap in the results between the tools. Furthermore, while Red Panda shares a significant number of variants with those identified by other tools, it also identifies the most unique variants. Because of this, Red Panda provides both the highest PPV (45%) of any of the tools as well as the highest number of variants in concordance with the exome (913 on average). In comparison, on average FreeBayes identifies 65 variants, GATK HaplotypeCaller 705, GATK UnifiedGenotyper 222, and Platypus 386.

Unexpectedly, this superiority is not entirely from the heterozygous variants found in each sample. Instead, it appears that Red Panda gains an advantage against other tools by intentionally separating homozygous-looking variants from those variants that are heterozygous and then processing them differently. From the heterozygous data, on average Red Panda identifies 154 variants in agreement with the exome as compared to 31 for FreeBayes, 136 for GATK HaplotypeCaller, 118 for GATK UnifiedGenotyper, and 36 for Platypus.

PPV was used as the main metric to determine how well each tool performed because calculating sensitivity provides results that are difficult to interpret due to the extremely high number of True Negatives identified in variant analysis. Red Panda's average PPV was the highest at 45% followed by GATK-HC with 32%. The rest of the tools all had a PPV of <10%. Despite having the highest PPV of any of the tools, 45% is still far lower than what we would expect in a variant calling experiment using bulk sequencing[32]. This is likely due to much higher quality sequencing that is acquired in traditional NGS.

Variant calling on mouse embryonic fibroblasts

As with the articular chondrocyte results, Red Panda performs better than the four bulk variant callers assessed with the MEF results. After VCF files were generated for each cell by all five tools, the files were compared by looking at the total number of variants identified by each tool as well as the percentage shared in every pairwise comparison of each cell.

Red Panda identifies, on average, the highest number of variants per cell, surpassing all of the other tools: 1,071 on average by Red Panda, 865 by FreeBayes, 611 by GATK HaplotypeCaller, 574 by GATK UnifiedGenotyper, and 315 by Platypus. Assuming we can extrapolate from the PPV results from the articular chondrocyte data, this also means that Red Panda is identifying the highest number of True Positives as well. Surprisingly, FreeBayes identifies the second highest number of variants of the four tools. This is unexpected due to the results from the human articular chondrocyte data where FreeBayes had the fewest number of variants shared between the scRNA-seq results and the exome. One explanation from this is that while FreeBayes identifies a high number of variants, the majority of those are False Positives. This idea is supported by the PPV numbers seen in **Figure 20** for FreeBayes.

Since a paired exome to which we could compare our scRNA-seq results was not generated for MEFs, every possible pairwise comparison between each cell for every tool was performed instead to attempt to assess the quality of the variant calls. The MEFs sequenced are presumed isogenic, so the variants identified in each cell should theoretically exist in every other cell. Given this, these pairwise comparisons helped assess whether each variant caller performed well based on the consistency of their calls or if they performed poorly, randomly identifying variants in each cell. Comparisons were split up into three groups: total variants, exclusively homozygous-looking variants, and exclusively heterozygous variants.

These comparisons showed that each tool, especially Red Panda and FreeBayes, performs reasonably well when only the homozygous-looking variants are being compared, but they all suffer when assessing purely heterozygous variants. However it is important to note that, while GATK HaplotypeCaller and GATK UnifiedGenotyper both have a good fraction of overlap among homozygous-looking variants, it does not appear to affect the overall fraction of variants identified by these tools. This is because, for these two tools, homozygous-looking variants do not contribute much to the total list of variants shared. Also important is that Red Panda performs extremely well for homozygous-looking variants, but is average for heterozygous variants when compared to the other tools. This is due to the fact that, while Red Panda, in principle, confers an algorithmic advantage to identifying heterozygous variants, the monoallelic nature of gene expression and uneven sequencing coverage depth may preclude the tool from realizing its full potential. Due to this caveat, the majority of the heterozygous variants identified are actually picked up by GATK HaplotypeCaller since most of these are unsupported by a bimodal distribution as seen in **Figure 19**.

When looking at both the raw number of variants overlapping and the fraction of variants overlapping in these pairwise comparisons, Red Panda performs the best at both. This is important because, as is the case with FreeBayes, it is possible to have a large number of variants shared, but also have a small fraction of the total variants possible be shared. This indicates that there are potentially a lot of potential False Positives in the data generated by FreeBayes which fits with what was seen in the articular chondrocyte data.

It is clear that the tools with the highest fraction of variants shared pairwise between two cells comes from the homozygous-looking class. This makes sense as it is less likely that allelic dropout will occur in this class as a result of allele-specific expression making for a more stable population of variants in the scRNA-seq data. Tying into this is the fact that it is rare for any tool to have more than 100 heterozygous variants shared between two cells. This is likely due to the stochastic nature of allele-specific expression.

Lastly, due to the results from the Sanger sequencing, it is difficult to say with certainty whether these tools were ineffective at identifying variation in scRNA-seq data or if there was simply not enough DNA from the cells containing the variant to be picked up in the Sanger sequencing. Out of forty variants tested, only three were confirmed to exist, all of which were exclusively identified by Red Panda. However, this is likely due to these variants having been found in a larger proportion of cells than those variants being tested from the group that were identified by all five variant callers. Ideally Sanger validation would have been performed for locations supported by all five tools and found in more than 10 cells, but no such variants existed to be tested.

It is clearly possible to detect significant variation at the single-cell level, but due to the challenges in proving its existence with corroborative orthogonal sequencing it is difficult to know with certainty what software performs the best. Instead, we must rely on paired genomic sequencing as done with the human articular chondrocytes and simulated data to assess quality.

Evaluation using simulated data

Ideally, orthogonal sequencing would be able to validate variants that only appear in a small population of cells, but in the absence of such data, simulations can provide valuable insight. For our purposes, this involves inserting random variation distributed throughout the transcriptome and then using that as a master list of True Positives against which each tool can be measured. After adding ~1,000 simulated variants to the alignments from the MEFs, we were able to evaluate how well FreeBayes, GATK-HC, GATK-UG, Platypus, and Red Panda performed. Based on these results, Red Panda proves its advantage in identifying bimodally-distributed variants as well as homozygous variants, a class of variant that saw other tools struggle in comparison. When assessing total heterozygous variants, FreeBayes is superior to the other tools. This is counterintuitive to what was seen in the results from the human articular chondrocyte experiment where FreeBayes identified very few variants in concordance with the list obtained from exome sequencing.

Both GATK-HC and GATK-UG perform similarly in the simulation with the latter consistently performing slightly better than the former. However, it is because of this similarity in results that might offer an explanation for why FreeBayes seemingly performs so poorly in the chondrocyte data. When variants were called in the exome to generate a master list against which the variants from the scRNA-seq data could be compared, variants were only retained if they were identified by at least two of the following three tools: FreeBayes, GATK-HC, and Platypus. However, if the variants in this master list were consistently only supported by the latter two, then it follows that variants identified by FreeBayes in the scRNA-seq experiment would be filtered out and make it appear as though FreeBayes identified a low number of True Positives as seen in **Figures 17-20**.

The simulated data seems to indicate that FreeBayes has good sensitivity, but identifies a large set of variants different from both GATK-HC and Platypus. This is corroborated by data in **Figure 33**. Given this, in order to improve the accuracy of Red Panda, it might be wise to switch to only using FreeBayes or using a combination of both FreeBayes and GATK-HC to evaluate heterozygous variants that do not follow a bimodal distribution. For example, when searching for variants that are assumed to be private to a small number of cells—e.g., tumor cells—Red Panda could switch to using FreeBayes under the hood and then switch to using GATK-HC in situations where variants are expected to be shared across the majority of cells.

When to use Red Panda

Red Panda consistently performs better than other variant callers based on a number of metrics. Whenever it is necessary to analyze variation specific to a single cell—for example, looking at clonal tumor cell populations—, Red Panda will likely provide the best results. However, there are instances where using a different variant caller makes sense. Specifically, one should use GATK HaplotypeCaller for variant calling when it is preferred to pool together the reads from all of the cells. As seen in **Figure 29e-f**, this allows for greater sensitivity in identifying heterozygous variants. After identifying this specific class of variant, results can be pooled with those generated by Red Panda.

Lastly, it is important to note that the advantages conferred by Red Panda are currently limited to scRNA-seq generated by library-preparation methods that generate cDNA from full length transcripts such as Smart-seq2 and Holo-seq[104], although the latter has not been tested.

Final remarks

Based on the human articular chondrocyte and MEF data, Red Panda can provide a distinct advantage over other available software. However, this improvement is not entirely from the heterozygous variants as expected. Instead, Red Panda gains its major advantage in predicting homozygous-looking variants over other tools by intentionally separating heterozygous variants from variants that are homozygous-looking and then processing them differently. Due to the unique nature of scRNA-seq data, one must treat heterozygous variants with special consideration. Red Panda does provide a custom approach to this class of variant, but the number of variants that are specific to its method of dealing with bimodally-distributed heterozygous variants is, as seen in **Figure 19**, limited.

From these results it is clear that due to the inherent nature of RNA expression patterns in single cells, it is difficult to assess what variants exist in the genome with the same accuracy that we can with standard exome or WGS. Despite this, Red Panda provides a novel method of identifying variants in scRNA-seq and performs this function better than variant callers designed for bulk NGS datasets in certain categories. Future work includes creating a select dataset of genes that show consistent biallelic expression and testing the performance of Red Panda on this dataset.

BIBLIOGRAPHY

1. Navin, N. *et al.* Tumour evolution inferred by single-cell sequencing. *Nature* **472,** 90–94 (2011).
2. Francis, J. M. *et al.* EGFR variant heterogeneity in glioblastoma resolved through single-nucleus sequencing. *Cancer Discov.* **4,** 956–971 (2014).
3. Suzuki, A. *et al.* Single-cell analysis of lung adenocarcinoma cell lines reveals diverse expression patterns of individual cells invoked by a molecular target drug treatment. *Genome Biol.* **16,** 66 (2015).
4. Wang, Y. *et al.* Clonal evolution in breast cancer revealed by single nucleus genome sequencing. *Nature* **512,** 155–160 (2014).
5. 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).
6. McConnell, M. J. *et al.* Mosaic Copy Number Variation in Human Neurons. *Science* **342,** 632–637 (2013).
7. Ramsköld, D. *et al.* Full-length mRNA-Seq from single-cell levels of RNA and individual circulating tumor cells. *Nat. Biotechnol.* **30,** 777–782 (2012).
8. Ting, D. T. *et al.* Single-cell RNA sequencing identifies extracellular matrix gene expression by pancreatic circulating tumor cells. *Cell Rep.* **8,** 1905–1918 (2014).
9. Ni, X. *et al.* Reproducible copy number variation patterns among single circulating tumor cells of lung cancer patients. *Proc. Natl. Acad. Sci. U. S. A.* **110,** 21083–21088

(2013).

10. Lohr, J. G. *et al.* Whole-exome sequencing of circulating tumor cells provides a window into metastatic prostate cancer. *Nat. Biotechnol.* **32,** 479–484 (2014).

11. Gawad, C., Koh, W. & Quake, S. R. Dissecting the clonal origins of childhood acute lymphoblastic leukemia by single-cell genomics. *Proc. Natl. Acad. Sci. U. S. A.* **111,** 17947–17952 (2014).

12. Jan, M. *et al.* Clonal evolution of preleukemic hematopoietic stem cells precedes human acute myeloid leukemia. *Sci. Transl. Med.* **4,** 149ra118 (2012).

13. Yu, C. *et al.* Discovery of biclonal origin and a novel oncogene SLC12A5 in colon cancer by single-cell sequencing. *Cell Res.* **24,** 701–712 (2014).

14. Lodato, M. A. *et al.* Somatic mutation in single human neurons tracks developmental and transcriptional history. *Science* **350,** 94–98 (2015).

15. Grün, D. *et al.* Single-cell messenger RNA sequencing reveals rare intestinal cell types. *Nature* **525,** 251–255 (2015).

16. Jaitin, D. A. *et al.* Massively parallel single-cell RNA-seq for marker-free decomposition of tissues into cell types. *Science* **343,** 776–779 (2014).

17. Zeisel, A. *et al.* Brain structure. Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. *Science* **347,** 1138–1142 (2015).

18. Fluidigm | Products | C1. Available at: https://www.fluidigm.com/products/c1-system. (Accessed: 1st May 2018)

19. Picelli, S. *et al.* Full-length RNA-seq from single cells using Smart-seq2. *Nat. Protoc.* **9,** 171–181 (2014).

20. Tennessen, J. A. *et al.* Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science* **337,** 64–69 (2012).

21. Gill, R. *et al.* Whole-exome sequencing identifies novel LEPR mutations in individuals with severe early onset obesity. *Obesity* **22,** 576–584 (2014).

22. Ku, C.-S., Tan, E. K. & Cooper, D. N. From the periphery to centre stage: de novo single nucleotide variants play a key role in human genetic disease. *J. Med. Genet.* **50,** 203–211 (2013).

23. Lek, M. *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536,** 285–291 (2016).

24. Garrison, E. & Marth, G. Haplotype-based variant detection from short-read sequencing. (2012).

25. Poplin, R. *et al.* Scaling accurate genetic variant discovery to tens of thousands of samples. (2017). doi:10.1101/201178

26. DePristo, M. A. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* **43,** 491–498 (2011).

27. Rimmer, A. *et al.* Integrating mapping-, assembly- and haplotype-based approaches for calling variants in clinical sequencing applications. *Nat. Genet.* **46,** 912–918 (2014).

28. Macaulay, I. C. *et al.* G&T-seq: parallel sequencing of single-cell genomes and transcriptomes. *Nat. Methods* **12,** 519–522 (2015).

29. Angermueller, C. *et al.* Parallel single-cell sequencing links transcriptional and epigenetic heterogeneity. *Nat. Methods* **13,** 229–232 (2016).

30. Macaulay, I. C., Ponting, C. P. & Voet, T. Single-Cell Multiomics: Multiple Measurements from Single Cells. *Trends Genet.* **33,** 155–168 (2017).

31. Borel, C. *et al.* Biased allelic expression in human primary fibroblast single cells. *Am. J. Hum. Genet.* **96,** 70–80 (2015).

32. Cornish, A. & Guda, C. A Comparison of Variant Calling Pipelines Using Genome in a Bottle as a Reference. *Biomed Res. Int.* **2015,** 1–11 (2015).

33. Data Sciences Platform @ Broad Institute. GATK | Doc #3891 | Calling variants in RNAseq. Available at: https://software.broadinstitute.org/gatk/documentation/article.php?id=3891. (Accessed: 1st May 2018)

34. Wang, Y. & Navin, N. E. Advances and applications of single-cell sequencing technologies. *Mol. Cell* **58,** 598–609 (2015).

35. de la Chapelle, A. Genetic predisposition to human disease: allele-specific expression and low-penetrance regulatory loci. *Oncogene* **28,** 3345–3348 (2009).

36. Raval, A. *et al.* Downregulation of death-associated protein kinase 1 (DAPK1) in chronic lymphocytic leukemia. *Cell* **129,** 879–890 (2007).

37. Calling variants in RNAseq. *GATK-Forum* Available at: https://gatkforums.broadinstitute.org/gatk/discussion/3891/calling-variants-in-rnaseq. (Accessed: 1st July 2018)

38. Internals — bcbio-nextgen 1.0.9 documentation. Available at: https://bcbio-nextgen.readthedocs.org/en/latest/contents/internals.html. (Accessed: 1st July 2018)

39. Chiara, M. *et al.* CoVaCS: a consensus variant calling system. *BMC Genomics* **19,** 120

(2018).

40. Ewels, P., Magnusson, M., Lundin, S. & Käller, M. MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics* **32,** 3047–3048 (2016).

41. Li, H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* **27,** 2987–2993 (2011).

42. Babraham Bioinformatics - FastQC A Quality Control tool for High Throughput Sequence Data. Available at: https://www.bioinformatics.babraham.ac.uk/projects/fastqc/. (Accessed: 1st July 2018)

43. Kim, K.-T. *et al.* Single-cell mRNA sequencing identifies subclonal heterogeneity in anti-cancer drug responses of lung adenocarcinoma cells. *Genome Biol.* **16,** 127 (2015).

44. Kim, K.-T. *et al.* Application of single-cell RNA sequencing in optimizing a combinatorial therapeutic strategy in metastatic renal cell carcinoma. *Genome Biol.* **17,** 80 (2016).

45. Chu, L.-F. *et al.* Single-cell RNA-seq reveals novel regulators of human embryonic stem cell differentiation to definitive endoderm. *Genome Biol.* **17,** 173 (2016).

46. Braune, E.-B. *et al.* Loss of CSL Unlocks a Hypoxic Response and Enhanced Tumor Growth Potential in Breast Cancer Cells. *Stem Cell Reports* **6,** 643–651 (2016).

47. Lawlor, N. *et al.* Single-cell transcriptomes identify human islet cell signatures and

reveal cell-type-specific expression changes in type 2 diabetes. *Genome Res.* **27,** 208–222 (2017).

48. Dueck, H. R. *et al.* Assessing characteristics of RNA amplification methods for single cell RNA sequencing. *BMC Genomics* **17,** 966 (2016).

49. Picelli, S. *et al.* Smart-seq2 for sensitive full-length transcriptome profiling in single cells. *Nat. Methods* **10,** 1096–1098 (2013).

50. Lin, Z., Willers, C., Xu, J. & Zheng, M.-H. The Chondrocyte: Biology and Clinical Application. *Tissue Eng.* **0,** 060802052515066 (2006).

51. Bras, J. *et al.* Exome sequencing in a consanguineous family clinically diagnosed with early-onset Alzheimer's disease identifies a homozygous CTSF mutation. *Neurobiol. Aging* **46,** 236.e1–236.e6 (2016).

52. Ellingford, J. M. *et al.* Molecular findings from 537 individuals with inherited retinal disease. *J. Med. Genet.* **53,** 761–767 (2016).

53. Okonechnikov, K., Conesa, A. & García-Alcalde, F. Qualimap 2: advanced multi-sample quality control for high-throughput sequencing data. *Bioinformatics* **32,** 292–294 (2016).

54. Anders, S., Pyl, P. T. & Huber, W. HTSeq--a Python framework to work with high-throughput sequencing data. *Bioinformatics* **31,** 166–169 (2015).

55. Li, B., Ruotti, V., Stewart, R. M., Thomson, J. A. & Dewey, C. N. RNA-Seq gene expression estimation with read mapping uncertainty. *Bioinformatics* **26,** 493–500 (2010).

56. Visualization of a Correlation Matrix [R package corrplot version 0.84].

57. Create Elegant Data Visualisations Using the Grammar of Graphics • ggplot2. Available at: https://ggplot2.tidyverse.org/. (Accessed: 2nd July 2018)

58. Ecker, J. R. *et al.* The BRAIN Initiative Cell Census Consortium: Lessons Learned toward Generating a Comprehensive Brain Cell Atlas. *Neuron* **96,** 542–557 (2017).

59. Genome in a bottle—a human DNA standard. *Nat. Biotechnol.* **33,** 675–675 (2015).

60. Krawitz, P. *et al.* Microindel detection in short-read sequence data. *Bioinformatics* **26,** 722–729 (2010).

61. Rödelsperger, C. & Moreno, E. Differential Variant Calling In Mutants From Diverse Genetic Backgrounds: A Case Study In The Nematode Pristionchus pacificus. (2017). doi:10.1101/138479

62. Huang, M. N. *et al.* MSIseq: Software for Assessing Microsatellite Instability from Catalogs of Somatic Mutations. *Sci. Rep.* **5,** 13321 (2015).

63. Wyman, D. & Mortazavi, A. TranscriptClean: Variant-aware correction of indels, mismatches, and splice junctions in long-read transcripts. *Bioinformatics* (2018). doi:10.1093/bioinformatics/bty483

64. Cutcutache, I. *et al.* Exome-wide Sequencing Shows Low Mutation Rates and Identifies Novel Mutated Genes in Seminomas. *Eur. Urol.* **68,** 77–83 (2015).

65. Hwang, S., Kim, E., Lee, I. & Marcotte, E. M. Systematic comparison of variant calling pipelines using gold standard personal exome variants. *Sci. Rep.* **5,** 17875 (2015).

66. Kumaran, M., Subramanian, U. & Devarajan, B. Performance Assessment of Variant Calling Pipelines using Human Whole Exome Sequencing and Simulated data.

(2018). doi:10.1101/359109

67. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. (2013).

68. Data Sciences Platform @ Broad Institute. GATK | Tool Documentation Index. Available at: https://software.broadinstitute.org/gatk/documentation/tooldocs/3.8-0/org_broadinstitute_gatk_tools_walkers_haplotypecaller_HaplotypeCaller.php. (Accessed: 2nd August 2018)

69. Quinlan, A. R. BEDTools: The Swiss-Army Tool for Genome Feature Analysis. *Curr. Protoc. Bioinformatics* **47,** 11.12.1–34 (2014).

70. Conway, J. R., Lex, A. & Gehlenborg, N. UpSetR: An R Package For The Visualization Of Intersecting Sets And Their Properties. (2017). doi:10.1101/120600

71. Deng, Q., Ramsköld, D., Reinius, B. & Sandberg, R. Single-cell RNA-seq reveals dynamic, random monoallelic gene expression in mammalian cells. *Science* **343,** 193–196 (2014).

72. Reinius, B. & Sandberg, R. Random monoallelic expression of autosomal genes: stochastic transcription and allele-level regulation. *Nat. Rev. Genet.* **16,** 653–664 (2015).

73. MLEAF and MLEAC. *GATK-Forum* Available at: https://gatkforums.broadinstitute.org/gatk/discussion/1283/mleaf-and-mleac. (Accessed: 4th August 2018)

74. Sandmann, S. *et al.* Evaluating Variant Calling Tools for Non-Matched Next-Generation Sequencing Data. *Sci. Rep.* **7,** 43169 (2017).

75. Roychoudhury, S. *et al.* Human Apurinic/Apyrimidinic Endonuclease (APE1) Is Acetylated at DNA Damage Sites in Chromatin, and Acetylation Modulates Its DNA Repair Activity. *Mol. Cell. Biol.* **37,** (2017).

76. Bhakat, K. K., Izumi, T., Yang, S.-H., Hazra, T. K. & Mitra, S. Role of acetylated human AP-endonuclease (APE1/Ref-1) in regulation of the parathyroid hormone gene. *EMBO J.* **22,** 6299–6309 (2003).

77. Chattopadhyay, R. *et al.* Regulatory role of human AP-endonuclease (APE1/Ref-1) in YB-1-mediated activation of the multidrug resistance gene MDR1. *Mol. Cell. Biol.* **28,** 7066–7080 (2008).

78. Quadros, R. M. *et al.* Easi-CRISPR: a robust method for one-step generation of mice carrying conditional and insertion alleles using long ssDNA donors and CRISPR ribonucleoproteins. *Genome Biol.* **18,** 92 (2017).

79. Sanger, F., Nicklen, S. & Coulson, A. R. DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci. U. S. A.* **74,** 5463–5467 (1977).

80. Yan, H. Allelic Variation in Human Gene Expression. *Science* **297,** 1143–1143 (2002).

81. Gregg, C. *et al.* High-resolution analysis of parent-of-origin allelic expression in the mouse brain. *Science* **329,** 643–648 (2010).

82. Marinov, G. K. *et al.* From single-cell to cell-pool transcriptomes: stochasticity in gene expression and RNA splicing. *Genome Res.* **24,** 496–510 (2014).

83. Gimelbrant, A., Hutchinson, J. N., Thompson, B. R. & Chess, A. Widespread monoallelic expression on human autosomes. *Science* **318,** 1136–1140 (2007).

84. Xu, J. Preparation, culture, and immortalization of mouse embryonic fibroblasts.

Curr. Protoc. Mol. Biol. **Chapter 28,** Unit 28.1 (2005).

85. bcbio. bcbio/bcbio-nextgen. *GitHub* Available at: https://github.com/bcbio/bcbio-nextgen. (Accessed: 7th October 2018)

86. Dewey, F. E. *et al.* Distribution and clinical impact of functional variants in 50,726 whole-exome sequences from the DiscovEHR study. *Science* **354,** (2016).

87. 1000 Genomes Project Consortium *et al.* A map of human genome variation from population-scale sequencing. *Nature* **467,** 1061–1073 (2010).

88. Choi, M. *et al.* Genetic diagnosis by whole exome capture and massively parallel DNA sequencing. *Proceedings of the National Academy of Sciences* **106,** 19096–19101 (2009).

89. Barnett, D. W., Garrison, E. K., Quinlan, A. R., Stromberg, M. P. & Marth, G. T. BamTools: a C API and toolkit for analyzing and managing BAM files. *Bioinformatics* **27,** 1691–1692 (2011).

90. Li, W. V. & Li, J. J. A statistical simulator scDesign for rational scRNA-seq experimental design. (2018). doi:10.1101/437095

91. Zappia, L., Phipson, B. & Oshlack, A. Splatter: simulation of single-cell RNA sequencing data. *Genome Biol.* **18,** 174 (2017).

92. Risso, D., Perraudeau, F., Gribkova, S., Dudoit, S. & Vert, J.-P. A general and flexible method for signal extraction from single-cell RNA-seq data. *Nat. Commun.* **9,** 284 (2018).

93. Severson, D. T., Owen, R. P., White, M. J., Lu, X. & Schuster-Böckler, B. BEARscc determines robustness of single-cell clusters using simulated technical replicates.

Nat. Commun. **9,** 1187 (2018).

94. Vieth, B., Ziegenhain, C., Parekh, S., Enard, W. & Hellmann, I. powsimR: power analysis for bulk and single cell RNA-seq experiments. *Bioinformatics* **33,** 3486–3488 (2017).

95. Zhang, X., Xu, C. & Yosef, N. SymSim: simulating multi-faceted variability in single cell RNA sequencing. (2018). doi:10.1101/378646

96. Papadopoulos, N., Parra, R. G. & Soeding, J. PROSSTT: probabilistic simulation of single-cell RNA-seq data for complex differentiation processes. (2018). doi:10.1101/256941

97. Build software better, together. *GitHub* Available at: https://github.com. (Accessed: 8th October 2018)

98. The MIT License | Open Source Initiative. Available at: https://opensource.org/licenses/MIT. (Accessed: 8th October 2018)

99. Home | Read the Docs. Available at: https://readthedocs.org/. (Accessed: 8th October 2018)

100. bedtools: a powerful toolset for genome arithmetic — bedtools 2.27.0 documentation. Available at: https://bedtools.readthedocs.io/en/latest/. (Accessed: 8th October 2018)

101. Danecek, P. *et al.* The variant call format and VCFtools. *Bioinformatics* **27,** 2156–2158 (2011).

102. Picard Tools - By Broad Institute. Available at: https://broadinstitute.github.io/picard/. (Accessed: 8th October 2018)

103. The Comprehensive Perl Archive Network - www.cpan.org. Available at: https://www.cpan.org/. (Accessed: 8th October 2018)

104. Xiao, Z. *et al*. Holo-Seq: single-cell sequencing of holo-transcriptome. *Genome Biol.* **19,** 163 (2018).

APPENDIX A: SANGER SEQUENCING

Sequencing primers

#	Primer name	Forward primer	Reverse primer	Variant location	Expected length
1	chr1:43954201-43955201	TGGATTCACTATGGCAGCAA	GGTCACAATGGAGAGCAGGT	chr1:43954701	510*
2	chr1:181175375-181176975	TGCTGACTCCGATCTGTCAC	CATTCAAACTGGTGCTGTGG	chr1:181176175	484
3	chr2:3327701-3329301	CAGCAAGTGGAACAAAGTGG	TCTAGCAAGGTGGGTGAAGC	chr2:3328501	606
4	chr2:22939805-22941405	TGCTTCATGTGCAGAAAACC	GATTCAAGGGAGGGTGTGAG	chr2:22940605	690
5	chr2:33245975-33247575	TGCTCAGTTCTCAGGTGCTG	CCTATTGGCAGCGACTTCTC	chr2:33246775	585
6	chr2:39194566-39196166	TAGCAGATTCCCTCGCCTAC	TGGTGTGGTTTTTGAACAGC	chr2:39195366	495
7	chr3:19133119-19134719	TGAGGCTGGAGGAGAAAATG	AGAATGAGGATGTGGCTTGG	chr3:19133919	619
8	chr4:43976853-43978453	TTTCCCACAGGGCACTTTAC	CAGAATCCTCAAAGCCCAAG	chr4:43977653	676
9	chr6:83801689-83803289	AGCAGGCACAGAACTCCTTC	TATAACCAGAGCCGGGTGAG	chr6:83802489	691
10	chr8:36567023-36568623	AAGCAAGGATGGAAACGATG	ACTCACCCACCAACAGGAAG	chr8:36567823	520
11	chr8:71359179-71360779_7	AGGAGGCTGTTGTTCCAGTG	GCCCATGTCCAGGTTACAAG	chr8:71359979	549
12	chr9:44741870-44743470	CTATCCCAGTCCCCTTCCTC	CACCTCCCTCTCTGTCCTTG	chr9:44742670	461
13	chr10:112925393-112926993	TCCCTTTCATGTTTCCCAAG	ATCTCTCATGGCTCCCTCTG	chr10:112926193	607
14	chr11:73175160-73176760	TTACCCAATCCAGCAAAAGc	CTCATTCTCAAAGCGGGAAG	chr11:73175960	690
15	chr13:75771143-75772743	GTTGGTGTGTGTTTGCTTGG	ATGCTTCCCTTTTCAACTGG	chr13:75771943	515
16	chr13:90104423-90106023	CCCAAAGGTGGTATTTGTGG	TTCAAGCACGATGTCAAAGC	chr13:90105223	473
17	chr16:49867208-49868808	AAACTGTGGTCATCCCTTGC	AACACGAGTGCCAGACTCAC	chr16:49868008	455
18	chr16:58465697-58467297	GGTCTCAGCTCTGCTCATCC	ACTTGGGTCAGTTGGGATTG	chr16:58466497	407

19	chr17:12683139-12684739	AGGCACCACGGAGTTAGATG	CAGCAAGGAGGAGGAGACAG	chr17:12683939	428
20	chr18:43320998-43322598	AGCAATAACTTGGCGTTTGG	CTGTAATTCTGCGCCTCCTC	chr18:43321798	506

Table. Primers for variants identified by all five variant callers

#	Primer name	Forward primer	Reverse primer	Variant location	Expected length
1	chr2:120515174-120516774	GAAAGAGAAATGGCGCAGTC	TCAGAACAACAAGCGACAGG	chr2:120515974	602
2	chr3:19133119-19134719	TCTGTCTTGTGCGGAAAATG	CCACGGATGTGTTCAAAGTG	chr3:19133919	598
3	chr3:95734076-95735676	CTTCCTTTCCACAGCAGGAC	TGCCTATCCACAACCTCCTC	chr3:95734876	611
4	chr4:130165017-130166617	TAGGAGGGTGATGAGGTTGG	TGCGATCCAGATGTTGAGAC	chr4:130165817	658
5	chr4:132832255-132833855	GGCAATGTCTGAGGCTTCTC	TCAGGAAGATGAGGCAGGAG	chr4:132833055	523
6	chr5:104434320-104435920	GAAAGTTCTGCCGAGACAGC	TGAAAATACCGGGAAACCTG	chr5:104435120	688
7	chr7:27204354-27205954	TATGTGGTTGCTGGGAATTG	GGTGTGGAATATGGGCTGTC	chr7:27205154	485
8	chr7:27204354-27205954_2	GGACAGCCCATATTCCACAC	AGCACAGCGGTCAGGTAGAC	chr7:27205568	689
9	chr8:85260471-85262071	TTCATAGATTGGCCCCTCAG	TTTCCAGAAGACCTGGGTTC	chr8:85261271	559
10	chr10:40250385-40251985	AGCTCACTCTGGCCTTGAAC	CTTCATTTGGGCGATAGGAC	chr10:40251185	628
11	chr11:72777065-72778665	GGCAGGTGGATTTCTGTGAG	GCTGGTACTTGGAGCAGGAC	chr11:72777865	540
12	chr12:54782625-54784225	GTCTCGCTGGTCCTTGAGAG	TGGACTGCTGGGATTAAAGG	chr12:54783425	525
13	chr13:31630105-31631705	ACTGCAACGGACTCACACTG	GGCACCTGTATCCGAAGAAG	chr13:31630905	429
14	chr14:54541419-54543019	GTTCTGCCTCCACTCAGCTC	GCTGGCCCCTAAACTCTTTC	chr14:54542219	428
15	chr16:52269942-52271542	TTCCTCTCCTGGGAAAAGTG	TGCCCTGTGTCATCTACCAC	chr16:52270742	465
16	chr16:94468034-94469634	GCTCTCAGCCTCCTCAGTTC	CAGGGACACCACAGACAATG	chr16:94468834	680
17	chr19:60770223-60771823	CTCCCGAATGTCCTGAGTTC	CTGCAAAATACAGGGGAAGG	chr19:60771023	662
18	chrX:101403719-101405319	CTACATCTCCAGCCCCTGTC	TCCCCATCTTACCTTTGTGG	chrX:101404519	600

Table. Primers for variants only identified by Red Panda

Summary of the quality for the first round of sequencing

Primer Name	QS	CRL	Failure Reason
1-SR_chr1_43954201-43955201	50	458	Poor Quality
2-SR_chr1_181175375-181176975	54	427	
3-SR_chr2_3327701-3329301	56	555	
4-SR_chr2_22939805-22941405	38	385	Poor Quality
5-SR_chr2_33245975-33247575	57	532	
6-SR_chr2_39194566-39196166	53	446	Poor Quality
7-SR_chr3_19133119-19134719	32	244	Poor Quality
8-SR_chr4_43976853-43978453	37	529	Poor Quality
9-SR_chr6_83801689-83803289	54	643	
10-SR_chr8_36567023-36568623	40	439	Early Termination
11-SR_chr8_71359179-71360779_7	56	495	
12-SR_chr9_44741870-44743470	41	402	Non-specific
13-SR_chr10_112925393-112926993	54	527	Homopolymeric or Repetitive Region
14-SR_chr11_73175160-73176760	57	637	
15-SR_chr13_75771143-75772743	11	1	No Priming
16-SR_chr13_90104423-90106023	51	422	Early Termination
17-SR_chr16_49867208-49868808	52	395	
18-SR_chr16_58465697-58467297	56	356	
19-SR_chr17_12683139-12684739	53	377	
20-SR_chr18_43320998-43322598	57	453	
21-SR_chr2_120515174-120516774_1_F	44	528	Poor Quality
22-SR_chr3_19133119-19134719_1_F	43	528	Poor Quality
23-SR_chr3_95734076-95735676_7_F	23	215	Poor Quality
24-SR_chr4_130165017-130166617_2_F	49	482	Poor Quality
25-SR_chr4_132832255-132833855_7_F	18	14	Poor Quality
26-SR_chr5_104434320-104435920_F	34	220	Poor Quality
27-SR_chr7_27204354-27205954_7_F	13	11	No Priming
28-SR_chr7_27204354-27205954_F	25	99	Poor Quality
30-SR_chr8_85260471-85262071_F	20	17	Early Termination

31-SR_chr10_40250385-40251985_8_F	55	567	
32-SR_chr11_72777065-72778665_5_F	44	269	High Background
33-SR_chr12_54782625-54784225_1_F	24	235	Poor Quality
34-SR_chr13_31630105-31631705_F	43	303	Early Termination
35-SR_chr14_54541419-54543019_1_F	41	374	Poor Quality
36-SR_chr16_52269942-52271542_F	55	422	Poor Quality
37-SR_chr16_94468034-94469634_3_F	26	507	Homopolymeric or Repetitive Region
38-SR_chr19_60770223-60771823_F	33	366	Homopolymeric or Repetitive Region
40-SR_chrX_101403719-101405319_6_F	57	551	

Table. Sequencing results from first round of Sanger sequencing

Summary of the quality for the second round of sequencing

Primer Name	QS	CRL	Failure Reason
SR-21-F_chr2_120515174-120516774_1_F	45	532	Non-specific
SR-21-R_chr2_120515174-120516774_1_R	44	538	Non-specific
SR-22-F_chr3_19133119-19134719_1_F	56	545	
SR-22-R_chr3_19133119-19134719_1_R	54	545	Poor Quality
SR-23-F_chr3_95734076-95735676_7_F	18	70	Non-specific
SR-23-R_chr3_95734076-95735676_7_R	31	330	Non-specific
SR-24-F_chr4_130165017-130166617_2_F	14	11	Poor Quality
SR-24-R_chr4_130165017-130166617_2_R	50	600	
SR-25-F_chr4_132832255-132833855_7_F	38	338	Non-specific
SR-25-R_chr4_132832255-132833855_7_R	39	420	Poor Quality
SR-26-F_chr5_104434320-104435920_F	19	43	Non-specific
SR-26-R_chr5_104434320-104435920_R	31	220	Poor Quality
SR-27-F_chr7_27204354-27205954_7_F	12	24	No Priming
SR-27-R_chr7_27204354-27205954_7_R	13	1	Poor Quality
SR-28-F_chr7_27204354-27205954_F	51	631	
SR-28-R_chr7_27204354-27205954_R	50	627	
SR-30-F_chr8_85260471-85262071_F	22	77	Non-specific
SR-30-R_chr8_85260471-85262071_R	15	28	Poor Quality
SR-31-F_chr10_40250385-40251985_8_F	52	554	
SR-31-R_chr10_40250385-40251985_8_R	21	82	Non-specific
SR-32-F_chr11_72777065-72778665_5_F	28	197	Non-specific
SR-32-R_chr11_72777065-72778665_5_R	23	84	Non-specific
SR-33-F_chr12_54782625-54784225_1_F	11	1	No Priming
SR-33-R_chr12_54782625-54784225_1_R	17	72	Poor Quality
SR-34-F_chr13_31630105-31631705_F	38	259	Early Termination
SR-34-R_chr13_31630105-31631705_R	24	132	Non-specific
SR-35-F_chr14_54541419-54543019_1_F	15	32	Poor Quality
SR-35-R_chr14_54541419-54543019_1_R	13	1	Poor Quality

SR-36-F_chr16_52269942-52271542_F	55	404	
SR-36-R_chr16_52269942-52271542_R	53	412	Poor Quality
SR-37-F_chr16_94468034-94469634_3_F	21	221	Poor Quality
SR-37-R_chr16_94468034-94469634_3_R	39	434	Non-specific
SR-38-F_chr19_60770223-60771823_F	23	320	Poor Quality
SR-38-R_chr19_60770223-60771823_R	30	321	Non-specific
SR-40-F_chrX_101403719-101405319_6_F	56	545	
SR-40-R_chrX_101403719-101405319_6_R	56	544	

Table. Sequencing results from second round of Sanger sequencing

Sequence from the first round of Sanger sequencing

Sanger sequence for sample 1-SR_chr1_43954201-43955201:
NNNNNNNNNGNNTAGANCCTTCNTGTGGAACCATCCTAAAACTAGTCTACACTCCAAGCTAA
GCTTTAGTATATACTTTTCACTTGCCCTGATGGTTCTCTGTATCATCTCTGACCATCCATGTCCT
TGTGTGTGGGTTAGCATTGACTTCTGGGGAACTTCTGACCTAAGCCTAACTGCTTACTCTGTGA
AACAAAGGAATGTAGTCATTGGATTTGAGTGGTGGGAGAGGAAAGCAATGAGAAGGTGACTG
AGAAACTGTATTGTGTGATCTTTAAAAAGGAGTGGGAGGATAAATTTTAATGCCTATTTCTCC
TTCCCAACAGAGCCTGTGTTGATTCAAATGAGAATGGGGACTTGAGTAAATGTGCCGTATTGA
GAAACTACAAAGAGGCCCAAGAGTACAGTTCTTTTGGCACAGCTGAGATGCTGAATTACTCTG
TGAACATTTATGACGATGGGAACCTGCTCTCCATTGTGACCA

Sanger sequence for sample 2-SR_chr1_181175375-181176975:
NNNNNNNNNNNNNNNNNTGGGANANGACCACTGCCTTCCTGGCTTTTCCCTCCCTGCGACCCCCA
CAATGGGAAGACCCCTCAGTACAAGCCTCTCTCTTCTCTCAGGATCCCGAGGGGCATATGATC
ATGGTGAATGCTATGGACTTGGCTTGGCCTCAAGGGCAAGTTGTACCAGATTTTTGTTGTTGTT
GTTCATCAGGATACATTGGAGTTAATTCCACTTTTCCTTCCAAGAGCTGTGGTCACCCTGGTTA
TCTCCTATTGGAAAACATGGATTTCAAGGGAGACTGGTTAGACCCAGCAATTATGGAGTTGAA
AACACCATGAACATCAATCAGGCTTATGTTAGATATAGGGTCTTCAAAATGACAAGTCACTTT
TTTTCCATAAAGGAAACATTCCCGTAAACTAAAAGGGGAGGGAGAGGGAAAGACTGTCCACA
GCANNNNTTTGAATGA

Sanger sequence for sample 3-SR_chr2_3327701-3329301:
NNNNNNNNNNNNNNNNNTCNAANCTTGGGANGAAAAGTTTTGCTATCCCACTCACCTCCTAAAT
TCCACAATGACTTCCAGACAGAGTGGATTAGTAAAGTCCTGACAGCTCCACTCCCAGCCAGAA
AGACCCATCACTGTCTTGTCTGTCCTCGTGCCACATGTCAGCTAAGTACTGCTGTAATAAGTTC
TGGCTTAGGTTTTGTTTTTAGAGTTGTTAGCTTTGATTTTTGTTTGTTTTTGGTGCTGGGGATCA
AGCCAGGGATAGTATGTGCTGGTAAGCACATATTCTGCCATAGAGCTCTGCCCCTAATGTACT
CTTGATAGATGTTATATATTACAAGGAAACTGATGATGCGCAGGGAGAGAATTCTTATGAAAC
AACCTTATCAGGCTTTTGTTCTGTATCAATTAAGCCTTTCTCCCAAGCCTGCCTTGATGCTTAC
CTACAAAGAAAGCCAAATTACCACAAAGTAAAAATGACATGCCGCTCTGAAGGCAGTGTACT
GCTTAACATTTAGTGTCTCTCTCTCTCTCCGTTGCAACTGAGGACTTTCTTCAGTTGCTTCACCC
ACCTTTGCTAGAA

Sanger sequence for sample 4-SR_chr2_22939805-22941405:
NNNNNNNNNNNNNNNATTTCANCTTATTGATGGGCAATTTTTATTGGCAAAGTTTTTCGGAAAAC
TTTTTAAATGTAATTAAACCAGTGTCATTATAGTCCTATAAATTCTAATCGAGGTATCCTGATG
GTTATATGTGGTATTGTTTACACTGTTAATGCCCACATGTAAAGCCATTACACAAATAAATAA
TCAACGTTAAAATTCAAGTGGTTTGTCTTTGTTTCACCATAGGATTAAAGGTCAGAGAATTTTG
AAGTCTGTACTATTTAAATCCACATTAGTTATACTTTAACAATATCCAAATTTTTCATATAGGA
GCATAGTTTATTATAAAAGACTAGATAAAATTTAGACAACAGGTTATTTACAAATGAAGAAA
ACATTTTGCTTCAAAAAGGAAATGCATAATAAAGAGCTATCNGATTGCTAATGNCATAGTACT
TCGAAAGTAGGANAAGATAANGGNTATCTGGAGTCNGTGTTNTTGGNGAACNNNGGTCTTNN
GTTTTTGNGAT

Sanger sequence for sample 5-SR_chr2_33245975-33247575:

NNNNNNNNCTNGCGCTNNNANATGTATGTCTCATTGTCTAAGGACAGAACAGGCATTGTAGC
CCTGATTTCAAAAACAGTCCTTGGCACTGCCTGGATTTTCCCAGAATGTCCTCAAGCTCATCTC
TCACATAGGGGCTCCTGGCCTTCTCTCCTTGAGCCCCACCTCCCCTCAGACTGTTGCACTTCCC
CTCTCAGGACGGCTCAGCATCCCTCCGTACAGTTACCCCTCAGCCTGCACCTCCTGTGCCTTAG
TCTCTGGCTGCTCACTGGAAGTCAAGTCCTCTTCTCCCTGCTCCCCTGGCCTCTCCTTTTCTGCC
ACACAGAGCTTTATTTCTGGCACAATTCGTTGGCCTCTGGGCAGGAAACAGTCTGGGCTCAGG
TCCTGGCTGAGAAGGGAAGGCCAGGCCAGAAGCCACAGAGGCAGCGGCATAGACCTGTATTC
AGTTCTGCACCTTCCATTCATACTTTAGCCTCCACAGAATTTTAACCTCTACACAAACAGTACC
CTGCTTTGCCAGAGACACCCCACTGGAGAGAAGTCGCTGCCCAATAGGANN

Sanger sequence for sample 6-SR_chr2_39194566-39196166:
NNNNNNNNNCNGNNNNNTCTCCCACAAGATTACTCTGTATAATACATAGCCCTTGTTTAGTAG
AGGGATCCAAATATTCTTTTTCAGGCTTACAAAGTCCAATACATTCATTCACTCTCTCTTTCCT
TCACAAGTCTAATAGCAAAAACTACTTTTTCCATGCCCCAAAGCCATTATCAGTGGAAGAATA
GTCAGGCAAAACAGAGATGGCAGTTAAGGAATGGACAGAATATTATTGGCACATGCCCAGCT
AGTGACAAACAAATGCAGTACACCATGACTTGAAAATAAGTCACATTACAAGGAGAATGAAA
ACAACTACATCAACTAAGCTAAGGAGTGTGAAGTGGAAAGGGGATTGAGAGTTACTGGTTTA
ACTGGTACAACTTAAAAGCAGGAGGGCAAGCACTTAAATACAATTCATGGTAACATGATCAA
GAAAATACTGCAGAGCTGTTCAAAAACCACACCAAGTNANNN

Sanger sequence for sample 7-SR_chr3_19133119-19134719:
NNNNNNNNNNNNNNNNTGTNNACAGAATAAAGCTACACATACATTCGGTTTTAACAGTGGACC
CATGTTTTTCANAACAAAATGTGAGGCAATTAAGCAAAGGAAAACAATACCAAAATACCTTC
ATAAACAAACTTGACCTGTAGCCTCTGTCTTGTGCGGAAAATGTCCAATAAGATGTAAAATGC
AGCATACATAGTCATGTAATCTAGAAACTTTCCATGATAATTAATTGCAAATTGCACTTGACA
GTTCACCACAACAATGGAAAGCTGGNCCCCTGTTGGTTTGCACAGTCCTTGAACCCCTAACTG
AANTTAAAGNCNCCAGNNAAATGATCTGGAAGTGTGTATGN

Sanger sequence for sample 8-SR_chr4_43976853-43978453:
NNNNNNNNNNNANNNNNNGANNNAGGCGTGGGGAAGGCATCTGCGAGTGACGGGTCCTCCT
TTGTGGTGGCTAGATACTTCCCAGCAGGGAATATTGTCAACCAGGGCTTCTTCGAAGAGAATG
TCCCACCTCCAAAGAAGTAGCTCTCGCCAGCAAGTGTAACGAGATGGAGATGCATATATGAA
GTGCCTATAGAACCACAAAAGCACCAGCGAGTGTCTGTCCCCGTGGGTGTGTGTATTTATGTG
TTTACCTGTGAGCATCTCTCGTGACAGCTACACCTGAGTACCTTCTGATTGAAGGATGGCCATC
TGAAACATTTGTCCAGAGGGCTAGCTAACCACAATTGGGTAAGATTAGTTCTGAAATCTTTTT
TCTGTTTTTGTTTGGTTGGGGTATTGTTTGTTTGTTTTGCTTTTATTTTTGTTTTGTTTTGTTTTTT
CAAGGCAGGGATTCTCGGTGTAGCCTTGGCTGTCCTAAAATTTGCTCTGCANACCAGGCTGGC
AGAGATGCACCTGCCTGTGCCTCCCAAGTGTGGGANTAAAGGCATGATCCACCACTGCCTGGN
TTTTGNTTTTGCTTTTTAAAACAAACTTNTTTTGAAACNGTCGTCCNTCTANTATCCNTCCNTG
G

Sanger sequence for sample 9-SR_chr6_83801689-83803289:
NNNNNNNNNNNNANNGCGGCATTCTTCTGCATTGGGGGGTGCCAGCCTGGGGGCCAGGCACA
TTGGATACCACCTTCCCATGGACTACAGTATCAATGCCATTGCCTTCTATTCCTATACCTTCTA
GAGTCTGTCCCCCTGCCCAACCAGCCAACACCGAGAGCTGGGAGACTTTCCTTTTTAAAAAAC
ACATATGGAAGAAAATAAATGCACTTTACTCCTTCCCTAGCAGGGTGTCATCTTCCATACATT

GCTGTGCATGTCCTTGGCACCTCCAGCAGGTTCACCTAGCAAGCTCTCAGTTTAAAGACCCAC
CAGGCTCTTAGGCATATAAGAGCCCAGACCTTCCCATTTCCACACTATATGCTTTGGCTGCCA
GAAGCTGGGACCCAAGTTTCCTCTTAGCAGTTATGATTCCTCATGACAGAACTAAGTGTGTCT
AAGCCTACCTCCTGGGGAATGACCCAAGGTCTATTGTCTCCCAAGCTCTGGGAGACAGTACAA
TCCTAGTCTCAATTTTGACTTAATTTTATTTGGTAGCTGTAGGTACACAGGGCAGTTTTTCTTTC
TCATGTGTGCATGAATCCCCTGAGGGCCAAGTAGGCATTGCTTAACTCAGTGGCCTGCCTGCT
TTCTGACTGCTGCTCACNCGGCTCTGGTTTNANAANN

Sanger sequence for sample 10-SR_chr8_36567023-36568623:
CGNNNNNNNNNTAGGGTCATGCTAAGCTACCTTTTTGTCTCTTAAAAAATAAGAGAAGCAAA
TTCAATCAGTGTTCTTTAGCAAATGATCAATTTATTAGGTGTCTACAAGTACAAAACACTGAC
AGCTGCTGTAGAGAAGTAGAAAGATGTGCAGGAGATGAGCCTGCCCTTGAAGAACTCACAAT
CTAGACCCAAGAATCATCACGGGGACAAATCGTGGCCACAGTACAAGATGTTAACACCAGTT
CTTGGAAAGTTCCGTGCTCTGAATTCTCTCCTCTGGCCAGACGTAGGTAATCTCTTCAACAGGA
AAGTCAAAAGTGGGAAAGAGACCTCTTGCGGACTCATCAGTGACCTTTCCTTTGCTTCTAAGC
ATCGTTCCGTTGCTGTTGCTCTGTTCTNGGAACCTCTGGGCATCCTTTCTGACCTTGGTCTGGG
GAGACTCTCCGACTTCCTGAACCTTCACCANGNCTTCCTGNTGGTGNNNNNANTANTCAANNC
CCTTGAANTG

Sanger sequence for sample 11-SR_chr8_71359179-71360779_7:
NNNNNNNNNNNNCGCAGANAGGACCAGGTACATTCCGTATACATCGCCCCTGGGGCTGACCT
GCCATCACAGAGTACACTGATAGCCCTGGACCATGATACCATACTTCCTGGGACCAAGCGCAG
GTATTCGGACCCCCCTACCTACTGCCTGCCCCCCAGCTCCGGCCAGGCCAATGGCTGAGGACC
ATGACTGGCAGTCTGCATCTCCTAACATCCCCGAACTGGCATCCCAGCTGTGGAGCTGGCCTT
CACTTTCTGAGAAGGATCTAGAATGAAAAGCTCCCAAAGGGATGCAGTGGCCAGCTCTGTGT
GTTGTGGAGACTGGGAGCTGCTGGCCAGGAGCCATCAGAGCCCCAACCTGCACAGCAGTGGC
TCCTTTGTCCTTTCAGTAACTGTTTCTCTTTTTGTGGTTTACATAACTTTTAAGTTCATAACAGC
CTTAATGGAGGACCAAACTTTTGTATTTGTATGTCTGAACTTTTATATTAACTCTGCACCCTTG
TAANTGNNNNNGGGGCAAN

Sanger sequence for sample 12-SR_chr9_44741870-44743470:
NNNNNNNNNNNANCTAGAGATANGGATTCTTCCCTGATGCANNNTAGGGAAAAAGGAAAGG
CTAGAAATTTCTTTGGCAAGCCATCCCAGCCACAGTCTTCTGTGGGACTGCCCTGCTTCATGGA
TAGTATACTCTTGCCAGGGAAGGCAGGTGTTCAGAGGGAGATCTCTGCTCTGCTCTGCTCTGC
AGCCACCTGAGAGAATGGAGGTCATCTACTGCTTCCCATTGCTACTGCTTGCAGCCTTCAGCA
ACTGATCTTCCGCCCCACCCAGTTCAGTGCTTGGCGGGTGGGATTGGCTGATTCAGCCTCTATT
GAAAAGGTAATAGATCAAAATGAGCTGAGAAACTCCTACAATTATTACATGATGACACCAAA
AAGCCAGAGGAGGANAAAAAAGGTTTTCANAAACAAGGACANANAGGNNNNNGTGNNNNN

Sanger sequence for sample 13-SR_chr10_112925393-112926993:
CNNNNNNNGNNCTCNCACTGCTTTTTTGNTTTTCCACAATTCACTGAAGGTTTTTAAAGAACTC
ATGTAGCAAGGTATCTTTTAAGTATTTTTTAGCTCCTGCCAAGGTTTTACTCACGTTCTCAAAA
TTCTTGGACTAAGGATCCAAACTGTAACTAGGCCCCAACCCTGAGCTGCAAAACTCTGAGAAG
GCAACAACAGACAGTCTCAAGGCTCAAGCACTATAATTGCCTCAACTCTTTGAGTATAATCCA
AATTAGTCCACACACTTCAAATAGCTCCTGGGAGAAGGAAGATACATACACAATAGGATTAG
AAATGTGGAGCAGAACCAACTCCTTCCTTGTGTATGCCAGCATTCCTGGCTCTGCAGTAAGTC

CTAAGAGGCCGTATCCAAGTAGCAATGGGAAGAAAGAACTCATGACAGGCAGGGAGATATAC
TAGCAAGGCAGCCTTATATAAGAGAAACACTTAGACATGACACTGCCTTGTAAATTTAAAAA
AAGGTTGTTTTTTTTTTTTTTTTTTTTTTTNNNNTTTTCCCCNTCCCNNCCCAAANNGGGGNNC
NNNNNNAAAAAAAN

Sanger sequence for sample 14-SR_chr11_73175160-73176760:
NNNNNNNNNNNNGNCTCGNNNNGCCCCCTTCCAGACGAGCCACTTCTGGCAGAGACAACTAA
CTTCTCTGTGCTTCAGGTTTTTCATTAACAGAAAGTTAGCAGTAACATCTCCATTGCTATGGCG
AGAATGAAAACACACAATGCCGGTACACAGTGAAAGCTTAGTAACACGAACTGTTACCAGTT
CTTAAGATGTTAACGTTAAATTCTCTCATGTATACTCTAGAGGTCCCTTTGGACTTACTAGGCG
TTAAATCCAAAAAAATATACAGGCTGTGGCCTCGTACTTGCTCTCCAGGGAGCAAATCTTGGC
GACTTCAAAGCTTGGCTAAGCAAGGAACGACAAGGTGTCACACTTCATTTAATCTGAAGAATA
ATATTACAGGCTCTGTCTTCAGATATAAATTATAACAGTACAGAACAAGCAGCGAATTCCAAC
AATTTAAAATCTAAGTAAGTCTACCCGGTGTTAATTCTGGCAAGAGCCTTGCCGATCTGTTTTA
AAGTCACCCCTGCCCTCATTTCAGTAGACGTGCACCATGCCATAGAGGAAGGTCCAAAAGAG
GACGTAGGTGAAGAGGCCTCCAATGAGGCCTCCCGTAAAGAGAGGTCTTCGTGACTTAAAAT
ATTTGTTCCACCTCCTTCCCGCTTGAANNAATGAGNA

Sanger sequence for sample 15-SR_chr13_75771143-75772743:
NNNNNNNNNNNNNNNNNNNNNNNACANNNNANNNNNNTNNNNNNNNNNNNCCCTGGNNTT
CTCNGCGNACNNNNNNNNNAGNNNNNNCGTGAAAACNACAACAAGGGANACGACACCTCTCG
AACCTCCCCACTNCNAGCCCCCCGGATNNNNTCACNTGTGGCGNNTTCNACNATNNTGNTGGT
TTCGTNTTTCTCTTNNCCACTATANNGATNCTGNTTNTTTTNCNCNATTNTCGTTTCTGTTGCTG
TTTATGAATATCNTGNCTTCCTTNAAANN

Sanger sequence for sample 16-SR_chr13_90104423-90106023:
NNNNNNNNNNNNCTTNAGTCTAGTGTATGTTCTGTCAGCTTGAACTGGAATCTCTCTTGTAACT
TTGTAGGTTATAAACATATCTCATATCTGCTTTAGTCTGGGTACTATGCTCTAAGTACATTTCA
GCTTTGACACAGAATGTGAATAGACGAATATCAAAGGATACTTACAAGTTTGTATCCAACATT
TCTTCAGGTTCAGCTGAAAATCAGTTACTGTTTCAAAACAAAGAGGAATTAAATCCTAGCTGA
AAACTATACATAGCATTTATTAATTAATTACTGGGTTTAACTGCTCTTTTTAAAAGTTTGAAAA
AGAAAAAAAAAATTTTTTTCTTAAAAGTGAAGTTTCTATAAAAACAAAGCCCTGAACTTGCAG
TCTTCACTGTGTAGCCCAAATGGCCACCAGAGCTAGCTAGACCATCAGCTTTGACATCGTGCT
TGAAANGN

Sanger sequence for sample 17-SR_chr16_49867208-49868808:
NNNNNNNNNNNNNNGCNCNNNCACCGAAGANNGTTTGTGAAGTGGAAGTTGAACAAATCGT
ATATTTTCATCTATGATGGAAATAAAAATAGCACTACTACAGATCAAAACTTTACCAGTGCAA
AAATCTCAGTCTCAGACTTAATCAATGGCATTGCCTCTTTGAAAATGGATAAGCGCGATGCCA
TGGTGGGAAACTACACTTGCGAAGTGACAGAGTTATCCAGAGAAGGCAAAACAGTTATAGAG
CTGAAAAACCGCACGGGTAAGTGACACAGTTTGCCTGTTTTGAAACGTGTGTTGAGATATGGT
TGCCACTGTGGGAGTGCTGTAAGGTGGAACCTTGCAGAAGTCACTAGGAGGAATTAAGGCTC
TTCTTGGGCAAGTGGGCTAGCCATCTGGATAGAAAGTGAGTCTGGCACTCGTGTTA

Sanger sequence for sample 18-SR_chr16_58465697-58467297:
NNNNNNNNNNNTGCTCACCTGTGTTCATAAACCCTGGCAGCTCCTATCCAGCAGCAGTCACGG

GAAGCCAGCCCTTCCCATCTTTAAGTATACTTCTCCACAGGGAACTGGGAGAGAAGCAACTTA
CGATGTAGAAAAGCTATAAAGGGAGTCCTCTGCGTCACTAAAGTAATCTGTAAATTACTTTAG
GGACAAAGTTAAGGGGGAACTAAAATTTAACCTTCTTTGCTTTCTTTACACACAAGCCCAAAA
ATTTGCAAAATTTTTTTAAATTTCAATTAAATTTTTAAATTTAATTTTAAACCATTGTAAGAA
AAAGAAAGCCAACATGAGCCAGGTCTGATGGCTCAAGCCAGCAATCCCAACTGACCCAAGAA
N

Sanger sequence for sample 19-SR_chr17_12683139-12684739:
NNNNNNNNNNNNCNCTGTCNTCATGGANGGACACCAGCTTGGCAGGGGCTGTTGGCTTGCGC
TGTCCGGGTCTGAACTTCCCTTTTCTGGCCTTCTCACCTCGGACCAGCCCCAGCCTCTCACCCT
CCCTTTCCTTCAGGACCTTGCGCTGTGGGTTTCTCAGGGGCTGTGAGCTTTCCACCTCCGCTCC
TCGGCCTGAGTGAACTTTCACCTCTGGAACGGTCAGGACCTCATCCTCACTGTCCTGGTCGTCC
CCATGCAGGGACGAGAGAGCTTCTGCTTTCACCGCCTTGGTGGTGATATGGCCATTTTCTTGCC
CGTCCTTGCCTAGTCTTGGGGCTGGCACCTGGATCTCTTCCATCAGCCATTCTGTCTCATTCTC
ATCTGTCTCCTCCTCCNCNNANGNNN

Sanger sequence for sample 20-SR_chr18_43320998-43322598:
NNNNNNNNNNNNTAGANNNNANTTGGTGGAAGCCCCGATGACAGAGCTACTTGAGAGAAGA
TGGTACCAGCACGCTCTCCCACCCTGGAATGCGGCTGTCTGCCTGTGAAGCAGAAGCAATTTG
CATTTTGATTTGTAGCTCAAGTAACATGTTGCTTAGAAAAATTCTCATGTTAAGAAAACAGTT
GGGGAGTTGCCAAGGGATTCTCAGGATTTACGAGGCTTGGGCACTACTCATAGTGAATGAAG
AAAGAGACGCATCCTGAGATGGCGCCTGCCCACCTTGAGGACTTCAAGAAGCTGTTCCTTGCA
CAGGAAAGGAACACTTGAGTGGAACATAAGGTTTCGGCACAGTGGTACACTTTGGAAATGCC
AAGTGCAGTGCAAATGACCCAAAGGAGTTTATTTTTCATGATCAGAGAGACCCTGGAAGGGC
TCAGGCATGGGGAACCACACAGTTGAGGAGGCGCAGAATTACAGA

Sanger sequence for sample 21-SR_chr2_120515174-120516774_1_F:
NNNNNNNNNNNNNNTGTCTCNCTGNNNTTGAAAGCACACTGGTAAATGTGAATTTCCTTCCTA
TTTTGAGATGAATGCAAACTGCATGAGATAATGTTTCCCAGACAATTTTCTAACTGAAAATAC
ATGCCTTACCACTCACCAGACACTGATGAACTTGCCAGGGCATTGCTGAGTTCATGGCCTATG
CAGGACAACTATGAACTATCCTAGAAGGGCACAGAGTGCGGCCACAATAAAATACATTTGCT
TCTTGCTAGCTTTGTGCTTTAACCCAGGCACAAAGGACTATGATCTCCCTGACTGTGCTCTTTA
CAGGATGCCTGTTCTAAGCCTCACCGCACAGCAGCAGAAACGCCATCAATCCCATCCGTTCCC
TCACCTTCATGCACAGAACAGTCTNGCTGTGGCCCTCCAGGGTTCGGAGCAGCAGTCCGTTCC
GTGCATCACGAACACTAATGGTGCAGTCGTAAGATCCTACAACCAGCAGTTTTCGGGCACCTT
CCTGAGCTGTTGCGAGGCANCTGACTGCCCNAGGGCCATGGCATTCAAANATCTCCTGTCNCT
TGTTTGTTNNNNAANNNANNNN

Sanger sequence for sample 22-SR_chr3_19133119-19134719_1_F:
NNNNNNNNNNNATGCAGCATACATAGTCATGTAATCTAGAAACTTTCCATGATAATTTATTGC
AAATTGCACTTGACAGTTCACCACAACAATGGAAAGCTGGCCCCCTGTTGGTTTGCACAGTCC
TTGAGCCCCAAACTGAAGTTGAAGTCACCAGTAAAATGATCTGGATAAAAAGTTTGCTGCTGT
GCTCAACCAGCTGGCTCCACCTTCAGGGTGCGATCCAACCCGGGACACTGCCTTGTCCTGCTC
CTTTGTGGGACTCTCATCCTCAGGCAGGTAGTCAGGTAGCTCTGTGTTCTCTTCCACTTCCACA
GGGGCATCCTGAAAAGGGACCAGCATCTGACACGGAGAAGCCGGAAGAGCCATTACAACCA
ACCTAACACGACACGTACATCCCCACAGCCACATCCTCATTCTATGCATCATCTTCATTCCCAA

GAGTCGCAAGGGCTCGCAATCTAACGCTTTCTTCTTTCATGCAGTCAGTCTGTCTGGGTTCTAG
GACAGGAGTAACCTATCNGTGTAGAAATGGGTTAAACGGGTGCAGGGCANCTTTGA

Sanger sequence for sample 23-SR_chr3_95734076-95735676_7_F:
NNNNNNNNNNNNNNNNNNTGCNNNNNANGGGGCCTTCNNNGTCCACACCTCTCACCCCCAAC
TCCCCCCACACACACACACCTTGTACTCTTTTCCCCTACAGACCCTGCTTTTACTCCGAGGCAT
CTGGAGGCTGGAGGCTCTGCCATCATCCGATGCTGGCTGGCTCATCAGCTGTTGCCATGGAGA
CCAGCCAGCAATCAGATCCATGCCCTCCTGCTGCTGGGCACAGTGGAGGTAAAGCCACAGAG
AACCTTGACAAATGAGGGAAAGNGCATACCTCCACGCCCACAGAAGACNATGNCTGNTTANC
TCCTCTTTCCTGCGGGAGGTGACCTCTGGAAGTCATTTCACCCCCCAAGGTCACTNCACTGGA
ATGNGGATGGGGCTGNCAACCCCTGCTGTTATATAGAGGTGAAATANNCCCCNGNNGTGTGT
GGGCCTGACAACACTTTTTGACGGNNTATGACTTCTGATTGA

Sanger sequence for sample 24-SR_chr4_130165017-130166617_2_F:
NNNNNNNNNNNNNGTCTTGGNTGGCTGACCAGGGACCTGGAATCTGTCTATACTCCCTGCAG
TTTCCAAACAGGGCCCTTCCCAGTTCCGGCTGTGATATCTGGGGGCTTCCGCCAGCCTGGTGC
ATCTTTGTCTCCTAGGTGGGATTCTGCCTTAGCGCGCACGCGCGGATCTGAGTCTGTAGAGATT
AGCACCTGGTCCCGAAGTGCACACAAGGCTTCATCGGGCATGGCCCCCAAGGCCTCTTCTGTG
GCTGTGGATCCCACCTCGCCTAGTGGCTGAGACTCAGTGGTGCCCCATGTCCTCCATTCCCAC
GCGACCCCAGACGCCCAGCACGAAGGTCTCGATGTCGCACTCGTTGGTGCCACGCACGATCCG
GAAGTGGCCCCTTTCACCCCACCATGGGCCCCACGAGTTGGCAGCAGTCTGGGAGTGAAAAG
GGGAGAGACATGAGGCAAAAGACAGATGATGGAGGGCTCAGAAGGGCAGGAACTTGGGACT
CGNCCTCGCACCAGANACTCACCCANTACTTAATGGTNCTTCCGTCTGGCAGCTTCTN

Sanger sequence for sample 25-SR_chr4_132832255-132833855_7_F:
NNNNNNNNNNNNNNNNNNNNNGCGCCNATGAGACCGTAGAATCGATGGG

Sanger sequence for sample 26-SR_chr5_104434320-104435920_F:
NNNNNNNNNNNANNAGTCTGANAGATCAAATTGTGTATCCATGTGGCCTTTATCTGTAACTTA
GATAGGAGAATCCATACCTTTCATCCCCATTGATGTTTTTCTACTAATTCAGTAACTATAAACA
AAGTCTCTGTGAGGGTGATCTACTCTTCCTTTCCTTATGGATCCCTGATGCTCTTCCGGGATTC
TAAATGCAGTCTATAAATGAAAAGGGTAGTTAATGACATCGTTCATCAGNAATGCTTTGTGTG
TGTTTCCTTTTCTTCCTTTTTTTTTTTTTTACCCACAAAACCAAAGGAGGAAGGTTAGGCNCTCT
NCCGCTTCCTCGCGCNCTAACTCNCTGCNCTCNACCNTTCNCCTGCGGCTGGCNNNATCNNCT
CNCTCAAANGCNCNNNATTCGGTTATCCNCNGAATCTNGNGATNACNCAGGAAANA

Sanger sequence for sample 27-SR_chr7_27204354-27205954_7_F:
NNNNNNNNNNNNANNNNCAGTGCTCTTANCTGCTGAGCCATCTNTCCANCCCCCCNANNTNA
TTCCANCNCNNANNGGTGNCGGGTTGCTNTTNNNNNNTACTGTTTGGGCGGTTTAAAANCGNC
CATGGGANTATGAAANCCCTGGGNCCCCCNGANNCATNNNNNCANACCCGNTTNNNTCCTTT
ACAGAAATTTTCCTCCNNTTCTNNTTCTTTACCTTNAANCTNNCTGAC

Sanger sequence for sample 28-SR_chr7_27204354-27205954_F:
NNNNNNANNNCTCNCNTNNATTTCTATGTTTTTCAAAGAAACCAAAATTTTTGCTACAGAGTC
ATGACCCCTTTGGGTGTCAAAAGACCCTTTCACAGGGATTGCCTAAGACCATCAGAAANAATA

TATATTTANTTTATGATTCNTANTANNANAGNATATTGTAGTNATAGCNTANGGATCTNNTGA
ANTATATTT

Sanger sequence for sample 30-SR_chr8_85260471-85262071_F:
NNNNNNNNNNNNNNNNNNGNNNNNGCCAGTATCTCAGCATTGTTTGAAGGAAAAA

Sanger sequence for sample 31-SR_chr10_40250385-40251985_8_F:
NNNNNNNNNNNNNNNNNNNNCTGCNNNNTCACTGGGANTTCAAGCATGTGCTATCACATCCC
TCTAGTACTAGGCACCGATCCCAGGGCCCCACACATGCTAGGCAAGTGCTCTTGACCACTTAT
CTACACTCACAGCCCAAGAGTGCCCCTCTCCAGAGAAGGAAGTGAACGAACCTATGTTTTCTT
CTCCTTCCTTCTTTTCTGTCAGAAGCGTCCTTGTCATGCTCAGCTTCTTCATCGTGTGGCACTTG
TACTTCAGCACCGTCTTATCAGTCCCCTCTGGATCCACTAGGACATGAAATGATCTTATTATAC
CGCTTCATGCCATCAGATGTTGAGGGAGTTAGTCTCAACTACTTCCTATGGGAACAGCACAGG
AAGTTCCCACTGCAATTGATCACTCCTGTCCCACGTGCAAAGAAATGACAGATAACACTGACA
TTTGAAGGCACAATAGGTCTCACACACAAAAAACTCTAAGGACAGAGACACGTAAACACAGA
GACTAACGCAACAAGGTAGGGTTCCTAACATGGGAAATGGGGCTTAGGAGTTGGCCCGCAGA
GCACTGTGTCTACAGTCCTATCGCCCAAATGAAGA

Sanger sequence for sample 32-SR_chr11_72777065-72778665_5_F:
NNNNNNNNNNANNNNANANTNNGCATGGTTGCTGTGCTCNGTAATTTCTGTAACTGATGACA
CAAACAATCAAGTTCTTTTAATGTGGGAAAGCTAATGTGGTAATTATCATGCAAGTAACATAG
CAAATCTTAAGACTAATGGTGGGAACTCAAAGATATGTGGGGATGAAACCTGCTTCAAGTTTC
TAGAGCTTCACAGAGGATTTTTAGTGGGTGAGCTTGAACATGTGAGGGCAAACATGTTTAATT
TCGATGAAACAGCTCAGTGTCCTGCTCAAACATTTTGGTGTCCTGCTCCAAGTACCAGCAN

Sanger sequence for sample 33-SR_chr12_54782625-54784225_1_F:
NNNNNNNNNNNNNNNNNNNNCACNNNNATTTCANTCAGACCCTTCCNGGTTCCAAATTAGGC
TTCACTATGAGCAANTCGTTGGTTCTCTGAGCTTCCACACTCTCAACTCTAAAGTGAANACAG
TGAAGTTAGGCTGGGCTGTGGGAAGCTATGAGGGATGCACAAGTCACGTGCTTAATACAGTG
GATGGCTAACCAAGTAAAACCACTGCTATATGGTACAAAGTCCTGANAAGGAAAAATGAACT
TANAATTATTTTTTTAATTCATAAACCTGACCAGGCAGTGGTGGTGTATGCCTTTAATCCCAG
CAGTCCAANNCCC

Sanger sequence for sample 34-SR_chr13_31630105-31631705_F:
NNNNNNNNNNCNNANNGCAGNNNNAGCACACCCATCACTTAGACAAATACCCAAGGGAGTT
CTGCTCACCGATATTTGCCCGGCCCCTGGAAGAGGAAACCTTTCGAAAGCTAATATCCCAGAA
GAGCGACAGACAGAGGAGGTGACTACATGTAAGACATATGTTACTGTGTGGAGGACATAAAA
CTTTTCAGTTCTGGGGTGGCCATTGCATTCACTAATCAGGGTCTGAAAAGGGAGGTGTGTGTG
TGTGTGTGTGTGTGTGTGTGTGTGTGTGCTTTTCAAAATTGCAGTGCTAAAAACACACTATT
TCAAGAAAGCCTTCTCTATCTCCCTGGCTCAGTAAGATTTTTCTTCCACCCACTCCCACTCCCC
CCACTCTCCTTTCTTCATATACGGGNGACA

Sanger sequence for sample 35-SR_chr14_54541419-54543019_1_F:
NNNNNNNNNNANGGNNNNNNCTTAGGTCTCTGGGCCTCACAGAGGCCCGAGATAACAGGTTC
GAAGTCTTGGACATCCCNGACACAGCTGGACACCGAGGAAAACATAAGAACAAAATGGGAG

```
CCCCAGGAACAGCTCAGTCAGCTCAGGGACCAAAGGGAGGAGAGGGCAGGAAGAGAGAGGG
CACTGGGTGGTTCCCATGCGGATGAGAGTCCTTGGTCCAGTCCTTGGCTGGAGCACAGGAAGG
CCTTCTGGCAGGAGGTTAGATGCTGCTCTTTAGGGGAAAGCCTATCCCACCGTTCAAGCATGG
TGGGTCTTGATGAGCAGAGACAGTTTATAATTTTAGAACTTTATTGTAGAAAGGCAAGGAGAA
AGAGAGAAGGNAGAAAGAGTTTANGGGCCAGCA
```

Sanger sequence for sample 36-SR_chr16_52269942-52271542_F:

```
NNNNNNNNNNNNGANTGACTTAACTGTGAAGACAAGGGTAGTAGAAGAATGGAAACCAGGT
AAAAAAGAAAACNNNAACCTCATGGCGATGATATCAAAATGCCAAACAAAAGCGTGAAACTT
ACCATGATGCAGTACAGGGAGATGGAACTGAAGATGTTAAGCTAATGAAGAATAACTTCAAA
TTTTCTTACGGTTTTAATGCCAGAAATGAAATACTCACTCGCAGAGACATTCAGGGAGTTTAC
TGTTCTCTCCAGTTGGTTTTCTGCTGTGCAAGTTAATGTTACATTCTCCTCAGGGGAAATGATA
ATTTTACTATAATACCTGCCATTAATATAAGGAGACTCCTCTGTCTGGAAGAAGAAGGAAGAA
GAGTTTAGAGATGGAGGTGCACTTTACAAGTCTGGGAAAGCATAGTGGTAGATGACACAGGG
CAANACNANN
```

Sanger sequence for sample 37-SR_chr16_94468034-94469634_3_F:

```
NNNNNNNNNNNNANCNNNTCNTAATTTTTAAGTGCTTATGATTCAATTAATCTACATTTTGGTA
ATCAATTATAAAATACTTTTTAAATTAGTATTGGCTATTTTTTTTTTTTTACCCTGGGAAAATGCT
GTCTGCTCCCAGGAGTCCCACCCTTGGCACTTACACTAAGGTTACCATGTAAGTCTCATGATTC
AGCAGTGCTTGGAAGTGCACAGCTGGCTAGGGCTGAAAACAGGACATATGGCCCAGCGCAGG
GCAACCACCCTCACCCTCCTTCCTTATCCTTCCACCAAAAAAAAAAATGTACCACAGCCCTGAT
GTGTTGGTGTTAACATTGTTCTGTTTGACTTTTGTGTAAAGTAATGCATGCAATCTTGTAATGG
GGCCTGAAAACAAGCTAACTGTATTAAAAACTATTCAAAAACTAGGTATTTTTAGTGACCTGT
AGGGAGTGGCAAATACAGACATGGGAACCTTGAATACATTACATTTCTCTCAAACACAAAAA
AAAAAACAAAAAACTTTTCCACCGTGTTCTCTGGTGCCCTGGAAATGCCNTCACCTTCCACCC
GTGTCTGTGTAANTTGTGGTGCCTTAGNGACTTANNATGNGACTCACTCCTTATTTNGAAANG
NTGTCTGTGGNGTCCCNGN
```

Sanger sequence for sample 38-SR_chr19_60770223-60771823_F:

```
NNNNNNNNNNNNNNNNNCTCNCAACCATCTGTAATGATATTTAACCACCTCTTCTGGTGTGTCT
GAAGACAGTTACAATGCACTTACATATAATAATAAATAAATCTTTAAAAAAGAAAAAAAACT
TAGGGAAAAAAAAAAGACTCTAGAGGTAGCTATCTGGTAGGCCTGAAATTCCATCCTGCACT
GCCCCCAAAATCCCACACTTCAAAAAGATCAAAAAAAATAAACTTTCTTCTTTGAAACACTTG
AAAAGATTAAAATATCTCCCAACTCATCTATACAAATACTGAGTAATTACCTGTTAACCTTTA
ACCTGTGTTAAGGGAAAATCCTCAAAAAAAAACTAAAATTTCACAAATTAAACTTTTCCCTAA
ACAAAACAAACAAAAANNANCNCATTANTGAGCTTACCTCNTNNGCATATGGCAGGGCATNT
T
```

Sanger sequence for sample 40-SR_chrX_101403719-101405319_6_F:

```
NNNNNNNNNNGTCTGTCNNAGGGAGCACCCAGTTCTTTCCCTGTTGGCTTTGCTGTTCCCCAG
CCTTCTTTTTGTGTTTTTATAACTGTCCTCAGTTTAGCCACTGTTAAAATGTATATATTGTACTG
AGGTGCCTGGCCTGTTCCTTCAGTGAGCCATGCCCACCCTTGTGTTGTAGTGAGAAACTGTTGT
CACAACTAACTTGTCTCTGGAATTGTTTCAAATAAAGAGTTAAAATTGTTCTTTGCTTTCTCTG
GGGGAGGTAGAGCTGGCGTTGAAGAGTGGAAGAGAAGAGAAAGAGCACCCACTGTGGGTCC
CTGAAGATTAGTCTTCCCTCAGTCAATGAATATCACAACGTTGGTCCTCTTCTCATACATTTTC
AGATACATCAGAAAAAAATATTTTTCAATAGCCATTTATTGAGCTAGAGTTGCTTATGTCTAT
```

AATCCCAGTAGTTGGGAGGTGGAGGCAGGAAGTTAGGAATTCAGTCATCCTTGGCTACATGTG
GAGTTCAGAGCCAACCTAGGTTATGTGAAACCCTGTCACAAAATGGGGACAGGGGCTGGAGA
TGTAGAA

Sequence from the second round of Sanger sequencing

Sanger sequence for sample 21-F-chr2_120515174-120516774_1_F:

```
NNNNNNNNANTCTGNNTNTCTGTNACTTNNNNNCNCACTGAATAAATGTGAATTTCCTTCCTA
TTTTGAGATGAATGCAAACTGCATGAGATAATGTTTCCCAGACAATTTTCTAACTGAAAATAC
ATGCCTTACCACTCACCAGACACTGATGAACTTGCCAGGGCATTGCTGAATTCATGGCCTATG
CAGGACAACTATGAACTATCCTAGAAGGGCACAGAGTGCGGCCACAATAAAATACATTTGCT
TCTTGCTAGCTTTGTGCTTTAACCCAGGCACAAAGGACTATGATCTCCCTGACTGTGCTCTTTA
CAGGATGCCTGTTCTAAGCCTCACCGCACAGCAGCAGAAACGCCATCAATCCCATCCGTTCCC
TCACCTTCATGCACAGAACAGTCTTGCTGTGGCCCTCCAGGGTTCGGAGCAGCAGTCCGTTCC
GTGCATCACGAACACTAATGGTGCAGTCGTAAGATCCTACAACCAGCAGTTTTCGGGCACCTT
CCTGAGCTGTTGCGAGGCAGCTGACTGCCCGAGGGCCATGGCATTCAAAGATCTCCTGTCGCT
TGNNNNNTCTNAAA
```

Sanger sequence for sample 21-R-chr2_120515174-120516774_1_R:

```
NNNNNNNNNNNNNNNNCTCNGGCNGTCNGCTGCCTCNCTCANCTCAGGAAGGTGCCCGAAAAC
TGCTGGTTGTAGGATCTTACGACTGCACCATTAGTGTTCGTGATGCACGGAACGGACTGCTGC
TCCGAACCCTGGAGGGCCACAGCAAGACTGTTCTGTGCATGAAGGTGAGGGAACGGATGGGA
TTGATGGCGTTTCTGCTGCTGTGCGGTGAGGCTTAGAACAGGCATCCTGTAAAGAGCACAGTC
AGGGAGATCATAGTCCTTTGTGCCTGGGTTAAAGCACAAAGCTAGCAAGAAGCAAATGTATTT
TATTGTGGCCGCACTCTGTGCCCTTCTAGGATAGTTCATAGTTGTCCTGCATAGGCCATGAACT
CAGCAATGCCCTGGCAAGTTCATCAGTGTCTGGTGAGTGGTAAGGCATGTATTTTCAGTTAGA
AAATTGTCTGGGAAACATTATCTCATGCAGTTTGCATTCATCTCAAAATAGGAAGGAAATTCA
CATTTACCAGTGTGCTTTTCAAGTTACAGAGGAGACAGAACTTCCGCAGAAGTGACTGCNCNT
NNTTCTCTTTTCN
```

Sanger sequence for sample 22-F-chr3_19133119-19134719_1_F:

```
NNNNNNNNNNNNATACNTAGTCATGTAATCTAGAAACTTTCCATGATAATTTATTGCAAATTGC
ACTTGACAGTTCACCACAACAATGGAAAGCTGGCCCCCTGTTGGTTTGCACAGTCCTTGAGCC
CCAAACTGAAGTTGAAGTCACCAGTAAAATGATCTGGATAAAAAGTTTGCTGCTGTGCTCAAC
CAGCTGGCTCCACCTTCAGGGTGCGATCCAACCCGGGACACTGCCTTGTCCTGCTCCTTTGTGG
GACTCTCATCCTCAGGCAGGTAGTCAGGTAGCTCTGTGTTCTCTTCCACTTCCACAGGGGCATC
CTGAAAAGGGACCAGCATCTGACACGGAGAAGCCGGAAGAGCCATTACAACCAACCTAACAC
GACACGTACATCCCCCAAGCCACATCCTCATTCTATGCATCATCTTCATTCCCAAGAGTCGCA
AGGGCTCGCAATCTAACGCTTTCTTCTTTCATGCAGTCAGTCTGTCTGGGTTCTAGGACAGGAG
TAACCTATCAGTGTAGAAATGGGTTAAACGGGTGCAGGGCCACTTTNANNNNNTCCGTGGGA
```

Sanger sequence for sample 22-R-chr3_19133119-19134719_1_R:

```
NNNNNNNNNNNNNNNTTCTACNCTGATAGGTTACTCCTGTCCTAGAACCCAGACAGACTGACTG
CATGAAAGAAGAAAGCGTTAGATTGCGAGCCCTTGCGACTCTTGGGAATGAAGATGATGCAT
AGAATGAGGATGTGGCTTGGGGGATGTACGTGTCGTGTTAGGTTGGTTGTAATGGCTCTTCCG
GCTTCTCCGTGTCAGATGCTGGTCCCTTTTCAGGATGCCCCTGTGGAAGTGGAAGAGAACACA
GAGCTACCTGACTACCTGCCTGAGGATGAGAGTCCCACAAAGGAGCAGGACAAGGCAGTGTC
CCGGGTTGGATCGCACCCTGAAGGTGGAGCCAGCTGGTTGAGCACAGCAGCAAACTTTTTATC
CAGATCATTTTACTGGTGACTTCAACTTCAGTTTGGGGCTCAAGGACTGTGCAAACCAACAGG
GGGCCAGCTTTCCATTGTTGTGGTGAACTGTCAAGTGCAATTTGCAATAAATTATCATGGAAA
GTTTCTAGATTACATGACTATGTATGCTGCATTTTACATCTTATTGGACATTTTCCGCACAAAG
ACAGAA
```

Sanger sequence for sample 23-F-chr3_95734076-95735676_7_F:

```
NNNNNNNCNNNNNNNNGNGTNNTCTGNNANNGCTNNAGGCTTCTCGGGACGGCNGNACTGAG
GTTCTCCAAGAAGGATTNNCCTTCTTTAACTCCCANAAATGGTCTATTCTCCTCNACCTGACTT
CTGCGAACAGGCTTGAGGTACTAAAAAGAGGGGTAACCTCTACATCTANCCTTCTGATTCCCT
GGAGCCTTGGAATCCCAGGTCACACGCACTCACCTCCTCTTCACCGCAACACGCCTGTTCTGG
```

ATATGGAAGCTCGAAACAGCGGACGTCCATATTCTGGATCAGCCCCGGAATCTGTTTACTGTG
GTTGGTATGAAANAGGNNCNNTCNNGNTTTCTGTCCCTTCCGCANCATCCANGGNCTGCTGGG
GGCGGTGGCTTATGCCGGTAATCTCTTTACTTGAAGTACNNGGGANAGAAGATTGGTCCNGTT
NNAAACNNCCTTGACNACNNANTGAAACTTTCANNATATNTNANNNTCTCNTNNNNNCANCA
ACTCAAATGGAACTGGGAAGTCATGATTTTGACCTCCCCAAGGGAACCCCCGGGGAGGAGGA
GGTTGGGGGATNAGGCAA

Sanger sequence for sample 23-R-chr3_95734076-95735676_7_R:
NNNNNNNNNNNGNNGGNNNNCNNNNNATGACTTCNNNNTTCCGTTTGANGCTGGTGGGGTG
AGTGAGATGTTTATTTATTGTGAAAGTTTCACTCCGTGGTCAAGGCTGTTTTGAACTGGACCAA
TCTTCTTTCCCCTGTACTCCAAGTAATGAGATTACAGGCATAAGCCACCGACCCCAGCAGGCT
TTGGGTGCTGTGGAAGGGACAGATACCCTGAGCTCTTCCTCTTTCATACCGACCACAGTAAAC
AGATTCCGGGGCTGATCCAGAATATGACCGTCCGCTGCTGCGAGCTTCCATATCCAGAACAGG
CCTGCTGCGGCGAAGAGGAGGTGAGTGTGTGTGACCTGTGATTCCAAGGCTCCAGGGAATCA
NAAGGCTAGATGTAAAGGTTACCCCTCTTTTTAGTACCTCAAGCCTGTTCGCAGAAGTCAGGT
TGAGTGGAATAGATCATTTCTCGGAGTTCTAGAAGGCTTCTTCCTTCCCGAGAACCTCTTGTCC
CTGCCCTGGGCCTCAAGCTTCTAGCATTTCCAGATGTACACATCTGTCAGGGACAAGCAGGCA
GTCCTGCTGTGGAAAAGGAAGA

Sanger sequence for sample 24-F-chr4_130165017-130166617_2_F:
NNNNNNNNNNNNNNNNNTNNNNNNNNNCTGANCNGGNACCTGGNATCTGTCTATACTCCCTG
CAG

Sanger sequence for sample 24-R-chr4_130165017-130166617_2_R:
NNNNNNNNNNNNTCNNNNNTTGCTTTGCCGCTAGGTGGGGAGAAGAGACGCTGCCAGACGG
AAGGACCATTAAGTACTGGGTGAGTCTCTGGTGCGAGGCTGAGTCCCAAGTTCCTGCCCTTCT
GAGCCCTCCATCATCTGTCTTTTGCCTCATGTCTCTCCCCTTTTCACTCCCAGACTGCTGCCAAC
TCGTGGGGCCCATGGTGGGGTGAAAGGGGCCACTTCCGGATCGTGCGTGGCACCAACGAGTG
CGACATCGAGACCTTCGTGCTGGGCGTCTGGGGTCGCGTGGGAATGGAGGACATGGGGCACC
ACTGAGTCTCAGCCACTAGGCGAGGTGGGATCCACAGCCACAGAAGAGGCCTTGGGGGCCAT
GCCCGATGAAGCCTTGTGTGCACTTCGGGACCAGGTGCTAATCTCTACAGACTCAGATCCGCG
CGTGCGCGCTAAGGCAGAATCCCACCTAGGAGACAAAGATGCACCAGGCTGGCGGAAGCCCC
CAGATATCACAGCCGGAACTGGGAAGGGCCCTGTTTGGAAACTGCAGGGAGTATAGACAGAT
TCCAGGTCCCTGGTCAGCCAGGCCAAGACCACAGGAGCTAAGACACCCCAACCTCNNNNCCC
CTCCTAAAN

Sanger sequence for sample 25-F-chr4_132832255-132833855_7_F:
NNNNNNNNNNNNNNNGGACNCGCCANTGANACNNNGTAATCAATGACCAATCCGTTGAGGC
TTGATGAGGTTCCNNTTTAAATGGGACCCTCTAGTTTGCCATGTGCTGCAAAGGCGAGTCTTA
AGTGTAAAGGGAGATACCAGCGCATAGGTAGCCAGTGCTCCCTTGCAGGATGTCAGGCAGCT
CCCACAGTTCATCAAGTTAGCCANCCTAACTGTTAGTGTTTAATGATAATAAGTGTCATATTCA
GGATCACTGACACAAAAGCACCTTTTTTGTTTCTTTTTTCTTTTACAGTACTAGGGTCAACCTT
AGGGTTTTGCCCATCCTAGGCAAGCACCCTACCATTGAGGTACAACCCAGCCCTTGGTTTCTG
AGTCAGAGTCTCACTATATAGTTAGGGCTGGCTTCCAACTTACTATGTAGCCAGTTTGGAGTC
AAACAAACTATGTCAACCAGACTGACTGAACTCATAATTCTCCTGCCTCATCTTCCTGAA

Sanger sequence for sample 25-R-chr4_132832255-132833855_7_R:
NNNNNNNNNNGNNNCGGNNGNATATTNTGATTGACTCCCACTGGTTACATATCAAGTTGGAC
ACCTTCCCTAACTATATAGTGAGACTCTGACTCNAAACCAAGGGCTGGGTTGTACCTCAATGG
CANGGTGCTTGCCTAGGATGGGCAAAACCCTAAGGTTGACCCTAGTACTGTAAAAGAAAAAA
GAAACAAAAAAGGTACTTTTGTGTCAGTGATCCTGAATATGACACTTATTATCATTAAACACT
AACAGTTAGGGAGGCTAACTTGATGAACTGTGGGAGCTGCCTGACATCCTGCAAGGGAGCAC
TGGCTGCCTCTGCGCTGGTCTCTCCCCTTACACTTAAGACTTGCCTTTGCAGCACATGGCACTT

TCTTGGGTCACATTCGGCTGGAACCTCACAAGCCTCAACAGATTCCCATCGATTCTACGGTCTC
ATTTGGCGCGTCCACGAGGGCATACACTCTGCGAGAGAAGCCTCAGACATTGCCA

Sanger sequence for sample 26-F-chr5_104434320-104435920_F:
NNNNNNNNNNNNNNANTNATANAGGTCACTGCCCTGTGGAAAGGTGTGCTCTCTTCTGGGCTA
GCCACCCTTGGCTATCACTCCACACAGCTGGGATGGGGGCTGGTCACTCTGCTGTGCCTCCCA
ATTTCCCCCAGGTTTCCCGGTATTTTCAAN

Sanger sequence for sample 26-R-chr5_104434320-104435920_R:
GNNNNGNANNNANNNNANTGACCAGACCNNNTCCCAGCTGTGTGGAGTGCTAGNNNAGGGT
GGCTAGCCCACAANAGAGCACACCTTTCACCACGGCAGTGACCTTTATGAGTCTTACTTGTTG
GTTTAGGCTCCNGGGCTGTCTCGNCANAACTTTCANCTGCCGCAGAAGACTGCAAACCCAAGC
AAGGATGCTTCTTCAGTGTGAGCTGCTGGTGGCTCAGACCTCCCAGAATTTAAATGCTGGTCC
AGACAGTCTCCACCAATCAGGAGGTGGAGTGATGTGTCATGAGGTTTTTGCCACTACCCGGCC
CACCTGCTCCTACACTTCCTCCTCTGGTTTTGTGGTTAAAAAAAAAAAAAAAGGAAAAAAAAGGA
AACCCCCCCAANNCTTTACTGATGAACNATGTCTTTANCTACCCTTTTCTTTTATAAACTGNNT
TTAAAATCCCGNAAAANCATCNGGGNTCCNTAAGGAAAGGAAAAGTAAATCNCCCCCACANA
AACTTTGTTTATANTTACTGANTTANTANAAAAACATCNNTGGGGATNAAAGGTATGGATTCT
CCTATCTAAGTTACANATAAAGGCCACATGGATACACAATTTGATTCTCTCNNACTTACTTAA
ATCTANAAAACTGCTGTCTCGNNAAAAATTTTANN

Sanger sequence for sample 27-F-chr7_27204354-27205954_7_F:
NNNNNNNNNANNNNNGTCNGTGCTCTTNNNGNTGAGCCATCTCTNCAGCCCNTATTCTNCAC
CATATTNTCANTGNTCAGTGNTGCGNNTNNNNNGGAATCGTTTGTTGATATTTNGGCANNGCA
NNTGNNANNNNATAACGTNNCTNNGGNNANAANAAANGAATNNNCANAAGGNCNNGANCCN
GGNAGNGGNGANTAGTGCACTAAGGGTGTAGGGTATTAAAGCTCAAAATTTNATTCATTTTGC
TTCATTGATANGCCTTACTNAANNCCCCNAATCCACAACACTCCNNNCNCGANGACTGGGGN
ACTTTTTTTTTTTAACATCAGAAAACCCCCNAAATTACNNAGGCTCCANCCTGGNAAAGCTTTT
TGAAAATGGANAANNCACCCCTTNCTAATNAAGAAACTTTCCCTCGGGGGGGGGNANNCCAT
TTTNCNCNCAAANTNGTATCTNCCTTTTNCCNCGNANNAGACGGCCTGGNGCTNCNTGCCTCC
GNCNGGACGTNTTCCTANGAGTNGANGGATCNNTGNGANTTGTCTGTCCNTCGGNGANTGTG
CCNTCGNTTATTNNTTCATANGGGNAGTGGTCGGNGTTACTGNNGNTTTNCANANCGTTCGGT
TCTNCTNNATCNTNGCCNTGTGGGAGTGANACGAGCGGNTNTACTGNAGTTGNNGAGTAGAG
NNGNGAGNGGCNNACGNNANGTNNAGTNNCNNNTGNTNNGNAGNNACNNAGNNGTNNNNG
NNGNNAANNNNNNNNNNGNNTNNNNTNAGNNGTNNTAGANNGNAGNAATNGNNNNNTNGA

Sanger sequence for sample 27-R-chr7_27204354-27205954_7_R:
NNNNNNNNNNNNNNNANNNNNNNCTGCTCTTCTGANNTCCTGAGTTCTNTTCCCANCNACCACA
TAAAAGGAACCTTTGGGGCTGGAGNNATTCCNGGTCTGCTAACTGTGGTTAGTGGANNCCGG
AACAGACATANGGNNTCCTACGAGGTTAAGNTGGGNNAGCGGTNNCGTGCTTGGGGAGGATC
CCCGTTAACTTCCAAGAAGCTGCCTAANTGGCTGCANGGNGNNAAGCCCTGNCTGCTCTTTGT
GAGGTCCTGAGTTCCNTTCCCGCTGACTTCATAANGNNACANAACTAACCCTTGTCTCGAGAT
CTATTCCNAGCGACGACNTAAAANCACANGNNACGCACAAACAAANCAATGGNTTGNGGTGN
CCGTAAAAANAGCATTNCNTTGCAACCACCGTGCTGGTCTGNCGGATCACCGGAATTCACTTC
CCACGGTCCTTACAAAGTNCCGAGTACCTTGCAACGATCATGATACATNNNTTNANATTTATT
TTTTNGTGATGNNCTTTAGGCGGTCCCTGNNATNNGNNCNGNNNNNNNCNTAAGGNGTCTGN
NCCCTNNANNAAAATTTTNGNTGTGC

Sanger sequence for sample 28-F-chr7_27204354-27205954_F:
NNNNNNNNNANCTNCTNNNNTAATATTTCTATGTTTTTCAAAGAAACCAAAATTTTTGCTACA
GAGTCATGACCCCTTTGGGTGTCAAAAGACCCTTTCACAGGGATTGCCTAAGACCATCAGAAA
AAATATATATTTACTTTATGATTCATAATAGTAGCAAGATTATAGTTATAAAGTAATGATGAA

AATAACTTTATGGTTTGGGGTCACCACAACATGAGGTTATTATTCAAGGGTTGCAGCATTACA
AAGGTTGAGAACCACTGAGATAGAGGTAGCTGATCTGTGAGTTTGGAACCAGCCTGGTCTAC
ATAGCAACATCTAAGCCAGCCAGGGCTACATAGTGAGAACCTGCCTCAAAAGCAAAACAGCA
GTCTTACTATGTAGCCCTAGATGGCCTAGAACTCCCTATGCAGTAGGCCAGGTTGGCCTTTGCT
TCCTGTGTCCTGGGATTAAAGGCCTGTTCCTACATGCCTAGCCTAGAAGGCTCTTGGTAGTAA
CTCAGACCCTAGTCGTGGCCCTCCCACCTGAATGTGACATCCACTGGTCTTTCTTTGGGGGAAT
TGTTAAAGACACTGTCAGTTCACAGAGCAGGACTTAGCCCAGCTTCCTTCTCCTTGGCACACT
CAGACTACCCCGTCTNNNNNNNNCNNNNNNGNNNNNNN

Sanger sequence for sample 28-R-chr7_27204354-27205954_R:
NNNNNNNNNNNNNNNNANGAGCTGGGCTAAGTCCTGCTCTGTGAACTGACAGTGTCTTTAAC
AATTCCCCCAAAGAAAGACCAGTGGATGTCACATTCAGGTGGGAGGGCCACGACTAGGGTCT
GAGTTACTACCAAGAGCCTTCTAGGCTAGGCATGTAGGAACAGGCCTTTAATCCCAGGACACA
GGAAGCAAAGGCCAACCTGGCCTACTGCATAGGGAGTTCTAGGCCATCTAGGGCTACATAGT
AAGACTGCTGTTTTGCTTTTGAGGCAGGTTCTCACTATGTAGCCCTGGCTGGCTTAGATGTTGC
TATGTAGACCAGGCTGGTTCCAAACTCACAGATCAGCTACCTCTATCTCAGTGGTTCTCAACCT
TTGTAATGCTGCAACCCTTGAATAATAACCTCATGTTGTGGTGACCCCAAACCATAAAGTTAT
TTTCATCATTACTTTATAACTATAATCTTGCTACTATTATNAATCATAAAGTAAATATATATTTT
TTCTGATGGTCTTAGGCAATCCCTGTGAAAGGGTCTTTTGACACCCAAAGGGGTCATGACTCT
GTAGCAAAAATTTTGGTTTCNNNNAAAAACATAGAAATATTATGGGAGTATGTTTTCATTTTA
ATCCCAGGTGTNNNNNNGGGNNNNNNNNCNAN

Sanger sequence for sample 30-F-chr8_85260471-85262071_F:
NNNNNNNNNNNNNANCNCNNTGCTCTNTGTAGTTGTGCAATGCTTGGTGCTATGCCCCCCCTG
NTCNCTGGTCTCGTGAAACCCNGATCCTTCTCATNTCTTTTGAATAGGTTGGTTCGAGACTTGA
ATGGTTACCTACAAAAAGAATAATTGTGTAGANGGCNTGTATCACCAAACAACATTTATCAGC
ACTGTTTTAAGGAACTTGCGTTACTTTCATCCCTACTTGAGTAATTATGAAAAAGAGATGAGA
ATATGATTCTGAGTTTAAGCAACGTCTATTACCTATTCATTAAGGAAAAAACAATTGTAAAAA
CTCAAAATATANAGCAACATAAACCAGCAAAAGAACAAAAAAACAAAACTGCCTTTAACTTA
AAAATCTTTTAAAAATGTATTTACTTCTATTTTATGTGCATTGGTATCTTGTCTGAATGTATGCC
TGGGTGAGGGGGTTGGATCCCGGAGTTACAGACAGTTGTGAGCCGCCATGTGGGTGCTGGGT
ATTGAACCCAGGTCTTCTGGAAAA

Sanger sequence for sample 30-R-chr8_85260471-85262071_R:
NNNNNNNNNNNNNNNTNGNNNNTTTTGACCCGGGAGAATCCCCATCAGGGGGGGGAGACA
GAGATTTGGGGCGNGCGAAGAGCCTTACTATAAAAAGCTTTCTCNGATTGTTGNGTTGNNCAG
CGCTCGTTTCCTNCCTCTTTAGTTGGTTNACGTTGATCTATATTTTGAGTTTGCCGGTTCTTTTT
TCCNNAATGAAAAGCTATGACANGTTGCTTGAAATCGGAANCATATACTCATCTCATTTTCNT
AGTTACTTCGTATTTGATGANNGTCNTGCCNTTCCTTAAANAGNGCTGCTAAATTTTCTTTTGG
GATCCATGCCTGCTACCCAATTATTCTTGTTGCCTTTAACCTTTCCTTTCTTGAAGCTGCTTATT
CNAAAAATTCGAGAANATCTCCCAAATCNNGANTCCAGNGAANAANCNNAAAANATAAACA
AGCATTTTCNCAACNCATNGNAGCACTTGCGCTCAANANNNNAATCACCAGAGTCTGAGGGG
CCAATCTATGAAA

Sanger sequence for sample 31-F-chr10_40250385-40251985_8_F:
NNNNNNNNNNTNNNNNNNTGNNNNANCACTGGNAGTTCAAGCATGTGCTATCACATCCCTCTA
GTACTAGGCACCGATCCCAGGGCCCCACACATGCTAGGCAAGTGCTCTTGACCACTTATCTAC
ACTCACAGCCCAAGAGTGCCCCTCTCCAGAGAAGGAAGTGAACGAACCTATGTTTTCTTCTCC
TTCCTTCTTTTCTGTCAGAAGCGTCCTTGTCATGCTCAGCTTCTTCATCGTGTGGCACTTGTACT
TCAGCACCGTCTTATCAGTCCCCTCTGGATCCACTAGGACATGAAATGATCTTATTATACCGCT

```
TCATGCCATCAGATGTTGAGGGAGTTAGTCTCAACTACTTCCTATGGGAACAGCACAGGAAGT
TCCCACTGCAATTGATCACTCCTGTCCCACGTGCAAAGAAATGACAGATAACACTGACATTTG
AAGGCACAATAGGTCTCACACACAAAAAACTCTAAGGACAGAGACACGTAAACACAGAGACT
AACGCAACAAGGTAGGGTTCCTAACATGGGAAATGGGGCTTAGGAGTTGGCCCGCAGAGCAC
TGTGTCTACAGTCCTATCCNNNNAAANNNAANGAA
```

Sanger sequence for sample 31-R-chr10_40250385-40251985_8_R:

```
NNNNNNNTCNNNNNNNNNNNTNNTNNGCNCCATTTNCCATGTTANGAACCCTACCTTGTTGCGT
TAGTCTCTGTGTTTACGTGTCGCTGTCCTTANAGTTTTTGTGTGTGAGACCTATTGTGCCTTCA
AATGTGNNTGTTTTCTGTCATTTCTTTGGACGTGGGACAGGAGTGATCAATTGCATTGGGAAC
TTCCTGTGCTGTTCCCTTGAGTGNTAGTTGAGACTAACTCCCTCATCATCTGATGGNGTGAAGC
GGTATAATAATATCGTTTCATGTCCTANTGGATCCANAGGGGACTGATAANAANGTGCTGAAG
TACAAGTGCCACACGATGAAGAAGCTGAGCATGACTAGGACGCTTCTGACACAAAAGAAGGA
AGGANAATAAAACCTACGTTCTTTCACTTCCTTCTCTGGATAGGGGAACTCTTGGGCTGTGAG
TGTANATAANNGGTCAANAGCACTTGCCTANNATGTGTGGGGNCCTGNGATCGGTGNNNNNN
ACTNNNNANNTGTGATAGCACATGCTTGAAATCCCAGTGATTTGGGCAGTTGGGGAAGGCAG
CTGAAGATCATGAGTTCAAGGCCNGANNTGAGCTA
```

Sanger sequence for sample 32-F-chr11_72777065-72778665_5_F:

```
NNNNNNNNNNNNNNNNNNNNTNNGANNCTGNNCCNCNTTTTGGTACCCGNTCTGGAAAAAAA
NNAATATTTTAATGGAGANGAGCTNATGTGGTAGTGTGCATCAAAGAACGGNATTCTGTCTTG
GGAGTGNTGNNGCAGCNTCGAAGACTCGTGGNGANGTGGCTGCTTAAAGGGATNCTTTTNCT
AAGGAATGAGTAAGGAAGAGTGTTTGATGGACACGNCGGAAGGGGGATTATGGTGTGATCCA
TCTGTGTGTCCTGCTCCCCCAAATCGGCGACTCGATCTTATTATCTCCACTCACTGGACCATCC
CAGTTCTCATGGAGTCAGTTTGCCCTGGAGTAGAGGTTTTAGAAATGTCACAAATGTCAGAAG
TTAAGAGCTCAGTCAATTACCCTGTGACTTAGAACAGTTACCTTCATTTTACTAAGGTGCAAA
ATCTTCTTGGGAAATGTCTGATTTTCCACTGAGAACCACAGTCCTGCTCCNANNTACCANN
```

Sanger sequence for sample 32-R-chr11_72777065-72778665_5_R:

```
NNNNNNNCNGACNCNNNNTNCNNNNNCCATTTCCCATGTTAGGAACCCTACCTTGTTGCGTTA
GTCTCTGTGTTTACGTGTCTCTGTCCTTANAGTTTTTGTGTGTGAGACCTATTGTGCCTTCAAA
TGTGNNTGTTTTCTGTCATTTCTTTGGACGTGGGACAGGAGTGATCAATTGCATTGGGAACTTC
CTGTGCTGTTCCCTTGAGTGNTAGTTGAGACTAACTCCCTCATCATCTGATGGNGTGAAGCGG
TATAATAATATCATTTCATGTCCTANTGGATCCAGAGGGGACTGATAANAAGGTGCTGAAGTA
CNAGTGCCACACGATGAAGAAGCTGAGCATGACAAGGACGCTTCTGACACAAAAGAAGGAA
GGAGAATAAAACATACGTTCTTTCACTTCCTTCTCTGGAGAGGGGAACTCTTGGGCTGTGAGT
GTANATAANTGGTCAANAGCACTTGCCTANCATGTGTGGGGCCCTGGNATCGGTGNNNNNNA
CTANNNNNNTGTGATAGCACATGCTTGAAATCCCAGTGATTTGGGCAGTTGGGGAAGGCAGC
TGAAGATCATGAGTTCAANGNNNAAAGTGAGCCNN
```

Sanger sequence for sample 33-F-chr12_54782625-54784225_1_F:

```
NNNNNNNNNNNTNNNNNNNNNNNTCGNNNNGTCTTTCNNNNNNNNCATGCCTTTNATCCCNGCNN
TCCAANNNNNNNNNNNNNGNNNGNTNNCNGCTTCTCCAAGNNCCGTTTTTTCTTCTGGNCTCTG
TGGNCNCTGCNGNNGGCTCATGCCTNTATTNCNNNCTGTCCNNGNAGNNANGCAGGANATGG
TGAANGAAGGAAAGCCAAAGCTANNCNNGCACCNNACTTNNAAAGGAACCTGNATAGTCNN
CTTATCNTTTTAAAGAATTATGACNGNACATGAATTGATGNNGTNTGCNAACNATCGCAAATT
CTCATTTCTTANNATTTGGCAATGATTNCTGTCTANACNNATCTTTCTTGANNNNAGCANCCN
NNTCCTCANATCTTNGNNTNNNNANNNANNNNNNNNNTTATATATTGNANNGNNNNNCNTNNN
NNNNNNNNTNCTTTNNNNNNNNNNTCNTNNNNNNNNNNCCTTNNATCNTNNNNNNNCCANNNNTN
NTNTGNNNNNNNNNNN
```

Sanger sequence for sample 33-R-chr12_54782625-54784225_1_R:

NNNNNNNNNNNNNNNNTAGCTGNTTTTNACCCAGGAGAATCCTCCTCAGGTGGCGTANAANAG
ATTTGATGTTGNGAGAAGAGACTAAATATAAAAAGNTNTCTCNGATTTTTANGTTGCCCCACG
CTCGTTTTTTTGCTCTTTTGTTGGTTNACGTTGATCTATATTTTGAGTTTTTACGATTCTTTTTTC
CTTAATGAATAGCTAATACATGTTGCTTAAAATCNGAAGCATATACACATCTCATTTTCATAGT
TACTCNNGTATTTGATGAAAGTAATGCAAGTTCCTTANAANNGTGCTGCTAAATTTTCTTTTGN
GATCCATGCCTGCTACCCAATTATTCTTTTTGTNTTTAACCTTTCNNGTCTTGAACCNACTTATT
CAAANGATTCGAGAANATCTCCCAAATCTTGANTCCAGTGAANAAACATAAAAAATAAACAA
GCATTTTCACAACACATTGGAGCACTTGCGCTCAAGACNNNAATCACCAGAGTCTGAGGGGC
CNATTCTATGAAA

Sanger sequence for sample 34-F-chr13_31630105-31631705_F:
NNNNNNNNNATANCNGTAAGTANCACACCCATCACTTATACAAATACCCAAGGGAGTTCTGC
TCACCGATATTTGCCCGGCCCCTGGAAGAGGAAACCTTTCGAAAGCTAATATCCCAGAAGAGC
GACAGACAGAGGAGGTGACTACATGTAAGACATATGTTACTGTGTGGAGGACATAAAACTTT
TCAGTTCTGGGGTGGCCATTGCATTCACTAATCAGGGTCTGAAAAGGGAGGTGTGTGTGTGTG
TGTGTGTGTGTGTGTGTGTGTGTGTGTTTTTCAAAATTGGAGTGAAAAAANCACACTATTNAN
ANANNNCTNTCTCTATCNCNNTGNCTCANNNAGATTTTTCNTNCNCNCTCTCCCACCCCCCCC
ACTCTCNNTTNTTCANATANGGGANANAN

Sanger sequence for sample 34-R-chr13_31630105-31631705_R:
NNNNNNNNNNGNNANNNTGNNNAATGTCCTACTGAGCCAAGGAACTAAGAAGGCTTTCTTGA
AATTGTGTGCTTTTAGCACTGCAATTTTGAAAAGCACACACACACACACACACACACACACAC
ACACACACACCTCCCTTTTCACACCCTGATTAGTGAATGCAATGCCCCCCCCAAAACTGAAAA
TTTTTATGCCCTCCACACAAAAACATATGTCTTACATGTACTCACCTCCTCTGTCTGTCTCTCTT
CTGGGATATTATCTTTCNAAAGGTTTCCTCTTCCGGGGGCCGGGCAAATATCGGAGAACAAAA
CTCCCTTGTGTATTTGTCTAAGTGATGGGTGTGCTACTTACTGCTATTCTCTATCCAGCGTGCA
CAGTGAGANTCNNTTTGAAATANA

Sanger sequence for sample 35-F-chr14_54541419-54543019_1_F:
NNNNNNNNNNNNNNNNNNNNACTCTCNNNNNCTNCNGTGTGCTCGNGACTGGCTGGTACTGNA
ATCATTCCCCGGGCCCCCATGTAGCCACTCCTCCTGGANATTGTAGTCCCACCNACTGAACTC
AATGNNCCTANNCCCTCTGGGNAGGTCGATTTCATTTTCACAGGAGAGGCTAGGTGGGGGAG
GGATCAGCAGAAAAAGTTTAGGGGCCAANAATGGNTGGAGCACGGGAANGCATTCTGGCAA
GAGGAAAGATGCCGCTCTTTANGGGAAAGCCTCTCCCACCGTTCTCTCATGGTGGGTCTTGAG
GAGCAGAGACAGTTTATAATTTTAAAACTTTATTGNAGAAAGGCAAGGAGAATCTNNGAAGG
NTGAAAGAGTTTAGGGGCCNANNNNNGATATGCCGAAGGAGTTTAGGGGCCANNGNTAGTTA
CAACCTACCGCTAACAGANCTTNNATCCTCTANAACATCTACTCGCCTNNAANNTTTNCCNNC
GCCNTCACAACTACTNNTTANAAGCAGANNNCGTAACAGCATTAAGGGAGGACTCTTAGTTA
CTTTTCTATTGCTGTNNAAGGNTGATCATGNCCNGGCATCTTATAGNNNAAAGAATTTAGGGG
CCCANCANANAGNAGTANGGGGGGNGNNNNNTTNNNNNNTGGNNACCANNNTCTGCAATNNA
NGNNNNATNGGNNATNNCCGTTTCNGTNNNNNNTNNNNNGANNNNNN

Sanger sequence for sample 35-R-chr14_54541419-54543019_1_R:
NNNNNNNNCNNNNNTCCNNATNNTNTNNGTATACACTTCTGACGGTGTTGTAAGACACCGCT
GATNGAGGCTAAACAGATCCTTTAAGCGTGGCTAAAGTGCTCCCGGGGGCCAGATACNCTGN
CTGNGATCNNCTATCAGACTGGACNNNGTNTCTCTTCCTGTGATAACGNCCATGAACCTATCN
CTTCCTGCCCTAGGCTCNNNNNGGTCCCTGANCTGACTGANNTGTNCCTGGGGCTCCCATTTT
GTTCTTATGTTTTCCTCGGTGTCANNNTGTGTCTGTNATGTCCATTACTTCGAACCTGTTATCTC
GGNCCTCTGTGAGGCCCNNNNCCTAANACTGCCCCTTGCCCACCCCCCTTGANCTGANNGNNNG
NNNANNNNNNNGT

Sanger sequence for sample 36-F-chr16_52269942-52271542_F:
NNNNNNNNNNNANTGACTTANNTGNNNNACAAGGGTAGTAGAAGAATGGAAACCAGGTAAA
AAAGAAAACGGTAACCTCATGGCGATGATATCAAAATGCCAAACAAAAGCGTGAAACTTACC
ATGATGCAGTACAGGGAGATGGAACTGAAGATGTTAAGCTAATGAAGAATAACTTCAAATTT
TCTTACGGTTTTAATGCCAGAAATGAAATACTCACTCGCAGAGACATTCAGGGAGTTTACTGT
TCTCTCCAGTTGGTTTTCTGCTGTGCAAGTTAATGTTACATTCTCCTCAGGGGAAATGATAATT
TTACTATAATACCTGCCATTAATATAAGGAGACTCCTCTGTCTGGAAGAAGAAGGAAGAAGA
GTTTAGAGATGGAGGTGCACTTTACAAGTCTGGGAAAGCATAGTGGTAGNNNNNNCAGGGCA
A

Sanger sequence for sample 36-R-chr16_52269942-52271542_R:
NNNNNNNNNNNNNAAAGTGCNCCTCCNTCTCTNNNCTCTTCTTCCTTCTTCTTCCAGACAGAG
GAGTCTCCTTATATTAATGGCAGGTATTATAGTAAAATTATCATTTCCCCTGAGGAGAATGTA
ACATTAACTTGCACAGCAGAAAACCAACTGGAGAGAACAGTAAACTCCCTGAATGTCTCTGC
GAGTGAGTATTTCATTTCTGGCATTAAAACCGTAAGAAAATTTGAAGTTATTCTTCATTAGCTT
AACATCTTCAGTTCCATCTCCCTGTACTGCATCATGGTAAGTTTCACGCTTTTGTTTGGCATTTT
GATATCATCGCCATGAGGTTCCGTTTTCTTTTTACCTGGTTTCCATTCTTCTACTACCCTTGTC
TTCACAGTTTAAGTCACTTCATATATTAGCACTCAGACACTTTTCCCAGGAGAGGAAAN

Sanger sequence for sample 37-F-chr16_94468034-94469634_3_F:
NNNNNNNNNANNNNNNNCTAATTTTTAAGTGCTTATGATTCGTTATTCTACATTTTGGTAATC
AATTATAATATACTTTTTAAATTATTATTGGCTATTTTTTTTTTTTACCCGGGGAAAATGCTGTG
TGATCACGAGGATCCCNCACTTGNNCCTTACCCTAAGGTTACCATGCTGCNCTCATGAGTNNT
CCTAGCTTGGAAGTGCACGTNNCTGNTAGGGCTGAAAACAGGACATAGGGCCCAGCGCAGGG
CACCCACCCTCACCCTCCTTCCTTATCCTTCCACCAAAAAAAAAAATGTACCACAGCCCTGATGT
GTTGGGGTTAACATTGTTCTGTTTGACTTTTGTGTAAAGTAATGCATGCAATCTTGTAATGGGG
CCTGAAAACAAGCTAACTGTATTAAAAACTATTCAAAAACTAGGTATTTTTAGTGACCTGTAG
GGAGTGNCAAATACAAACATGTGAACCTTGAATACATTACATTTCTCTCAAACACAAAAAAA
CAAACAAAAAACTTTTCCAGCGTGTTCTCTGCTGCCCTGGANATGCCGTCACCTTCCACCCGT
GTCTGTGTGAGTTGTGGTGTCTTAGTGACTTAGTATGTANCTCACTCNTTATTCTGAAACATTG
TCTGNGGGTGNNCNNNGN

Sanger sequence for sample 37-R-chr16_94468034-94469634_3_R:
NNNNNNNNNNNNNNNNNNNCTNNNTCACTANNACNCCNCAACTCACACAGACACGGGTGGA
AGGTGACGGCATCTCCAGGGCAGCANAGAACACGCTGGAAAAGTTTTTAGTTTGTTTCTTTGT
GTTTGAGAGAAATGTAATGTATTCAAGGTTCACATGTCTGTATTTGCCACTCCCTACAGGTCAC
TAAAAATACCTANTTTTTGAATAGTTTCTAATACAATTAGCTTGTTTTCAGGCCACATTACAAG
ATTGCATGCATTACTTTACACAAAAGTCAAACAGAACAATGTTAACACCAACACATCAGGACT
GTGGTACATTTCTCTTCTGGTGGAAGGATAAGGAAGGAGGGTGAGGGTGGCTGCCCTGCGCTG
GGCACTATGTCCTGTTTTCAGCACTAGCCAGCTGTGCACTTCCAAGCACTGCTGACTCATGAG
ACTTACATGGTAACCTTAGTGTAAGTGCCAAGTGTGGGACTCCTGTGAGCAGACAGCATCTTC
CCAGGGTAAAAAAAAAAAAATANCCAATACTAATTTAAAAAGTATTTTATAATTGATTACCAA
AATGTANATTAATTGAATCATAAGCACTTTAAAAATTTATGACGGGGATGCTACAGTGCCACA
TGAGGAACTGAGANGGNNNNNNNNNNNNN

Sanger sequence for sample 38-F-chr19_60770223-60771823_F:

NNNNNNNNNNNNGNNNNNNNNCTNCNCNNGNGTNNGTAANGATCTCTCCGGGTGTCCCTGA
AGGTTGNCTCCTGCACTTGAATTTTGCAGTAAATAAATCTTTAACACACAAAAAAAACTTAGG
GAAAAAAAAAAGACCCTANAGGATTATATCTGGTATGCCTGAAATTGCACCCTGCATTGCCCC
CAAAANCCCNCNCTTNAAAAGGATCGAAAGGAAGANCTTTNATCTTTGAAANNCTTGAAAAG
ATTAAAATATCTCCAAACTCTTCGACACTACCTTCGANTGNTTACCTGTGAACCTTTAACCTGT
GTTAAGGGAAAATCCTCAAAAAAAAACTAAAAGTTCACAAATTAAACTTTTCCCTAAACAAA
ACAAACAAAAAAAAACCCCATAGTGAGCTTACCTCGTAAACAAACTGCCGGGCAGCTTTGAG
GCGCATCACAAATAAATCTCTGTCTTCCAACATTCNTGACATCCTATTCTTATGCTCCAAAGCT
TTTTCTCGTTCTANCTGCATTGTANTGATCTGAAAGTACACAGCACTAGGTAATCTCCAAAGC
ATCACTCCAGGAAAATGCTCACCTTCCTTACACTGANAAGCTTCTCCTTCNNGNNTTTTTTNNA
GN

Sanger sequence for sample 38-R-chr19_60770223-60771823_R:
NNNNNNNNNNNNCTGANGNNAGCANCTTNCATGNCACTGGTGNANGACACNGTTCNCAGAG
GCTGTGTCTGGGAAAGGANGACNAGACATTCNGAAGAAGAAATGCTCGGGAACATCTTAATA
GGATGTCACAAATGTTGTAAGACAAAAATCTATTGGAGAGGCGCCTCAAAGCTGCCCGGCAG
TCTGTCTCNNTTCCTGCTCACTATGGGGTCTTTTTTTGTTTGTTTTGTTTAGGGAAAAGTTTAAT
TTGTGAACTTCTANTTTTCTTTTGAGGATTTTCCCTTAACACNGGTTAAAAGTTAACAGGTAAT
TACTCAGTATTTGTATAGATGAGTTGGGAGATATTTTAATCTTTTCAAGTGTTTCAAAGATGAA
AGTTTATTCTTTTTGATCTTTTTGAAGTGTGGGATTTTGGGGGCAGTGCAGGATGGAATTTCAG
GCCTACCAGATAGCTACCTCTAGAGTCTTTTTTTTTTCCCTAAGTTTTTTTTCTTTTTTAAAGAT
TTATTTATTATTATATGTAAGTGCATTGTAACTGTCTTCAAACACACCAAAAAAGGTGGTTAA
ATATCATTACANATGGTTGTGAGCCACCGTGTGNTTGCTGGGATTTGAACTCNNNNNTTTCGG
GAGN

Sanger sequence for sample 40-F-chrX_101403719-101405319_6_F:
NNNNNNNNNNNNNNNAGGGAGCACCCNGTTCTTTCCCTGTTGGCTTTGCTGTTCCCCAGCCTT
CTTTTTGTGTTTTTATAACTGTCCTCAGTTTAGCCACTGTTAAAATGTATATATTGTACTGAGGT
GCCTGGCCTGTTCCTTCAGTGAGCCATGCCCACCCTTGTGTTGTAGTGAGAAACTGTTGTCACA
ACTAACTTGTCTCTGGAATTGTTTCAAATAAAGAGTTAAAATTGTTCTTTGCTTTCTCTGGGGG
AGGTAGAGCTGGCGTTGAAGAGTGGAAGAGAAGAGAAAGAGCACCCACTGTGGGTCCCTGA
AGATTAGTCTTCCCTCAGTCAATGAATATCACAACGTTGGTCCTCTTCTCATACATTTTCAGAT
ACATCAGAAAAAAATATTTTTCAATAGCCATTTATTGAGCTAGAGTTGCTTATGTCTATAATCC
CAGTAGTTGGGAGGTGGAGGCAGGAAGTTAGGAATTCAGTCATCCTTGGCTACATGTGGAGTT
CAGAGCCAACCTAGGTTATGTGAAACCCTGTCACAAAATGGGGACAGGNNNGNAANAAATGT
AAGAA

Sanger sequence for sample 40-R-chrX_101403719-101405319_6_R:
NNNNNNGNNNNNNNNNNCTAGGTTGGCTCTGAACTCCACATGTAGCCAAGGATGACTGAATT
CCTAACTTCCTGCCTCCACCTCCCAACTACTGGGATTATAGACATAAGCAACTCTAGCTCAAT
AAATGGCTATTGAAAAATATTTTTTTCTGATGTATCTGAAAATGTATGAGAAGAGGACCAACG
TTGTGATATTCATTGACTGAGGGAAGACTAATCTTCAGGGACCCACAGTGGGTGCTCTTTCTCT
TCTCTTCCACTCTTCAACGCCAGCTCTACCTCCCCCAGAGAAAGCAAAGAACAATTTTAACTCT
TTATTTGAAACAATTCCAGAGACAAGTTAGTTGTGACAACAGTTTCTCACTACAACACAAGGG
TGGGCATGGCTCACTGAAGGAACAGGCCAGGCACCTCAGTACAATATATACATTTTAACAGTG
GCTAAACTGAGGACAGTTATAAAAACACAAAAAGAAGGCTGGGGAACAGCAAAGCCAACAG

GGAAAGAACTGGGTGCTCCCTCTGACAGACACTAACCTTTCTGGGCCCACAANNGNNNNNTG
GGGGAA