Structural Variants Lab Report

By MSGDA Cohort
One Team, One Dream

Introduction

Single nucleotide polymorphism (SNPs) are believed to be to the main contributors to genetic and phenotypic human variation. The advent of genome scanning technologies and commercial hybridization tools such as microarray-based comparative genomic hybridization (aCGH) and multiplex ligation-dependent probe amplification (MLPA) has allowed research into structural variants (SV). SV analysis has shown that these genomic variations are the drivers of human diversity and disease susceptibility. [2] Detection and proper classification of these SV pathologies are important for clinicians to make proper decision.

Structural Variation. Structural variants are DNA segments which range from single base changes to large chromosomal-level variation. These variants result in insertions or deletions (indels), duplications, inversions, and translations. SV affects large continuous regions of the genome that are at least 50 bp long. Because of the large changes caused, SV can distinctly affect phenotype in many ways, such as modification of open reading frames, production of alternatively spliced mRNA, alterations of transcription factor binding sites, and changes in chromatin structure. For this reason, SV has implications in many different diseases, including chromosomal microdeletion disorders, complex multifactorial diseases like autism and schizophrenia, and cancer. [4,5,6]

Cell Line. In order to identify SVs, this experiment used the MCF-7 cell line. These are ER-positive, progesterone-receptor positive, and moderately EGFR-expressing cells with no HER2 amplification that were originally obtained from the chest wall of a metastatic breast cancer patient in the early 1970s. This line was created in 1973 by Dr. Soule at the Michigan Cancer Foundation and it continues to be a commonly used breast cancer cell line today. [1,3] Because of their decades of use, MCF-7 cells are well characterized in the literature through 25,000 experiments and written reports, which make them an advantageous cell line for researchers and important to clinical diagnostics. [3] This is exemplified in MCF-7's use in cancer care and research. These cells have provided more knowledge than any other breast cancer cell line and continue to be used to make discoveries for clinical care applications worldwide. [1,3]

The MCF-7 cell line used in this experiment has a unique lentiviral insertion within it. This insertion works as a fluorescent reporter for the Ki67 promoter and signifies cell cycle entry and exit points within each cell. [7] For our experimental purposes we will not be looking at fluorescence within single cells, though we do expect to see this insertion to appear in our analysis of large structural variants for our data.

Workflow. The objective of the lab was to find SVs, which requires a variant call format (VCF) file to perform downstream analysis. The workflow begins with extracting the DNA, then using Illumina's rapid sequencing kit to fragment the DNA. We then conducted nanopore sequencing using a MinION FLO-MIN106 flow cells for the base-calling. We used minimap2 to generate a SAM file that was aligned to the reference genome GRCh38 and used SAMtools to generate the corresponding BAM file. This BAM file was converted to a VCF file using pre-established bioinformatics tools.

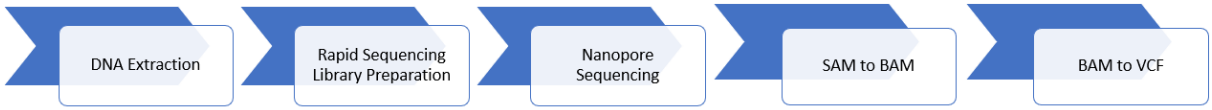

Figure 1. Workflow Diagram. First, we completed DNA extraction using MCF-7 cell line. Our quality check was conducted using Qubit and NanoQuant. Then we performed rapid sequencing library preparation using MinION library preparation kit from Illumina. Then we conducted DNA analysis using minimap2 to align DNA sequence to the reference genome and generate a SAM file. This file was converted to a BAM file using SAMtools and the BAM output was used variant calling using Sniffles. Analysis and visualization used IGV, SVRibbon, BLASTn, and UCSC Genome Browser.

Methods and Materials

DNA Isolation
MCF-7 breast cancer cells were collected for DNA isolation. In a tube, 72 mL of C1 buffer and 217.5 mL of ice-cold distilled water was added to $7.25x10^5$ of MCF-7 cells. After gentle mixing, the tube was incubated for 10 minutes on ice and spun in a centrifuge for 15 minutes at 4°C at 1300 x *g*. The supernatant was extracted, and 250 µL of ice-cold C1 buffer and 750 µL of ice-cold distilled water were added. This mixture was resuspended and vortexed again to form a pellet for another 5 minutes at 4°C at 1300 x *g*. The supernatant was discarded again. 1 mL of G2 buffer was added, and the pellet was resuspended on a vortex. 25 µL of Proteinase K stock solution was added and the solution was incubated at 50°C for 30 minutes.

Using Genomic-tips, isolation of the DNA was used to check the quality of DNA with Nanoquant and Qubit. With a 100/G Genomic-tip, we added 1 ml of QBT buffer into the tip and allowed the buffer to flow through by gravity. The pellet was resuspended and flowed through the Genomic-tip when added in. Three measurements of 1 mL of QC buffer (equally 3 mL of QC buffer total) was added to the Genomic-tip and allowed to flow through by gravity. Finally, the 2 aliquots of 1 mL of QF buffer (equaling 2 mL of QF buffer total) was used to elute DNA from the resin in the Genomic-tip. Collect this DNA in a collection tube and add 1.4 mL of room-temperature isopropanol. Centrifuge this mixture for 15 minutes at 4°C at 5000 x *g* to precipitate the DNA. Add 1 mL of cold 70% ethanol to the DNA pellet and centrifuge for 10 minutes at 4°C at 5000 x *g*. Remove the supernatant and add 2 mL of TE buffer at pH 8.0 to resuspend the DNA. [8]

DNA Quality Check with Nanoquant and Qubit
NanoQuant Plate
The NanoQuant Plate by Tecan is a laboratory instrument intended to measure nucleic acid concentration in small volumes of liquid using an Infinite multimode reader. The plate includes 16 slots for 1ul-1ul droplets of sample. The plate was first blanked with distilled water on four wells. Once blanked each well was used to measure each of the 4 samples prepared. 25 flashes of light were used to calculate nucleic acid concentration in ng/ul along with 260/280 ratios. [9] The concentrations are shown in Table 1.

Qubit
4 samples and 2 standards were used to measure nucleic acid concentration on the Qubit instrument. Six 0.5-mL tubes were labeled and prepared. To create the working buffer 7 µL of the Qubit dsDNA 1:200 HS reagent, and 1393 µL of the Qubit dsDNA HS buffer were mixed together. In each of the standard tubes, 190 µL of working solution and 10 µL of the appropriate Qubit standard was added to each respective tube. In each sample tube, 199 µL of working

solution and 1 µL of sample was added. Samples were first diluted using a 1:10 ratio before mixing with the prepared working solution. All tubes were gently mixed, avoiding bubbles. These tubes were incubated at room temperature for 2 minutes. Tubes were inserted in the instrument and read by the Qubit 3.0 Fluorometer, starting with 2 standards, followed by 4 samples. [10] The concentrations of the samples are shown in Table 1.

MinION Library Preparation/Sequencing
Sample 1 has a 260/280 ratio of 1.79, which is closest to our desired 1.80. In a Lo-Bind vial, we combined 0.96 µL of sample 1, 6.54 µL of water, and 2.5 µL of FRA (fragmentation buffer) for a total of 10 µL. This was then transferred into a PCR tube. This solution was then mixed and incubated at 30°C for 2 minutes on a heat block and then for 80°C on a heat block. 1 µL of RAP was then added to the tube, and the tube was incubated at room temperature for five minutes after the solution was gently mixed.

We used sample 4 because its ratio value of 1.87 is the second closest to the ideal ratio of 1.80. In a PCR tube, we combined 0.683 µL of sample 4, 6.817 µL of water, and 2.5 µL of FRA for a total of 10 µL. This solution was then mixed and incubated at 30°C for 2 minutes on a heat block and then for 80°C on a heat block. 1 µL of RAP was then added to the tube, and the tube was incubated at room temperature for five minutes after the solution was gently mixed. [11]

Flow Cell Priming
The flow cell priming mix was prepared by mixing 30 ul of Flush Tether with Flush Buffer. Next, 800 ul of the priming mix was loaded into the MinION SpotON flow cell (FLO-MIN106) via the priming port. Afterward, the loading sample was prepared by mixing 11 ul of the DNA library prep with 34 ul Sequencing Buffer, 25.5 ul Loading beads, 4.5 ul nuclease-free water. In the priming port, 200 ul of the priming mix was loaded into the flow cell followed by 75 ul of the sample loaded into the flow cell via the SpotON sample port in a dropwise fashion. [11]

Sequencing Run
The MinKNOW base calling platform was used to generate a FASTQ file. Following the sequencing run, the EPI2ME platform collected run statistics including cumulative throughput, duty time, read lengths, and local base calling.

Data analysis

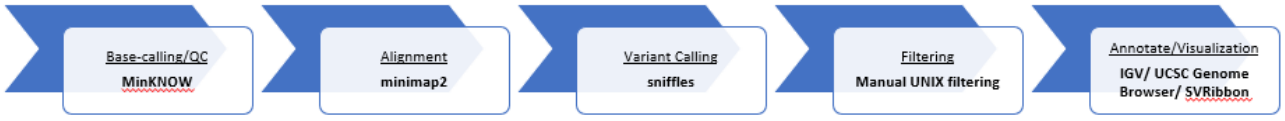

Figure 2: Workflow of data analysis following MinION sequencing
Sequencing reads with quality scores <7 were filtered out. We used minimap2 to map the DNA sequences to the human reference genome GRCh38. The output SAM file was then converted to a BAM file for variant calling using Sniffles. The variant calls were filtered to include only a read depth ≥ 3. A VCF file was generated and uploaded to Galaxy for further analysis. Through the Galaxy platform, the VCF file was filtered to include only read lengths with magnitude > 49. Annotations were exported from the UCSC Genome Browser as a BED file. The VCFannotate tool was used to annotate the VCF file and the final VCF file was visualized using either Integrated Genome Viewer, UCSC Genome Browser or SVRibbon (Fig 2). The following code allowed for the generation of the VCF file from fastq format (Fig 3-7). Software used were

minimap2 v2.16, SAMtools v1.4, Sniffles v1.011, and VCFannotate Galaxy version 1.0.0_rc1+galaxy0.

```
conda install -c bioconda minimap2
conda install -c bioconda/label/cf01901 minimap2
Reference genome download: http://hgdownload.cse.ucsc.edu/goldenPath/hg38/bigZips/hg38.fa.gz
gunzip hg38.fa.gz

minimap2 -d hg38.mmi hg38.fa
minimap2 -ax FAJ*.fastq hg38.fa > alignment.sam
```

Figure 3: Code for Mapping (minimap2)

```
samtools view -S -b alignment.sam > unsorted.bam
samtools sort unsorted.bam -o mapped.bam
samtools index mapped.bam
```

Figure 4: Code for Alignment (SAMtools)

```
cd Sniffles-master/
mkdir -p build/
cd build/
export PATH=$PATH:/usr/local/bin/gcc-8
export PATH=$PATH:/usr/local/bin/g++-8
export PATH=/usr/local/bin:$PATH
cmake -D CMAKE_C_COMPILER=/usr/local/bin/gcc-8 -D CMAKE_CXX_COMPILER=/usr/local/bin/g++-8 ..
make
cd ../bin/sniffles*
./sniffles

sniffles -t 5 -s 3 -r 2000 -q 20 -d 1000 —genotype -l 30 -m mapped.bam -v sniffles.vcf
```

Figure 5: Code for Variant Calling (sniffles)

```
awk '{ gsub(/SVLEN=/,"",$8); print }' sniffles.vcf > edited.vcf
sed 's/;/\'$'\t/g' edited.vcf > tabs.vcf
awk '$18 > 49 || $18 < -49 {print $0}' tabs.vcf > new2.vcf
```

Figure 6: Code for Filtering (manual UNIX)

```
vcfanno -ends -p 10 vcfanno_hg.conf sniffles.vcf > sniffles_annotated.vcf
```

Figure 7: Code for Annotations/Visualizations

Results

Samples	Sample Name	Concentration (ng/μL) from Nanoquant	260/280 Ratio	Concentration (ng/μL) from Qubit
1	RTV	550.9	1.79	38.4
2	LEIA	3	3	0.918
3	LEIA	366.9	1.9	49.8
4	RTV	252.9	1.87	58.6

Table 1. Concentration of DNA from Nanoquant and Qubit.

Before samples are ready for base-calling with the MinION SpotON flow cells using the MinKNOW software, the quality of the DNA and the concentrations need to be accurately assessed. Samples 1 and 4 were used in our experiment for SV analysis because their 260/280 ratios were the closest to 1.80 (Table 1). Concentrations of our samples were analyzed with Nanoquant and Qubit to determine if sufficient material was available for the sequencing runs.

Minion Sequencing Run Reports
Using the Minion Rapid Sequencing Kit, a total of 1.2 million reads (12.39 GB of total data) were analyzed and a base call metrics report was generated (Table 2). The run reports for Run 1 and Run 2 (corresponding with the two separate SpotON flow cell runs) estimated that the longest segment analyzed was in the range of 129.6 KB (Figure 8). The majority of reads analyzed were around 16 KB long with over 600 MB of data while the shortest reads were less than a 1,000 bases and made up around 200 MB of our total 12.39 GB of bases called (Table 2, Figure 8).

Base Calling Metrics	**Run 1**	**Run 2**	**Combined Runs**
Reads Analyzed	543,840	630,710	1,174,590
Estimated Bases	7.09 GB	6.8 GB	13.89 GB
Bases Called	6.06GB	6.33GB	12.39 GB
Skipped Reads	1	0	1

Table 2. Sequencing Base Calling Statistics

Base call metrics using the MinION Rapid Sequencing Kit (SQK-RAD004) produced a total of nearly 1.2 million reads and 12.39 GB of total data for bases called (Table 2). The most reads were generated from run 1 using sample 1 with one skipped read that did not meet the base call quality score (QC) required (Table 2).

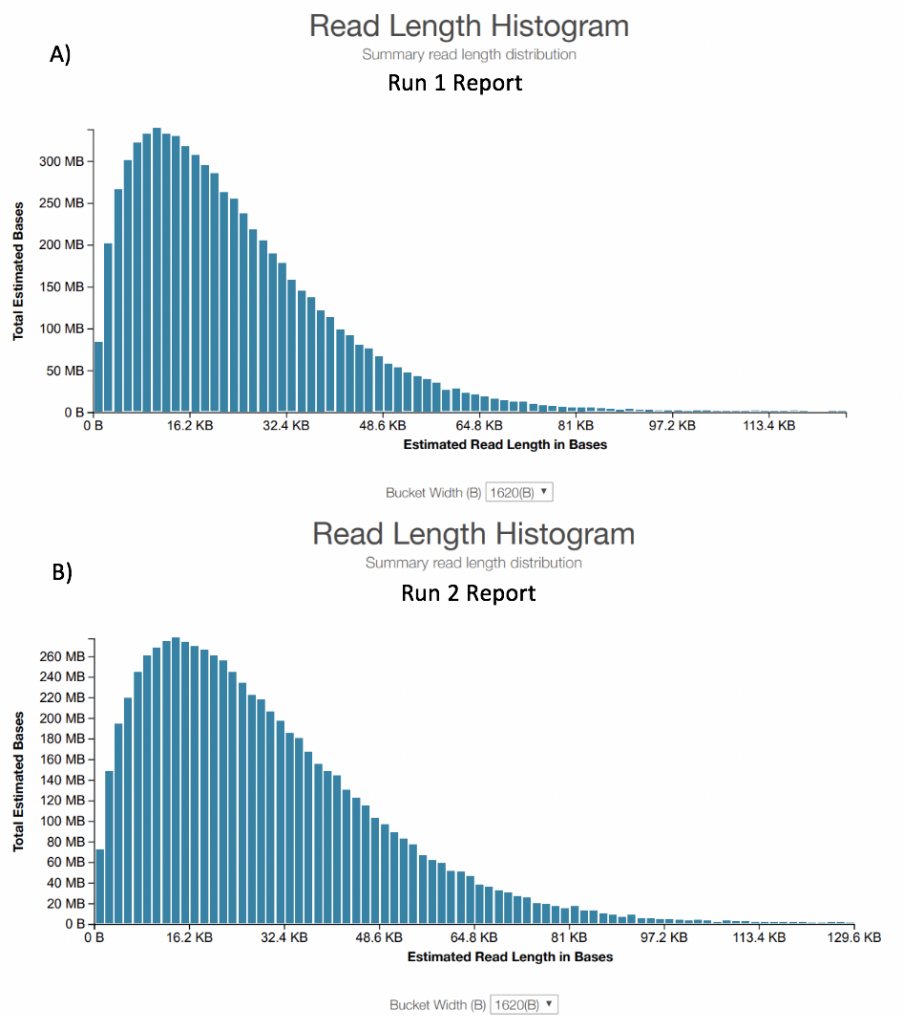

Figure 8: Sequencing Run Read Length

Figure 8 shows that the longest obtained segments in both Run 1 Report and Run 2 Report were around 129.6 KB. The majority of reads that were base-called in our two runs with separate SpotON flow cells were around 16 KB long with over 600 MB of data for those reads. The shortest reads were less than a 1,000 bases and made up around 200 MB of our total 12.39 GB (Fig 8).

EPI2ME Report Statistics

Of the 1.2 million reads analyzed by the minion, the EPI2ME real-time cloud-based analysis platform analyzed a total of 895,366 reads. The ~305, 634 bases not analyzed were filtered out by application of stringent quality control (QC) filters. The EPI2ME reads were aligned to the reference human genome (GRCh38). A total of 9.9 Gbp were aligned to the human genome at an average alignment accuracy of 89.2% (Table 3). The overall genome was sequenced and analyzed at ~2.3x coverage (Table 4). Individual Chromosome statistics generated by the EPI2ME analysis platform provided statistics including alignments, aligned yield, percent average accuracy, percent average ID, and average coverage. It was found that chromosome 3 had a total of 83,138 reads making it the chromosome with the highest number of aligned reads

yielding 970.5 M bp at a 89.5% accuracy rate. However, based on structural variant statistics chromosome 1 was found to have the highest number of insertions and deletions (Table 5, Figure 9). Chromosome 22 was found to have the lowest number of insertions and deletions. Structural variants including deletion-inversion (Del/Inv), duplication-inversions (Del/Inv) and other structural variants were also analyzed and are depicted in Figure 9b.

Report Statistics	**Values**
Reads Analyzed	**895,366**
Reads Aligned	**846,031**
Aligned Yield	**9.9 Gbp**
Average Alignment Accuracy	**89.2%**

Table 3: Epi2me Overall Report Statistics

We analyzed 895,366 reads from the nearly 1.2 million reads produced. 846,031 reads were aligned to the reference human genome (GRCh38), yielding 9.9 Gbp of aligned sequenced data. The average accuracy for the sequencing run was 89.2% (Table 3).

Chromosome	**Alignments**	**Aligned Yield**	**Average Accuracy (%)**	**Average ID (%)**	**Average Coverage**
chr1	67,559	804.4 M bp	89.2%	96.9%	3x
chr2	73,720	869.8 M bp	89.4%	97.0%	3x
chr3	83,138	970.5 M bp	89.5%	97.0%	4x
chr4	46,658	540.1 M bp	89.4%	96.9%	2x
chr5	48,742	568.5 M bp	89.5%	97.0%	3x
chr6	39,885	472.4 M bp	89.4%	96.9%	2x
chr7	55,182	648.8 M bp	89.3%	96.9%	4x
chr8	34,935	551.7 M bp	89.5%	97.0%	3x
chr9	30,497	361.9 M bp	89.2%	96.9%	2x
chr10	33,884	401.8 M bp	89.2%	96.9%	3x

chr11	29,103	346.3 M bp	89.2%	96.9%	2x
chr12	41,294	490.2 M bp	89.3%	96.9%	3x
chr13	20,743	238.5 Mbp	89.5%	97.0%	2x
chr14	27,778	327.1 M bp	89.3%	96.9%	3x
chr15	27,348	326.4 M bp	89.2%	96.9%	3x
chr16	25,941	310.4 M bp	88.3%	96.4%	3x
chr17	33,842	399.8 M bp	89.0%	96.9%	4x
chr18	15,068	177.5M bp	89.4%	97.0%	2x
chr19	12,320	146.4 M bp	88.5%	97%	2x
chr20	34,935	412.4 M bp	88.9%	96.8%	6x
chr21	10,542	124.8 M bp	88.9%	96.8%	2x
chr22	7,075	76.7 M bp	86.1%	94.7%	1x
chrX	28,169	334.5 M bp	89.5%	97.0%	2x
chrY	170	1.7 M bp	78.8%	90.9%	0x

Table 4. Epi2me Report Statistics per Chromosome

The statistical analysis metrics produced from using the EPI2ME platform aligned the sequence read data using the Fastq Human Alignment to the reference human genome GRCh38. The overall average coverage for the sequencing run was at 2.6x. Chromosome 3 had the highest number of aligned reads (83,138 reads) yielding 970.5 M bp at a 89.5% accuracy rate. Chromosome Y had the lowest number of aligned reads (170 reads) yielding 1.7 M bp at a 78.8% accuracy rate.

	Structural Variants						
Chromosome	Insertions	Deletions	Inversions	Duplications	Del/Inv	Inv/Dup	Other
chr1	252	277	12	9	-	-	30
chr2	211	255	3	3	-	-	2
chr3	219	272	9	9	-	-	17
chr4	127	139	2	4	-	-	19
chr5	116	171	1	6	-	-	-

chr6	127	127	2	2	-	-	-
chr7	217	246	5	5	-	-	-
chr8	142	150	-	3	-	-	3
chr9	129	107	1	1	-	-	1
chr10	118	161	1	5	1	-	4
chr11	107	91	2	4	-	-	1
chr12	140	137	3	1	-	-	1
chr13	70	73	-	1	-	-	-
chr14	110	129	1	7	-	-	21
chr15	95	97	-	-	-	2	4
chr16	94	131	6	5	-	-	2
chr17	94	163	7	5	-	-	34
chr18	68	61	1	1	3	-	-
chr19	76	73	1	1	-	-	-
chr20	185	232	23	6	1	1	26
chr21	79	99	3	1	-	-	15
chr22	35	29	1	4	-	-	20
chrX	62	65	-	1	-	-	-
chrY	17	16	-	-	-	-	-
chrM	-	2	-	1	-	-	-
chrUn	48	61	3	5	-	-	16

Table 5. Structural Variants per Chromosome Statistics

Over 6,703 structural variants were detected from the two sequencing runs of the MCF-7 cell line (Table 5). The most prevalent type of structural variant identified throughout the genome were insertions and deletions (94%) with over 4 times the amount of insertions/deletions to inversions/duplications. The highest number of a single structural variant for a given chromosome was 277 deletions on chromosome 1. There were over 200 structural variants that did not map to a specific subtype and about 115 structural variants not mapped to a chromosome. Over 120 reads mapped to the X-chromosome but around 30 reads mapped to the Y-chromosome.

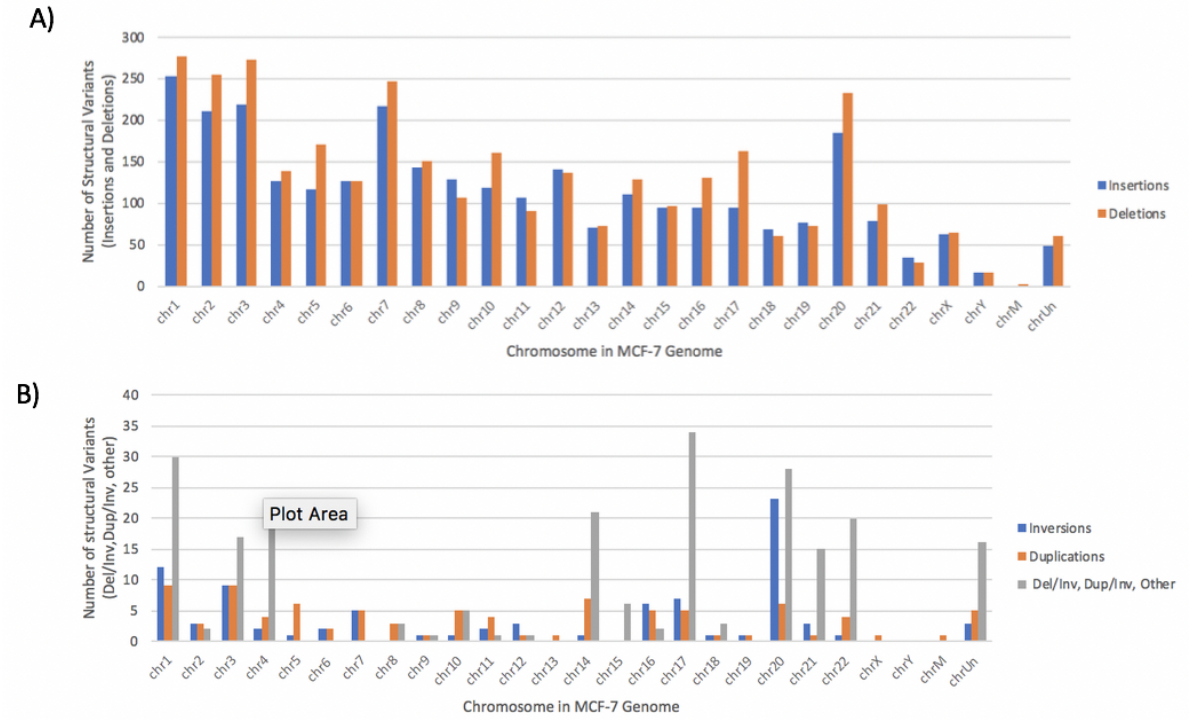

Figure 9. Summary of Structural Variants per Chromosome

The bar chart above graphical represents the total number of structural variants that were detected for each of the individual chromosomes. These SVs include inversions, duplications, insertions, deletions, and some more complex subtypes such as Deletion/Inversions, Duplication/Inversions, amongst others. The insertions and deletions had an overall higher quantity of variants whereas much less was identified for inversions, duplications, and the other complex subtypes (Fig 9). Chromosomes 1, 2, and 3 had the highest number of total SVs with very minimal variants detected for the Y chromosome and mitochondrial chromosomes.

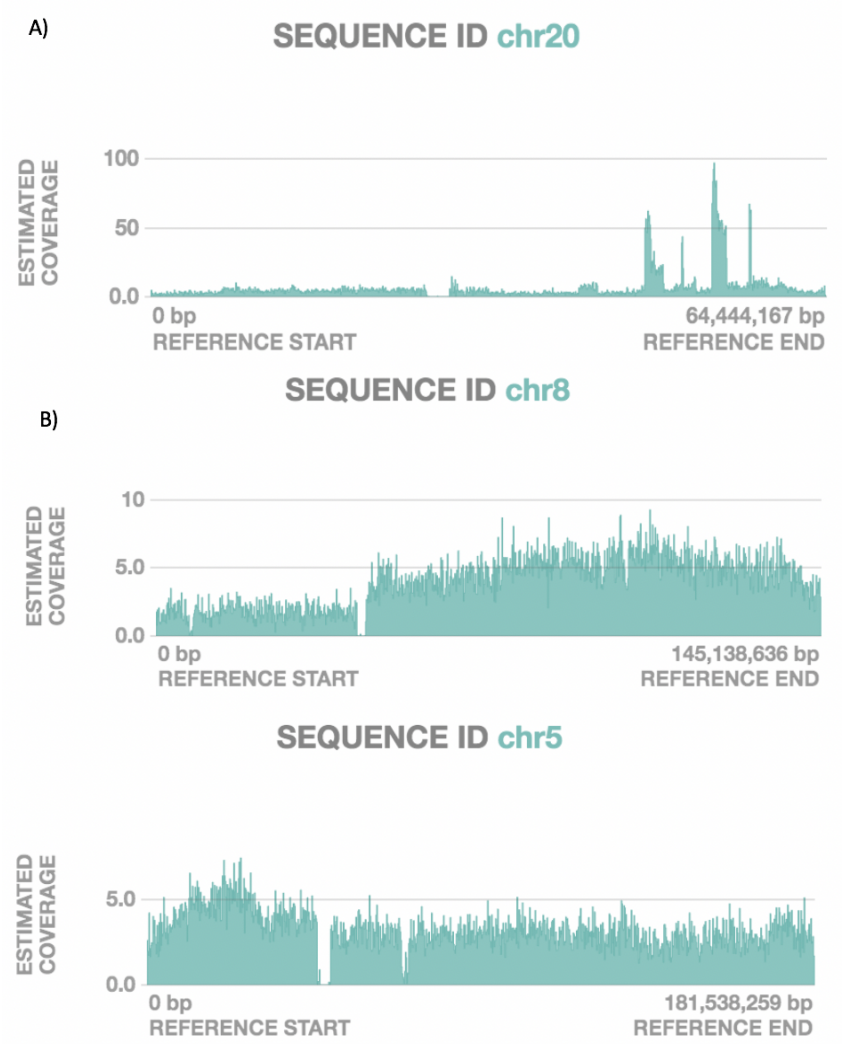

Figure 10: EPI2ME Sequence ID Chromosome 20, 8, and 5

The EPI2ME analysis software estimated coverage identified some peaks in the short arm (q) of the chromosome which had above average coverage. The peaks seen have an estimated coverage ranging from 9x-84x which can be indicative of regions where there may be a duplication event or other structural variants. Overview of overall coverage as seen in chromosome 8 and 5 are in accordance with coverage analysis in previous studies. The low coverage in the long arm (p) and high coverage in the short arm(q) of chromosome 8 as well as the observed high coverage in the long arm and low coverage in the short arm of chromosome 5 were also observed by Hampton et al. [12]

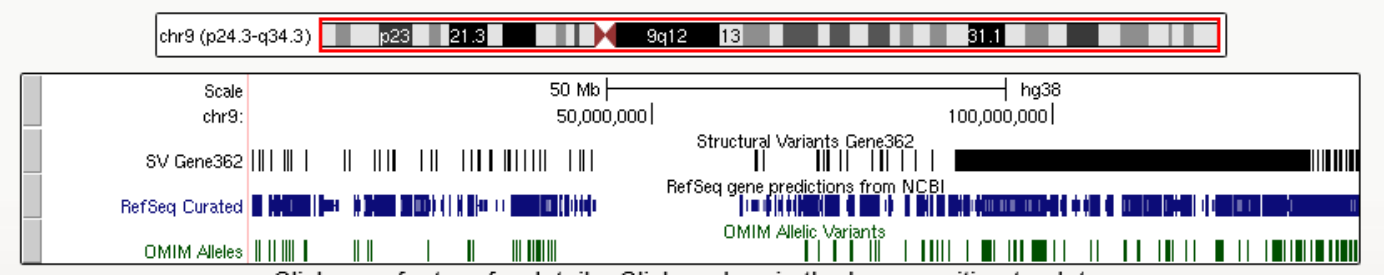

Figure 11: Visualization of Structural Variants at Chromosome 9 using UCSC Genome Browser

The annotated VCF file was uploaded to UCSC Genome Browser to create the custom track 'SV Gene 362' which can be used to view structural variants, such as in chromosome 9 as illustrated above (Fig 11).

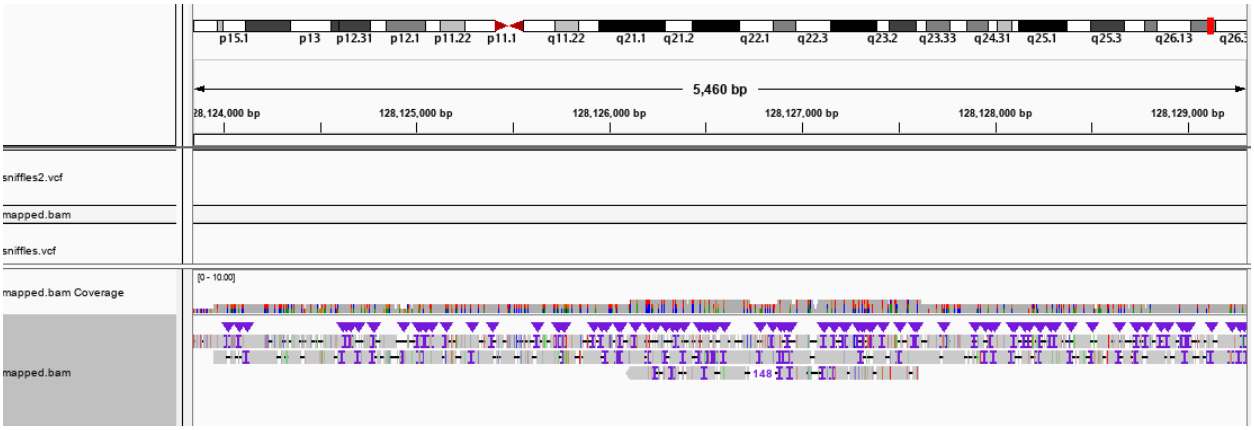

Figure 12: IGV Browser Window of Ki67 lentiviral gene insertion

The IGV browser window of the Ki67 cassette region demonstrates alignments to several regions of the genome outside of chromosome 2 and some regions that do not map to anywhere on the human genome (Fig 12).

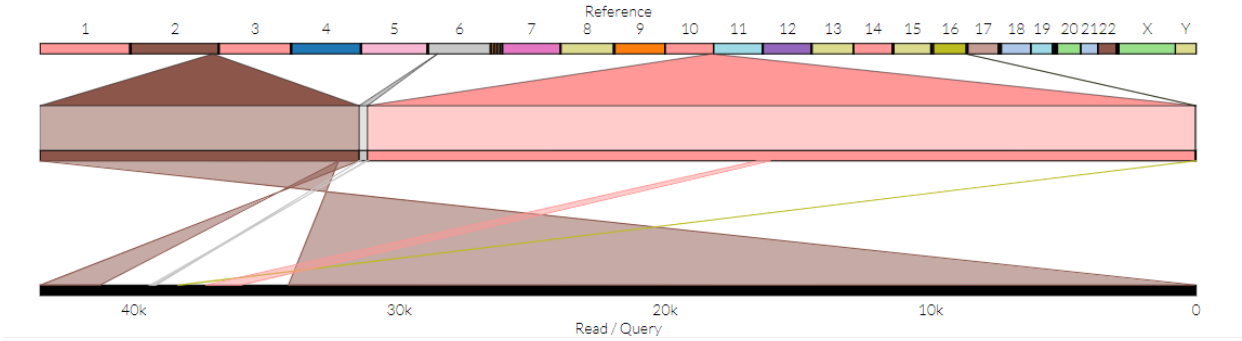

Figure 13: Visualization of insertion of Ki67 gene using Ribbon

The lentiviral insertion of Ki67 demonstrates mapping of the read to chromosome 2 (brown), 6 (grey), 10 (pink), and 17 (tan->olive green) across the genome and some regions are shown to be inverted (Fig 13). A cassette region appears in the sequence which does not align to the reference genome.

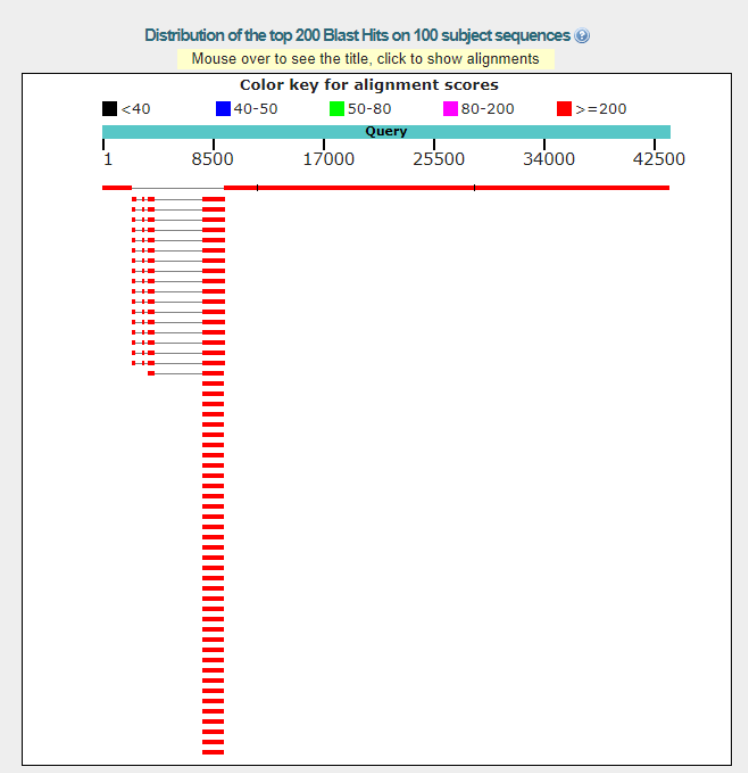

Figure 14: BLASTn

Regions of similarity were observed to be mostly in chromosome 2 and some in chromosome 10 indicated by the long region of aligned reads in the 8500 mapped regions (Fig 14). We also have genomic sequence from the lentivirus and cloning vectors seen with fewer aligned reads in the region to the left of the 8500 mapped regions.

The filtered VCF file was observed using various visualization tools. A custom track was created in UCSC Genome Browser to observe the chromosome location where the structural variants occur (Fig. 1). The MCF-7 cell line was transfected with a Ki67 lentiviral insertion which was mapped onto Integrative Genomics Viewer and Ribbon (Fig. 12,13). As illustrated, the regions of the cassette were aligned to chromosome 2 and chromosome 10 sequences. There were also regions that did not align to anywhere on the reference sequence, suggesting that these regions are lentiviral DNA. These results were later confirmed by obtaining the DNA sequence from Genome Ribbon and using the alignment tool BLASTn (Fig. 14).

Discussion

Analysis of MCF-7 revealed 6703 structural variants. Despite being a relatively small number, it is close to the range expected of human breast adenocarcinoma cells originally isolated in 1970. Cancerous tissue has properties of genomic instability and genome-wide alterations were sufficiently detectable in our analysis with less than 5x coverage. Despite having low coverage per chromosome (due to the amount or reads collected from sequencing Epi2Me analysis) overall coverage for the p and q chromosome arms was in accordance with data from a circos plot visualization of MCF-7 by Hampton et al (Fig 10).[12] The larger chromosomes 1, 2 and 3 are expected to have the greatest number of structural variants and our data confirmed that chromosome 1 was found to have the highest total number of variants. [13]

The MCF-7 cell line is well-characterized to have large-scale genomic aberrations from normal cells and our SV analysis confirmed this phenomenon. A Ki67 lentiviral insertion characteristic of the MCF-7 cell line was detected on chromosome 2 and shown to align to various regions of the genome. However, some regions of the Ki67 lentiviral gene did not map to any region on the MCF-7 genome, which is expected of a lentiviral gene that does not naturally occur in human genomes. Another explanation for the unaligned regions of in the Ki67 lentiviral gene could be due to the GFP reporter region inserted in the lentiviral gene. Overall, our mapped reads demonstrated similar results regarding cancer genome structural variant totals to prior studies examining the MCF-7 and other cancer cell lines. [14]

For many of our reads, the structural variants mapped to multiple regions of the genome instead of mapping to individual chromosomal locations. Some of the complex variants detected in the genome were comprised of duplication-inversions and deletions-inversions. The low number of inversions/duplications relative to insertions/deletions and the detection of multiple alignments per read may have resulted from the low coverage of our samples produced with only two flow cells. Similarly, low read coverage and alignments to multiple genomic regions may have led to small number of reads that were mapped to the Y-chromosome for cells that were extracted from a female patient. Low read depth may have hindered sufficient quality calling of these structural variant subtypes and led to over 25% of our reads being skipped during filtering. Less stringent filtering standards and extra flow cells for increased throughput in further studies could generate a higher total number of structural variants expected of MCF-7 cell lines for further analysis. Though, samples 1 and 4 were chosen for sequencing and analysis because of their 260/280 ratios closest to 1.8, demonstrating minimal RNA contamination, sample 3 could have been used if extra flow cells were available.

In conclusion, the identified structural variants can provide an insight into the underlying mechanisms in ER positive breast cancer. The general profile includes ratios of structural variants types present in the MCF-7 cell line and chromosomal locations. Identifying the chromosomal location of the Ki67 lentiviral insertion is of importance because the inserted lentivirus contains a GFP genetic based reporter. The GFP reporter is used to fluorescently label and identify the cell in cell cycle transition of the MCF-7 breast cancer cell line. [7] Thereby, providing further understanding as to why the use of the MCF-7 cell line is for breast cancer research. Future, studies could identify other structural variants found in the MCF-7 genes we identified via long read sequencing.

References

[1] Ş. Comşa, A. M. Cîmpean, and M. Raica, “The Story of MCF-7 Breast Cancer Cell Line: 40 years of Experience in Research,” *Anticancer Res*, vol. 35, no. 6, pp. 3147–3154, Jun. 2015.
[2] L. Feuk, A. R. Carson, and S. W. Scherer, “Structural variation in the human genome,” *Nat. Rev. Genet.*, vol. 7, no. 2, pp. 85–97, Feb. 2006.
[3] A. V. Lee, S. Oesterreich, and N. E. Davidson, “MCF-7 Cells—Changing the Course of Breast Cancer Research and Care for 45 Years,” *J Natl Cancer Inst*, vol. 107, no. 7, Jul. 2015.
[4] A. Sanchis-Juan *et al.*, “Complex structural variants in Mendelian disorders: identification and breakpoint resolution using short- and long-read genome sequencing,” *Genome Med*, vol. 10, Dec. 2018.
[5] P. Stankiewicz and J. R. Lupski, “Structural Variation in the Human Genome and its Role in Disease,” *Annual Review of Medicine*, vol. 61, no. 1, pp. 437–455, 2010.
[6] K. Yi and Y. S. Ju, “Patterns and mechanisms of structural variations in human cancer,” *Experimental & Molecular Medicine*, vol. 50, no. 8, p. 98, Aug. 2018.
[7] A. C. Zambon, “Use of the Ki67 promoter to label cell cycle entry in living cells,” *Cytometry A*, vol. 77, no. 6, pp. 564–570, Jun. 2010.
[8] Qiagen. (2015). *Qiagen Genomic DNA Handbook* [PDF file]. Retrieved from https://sakai.claremont.edu/access/content/group/34bb141f-10a2-4ca2-b8ab-cc8662220cd6/Nanopore/HB-1163-002-1095823-HB%20Genomic%20DNA%200615%20WW.pdf.
[9] Tecan. (2011). *NanoQuant Plate - Tecan* [PDF file]. Retrieved from https://ww3.tecan.com/mandant/files/doc/219/NanoQuant_FAQ_Sheet_LayV1.pdf
[10] Molecular Probes Life Technologies. (2015). *Qubit dsDNA HS Assay Kits* [PDF file]. Retrieved from https://sakai.claremont.edu/access/content/group/34bb141f-10a2-4ca2-b8ab-cc8662220cd6/Nanopore/Qubit_dsDNA_HS_Assay_UG.pdf
[11] Nanopore Protocol. (2017). *Rapid Sequencing (SQK-RAD004)* [PDF file]. Retrieved from https://sakai.claremont.edu/access/content/group/34bb141f-10a2-4ca2-b8ab-cc8662220cd6/Nanopore/rapid-sequencing-sqk-rad004-RSE_9046_v1_revC_17Nov2017.pdf.
[12] “A sequence-level map of chromosomal breakpoints in the MCF-7 breast cancer cell line yields insights into the evolution of a cancer genome.” [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2652200/. [Accessed: 25-Apr-2019].
[13] Eisfeldt J, Pettersson M, Vezzi F, Wincent J, et al, “Comprehensive structural variation genome map of individuals carrying complex chromosomal rearrangements,” *PLoS Genetics*. Vol 15, no. 2, e1007858
[14] A. M. Hillmer, F. Yao, K. Inaki, W. H. Lee et al, “Comprehensive long-span paired-end-tag mapping reveals characteristic patterns of structural variations in epithelial cancer genomes,” *Genome Res*. vol. 21, no. 5, pp. 665-675.