Human Genome Sequencing with Illumina and Nanopore to Identify Single Nucleotide Polymorphisms (SNPs)

Zan Chaudhry Partners: Andres Parra & Anjan Singh
Methods in Nucleic Acid Sequencing Lab, Spring 2023 10:30 AM

Introduction

With the fall in nucleic acid sequencing costs and the rise in sequencer availability, sequencing technology has become a widespread and important tool for clinical medicine. Human genome sequencing presents an opportunity to isolate the genetic determinants of disease, guiding innovative therapeutic efforts to target the source rather than the symptoms, such as *in vivo* gene editing, and providing important information on predispositions to develop certain diseases (i.e. cancer gene screenings) so that patients and clinicians are on alert and can catch these conditions in the early stages. However, sequencing a whole genome to search for these disease markers/variants is unnecessary, costly, time-consuming, and the excess off-target DNA dilutes the signal from the desired sequences.

Hence, approaches to select for particular sequences of interest have been developed alongside sequencing technology. One common mechanism involves the design of oligonucleotide probes that are complementary to desired sequences. These probes can be designed to include an anchor point (such as attachment to beads) to remove the desired sequences from solution, the remaining excess DNA can be discarded, and the enriched target DNA can be amplified for sequencing.[1] Another specialized method for sequence selection is adaptive sampling, which is specific to the Nanopore method. Nanopore sequencing involves passing a strand of DNA through a proteinous pore, leading to the creation of an ionic current signal particular to each nucleotide. In adaptive sampling, the sequence of the strand being passed through the pore is compared to the target, and if it does not match, then a reversed current is applied to eject the strand out of the pore.[2]

The following report details a human genome sequencing experiment from a line of human cells, utilizing both Illumina and Nanopore methods, involving the enrichment of extracted DNA for relevant loci from the Illumina exome panel (using oligonucleotide probes for Illumina and adaptive sampling for Nanopore), and analyzing the data for the single nucleotide polymorphism (SNPs) present. SNPs are genetic variants involving a change in the identity of one nucleotide (such as a G to a C) along a particular sequence. Thus, SNPs correspond to alleles of a particular sequence and can produce different phenotypes, such as the difference between a normal and diseased state. In the presented data, 47,161 SNPs were identified by Illumina sequencing. Nanopore sequencing failed variant calling, so it was excluded from further analysis. Within the variants, 13 were identified as clinically relevant, including a variant of the gene ITPKB, which has been implicated in certain blood cancers. This variant was chosen for further investigation and is discussed in detail due to the gene's high probability (1.0) of being intolerant of a loss-of-function. Additionally, both Nanopore adaptive sampling and Illumina probe hybridization enrichment methods are compared in terms of on-target percentage, with Illumina (61.32%) achieving a higher rate percentage than Nanopore (48.49%).

Methods[1]

High Molecular Weight DNA Extraction[2]

Cells pellets were thawed on ice in preparation for DNA extraction. Then nuclei prep solution, consisting of Prep Buffer (to resuspend the cells) and RNase A (to digest and remove RNA present in solution) was added to the cell pellet tube. Cells were resuspended and incubated at room temperature for 2 minutes. Nuclei lysis solution, consisting of Lysis Buffer (to lyse cells/nuclei and free the gDNA) and Proteinase K (to digest proteins in solution and particularly to deactivate DNAses in solution that would degrade the gDNA) was then added to the cell pellet tube. The sample was then transferred to a Monarch 2 mL round bottom tube (to prevent beads from sticking to the tube bottom in later steps) and the sample was incubated for 10 minutes at 56 ℃ in a thermal mixer with agitation (2000 RPM) to facilitate the lysis reaction. After incubation, precipitation enhancer was added to increase the efficacy of DNA extraction. 2 DNA capture beads were then added to the sample tube to bind gDNA, along with isopropanol to precipitate the DNA/bind it to the beads. DNA is insoluble in isopropanol, so its addition increases gDNA binding to the beads. The sample was then mixed on a vertical rotating mixer at 10 RPM for 4 minutes to bind the DNA.

Liquid was removed from the sample by pipetting and wash buffer was added to purify the DNA by washing away molecules not attached to the beads. The wash buffer was then removed by pipetting and an additional wash buffer addition/removal step was performed to further purify the sample. The beads were transferred to a Monarch Collection Tube II containing a bead retainer. The sample was then spun down to remove remaining wash buffer, and the flow through was discarded. The beads were transferred to a new Monarch 2 mL tube and the bead retainer was placed in an additional tube for later use. Elution Buffer II was then added to the sample tube containing the beads to remove the gDNA from the beads. The sample was incubated for 5 minutes at 56 ℃ in a thermal mixer with agitation (300 RPM) to facilitate this elution. Halfway through the incubation, the tube was inspected and shaken manually to confirm the beads were not stuck to the bottom. The sample (beads and eluted DNA solution) was transferred to the bead retainer containing tube and centrifuged at 12000 × g for 30 seconds to push the eluate through the bead retainer. Afterwards, beads and retainer were discarded, and the flow-through consisting of purified gDNA solution was then pipetted to homogenize the mixture for later sequencing steps.

Nanodrop Quantification

High molecular weight (HMW) DNA content and purity were measured with a Nanodrop spectrophotometer. First, a baseline reading was acquired with 1 µL of ultra-pure water, followed by 1µL of the sample. After quantification, the HMW DNA was stored at 4 ℃.

Illumina Sample Preparation

eBLT (enrichment Bead-Linked Transposomes) and TB1 (Tagmentation Buffer 1) were brought to room temperature and vortexed, while the DNA sample, index adaptors, and EPM (Enhanced PCR Mix) were thawed on ice. A solution containing 500 ng of DNA with a volume of 30 µL was produced from the sample and nuclease-free water. eBLT and TB1 were then combined and mixed by vortexing to produce a tagmentation master mix. The DNA solution was then tagmented with the master mix, by adding the mix to the DNA solution and incubating for 5 minutes at 55 ℃. In essence, the DNA present in the sample is bound to beads that attach adapter proteins while fragmenting the proteins into manageable lengths for Illumina sequencing. The adapter proteins are essential for future binding of the sequences to the flow cell used for sequencing. ST2 (Stop Tagment Buffer 2) was then added to stop the tagmentation reaction.

After the tagmentation reaction, all non-bead-bound compounds were removed by placing the tubes with beads suspended on a magnet (the beads are magnetic). As the beads adhere to the bottom of the tube, the solution clarifies and can be discarded. This process was repeated multiple times, washing each time with TWB (a wash buffer). Finally, the purified tagmented DNA was amplified using PCR. A PCR master mix containing the necessary polymerases and nucleotides was produced from EPM and water. After adding this mix to the DNA sample and spinning down to resuspend the beads, index adapters were added. Index adapters are complementary to the previously added adapters and contain primers on the ends. Thus, these adapters are incorporated into the replicated fragments, generating strands with primer ends. Primers are essential because they allow the fragments to bind to the Illumina flow cell during sequencing.

After this amplification step, cleanup was performed to remove any remaining extraneous compounds in solution (such as unbound index adapters/polymerases/nucleotides). This involved a similar bead/magnet/wash cycle, though this time using AMPure XP beads. These are beads specifically formulated for post-PCR cleanup of barcoded products in the process of library preparation. EtOH was used for the wash steps, discarding the supernatant following magnetic collection of the beads each time (two total washes). Finally, the beads were resuspended in RSB (ReSuspension Buffer), which also elutes the purified, fragmented DNA (with primer ends) from the beads, with an additional magnet step to collect the beads. The supernatant, containing the DNA, was collected for probe hybridization.

Probe Hybridization[4]

EHB2 (Enrichment Hybridization Buffer 2), enrichment probe panel, and NHB2 (a mixture of IDT NXT Blockers and Hybridization Buffer 2) were brought to room temperature. EHB2 and the enrichment probe panel were vortexed to mix them prior to use. NHB2 was heated on a microheatins system set at 50 ℃ for 5 minutes and vortexed to resuspend. The blockers are used while warm to prevent precipitate formation. These blockers ensure that the probes only hybridize with the gDNA of interest by blocking hybridization of the adaptors added during tagmentation. The DNA sample from the previous step, NHB2, enrichment probe panel, and EHB2 were added to a new PCR tube and mixed by pipetting and centrifugation (280 ✕ g, 30 seconds). Hybridization was then achieved by placing the sample in a programmed thermal cycler. Finally, the hybridized DNA sample was stored until the next steps.

Probe Capture[4]

EEW (Enhanced Enrichment Wash), EE1 (Enrichment Elution Buffer 1), HP3 (2N NaOH), EPM (Enhanced PCR Mix), and PPC (PCR Primer Cocktail) were thawed on ice and mixed. SMB3 (Strepdavidin Magnetic Beads) and ET2 (Elute Target Buffer 2) were brought to room temperature and mixed. The DNA sample from the previous step was spun down with a minifuge. Then the sample was transferred to a new microcentrifuge tube. Afterwards, SMB3 was added to the tube and vortexed. Then the tube was placed in a preheated block set to 58 ℃ for 15 minutes. In this step, the beads capture sequences that are hybridized with the probes, selecting for the sequences of interest (exons from the Illumina exome panel). Simultaneously, EEW was preheated by taping the tube to the lid of the heating block. Heating the EEW makes it more effective at washing away all of the unbound contents. Following incubation, the sample was minifuged for 30 seconds to remix the beads in solution. Then the sample was placed on a magnet for 2 minutes to collect the bead, and the supernatant was discarded.

The sample was removed from the magnet and pre-heated EEW was added to the sample, followed by vortexing to resuspend the beads. Unused EEW was returned to pre-heat. The sample was incubated in the heating block for 5 minutes to increase wash effectiveness, followed by vortexing once more to mix. Then the sample was placed once more on the magnet for 2 minutes to collect the beads, and the supernatant was discarded. The wash step was repeated for 3 washes to further purify the DNA / select for only the desired sequences (exons).

After these 3 wash steps, the sample was washed once more with EEW with the magnet step/supernatant removal, but this time the sample was centrifuged (280 ✕ g, 30 seconds) to extract residual liquid, which was removed and discarded while the sample tube was on the magnet (to collect the beads). Immediately afterward, the sample was removed from the magnet. An elution mix was prepared by mixing EE1 and HP3 in a new tube. This mix was then added to the sample to remove the DNA from the beads. HP3 dehybridizes DNA, detaching the probes/freeing the single-stranded sequences of interest, while EE1 is a buffer solution that resuspends the DNA/aids in elution. The sample tube was incubated for 2 minutes at room temperature, followed by minifuging for 30 seconds to remix the solution, and the beads were collected by placing the sample tube on a magnet for 2 minutes. The supernatant, containing eluted DNA, was collected and transferred to a new PCR tube, and ET2 was added to neutralize HP3. The sample was mixed by pipetting and minifuging. PPC and EPM were then added, followed by mixing by pipette and microcentrifuge. Finally, the DNA was amplified by PCR using preprogrammed AMP 10x settings, and the amplified DNA was stored at 2-8 ℃.

After this amplification step, cleanup was performed to remove any remaining extraneous compounds in solution (such as unbound index adapters/polymerases/nucleotides). This involved a similar bead/magnet/wash cycle, though this time using AMPure XP beads. These are beads specifically formulated for post-PCR cleanup of barcoded products in the process of library preparation. EtOH was used for the wash steps, discarding the supernatant following magnetic collection of the beads each time (two total washes). Finally, the beads were resuspended in RSB (ReSuspension Buffer), which also elutes the purified, fragmented DNA (with primer ends) from the beads, with an additional magnet step to collect the beads.

An Illumina MiSeq was used for sequencing the purified plasmid DNA. DNA concentration was measured using a Nanodrop spectrophotometer, and library molarity was adjusted to the specifications of the sequencer. Libraries were then denatured with NaOH (to produce single strands) and diluted with HT1. The reagent cartridge and flow cell were adjusted to appropriate temperatures, followed by loading the cartridge, the sample, the flow cell, and the reagents (PR2). Finally, sequencing was run using specified parameters/protocols of the machine by the manufacturer.

Nanopore Rapid Barcoding

A solution containing 400 ng of the original DNA sample (from before the Illumina preparation steps) with a volume of 7.5 µL was produced from the purified sample and nuclease-free water. Fragmentation Mix (RB01) was then added to this adjusted sample to tagment the DNA. This step both breaks the DNA into manageable sequencing fragments and attaches barcode adapters, using a barcoded transposome complex. The fragment-containing solution was then purified using a similar bead/magnet/wash cycle as in the Illumina preparation (though this time using a 70% EtOH solution). Finally, the DNA was removed from the beads by resuspending the DNA-bound beads in 10 µL of a 10 mM Tris-HCl and 50 mM NaCl solution and pelleting the beads once more on a magnet. The eluate (containing purified, tagmented DNA) was collected and stored on ice until sequencing. The Nanopore flow cell was prepared by priming with a specialized Priming Mix and sequencing was run using specified parameters / protocols of the machine by the manufacturer. In particular, the sequencer was run using real-time enrichment (adaptive sampling) as described in the **Introdution** to select for the sequences of interest (exons from the Illumina exome panel), in much the same way as the probe hybridization steps for Illumina sequencing.

Sequencing Analysis

The generated sequencing data could not be used due to some failure in the sequencing process (the generated files were very small in size, showing an absence of sequence data), necessitating the use of another group's data (Illumina index F7 and Nanopore index RB01). Sequencing results (FastQ files) were analyzed with FastQC in Python to assess sequencing quality. For Illumina data, trimmomatic was used to trim the reads in these files and pair the two reads together. Trimming is performed to remove any remaining adaptor sequences in the reads that may disrupt the alignment. Illumina reads were then aligned against the template human genome (hg38) by using BWA (Burrows-Wheeler Aligner). Minimap2 was used for Nanopore alignment against hg38 (the alignment took excessively long, so provided .bam alignment files were used). For both the Illumina and Nanopore aligned data, processing was performed with samtools. Variants were called from the aligned data (.bam files) with FreeBayes (in essence, FreeBayes checks the sequences that vary between the reference hg38 genome and the aligned data). FreeBayes was successful for Illumina but failed for Nanopore data, so only Illumina data was used for variant analysis. Integrated Genome Viewer (IGV) was used to visualize coverage for variants of interest (in this case, an ITPKB variant).[5] On-target rates (percentage of reads that align with the enrichment exome panel) were calculated for Illumina and Nanopore data using bedtools. Variants were annotated with OpenCravat using ClinVar, gnomeAD gene, NCBI Gene, and gnomeAD3. In particular, pathogenic variants were explored.

Results

FastQC data (Fig. 1) is provide for both Illumina and Nanopore sequencing. Illumina sequencing produced high-quality reads, with a per-base quality score of over 34 throughout the sample. However, the distribution of per sequence GC content deviated slightly from the theoretical distribution, possessing a similar mean (49%) and standard deviation, but with a flattened peak rather than a normal distribution shape. Sequence lengths varied from 35 to 75 bp. Nanopore sequencing produced markedly lower quality reads, with an average per-base quality score around 19 (from a minimum of 5 to a maximum of 22). The distribution of per sequence GC content deviated slightly from the theoretical distribution, with a similar mean (43%) and normal distribution shape, though a much narrower and higher peak. Sequence lengths varied from 100-126494 bp.

Illumina and Nanopore produced on-target percentages of 61.32% and 48.49%, respectively (Table 1). As stated in **Methods**, FreeBayes failed for Nanopore, so it was excluded from variant analysis. A summary of the Illumina variant data is included in Table 2. 13 clinically relevant (pathogenic or likely pathogenic) variants were identified, and of these 13, 2 had PLI (probability that the gene is loss-of-function intolerant) scores above 0.9. The first, ITPKB, had a PLI score of 1.0, and the observed SNP (226735804G>T) is likely pathogenic and is associated with myeloproliferative neoplasm, a blood cancer originating in the bone marrow. The second, APP, had a PLI score of 0.967, and the observed SNP (25891787T>C) is pathogenic and is associated with Alzheimer's disease type I. The chromosomal locations of these clinically relevant variants are included in Fig. 2, and the ITPKB variant location, chosen as the main variant for discussion because of its high PLI score, is visualized in Fig. 3. Notice that the SNP only has a coverage of 3x.

Illumina Summary:

PASS	Basic Statistics
PASS	Per base sequence quality
PASS	Per tile sequence quality
PASS	Per sequence quality scores
FAIL	Per base sequence content
WARN	Per sequence GC content
PASS	Per base N content
WARN	Sequence Length Distribution
PASS	Sequence Duplication Levels
PASS	Overrepresented sequences
PASS	Adapter Content

Nanopore Summary:

PASS	Basic Statistics
FAIL	Per base sequence quality
FAIL	Per sequence quality scores
FAIL	Per base sequence content
WARN	Per sequence GC content
PASS	Per base N content
WARN	Sequence Length Distribution
PASS	Sequence Duplication Levels
PASS	Overrepresented sequences
FAIL	Adapter Content

Figure 1: Illumina and Nanopore FastQC summaries. *Notice that Nanopore produced much poorer quality, failing 4 tests while Illumina only failed 1. The sequencing quality is discussed further in the* ***Results*** *section body above. The poor quality of Nanopore data may have contributed to its later-stage failure in variant calling. Nevertheless, both passed in terms of basic statistics.*

Illumina On-Target Percentage:	Nanopore On-Target Percentage:
61.32%	48.49%

Table 1: Illumina and Nanopore on-target percentages. *Notice the higher percentage for Illumina data. This will be further discussed in* ***Discussion****.*

	Total Variants:	Heterozygous:	Homozygous:	Clinically Relevant:
Count:	47,161	15,419	31,742	13
Percentage:	100%	32.69%	67.31%	0.02757%

Table 2: Illumina variant statistics. *Clinically relevant variants are those classified by ClinVar as pathogenic or likely pathogenic. The clinically relevant variants include 8 missense variants, 2 stop gains, and 3 splice site variants. Some particular PLIs are commented on in the* ***Results*** *section body above.*

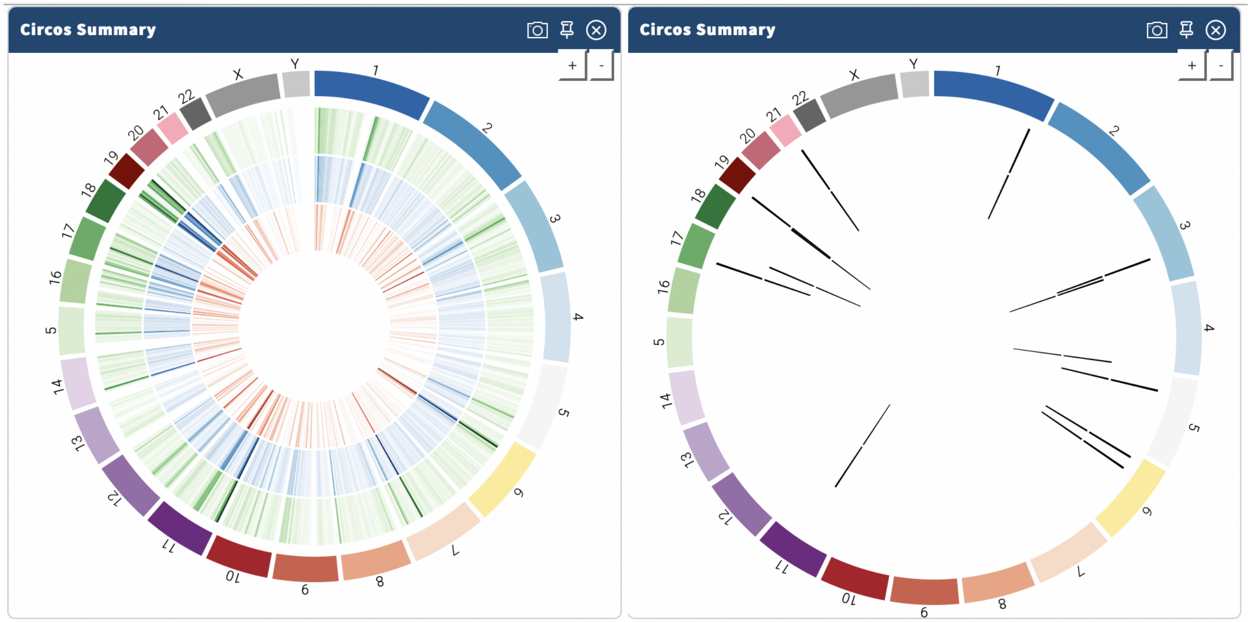

Figure 2: Chromosomal locations of all exome sequencing variants (left) and clinically relevant variants (right). *The three internal rings in both plots correspond to the activity of the particular variants observed. The outer ring indicates missense variants, the middle ring non-silent variants, and the inner ring inactivating variants.*

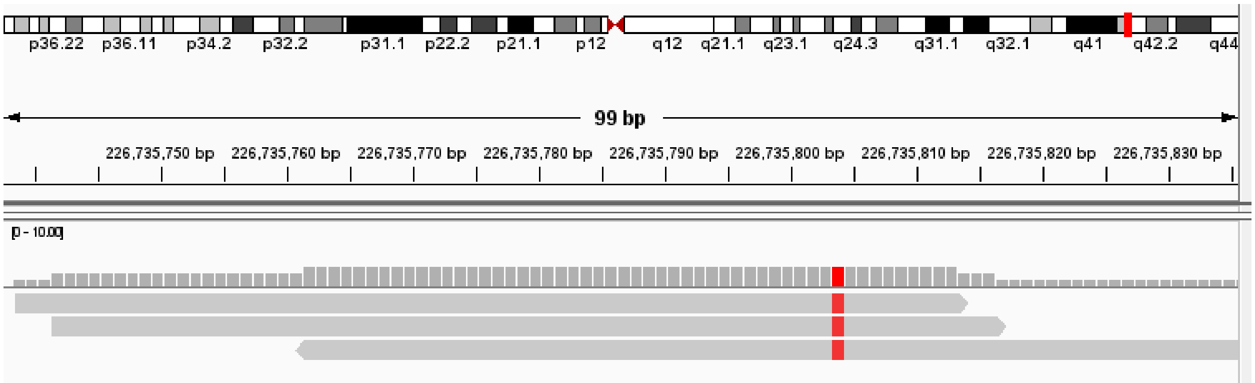

Figure 3: Integrated Genome Viewer (IGV) was used to visualize the coverage of ITPKB. *There was only a coverage of 3x for the SNP, which is highlighted with the red line. The gene is on chromosome 1, with the location annotated at the top of the figure (q42.12).*

Discussion

1. 61.32% of reads were on target for Illumina, indicating that 30.68% of the reads are extraneous/not from the Illumina exome panel. This may arise from a few mechanisms. First, the hybridization enrichment process likely failed to remove all of the off-target sequences. Additionally, the literature from Illumina discusses that the enrichment process still results in regions surrounding the target sequences to be captured (due to hybridization probes trying to capture the whole exome, including the edges of each

exon, so they also capture some base pairs up and downstream of the target in the attempt to cover the whole exon). In fact, in Illumina's studies on the subject, expanding the target region to include +/- 150 bp lead to an increase of roughly 15% in on-target percentage.[6] Thus, this is one method to improve the on-target percentage. Another method could be increasing the number of purification steps in the enrichment process (perhaps more washes, more time on the magnet to better collect the beads in each of the magnetic bead steps).

2. 48.49% of reads were on target for Nanopore, indicating that 51.51% (above half) of the reads are extraneous/not from the Illumina exome panel. Notice that the off-target percentage is higher for Nanopore than Illumina, likely due to a couple principles. First off, Nanopore generates far more errors than Illumina (see FastQC data in **Results**) and likely misidentifes certain sequences during adaptive sampling, rejecting on-target sequences and reducing their content. This is supported by investigation of adaptive sampling that suggested significant increases in enrichment efficiency would be achieved if false negative identifications and associated pore ejections were reduced.[2] Additionally, Nanopore's read length is much longer than Illumina (see FastQC data in **Results**), so it is likely to capture more of the regions up and downstream of the target sequences (the up/downstream capture effect is discussed in question 1). The on-target percentage could be improved by utilizing the hybridization enrichment methods used for Illumina, which clearly produced a higher on-target percentage. Additionally, if shorter reads were used, likely less sequencing error would be accumulated, leading to less false negative identifications.
3. The enrichment methods described in this lab report have both clinical and consumer relevance. In the clinic, these enrichment methods can be used for screening for pathogenic variants (as described in this report), which can lead to more personalized medicine, tailoring the patient's medical experience to match their genetic predispositions to disease. Additionally, the identification of particular variants that are pathogenic improves medical knowledge/provides insights into the genetic source of various diseases, especially cancers. If the source is known, then therapeutics that target it can better be developed. On the consumer side, of course the patient receives the benefits from these clinical screening tools, but the data can be used outside of a clinical context for ancestry information. Certain variants are particular to certain ethnicities/races, so enriching for the sequences that possess this variant content can thus be used to trace a genetic lineage and give people insights into their heritage. And another non-clinical usage is in forensics, where crime scene DNA can be examined to and enrichment can be used to screen for characteristics of the perpetrator that could help in identification. All of the described uses are already making their mark, with BRCA screenings for breast cancer risk being commonplace, thousands of people using services such as 23andMe, and suspect characterization with sequencing/enrichment being an essential tool in law-enforcement.

4. 32.69% of variants were heterozygous and 67.31% homozygous. In the human genome, two copies of every gene (excluding those on the X/Y chromosomes, which may differ) are present, one from each parent. Thus, in this case, roughly ⅓ of the variants contain two different alleles (i.e. the copies differ between maternal and paternal versions of the gene), and ⅔ have the same alleles. As seen in Table 2, the majority of variants are not clinically relevant, and thus are associated with normal variation between individuals, seen phenotypically, for example, as different hair, skin, and eye colors. Likely, the parents of the individual from whom this cell line was taken have similar genetic/hereditary backgrounds, leading to a high percentage of homozygous variants.
5. 31.06% of variants were coding and 68.94% were non-coding. Compared to the overall genome (1-2% coding), this is very high.[7] However, since this is an exome panel, the initial expectation would be that the entirety would be protein-coding. There are a large number of non-coding purposes for exons as well, such as all of the non-coding RNAs (tRNAs, rRNAs, miRNAs, etc.). The 31.06% figure is actually similar to numbers quoted in the literature for the coding percentage of the human exome (around 23.0% of exonic bases are protein-coding).[8] A reason the observed figure may be higher than the literature is that the panel used for enrichment may focus more on clinically relevant loci, which are much more likely to be in protein-coding regions since variants of protein-coding sequences are most likely to be responsible for disease.

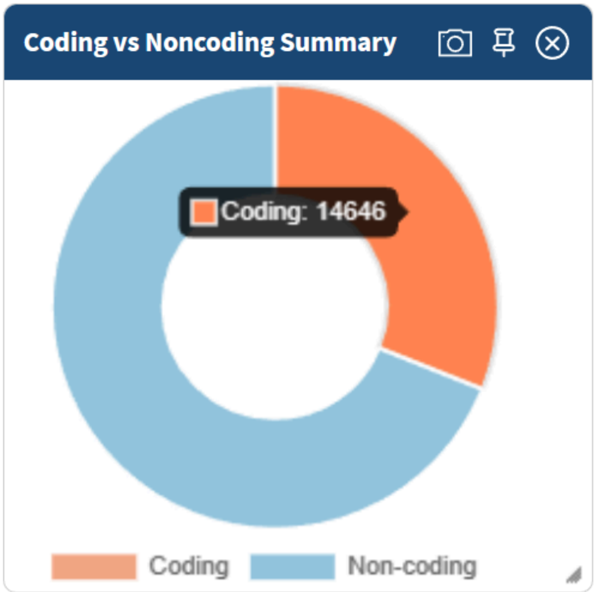

__Figure 4: Coding vs. non-coding variants.__ This figure displays the relative proportions of the total variants (47,161) that code for proteins vs. those that do not.

6. The ITPKB variant (226735804G>T) observed in the data is likely pathogenic because it is a non-silent missense mutation in a gene with a very high PLI (1.0), thus leading to inactivation of the ITPKB gene/the protein encoded by it (Inositol-Trisphosphate 3-Kinase B). This protein, IP3K B, is involved in MAPK signaling because it regulates the concentration of inositol polyphosphates that play a role in this cascade.[9] MAPK

signaling is widely used for cellular proliferation/division regulation, and disruptions of this pathway are widespread in cancer. In fact, mutations of ITPKB have been linked to certain cancers. In particular, a link has been suggested between the observed variant (226735804G>T) and myeloproliferative neoplasms. Myeloproliferative neoplasms are cancers effecting cells of the bone marrow that give rise to blood cells. These starting cells, known as hematopoietic pluripotent stem cells are responsible for self-renewal and production of terminal blood cells such as red and white blood cells (RBCs/WBCs). An example of a myeloproliferative neoplasm is chronic myeloid leukemia, which is an abnormal growth of the hematopoietic stem cells in the bone marrow (broadly, myeloid cells), leading to abnormal proliferation of WBCs (leukemia). These abnormal WBCs begin to consume resources needed to sustain RBCs and healthy WBCs and are also usually nonfunctional, reducing the body's ability to fight infection.[10] The location of the variant (1.q42.12) is visible in Fig. 3.

7. The variant had a coverage of 3x, meaning that 3 different reads covering that base position (226735804, chromosome 1) corroborate the changed identity of the base (G>T). Though this coverage is low, it should still be sufficient to identify this base with high certainty because the quality scores were so high. The average quality score for the Illumina reads was roughly 35, which corresponds to an error rate of approximately $1/10^{35/10} = 1/10^{3.5} = 0.000316$. Thus, the probability that any given base in a read is incorrect is 0.000316. If there are three overlapping reads for this variant that corroborate the base identity, the probability that all three are wrong is $0.000316^{3} = 3.16 \times 10^{-11}$. In other words, the certainty in the identity of this variant is 99.999999996%. Thus, we can say with high confidence that this human cell line contains the observed ITPKB variant (226735804G>T).

Conclusion

In the presented report, gDNA from a human cell line was enriched for the Illumina exome panel and sequenced with both Illumina and Nanopore methods. Two enrichment methods, Illumina probe hybridization and Nanopore adaptive sequencing were compared. Additionally, the sequencing results were aligned with the hg38 human genome, and variants were identified, with a particular focus on clinically relevant (pathogenic or likely to be pathogenic) SNPs. The Illumina data presented a higher on-target percentage, indicating better enrichment efficiency than Nanopore adaptive sequencing. For SNP analysis, variant calling failed for the Nanopore data, so only Illumina data was used. 47,161 variants were identified, 13 of which were clinically relevant. Of these, a variant (226735804G>T) in the ITPKB gene (1.q42.12) was selected for further analysis and visualization due to its high PLI (see **Results** and **Discussion**).

Future work could focus on investigating the issues with the Nanopore sequencing results that interfered with variant calling. Additionally, a continuation of this work would be an economic analysis of the costs and benefits of the two enrichment/sequencing methods, since Illumina provides higher quality results though with much higher costs (not just sequencing, but also labor and time), while Nanopore provides lower quality at lower costs. An interesting economic study would be quantifying the cost associated with cancer (perhaps for a specific set of variants) and then comparing the ability of the two sequencing/enrichment approaches to reduce this cost (the benefit of each method), compared to the costs associated with implementing either method. Ultimately, the presented study compares two popular enrichment methods for human genome sequencing and provides a framework for SNP analysis in sequencing experiments. This work has exciting applications in a wide variety of fields, including screening for disease markers, providing geneological information, advancing understanding of the genetic mechanisms of disease, aiding in forensic investigations, and facilitating the creation of novel methods for targeting genetic conditions.

References

1. Timp, Winston *Methods in Nucleic Acid Sequencing Lab: Module 3 Targeted Sequencing,* 2023.
2. Martin, Samuel, et al. "Nanopore Adaptive Sampling: A Tool for Enrichment of Low Abundance Species in Metagenomic Samples." *Genome Biology*, vol. 23, no. 1, 2022, https://doi.org/10.1186/s13059-021-02582-x.
3. "High Molecular Weight DNA Extraction." *New England BioLabs*, www.neb.com/monarch/high-molecular-weight-dna-extraction. Accessed 19 May 2023.
4. *Illumina DNA Prep with Enrichment Reference Guide*, Illumina, San Diego, CA, 2021, https://support.illumina.com/content/dam/illumina-support/documents/documentation/chemistry_documentation/illumina_prep/illumina-dna-prep-with-enrichment-reference-guide-1000000048041-07.pdf. Accessed 19 May 2023.
5. Robinson, James T et al. "Integrative genomics viewer." *Nature Biotechnology* vol. 29,1 (2011): 24-6. doi:10.1038/nbt.1754
6. *Optimizing Coverage for Targeted Resequencing*, Illumina, San Diego, CA, 2014, https://www.illumina.com/documents/products/technotes/technote_optimizing_coverage_for_targeted_resequencing.pdf. Accessed 20 May 2023.
7. Henninger, Jonathon. "The 99 Percent... of the Human Genome." *Science in the News*, 16 Mar. 2015, sitn.hms.harvard.edu/flash/2012/issue127a/.
8. Aspden, Julie L et al. "Not all exons are protein coding: Addressing a common misconception." *Cell Genomics* vol. 3,4 100296. 12 Apr. 2023, doi:10.1016/j.xgen.2023.100296
9. "ITPKB Inositol-Trisphosphate 3-Kinase B [Homo Sapiens (Human)] - Gene - NCBI." *National Center for Biotechnology Information*, 15 May 2023, www.ncbi.nlm.nih.gov/gene/3707.
10. Thapa B, Fazal S, Parsi M, et al. "Myeloproliferative Neoplasms." [Updated 2022 Aug 8]. In: StatPearls [Internet]. StatPearls Publishing, Treasure Island, FL, Jan. 2023.