Long-Read Trio Analysis with Oxford Nanopore Sequencing

MCDB 5520: Computational Genomics Group Project

Written component due 11/19/2025 12pm by Canvas

Graded out of 40 points, optional assignments can be completed for up to 5 pts extra credit

Project Overview

You are part of a **genomic diagnostics team**, tasked with better understanding the biology and disease risk from a "trio" of single family's genomes. Your group will perform a **comprehensive trio-based variant analysis** using Oxford Nanopore long-read sequencing data.

- Detect the genetic variants from the sequencing data
- Sort through and research variants that potentially influence biology/disease risk
- Trace inheritance patterns and detect *de novo* mutations unique to the child

This project will involve running the wf-human-variation, a Nextflow workflow for detecting genetic variants from human long-read whole-genome sequencing datasets. All students are expected to run the workflow on their own machines.

Deliverables: (i) a group writeup (due 11/19) and (ii) a 10 minute group presentation (final 3 lectures of class) where you will present your predictions and proposed future directions based on your variant analysis.

Chromosome Assignments (10 Groups)

Each group is assigned ONE chromosome from the family trio, which consists of **HG002 (son), HG003 (father), and HG004 (mother).** Each group member will analyze one individual, then collaborate to find inherited vs *de novo* variants.

Each chromosome contains 500-3,000 different genes with different functions. As you explore your assigned chromosome, research the genes predicted to be affected by variants.

Group	Chromosome	Examples of clinically relevant genes	Example biological theme
1	chr17	BRCA1, TP53, MAP2K3	Cancer predisposition
2	chr13	BRCA2, RB1, ATXN8OS	Cancer predisposition
3	chr19	APOE, XRCC1, APOC1	DNA repair
4	chr6	HLA-A, HLA-C, TNF	Immunity
5	chr4	HTT, SNCA, PARK2	Neurological/brain
6	chr15	HEXA, HERC2, FBN1	DNA repair
7	chr22	CYP2D6, COMT, TBX1	Pharmacogenomics/cardiac
8	chr3	CNTN4, MLH1, LAMB2	Cancer/neurodevelopment

9	chr2	SCN1A, MSH2, NRXN1	Neurological disorders
10	chr1	DPYD, GBP1, MTHFR	Immunity/metabolism

Phase 1: Individual Variant Calling and Analysis

Each group member will independently run **wf-human-variation** on their chosen trio family member.
You will be asked to document key findings to share with your group in **Phase 2.**
Before running, read through the workflow documentation and different options:
https://github.com/epi2me-labs/wf-human-variation

Setup workspace on Piel

```
cd /scratch/$USER

export SINGULARITY_TMPDIR="/scratch/$USER/singularity_tmp_human"
export NXF_SINGULARITY_CACHEDIR="/scratch/$USER/singularity_cache_human"

mkdir -p  "$SINGULARITY_TMPDIR"   "$NXF_SINGULARITY_CACHEDIR"
cp  /scratch/singularity_cache_human/*.img  singularity_cache_human/

mkdir -p /scratch/$USER/trio_project
cd /scratch/$USER/trio_project
```

Run wf-human-variation (each student runs the pipeline for ONE trio member)

Important: you will need to adjust the following flags for your particular individual and chromosome:

`--bam`, `--ref`, `--sample_name`, `--sex`, `--out_dir`, `-w`

Make sure to use the chromosome-specific FASTA file for your assigned chromosome, as below.

```
# View workflow help
nextflow run epi2me-labs/wf-human-variation --help

# Run the workflow on your individual and assigned chromosome
# E.g. student 1 analyzes HG002 (male), chr17
nextflow run epi2me-labs/wf-human-variation \
  -profile singularity \
  --bam /data/human_trios/family1/HG002/HG002_chr17.bam \
  --ref /data/human_trios/reference/hg38_chr17.fa \
  --override_basecaller_cfg dna_r10.4.1_e8.2_400bps_sup@v5.0.0 \
  --sample_name HG002_chr17 \
  --snp --sv --str \
```

```
  --sex XY \
  --bam_min_coverage 0 \
  --annotation true \
  --out_dir /scratch/$USER/trio_project/HG002_chr17_output \
  -w /scratch/$USER/trio_project/HG002_chr17_work

# Duration: 1-4 hours, depending on the chromosome
# Once the run completes successfully, remove the (temp) work directory
rm -r /scratch/$USER/trio_project/HG002_chr17_work
```

1. From your output directory, download all the report files (*.**html**), variant call files (*.**vcf.gz**), and their corresponding index files (***.vcf.gz.tbi**).

1a) Overall Quality control

Open the **alignment report** to assess sequencing and mapping quality. These metrics indicate how well the long reads cover your assigned chromosome.

- File: `SAMPLE_CHR.wf-human-alignment-report.html`
- Document: total read count, read N50 (median read length), mean coverage depth, etc. Note any concerns with the dataset, if any, and how they might influence your interpretation.

1b) View summary of detected variants

You can view results from three variant classes detected by the workflow. **Note: Group 1 can skip STR-related questions since chr17 does not have STR results.**

Variant Type	Description	Report File
SNP (Single Nucleotide Polymorphisms)	Single-base changes (e.g., A→G). These are the most common form of genetic variation.	`SAMPLE_CHR.wf-human-snp-report.html`
SV (Structural Variants)	Large-scale changes (>50 bp), such as insertions, deletions, inversions, or duplications.	`SAMPLE_CHR.wf-human-sv-report.html`
STR (Short Tandem Repeats)	Repeated DNA motifs (e.g., "CAGCAGCAG..."), which can expand or contract. Some expansions cause diseases like Huntington's or Fragile X.	`SAMPLE_CHR.wf-human-str-report.html`

For each variant type, open the report to see total counts, variant significance or consequences, short tandem repeat expansions within genes, etc.

1c) Inspect variant call files (VCF format)

You should see files like:

- `*.wf-human-snp.vcf.gz`: raw SNP and indel calls.
- `*.wf-human-sv.vcf.gz`: structural variant calls.
- `*.wf-human-str.vcf.gz`: short tandem repeat genotypes (if detected).
- `*.clinvar.vcf.gz`: variants annotated with ClinVar (e.g. known clinical variants).

- `*.vcf.gz.tbi`: index files required for viewing or uploading variant data to websites.

These are the raw variant files used to generate the .html reports. Explore the first 4 files (ignore the index) by unzipping and opening in a text editor on your computer, or by using the terminal to search within them on Piel.
Each line in the file specifies a different variant, with information on the position, type of variant, confidence, etc.
The VCF format is defined here: https://samtools.github.io/hts-specs/VCFv4.2.pdf

For example, you can use "grep" to filter VCF files to search for exonic structural variants:
(replacing the sample and chromosome as necessary for all following examples)

```
# Find SVs affecting exons
zcat HG002_chr17.wf_sv.vcf.gz | grep PASS | grep -v pseudo | grep -v LOC | grep
exon > HG002_chr17_exonic_SVs.txt

# View results
less HG002_chr17_exonic_SVs.txt
```

2. Investigate 3-5 interesting variants found in your dataset

Many "common" variants (frequency >1% in the population), particularly SNPs, are often already documented in online databases like ClinVar (with identifiers like *rs386833395* for BRCA1 cancer susceptibility). In the HTML report, the subset of variants that have already been annotated by ClinVar are reported.

- Use the search box to filter results by known disease genes (e.g., BRCA1 or TP53 for chr17). Click the links from the "Gene" and "ClinVar" page to see an example of the kind of online information available for these variants.
- Pay special attention to variant attributes like "Significance", "Type", and "Consequence" fields (you can sort the table by those). What do you notice about the "Significance" for most variance?
- For the next phase, identify 3-5 variants you think are the most interesting based on the gene function, ClinVar page, and variant attributes, noting:
    - Gene name, position, variant type, consequence, ClinVar significance, and any other notes from your research

Other types of variants, like SVs and STRs, are not automatically reported in ClinVar because they have been underreported prior to long-read sequencing. For these, look through the VCF files (eg the previous command to find SVs that intersect exons). Document any interesting structural or repeat variants that might affect genes, based on the VCF file. Record the gene name, variant type (DEL, DUP, INV, STR, etc.), approximate size (SVLEN), and predicted impact.

Phase 2: Trio Comparison and Inheritance Patterns (group)

Once all members have finished group 1, you will meet as a group and discuss variants of interest that may affect some or all of the family. Your task is to prepare a written genomic variation report summarizing your most important findings.

Select at least 3 genetic variants found in your trio analysis, grouped by a biological theme (e.g., see chromosome assignment table at the beginning of this document)**.** Many traits are known to be complex, meaning they are shaped by *multiple* variants each with a small effect. Your group should try to identify a collection of variants in your data predicted to impact a common pathway or trait in this family (e.g., immunity or cancer predisposition), based on the known function of the affected gene(s).

Create a 1-2 page comprehensive report for each genetic variant.

1. **Provide a high-level summary:** What is the gene, what kind of variant is it, how does it affect the gene, and what's the predicted impact?
2. **What are the family inheritance patterns of each variant?** Which variants are inherited (maternal/paternal) vs de novo? If the variant is common, what is its ancestry distribution?
3. **What is the variant predicted to do at the molecular level?** If the variant has a documented effect (eg by ClinVar), look up the reference and summarize the evidence for the effect (e.g., genetic association study, experimental evidence, etc). If the variant is "unknown" or N/A, what could be the range of possible effects?
4. **Describe in 1-2 paragraphs an experiment that could help you test and/or validate if the variant is pathogenic or benign.** Assume you have access to patient-derived IPS cells and all the equipment necessary for genome editing and functional genomics.
5. **What could be the implication of the variant for disease?** In addition to information provided by ClinVar/etc, look up recent research articles on the gene and speculate if there are other biological impacts that are possible.
6. **Include any other information conducted in "Additional analyses" section below**

Finally, create a summary report that synthesizes your predictions from the variants, and provide (non-official) recommendations for screening, risk assessment, counseling, etc.

Guidelines

- Make sure your report is clear and written at a level appropriate for a genetic counselor
- You are encouraged to use and/or create figures from online sources as long as they are clearly explained and cited.
- Cite all references and databases used to populate your report
- One final report will be due per group via Canvas on 11/19

Required analysis: Trio analysis

After all three group members finish **Phase 1**, work together to compare your variant-calling results and explore inheritance across the trio. Each person should have output folders containing SNP/indel (`*.wf_snp.vcf.gz`), structural variant (`*.wf_sv.vcf.gz`), and, if applicable, STR (`*.wf_str.vcf.gz`) results.

1. Gather your results

Combine the `.vcf.gz` files from all three individuals into a shared directory so you can examine them

together. As a group,

2. Explore inheritance and de novo candidates
Identify variants that are shared by all individuals (common/familial), shared between child + one parent (inherited), or unique to the child (potentially *de novo*). You can compare variants manually using your reports or use command-line tools such as bcftools.

As a simple first pass, you can use **bcftools** (included inside the wf-human-variation container image) to search for ClinVar variants only present in the child:

```
# bcftools is included in the wf-human-variation Nextflow container
# run it to find de novo variants in the child
singularity exec \
/scratch/$USER/singularity_cache_human/ontresearch-wf-human-variation-
sha8ecee6d351b0c2609b452f3a368c390587f6662d.img \
bcftools isec -n=1 \
HG002_chr17.wf_snp_clinvar.vcf.gz \
HG003_chr17.wf_snp_clinvar.vcf.gz \
HG004_chr17.wf_snp_clinvar.vcf.gz \
-p bcftools_output

# View results
ls bcftools_output/
# 0000.vcf      # ClinVar variants ONLY in HG002 (son) - DE NOVO CANDIDATES
# 0001.vcf      # ClinVar variants ONLY in HG003 (father)
# 0002.vcf      # ClinVar variants ONLY in HG004 (mother)
# sites.txt     # Tab-delimited: which file(s) contain each variant
# README.txt    # Description of the output files

# Count the number of de novo variants (excluding vcf header)
grep -v "^#" bcftools_output/0000.vcf | wc -l

# Search for pathogenic variants in the child
grep -i "pathogenic" bcftools_output/0000.vcf
```

Document any "de novo" variants in the child (not present in the mother or father), predicted to be pathogenic or otherwise. What else can you conclude or speculate based on those variants?

Additional analyses

(At least one required; multiple up to 5 pts extra credit)

Many databases, command line, and online tools exist to analyze human variant data, and can provide useful figures/analysis/screenshots for your report. You are encouraged to be creative and use online databases/websites or command line tools (feel free to use your own computer, or check with Dr. Chuong/Dr. Ivancevic prior to installing anything on Piel) to further analyze the variants and/or

genes and report what you learn. We are also more than happy to point you in the right direction or assist you if you get stuck in any of these analyses..

- **Conduct a more sophisticated trio analysis, e.g. installing/running:**
  - **Sniffles2**: jointly genotype SVs across the trio
  - **WhatsHap**: phase SNPs/indels using pedigree information
  - **Truvari**: compare SV sets quantitatively
- **Determine the likely ancestry of the VCF**
  - Can you determine the ancestry of your individuals? Try websites and tools such as **gnomAD**, **ADMIXTURE**, **AEon**, **G-Nomix**, or this PCA pipeline: https://github.com/laura-budurlean/PCA-Ethnicity-Determination-from-WGS-Data
- **Visualize the read evidence supporting the variant in a genome browser**
  - Visualizing variants and the read alignments used to predict the variant is a key step, especially for STRs and SVs, to determine the confidence in the variant. For at least 3 variants (ideally different types--eg SNV, STR, SV), visualize the read alignments from the cram file using a genome browser.
    - Tip: since the read alignment (*cram) files are large, you can use samtools to subset individual *.cram file into a smaller 5 kb window, and download those to your computer. These can be visualized on your computer using a tool like the Integrated Genomics Viewer, and there are also specialized tools such as **Samplot** or **SVTopo.**
- **Upload your VCFs to the UCSC Genome Browser (hg38 assembly)**.
  - Explore various tracks (e.g., dbSNP, GWAS Catalog, ENCODE regulatory elements, CRISPR tracks). Are there any interesting features directly overlapping or nearby the variant?
  - You can also examine your variants in genome browsers (e.g. UCSC Genome Browser) or using command line annotation tools.
- **Use online resources or command line tools to find out more about your variants.**
  - For example, you can import your `*.wf-human-snp.vcf.gz` file to various sites, including:
  - GenVue (https://genvue.geneticgenie.org). Browse the different tabs for details on possible genetic conditions, drug responses, rare or uncommon mutations, and other risks.
  - Enlis Genomics (https://www.enlis.com/import). This will generate an analysis report, sent to an email of your choice.
  - MySeq (https://skylight.middlebury.edu/~mlinderman/myseq/load). This allows you to explore variants or perform analyses, such as a PCA analysis to determine ancestry.
- **Model variant effects using modern AI prediction tools**
  - Use structural prediction tools such as **AlphaGenome**, **AlphaFold**, or **AlphaMissense** to model the potential consequences of deleting or altering one or more of your variants. Provide screenshots/data when relevant. Explain your interpretation of the model's output
- **Anything else!**

In-class Presentation

In class, each group will present their findings, going over the highlights of each variant and their final recommendations. The presentation is strictly limited to **10 minutes**, followed by a **5-minute Q&A** period.

Guidelines:

- Your group should upload the slide deck or presentation file in advance (by noon on presentation day).
- Each student will be responsible for discussing one variant
- Each student is responsible for presenting at least one key analysis or figure (e.g., ClinVar results, inheritance patterns, validation). This should be a clear explanation to non-experts of what the data shows and how to interpret it.
- Be sure to practice timing - your presentation ends at 15 minutes sharp.

Suggested format (~8 slides):

- 1 slide overview of the theme of the variants
- 1-2 slides for each variant
- 1 slide overall summary and recommendations

Good luck! Focus on explaining what makes your variants and associated genes important and biologically interesting.