CHOP take home assignment I Aayudh Das

Task

1. Write a script to identify strictly de novo variants (reference homozygous in parents and heterozygous in the proband). This will likely identify hundreds of *de novo* variants including false positives.

Script file attached- **CHOP_test.ipynb**

Total **343** variants were generated. File - **CDL-068-99.strict_denovo.tsv**

2. Write filters using variant quality attributes to arrive at the most likely true positive list of *de novo* variants and explain the rationale for the choice of filters.

The goal of this filtering strategy is to identify a high-confidence set of true germline de novo variants in a parent–child trio by systematically removing sequencing artifacts, mapping errors, and genotype misclassifications, while retaining biologically plausible heterozygous events in the proband.

- Proband genotypes must be heterozygous (0/1), and both parents must be homozygous reference (0/0) to confirm the variant is present only in the child and absent in both parents.
- Variants must have QUAL ≥ 30 and FILTER = PASS to exclude low-confidence calls and sequencing noise.
- When available, apply model-based recalibration (VQSR) and retain only variants with VQSLOD > 0 to improve specificity using machine learning.
- Variants must satisfy INFO-level quality metrics: QD ≥ 2.0, FS ≤ 60, SOR ≤ 3, MQ ≥ 40, MQRankSum ≥ −12.5, and ReadPosRankSum ≥ −8 to remove technical artifacts and mapping biases.
- All trio members must have high-confidence genotypes with DP ≥ 15 and GQ ≥ 30 to ensure accurate genotype calls.
- The proband must have at least 8 ALT-supporting reads and a variant allele fraction (VAF) between 0.35 and 0.65 to confirm balanced heterozygosity.
- Both parents must have ≤ 0 ALT-supporting reads and VAF ≤ 0.01 to confirm true absence of the variant and exclude mosaicism or contamination.

CHOP take home assignment I Aayudh Das

3. What is the number of true positive variants you expect to have? Explain your reasoning.

Filtered variant list file - **CDL-068-99.filtered_denovo.tsv**

- The filtered list **contains 88 candidate** de novo variants (**80 SNVs and 8 INDELs**), of which approximately 80–85 are expected to be true positives.
- This expectation aligns with biological data from large-scale trio WGS studies, which typically observe ~44–82 de novo SNVs per child, consistent with the 80 SNVs identified here.
- The count of 8 INDELs is also plausible, as germline de novo INDELs usually number around 3 per genome, but slightly higher counts are common due to short-read calling artifacts and retained borderline candidates.
- The applied filters—depth (DP), genotype quality (GQ), allele balance in the proband, and exclusion of parental ALT evidence—are known to produce high-precision de novo variant sets.
- Trio-based pipelines with similar filtering strategies often achieve precision rates near 99%, though real-world accuracy can vary with sequencing depth, alignment quality, and repetitive sequence content.
- Remaining false positives (approximately 3–8 out of 88) are most often due to mapping ambiguity in low-complexity regions, local alignment issues at multi-allelic sites, subtle parental mosaicism, or uneven sequencing depth.
- Overall, ~80–85 variants are expected to be genuine de novo events, with most of the few residual artifacts likely among the INDEL subset or variants in difficult genomic contexts.
- A “true positive likelihood” scoring approach based on quality metrics (e.g., MQ, FS, SOR, QD, and trio VAF patterns) can further prioritize 5–10 variants.
    - Additional analysis - **Top10 true positives.ipynb.** The Top-10 variants are those with the highest composite **“true-positive likelihood” score**, calculated by integrating multiple independent quality metrics into a composite likelihood measure for each variant. The scoring combines variant confidence (QUAL, VQSLOD, QD), mapping integrity (MQ, FS, SOR, MQRankSum, ReadPosRankSum), and genotype balance (VAF, DP, GQ across trio members). Variants showing high-quality signals, balanced heterozygosity (around 50% VAF) in the proband, and complete absence in

both parents receive the highest true-positive likelihood scores, ranking them among the Top-10 candidates.

Table 1: Top 10 variants based on true positives score

CHROM	POS	REF	ALT	TYPE	PROBAND_VAF	TP_SCORE
chr5	1.1E+08	G	T	SNV	0.5	102.0741
chr10	95296731	G	A	SNV	0.5	101.3843
chr6	11271313	A	C	SNV	0.4902	99.81068
chr14	41468025	G	A	SNV	0.5	99.67155
chr8	43348633	G	C	SNV	0.5	99.66545
chr1	9999315	G	A	SNV	0.5	99.6611
chr8	1.25E+08	A	G	SNV	0.4909	99.64562
chr1	90679453	T	C	SNV	0.5	99.40641
chr1	2.3E+08	A	C	SNV	0.5091	99.07637
chr16	7723068	A	C	SNV	0.5	98.83429

4. Explain important sources that may give rise to false positives.

False positive de novo variants arise when a variant appears present in the proband and absent in parents due to technical, biological, or analytical artifacts, rather than a true germline mutation. The most important sources are outlined below.

Source	Mechanism	Key Indicators
Sequencing error	Base miscalls	Low QUAL, skewed VAF
Mapping artifact	Misalignment	Low MQ, bad MQRankSum
Strand bias	PCR bias	High FS, SOR
Low parental depth	Allele dropout	Low parent DP/GQ
Proband imbalance	Noise	Low ALT reads
Parental mosaicism	True low-level variant	Small parent VAF
Multi-allelic sites	Parsing errors	Complex AD
Contamination	Read leakage	Low VAF noise
Reference issues	Misrepresentation	Recurrent loci