AVX-Parallelised GATK pipeline for Rare Disease Trios

Gabriele Oberti, **Michelangelo Rosa**, Università degli Studi di Milano

Next-generation sequencing (NGS) has revolutionised Mendelian disease diagnosis, yet challenges arise in distinguishing pathogenic variants from benign polymorphisms. Computational pipelines must balance between sensitivity for rare variant detection and specificity to reduce false positives, all while remaining computationally efficient given the increasing scale of sequencing datasets. This pipeline supports AR and de novo Mendelian diagnosis in trios by combining alignment, joint genotyping, variant filtration, inheritance-based prioritisation and annotation. To handle growing sequencing data volumes it employs multithreading and AVX-accelerated Pair-HMM steps available in GATK.

Methods

The pipeline begins by setting up a configuration script. At first the number of processors is accessed through the command `$(nproc)` and is saved in the variable `THREADS` to a maximum of 10, to balance speed and resource usage. Then reference files are loaded from a folder named Common: the reference genome (universe.fasta) and a BED file specifying the target region. A list of sample IDs representing the trio (child = 0, father = 1, mother =2) is defined to automate per-sample processing. Finally, a Docker command is stored in the variable `DOCKER_GATK` to run GATK tools inside a containerised environment.

1. Alignment

The first step consists of aligning the FASTQ files to the reference genome using BWA-MEM, with output sorted and indexed via Samtools. Before alignment, the script checks whether the reference genome (universe.fasta) is already indexed. If not, it runs bwa index and generates the corresponding sequence dictionary with GATK's CreateSequenceDictionary.

```
if [[ ! -f "Common/universe.fasta.bwt" ]]; then
      bwa index Common/universe.fasta
fi
if [[ ! -f "Common/universe.dict" ]]; then
      $DOCKER_GATK gatk CreateSequenceDictionary \
            -R /ref/universe.fasta \
            -O /ref/universe.dict
```

Each of the three samples (child, father, mother) is processed in a loop. For each sample:

1. A read group (@RG) header is added with the sample ID and name.
2. The alignment is performed using BWA-MEM with multithreading enabled (-t).
3. Output is directly piped into samtools sort to create a coordinate-sorted BAM file.
4. An index file (.bai) is then generated with samtools index.

This step ensures all BAM files are properly prepared for downstream processing.

```
for SAMPLE in "${SAMPLE_IDS[@]}"; do
      if [[ ! -f "out/${SAMPLE}.bam" ]]; then
            bwa mem -t $THREADS -R "@RG\tID:${SAMPLE}\tSM:${SAMPLE}" \
            $REF_FASTA "in/${SAMPLE}.fq.gz" | \
            samtools sort -@ $THREADS -o "out/${SAMPLE}.bam" -
            samtools index -@ $THREADS "out/${SAMPLE}.bam"
      fi
done
```

2. GVCF Generation with GATK HaplotypeCaller

GVCFs (Genomic Variant Call Format files) are intermediate outputs that provide genome-wide information on both variant calls and the confidence in homozygous reference positions. Unlike standard VCFs, which report only variant sites, GVCFs include non-variant regions, this allows joint genotyping: the simultaneous analysis of genetic data from multiple related individuals, which proves useful in family-based rare disease studies for reliably detecting subtle inheritance patterns.

To generate these files GATK's HaplotypeCaller performs local de novo assembly of sequencing reads and evaluates candidate haplotypes using a Pair Hidden Markov Model. This probabilistic model provides more accurate read-to-haplotype alignment than direct base matching, proving useful for detecting indels and SNPs within repetitive genomic regions. The model transitions between three hidden states: Match/Mismatch (M), Insertion (I), and Deletion (D), incorporating base quality scores from the sequencer to differentiate sequencing errors from true variants, ultimately contributing to more accurate genotyping by statistically weighing the evidence. For example

a consistent mismatch of high-quality bases across many reads is more likely to be a true variant (like a SNP) whereas a single mismatch of a low-quality base might just be a sequencing error.

The haplotype comparison step in the Hidden Markov Model (HMM) involves calculating emission and transition probabilities across millions of read-haplotype-position combinations. Each calculation relies on floating-point operations such as log-likelihood evaluations for mismatches or gaps.

The AVX instruction set on x86 CPUs accelerates these computations by processing multiple operations in parallel:

- Without AVX: A loop computes probabilities for one read-haplotype position at a time.
- With AVX: The same loop processes 8 single-precision (32-bit) or 4 double-precision (64-bit) positions simultaneously by packing data into 256-bit wide registers.

Since the release of the fourth generation of x86 CPUs in 2013 further optimisation is provided by FMA which enables the computation of a multiplication and addition in a single cycle, reducing latency and improving throughput. HMMs calculations benefit hugely from this enhancements since matrix-vector products (common in emission/transition calculations) often chain multiply-add operations.

Theoretically, AVX2 alone reduces the cycle count for linear algebra operations to ~1/8th (32-bit) or 1/4th (64-bit) versus scalar code.

FMA can provide an additional ~2x speedup for dependent operations by eliminating intermediate rounding and doubling throughput for multiply-add patterns.

Modern CPUs with AVX-512 have wider 512-bit registers and could theoretically double throughput again, but thermal/power constraints often limit sustained performance.

```
for SAMPLE in "${SAMPLE_IDS[@]}"; do {
      $DOCKER_GATK gatk HaplotypeCaller \
            -R /ref/$(basename $REF_FASTA) \
            -I /data/${SAMPLE}.bam \
            -O /data/${SAMPLE}.g.vcf.gz \
            -ERC GVCF \
            -L /ref/$(basename $TARGET_BED) \
            --native-pair-hmm-threads $((THREADS/2))
            } &   # The & allows for concurrent operation
      done
wait
```

3. Joint Genotyping with GATK GenomicsDB

Once individual GVCFs are created for each sample, joint genotyping is performed to integrate evidence across all samples.

GenomicsDBImport consolidates the individual GVCFs into a GenomicsDB workspace, a columnar data store optimized for efficient querying across genomic intervals and multiple individuals. This stage is primarily I/O bound, so performance benefits from fast SSD storage and sufficient RAM.

```
$DOCKER_GATK gatk GenomicsDBImport \
      -V /data/child.g.vcf.gz \
      -V /data/father.g.vcf.gz \
      -V /data/mother.g.vcf.gz \
      --genomicsdb-workspace-path /data/trio_db \
      -L /ref/$(basename $TARGET_BED)
```

GenotypeGVCFs performs joint genotyping by querying the workspace. At each genomic position, it examines genotype likelihoods from each sample and applies Bayesian inference to compute posterior probabilities. This Bayesian approach refines uncertain calls, highlights de novo mutations and reduces false positives from sequencing noise.

```
$DOCKER_GATK gatk GenotypeGVCFs \
      -R /ref/$(basename $REF_FASTA) \
      -V gendb:///data/trio_db \
      -O /data/trio_joint.vcf.gz \
```

4. Variant Filtering and Prioritisation

After joint genotyping the resulting VCF file is filtered to retain high-confidence variants. First, low-quality variants are excluded by applying a minimum quality threshold (QUAL ≥ 10) and a minimum read depth (DP ≥ 10). This ensures that only variants supported by sufficient evidence are considered.

```
bcftools filter -i 'QUAL>=10 && INFO/DP>=10' |
bcftools view -T $TARGET_BED -Oz -o out/trio_filtered.vcf.gz
```

With high-confidence variants filtered, the next step is to assess their inheritance patterns within the trio. Using the sample index convention established earlier (child = 0, father = 1, mother = 2), two inheritance models are considered:

Autosomal Recessive

Under the AR model, the proband is homozygous for the alternate allele (GT=1/1), while both parents are heterozygous carriers (GT=0/1). This pattern is consistent with biallelic inheritance, where one pathogenic allele is inherited from each parent.

```
bcftools view out/trio_filtered.vcf.gz | \
bcftools filter -i '(GT[0]="hom" && GT[1]="het" && GT[2]="het")' \
-Oz -o out/AR_candidates.vcf.gz
```

Autosomal Dominant De Novo

In the AD de novo model, the proband is heterozygous (GT=0/1) or homozygous alternate (GT=1/1) and both parents are homozygous reference (GT=0/0). This genotype pattern suggests a de novo mutation arising in the proband, absent in the parental genomes.

```
bcftools view out/trio_filtered.vcf.gz | \
bcftools filter -i '(GT[0]="het"||GT[0]="homalt")&&GT[1]="homref"&&
GT[2]="homref"' \
-Oz -o out/de_novo_candidates.vcf.gz
```

5. Annotation with VEP and Diagnosis Results

Variants were annotated using Ensembl VEP v113.0 with RefSeq transcripts. Pathogenicity was predicted using SIFT and PolyPhen-2, while phenotype links were assessed via Geno2MP, Mastermind and Phenotypes plugins. Population allele frequencies were sourced from gnomAD. Prioritization focused on rare variants (gnomAD AF <1%) within coding regions, predicted as deleterious (SIFT "deleterious" or PolyPhen-2 "probably damaging") and associated with relevant phenotypes in OMIM or Geno2MP. Final candidates were evaluated under both autosomal dominant and recessive inheritance models. For each candidate the proband's genotype and phenotype annotations were reviewed to assess whether the variant would be expected to be disease causing under the respective inheritance model. In all cases the probands were consistent with an affected status.

Case #	*Gene Affected*	*Location hg19*	*REF/ALT*	*Inheritance Type*	*Mutation Type*	*Disease*
624	PKD1 (5310)	16:2158684	-/C	AD	Frameshift Variant	Autosomal Dominant Polycystic Kidney Disease
655	ANKRD11 (29123)	16:89346629	-/C	AD	Frameshift Variant	KBG Syndrome
657	CYLD (1540)	16:50788249-50788253	GGAT/-	AD	Frameshift Variant	Familial Cylindromatosis
661	PKD1 (5310)	16:2140563	C/T	AD	Stop Gained	Autosomal Dominant Polycystic Kidney Disease
674	CREBBP (1387)	16:3807326-3807328	GG/-	AD	Frameshift Variant	Rubinstein-Taybi Syndrome
696	PKD1 (5310)	16:2140149-2140150	G/-	AD	Frameshift Variant	Autosomal Dominant Polycystic Kidney Disease
704	PKD1 (5310)	16:2161376	G/T	AD	Stop Gained	Autosomal Dominant Polycystic Kidney Disease
715	PKD1 (5310)	16:2140953	G/A	AD	Stop Gained	Autosomal Dominant Polycystic Kidney Disease
741	ANKRD11 (29123)	16:89349299	G/T	AD	Stop Gained	KBG Syndrome
746	RPGRIP1L (23322)	16:53682877	G/T	AR	Stop Gained, Splice Region Variant	Joubert Syndrome; Meckel-Gruber Syndrome

Quality Control and Visualisation of Disease-Causing Variants

In order to work with raw sequencing data it is necessary to assess its reliability: tools like FastQC evaluate read quality and adapter contamination, whereas BamQC and MultiQC generate comprehensive reports on alignment metrics, coverage depth and potential biases. These steps are critical to filter out low-quality data that could lead to false variant calls. For further validation tools like IGV and the UCSC Genome Browser enable interactive visualisation of aligned reads, allowing manual inspection of candidate variants within their genomic context. The following are the quality control reports of the case 657's trio followed by the visualisation on the USCS Genome Browser.

Sample Name	% GC	≥ 30X	Median cov	Mean cov	% Aligned
case657_child	46%	22.8%	5.0X	24.1X	99.8%
case657_father	52%	31.2%	18.0X	27.3X	99.9%
case657_mother	52%	31.0%	18.0X	26.3X	99.8%

***Figure 1**: Sample Summary Table*
Summary of sequencing quality metrics for the trio (child, father, mother), including GC content, coverage , and alignment rate.

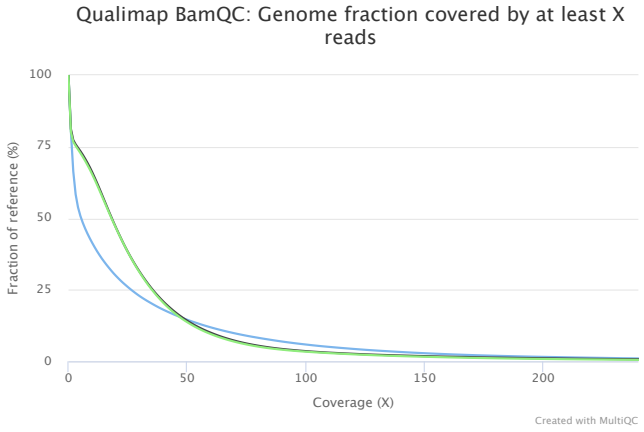

***Figure 2** - Qualimap BamQC: Genome Fraction Covered by ≥X Reads*
Cumulative coverage plot showing the fraction of the genome covered by at least X reads.
The child's curve, in blue, drops more steeply, indicating lower overall coverage depth compared to the parents.

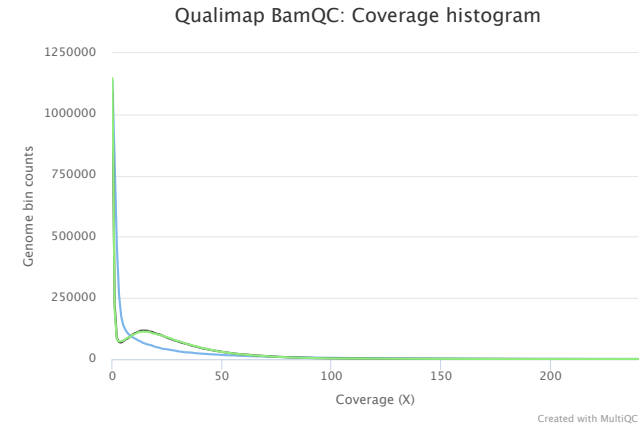

***Figure 3** - Qualimap BamQC: Coverage Histogram*
Distribution of coverage across the genome for each sample. The child exhibits a sharper peak at lower coverage.

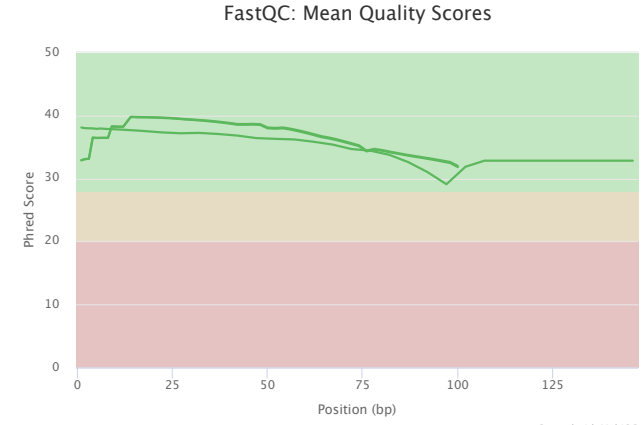

***Figure 4** - FastQC: Mean Quality Scores*
Base quality scores across read length for one or more samples. Quality remains high throughout most of the read, with a slight drop near the end

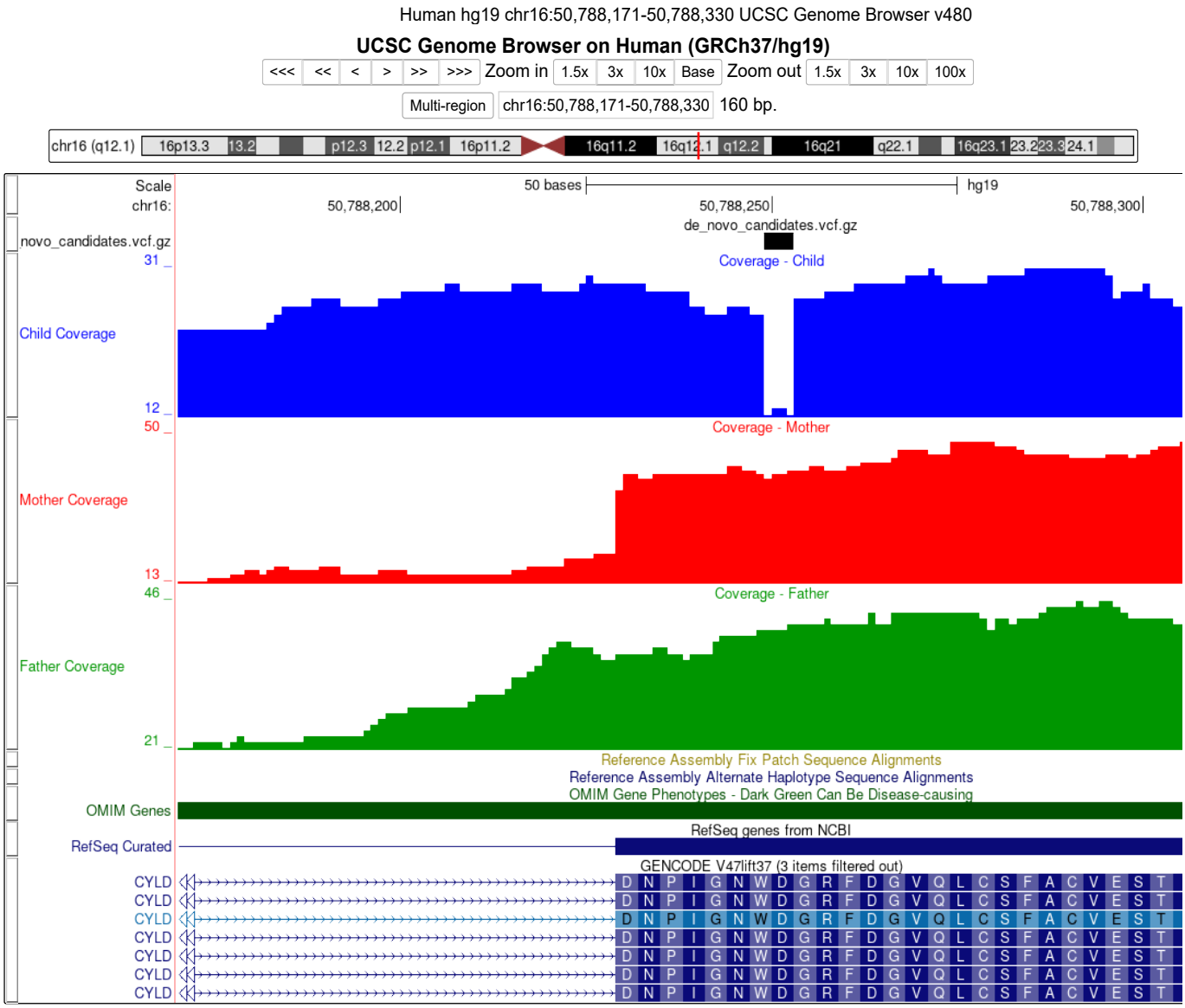

***Figure 5**: UCSC Genome Browser snapshot showing read coverage for trio 657 (child: blue; mother: red; father: green) aligned to the CYLD gene (chr16:50,788,249–50,788,253, GRCh37/hg19). The sharp coverage drop in the child (blue) reflects a de novo heterozygous 4-bp frameshift deletion (GGAT/–), absent in both parents. This pathogenic variant disrupts CYLD, consistent with Familial Cylindromatosis (OMIM: 132700), an autosomal dominant disorder caused by loss-of-function variants in CYLD. The coverage asymmetry arises from misaligned reads spanning the deletion.*

All the results presented in this study, along with the complete analysis pipeline and source code, are publicly availiable at: https://github.com/michelangelorosa/TrioAnalysis