Germline CNV Results — User Guide

Ultima Genomics Germline CNV Pipeline | UMGC

1. Overview

When the germline CNV pipeline finishes, the results consist of an interactive HTML report (one per sample) and several folders of per-sample data files. This guide describes each of them.

The output folder contains these subfolders, each with one file per sample:

- **bed/ —** final CNV calls — the recommended file for downstream analysis.
- **cnv_calls/ —** the same calls as a tab-separated table.
- **report/ —** interactive HTML report.
- **vcf/ —** the calls as a bgzipped VCF with full annotations.
- **raw/ —** the complete, unfiltered call set

Important: the report and the bed/, cnv_calls/, and vcf/ files contain HIGH-QUALITY calls only — those that passed the machine-learning quality filter (TREE_SCORE $\geq$ 31). Downstream analysis should focus on these high-quality calls.

Justification for focusing on high-quality calls

Read-depth CNV callers inevitably produce a large number of false positives, driven by sequencing noise, uneven coverage, low-mappability regions, and reference-genome gaps. The machine-learning filter is trained to separate genuine copy-number variants from these artifacts, so the high-quality set keeps the false-discovery rate low and the results interpretable and reproducible. Including low-quality calls would bury the real variants under noise and inflate the counts with artifacts — which is why they are excluded from the report and the standard output files.

The complete, unfiltered set (both high- and low-quality calls) is preserved in raw/<sample>.raw_dump.vcf.gz, from which every high-quality file can be regenerated (see section 6). This raw set is mostly noise — the large majority of its calls are low-quality artifacts — and is intended only for experienced programmers and bioinformaticians who need to apply a custom filtering strategy or examine specific borderline calls. It is strongly discouraged for routine downstream analysis; use it only if you understand the trade-offs and are prepared to do your own filtering and validation.

2. Interactive HTML Report

One report (<sample>.cnv_report.html) per sample is located in the report sub-folder. **All plots are interactive** (hover, zoom, click). It contains the following sections, in order: Metadata table, Genome Overview, CNV Confidence Score Distribution, CNV Size Distribution, Data, and Methods .

2.1 Metadata Table

A key/value panel at the top of the report. The keys are fixed; the values change per sample:

- **Project Name —** a free-text project label, set when the pipeline is launched.
- **Run Name —** a free-text run name label, also set when the pipeline is launched.
- **Sample —** the identifier derived from the input file name.
- **Sex —** male / female, determines which machine-learning model is used for chrX calls.
- **Deletions —** number of high-quality deletions.
- **Duplications —** number of high-quality duplications.
- **Total CNV calls —** total number of high-quality calls (deletions plus duplications).
- **CNV Report Guide —** Link to this document
- **Report Generated—** date and time the report was produced.

2.3 Genome Overview

An ideogram-style genome plot showing the high-quality CNV calls along all 22 autosomes plus the X and Y chromosomes. Deletions are red and duplications are blue.

Interactivity: hover over a call to see its chromosome, coordinates, length, type (DEL/DUP), confidence score, copy number, and calling source. Click a chromosome row to open a zoomed-in detail track for that chromosome; in the detail view, scroll to zoom and drag to pan. **Figure 1** below shows this detailed view.

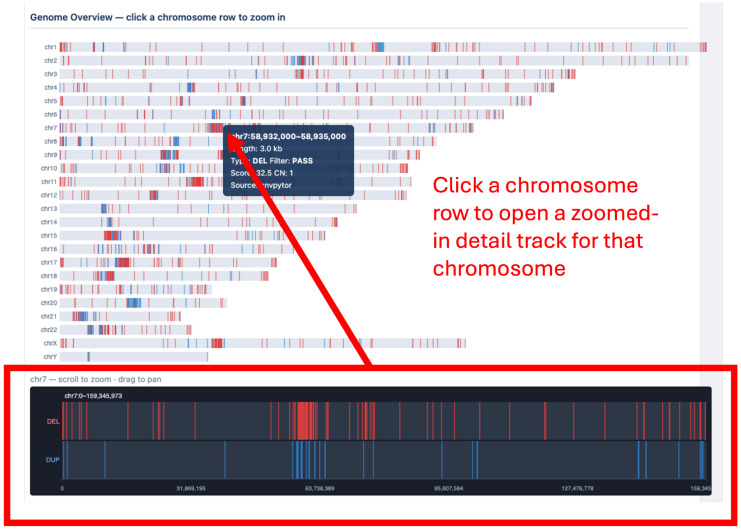

Figure 1: Zoomed in portion of the chromosome showing deletions and duplications. You can zoom in further to see detailed view of the CNVs

2.4 CNV Confidence Score Distribution

Each CNV call is scored from 0 to 100 by a machine-learning classifier (the "tree score") that estimates how confident the pipeline is that the call is a true variant. Higher is better: calls scoring 31 or above pass the filter (PASS, green), and those below are flagged LOW_SCORE (orange). The dashed line marks the threshold. This is the one plot that shows every call — both kept and filtered out — giving a quick visual quality check.

Interactivity: Hover over a bar to see its exact count. Click a legend entry (PASS or LOW_SCORE) to hide or show that series; double-click a legend entry to isolate it. Drag a box on the plot to zoom in, and double-click the plot to reset. A toolbar appears at the top-right on hover with zoom, pan, reset-axes, and download-as-PNG buttons.

2.5 CNV Size Distribution

Counts of the high-quality CNV calls are grouped into six size bins (<1 kb, 1–5 kb, 5–25 kb, 25–100 kb, 100 kb–1 Mb, >1 Mb), with deletions and duplications shown side by side. Useful for confirming the size profile of the calls — most germline CNVs are smaller than 100 kb.

Interactivity: Hover over a bar to see its exact count. Click a legend entry (DEL or DUP) to hide or show that series. Drag a box on the plot to zoom in, and double-click to reset. A toolbar appears at the top-right on hover with zoom, pan, reset-axes, and download-as-PNG buttons.

2.6 Data

A short section listing the output files that accompany the report — one file per sample in each of bed/, cnv_calls/, report/, vcf/, and raw/. These files are described in detail in sections 3–6 below.

2.7 Methods

Analysis was done using the published Ultima Genomic germline CNV caller modified to run on a HPC cluster. A one-paragraph summary of how the calls were produced: per-base coverage is normalized against a reference cohort and segmented with cn.mops (v1.55.0); read-depth CNVs are independently called with CNVpytor (v1.3.1) at 500 bp and 2500 bp resolution; calls from both callers are merged; breakpoints are refined by local re-alignment of split reads (jalign, ugbio_cnv v1.24.0); each candidate is scored by a machine-learning classifier (high-quality at TREE_SCORE $\geq$ 31); and redundant overlapping calls are collapsed into the final set.

3. bed/ — CNV BED Files

Filename: <sample>.cnv.bed. One file per sample, containing the high-quality calls. This is the recommended output for downstream analysis — annotation, intersection with gene lists, IGV visualization, etc.

The file is a standard BED with chrom / start / end in the first three columns and a semicolon-delimited INFO field in the fourth column. The INFO field carries the following keys:

Key	Meaning
SVTYPE	DEL (deletion) or DUP (duplication).
FILTER	PASS — every call in this file passed the quality filter. (The complete set, including low-quality calls, is in raw/.)
TREE_SCORE	Machine-learning confidence score (0–100). Higher is better; the high-quality threshold is 31.
SVLEN	Variant length in base pairs (always positive).
CNV_SOURCE	Caller(s) that produced the call: cn.mops, cnvpytor, or both.
CN	Estimated absolute copy number (e.g., 0, 1, 3, 4). "." if not estimated.

BED files are coordinate-sorted and use 0-based half-open intervals (standard BED convention).

4. cnv_calls/ — Tabular CNV Calls

Filename: <sample>.cnv_calls.tsv. One file per sample — the same high-quality calls as the BED file, with each INFO field broken out into its own column. Designed for easy loading into Excel, pandas, R, or any spreadsheet tool — no INFO-string parsing required.

Columns (header row included):

Column	Meaning
sample	Sample name.
chrom	Chromosome (chr1 … chr22, chrX, chrY).
start	Start coordinate (0-based).
end	End coordinate (half-open).
svtype	DEL or DUP.
filter	PASS (every call in this file is high-quality).
tree_score	Machine-learning confidence score (0–100, high-quality ≥ 31).
svlen	Variant length in base pairs.
source	Caller(s): cn.mops, cnvpytor, or both.
cn	Estimated copy number. "." if not estimated.

5. vcf/ — VCF Calls

Filename: <sample>.mrg.vcf.gz. One bgzipped VCF per sample — the high-quality call set in VCF 4.2 format, suitable for use with bcftools, GATK, IGV, or any standard VCF tool.

Compared to the BED file, the VCF retains additional caller-level metadata (sub-source records, breakpoint-refinement evidence, jalign split-read counts, raw caller scores) in standard VCF INFO fields. Use the VCF when you need full provenance for each call; use the BED or TSV when you need a flat, ready-to-analyze table. Key INFO fields:

- **SVTYPE —** DEL or DUP.
- **END —** variant end coordinate.
- **SVLEN —** variant length.
- **CNV_SOURCE —** caller(s) that produced the call.
- **TREE_SCORE —** machine-learning confidence score (also reproduced in QUAL).
- **CN —** estimated absolute copy number, where computed.

A companion .tbi index is not produced by default; create one with "tabix -p vcf <file>" if needed.

6. raw/ — Complete Unfiltered Calls

Filename: <sample>.raw_dump.vcf.gz. One bgzipped VCF per sample containing the COMPLETE, unfiltered call set — both high-quality (PASS) and low-quality (LOW_SCORE) — with all INFO annotations. Use this when you need the full results, including the calls the quality filter removed: for example, to apply a different stringency, or to inspect borderline calls near the threshold. To recreate the high-quality VCF:

```
bcftools view -f PASS -Oz -o <sample>.mrg.vcf.gz raw/<sample>.raw_dump.vcf.gz
```

The bed/ and cnv_calls/ files are projections of that high-quality VCF, so filtering the raw VCF to PASS reproduces all of them.

7. Which File Should I Use?

Output	Best for
report/<sample>.cnv_report.html	Visual QC of a single sample — confidence-score, size profile, and genome distribution.
bed/<sample>.cnv.bed	Downstream analysis of the high-quality calls: annotation, intersection, IGV (recommended).
cnv_calls/<sample>.cnv_calls.tsv	Spreadsheet-friendly tabular high-quality calls.
vcf/<sample>.mrg.vcf.gz	High-quality calls with maximum metadata, for standard VCF tooling.
raw/<sample>.raw_dump.vcf.gz	The complete unfiltered set; reproduce any filtered file, or analyze low-quality calls.