



# Copy Number Private Investigator (CNPI): making rare copy number changes usable at scale

## At a glance

- CNPI turns genome-wide copy-number estimates into actionable summaries (genes/regions, karyotype, chromosomal sex, QC, and plots).
- It is especially useful for finding large deletions/duplications (large CNVs) quickly.
- ICNS provides a “PRS-like” rare CN burden score to compare individuals and cohorts.
- Designed for cohort-scale pipelines (cloud or HPC).
- Best first step: run the workflow once using the repo scripts, then scale to batches / cluster execution
- On-premise compute: start with the GitHub repository.
- Cloud compute: Start with the CAVATICA demo, then apply to your cohort.

## 1) What problem CNPI solves

Whole-genome sequencing (WGS) is now common in research cohorts but copy number changes are still hard to operationalize at scale. Most pipelines focus on small “letter” changes in DNA; CNPI focuses on larger events where DNA segments are missing (deletions), repeated (duplications), or whole chromosomes are present in unusual counts (aneuploidy).

At cohort scale, teams repeatedly run into the same issues: copy number workflows can be slow or manual, outputs vary across tools, and it is hard to turn results into simple signals for QC, triage, and cohort comparisons.

## 2) A quick analogy

Think of the genome as a book. Single-letter variants are like typos. Copy-number changes are like missing pages (deletions), duplicated chapters (duplications), or extra/missing volumes (aneuploidy). CNPI helps you quickly answer: “What is missing or duplicated, where, and how unusual is this sample compared with others?”

## 3) What you need / What you get

What you need	What you get
<ul style="list-style-type: none"> <li>• Genome-wide copy-number estimates (per sample)</li> <li>• A region list (BED): genes / panels / intervals</li> <li>• Same genome build for both inputs</li> <li>• Optional: a reference cohort (for ICNS percentiles)</li> </ul>	<ul style="list-style-type: none"> <li>• Region CN summaries (per gene/interval)</li> <li>• Digital karyotype + chromosomal sex check</li> <li>• Plots</li> <li>• ICNS: “PRS-like” rare CN burden/outlier score</li> </ul>

## 4) What CNPI outputs mean

### A) Region summaries (genes/panels/intervals)

CNPI produces standardized copy number summaries for each region you care about (often genes). This makes copy number data easy to join with phenotype tables, dashboards, and downstream analysis.

## B) Digital karyotype + chromosomal sex assessment

CNPI provides chromosome-level summaries that help quickly spot broad chromosome shifts and basic QC anomalies (for example, unexpected sex chromosome patterns).

## C) Plots (big deletions / duplications)

CNPI is especially helpful for quickly identifying large copy number variants (CNVs): big deletions or duplications that span many regions. The digital karyotype can highlight chromosome-scale shifts, while plots help you spot and prioritize large events for review. In practice, this supports fast triage: “Is there a big event here, and where should we look first?” These can further be plotted with CNPI.

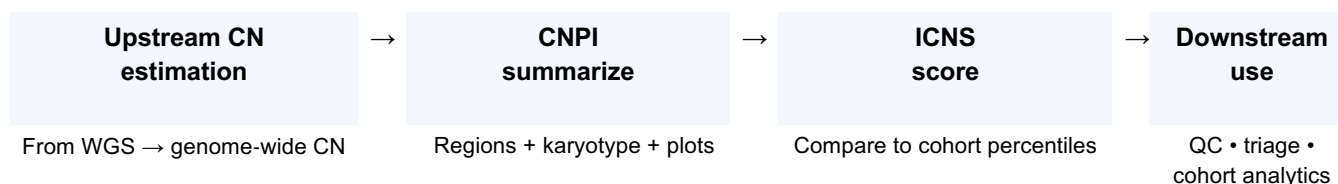
## D) ICNS: “PRS-like” scoring for rare copy number (key differentiator)

ICNS summarizes how unusual a person’s copy number profile is relative to a reference cohort. Like a polygenic risk score (PRS) aggregates many small signals into one number, ICNS aggregates rare copy number outlier burden across many regions into a single metric, with optional chromosome-level attribution.

### How to interpret ICNS

- Higher ICNS = more unusual copy number profile vs the reference cohort (useful for triage and cohort comparison).
- ICNS is best viewed as a burden/outlier score: not automatically a disease-risk probability.
- ICNS is reference-dependent: results are meaningful relative to how the reference cohort percentiles were built and QC’d.

## 5) Simple workflow diagram



## 6) Try it first

Checkout our GitHub to find the latest version of CNPI and steps for running it. If you work in the cloud, the fastest way is to run the CAVATICA demo end-to-end, then inspect the outputs: region summaries, a karyotype/sex overview, plots, and ICNS scores. After that, swap in your cohort inputs to reproduce the same workflow in your environment.

- GitHub / documentation: <https://github.com/tnturnerLab/cnpi>
- CAVATICA demo: <https://cavatica.sbgenomics.com/u/sevenbridges/quick-mer2-and-cnpi>
- Contact for collaboration: [tychele@wustl.edu](mailto:tychele@wustl.edu)

## 7) FAQ

**Q: What is copy number?**

A: How many copies of a DNA region a person has. Most regions are about 2 copies; deletions lower that and duplications raise it.

**Q: What is aneuploidy?**

A: A whole chromosome is present in an unusual count (extra or missing copies). The digital karyotype is designed to spot this quickly.

**Q: What does a high ICNS mean?**

A: A high ICNS means the sample's CN profile is more unusual relative to the reference cohort percentiles. It is useful for triage and cohort comparison, not automatically disease risk.

**Q: Do I need ICNS to get value?**

A: No. Many teams start with region summaries + karyotype/chromosome sex checks. ICNS adds a cohort-comparable burden feature.

**Q: What are the most common pitfalls?**

A: Genome build mismatch, noisy upstream CN estimates, and building ICNS references from cohorts that include strong outliers or batch artifacts.

## 8) Responsible use

- CNPI summarizes the CN estimates it is given, input quality matters.
- Ensure CN estimates and BED regions use the same genome build.
- ICNS is reference-dependent; interpret scores relative to how the reference percentiles were built and QC'd.
- Treat ICNS as a burden/outlier score unless calibrated and validated for a specific phenotype.

## References

- Ustanik J, Turner TN. CNPI: Rapid Analyses of Human Copy Number Data. Journal of Molecular Biology (2025). Oct 1;437(19):169313. doi: 10.1016/j.jmb.2025.169313. PMID: [40588120](#).
- CNPI repository: <https://github.com/tnturnerLab/cnpi>.