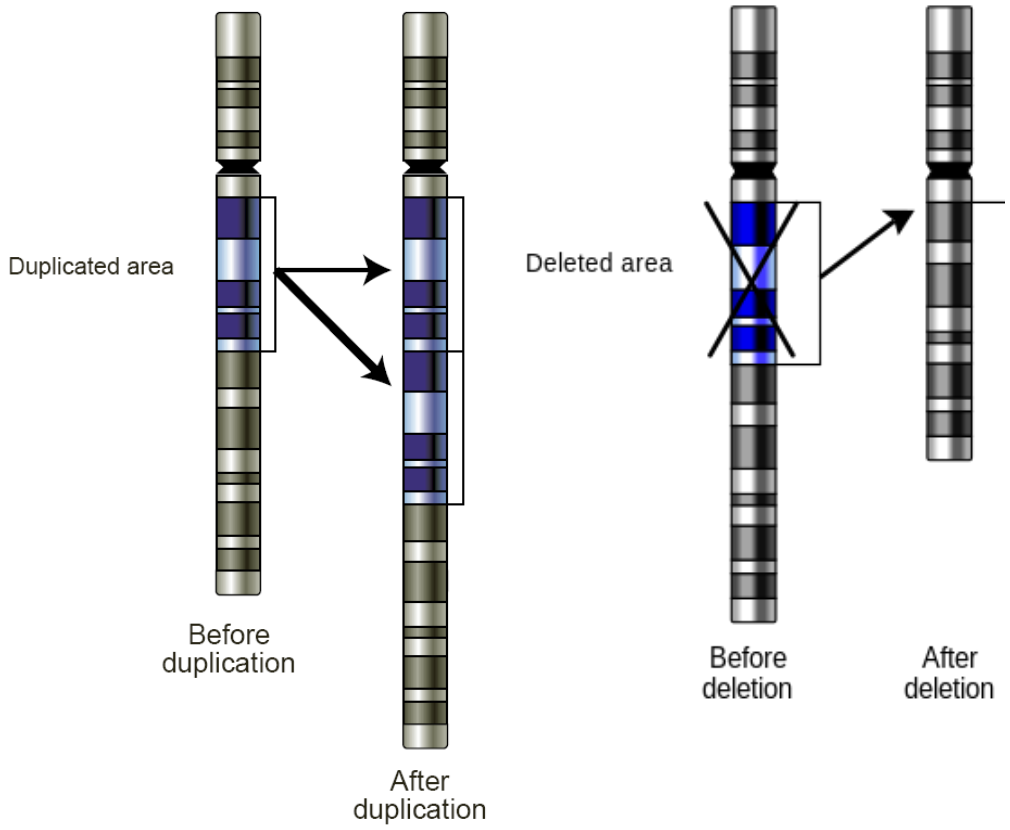
From Fixed Loci to Genome Wide

Automated CNVs validation, development and application in UKB

Simone Montalbano

What is a CNV?

- CNVs are a class of **structural variants**
- They consist of **deletions** and **duplications**
- The size spans **from 25-50kbp to 5-10Mbp**
- **Large** and (usually) **rare** variants
- CNVs are a classic focus of human genetic research



Leading Edge
Review
Cell

CNVs: Harbingers of a Rare Variant Revolution in Psychiatric Genetics

Dheeraj Malhotra[1,2] **and Jonathan Sebat**[1,2,3,4,*]
[1]Beyster Center for Genomics of Psychiatric Diseases
[2]Department of Psychiatry
[3]Department of Cellular Molecular and Molecular Medicine
[4]Institute for Genomic Medicine
University of California, San Diego, La Jolla, CA 1020103, USA
*Correspondence: jsebat@ucsd.edu
DOI 10.1016/j.cell.2012.02.039

How are CNVs detected?

- Genotype data from **SNP arrays** is the standard
- All methods rely on **DNA intensity measure** (LRR / read depth)
- Some also integrate a measure of the **allelic composition** (BAF for Illumina arrays)
- **PennCNV** is still considered the golden standard

CURRENT PROTOCOLS

PROTOCOL | Open Access

Accurate and Effective Detection of Recurrent Copy Number Variants in Large SNP Genotype Datasets

Simone Montalbano, Xabier Calle Sánchez, Morteza Vaez, Dorte Helenius, Thomas Werge, Andrés Ingason

First published: 05 December 2022 | https://doi.org/10.1002/cpz1.621

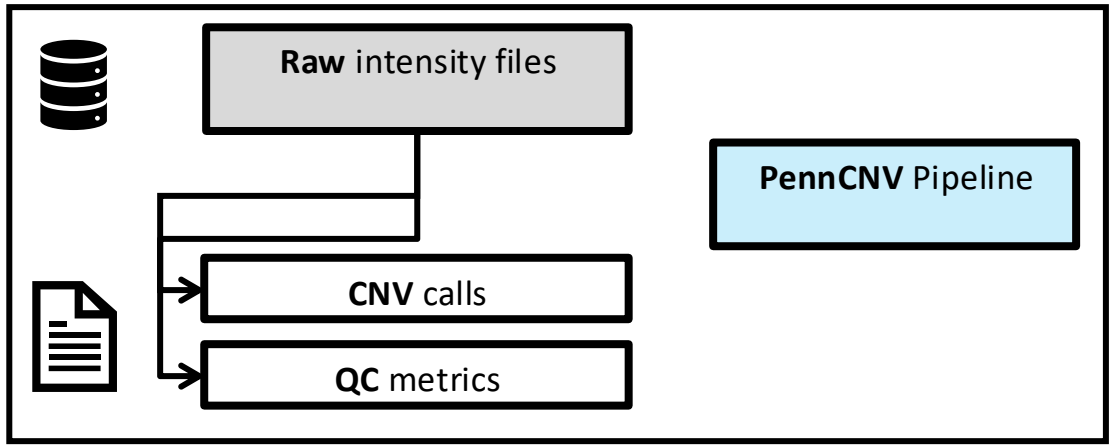
Published in the Bioinformatics section

Methods

PennCNV: An integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data

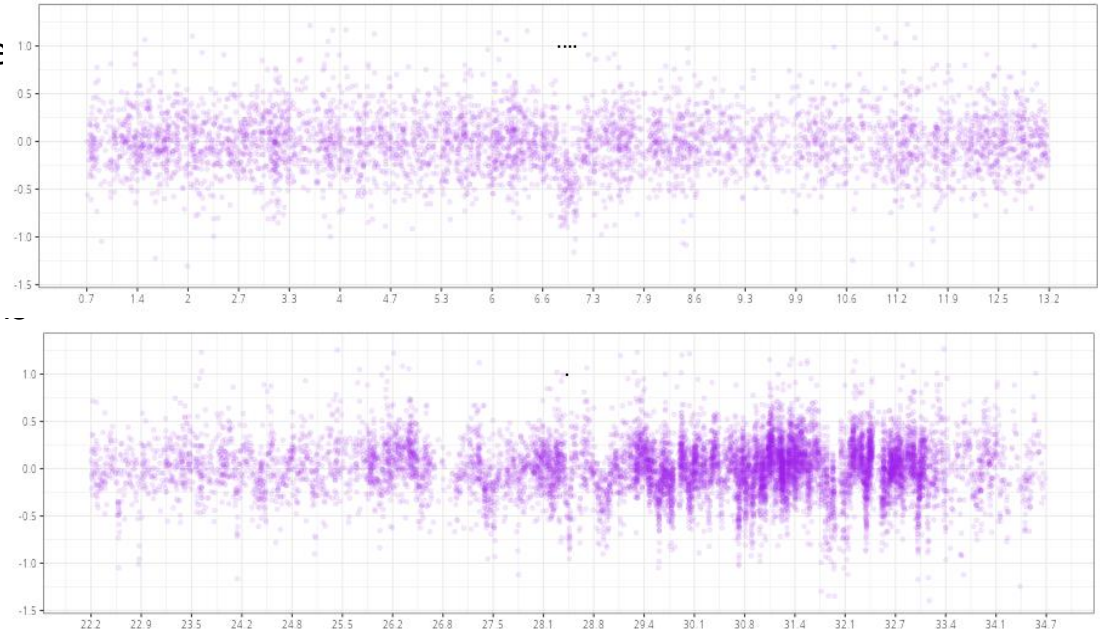
Kai Wang,[1] Mingyao Li,[2] Dexter Hadley,[1,3] Rui Liu,[1] Joseph Glessner,[4] Struan F.A. Grant,[4] Hakon Hakonarson,[4] and Maja Bucan[1,5]

[1]Department of Genetics, University of Pennsylvania, Philadelphia, Pennsylvania 19104, USA; [2]Department of Biostatistics, University of Pennsylvania, Philadelphia, Pennsylvania 19104, USA; [3]Department of Biology, University of Pennsylvania, Philadelphia, Pennsylvania 19104, USA; [4]Center for Applied Genomics, Children's Hospital of Philadelphia, Philadelphia, Pennsylvania 19104, USA



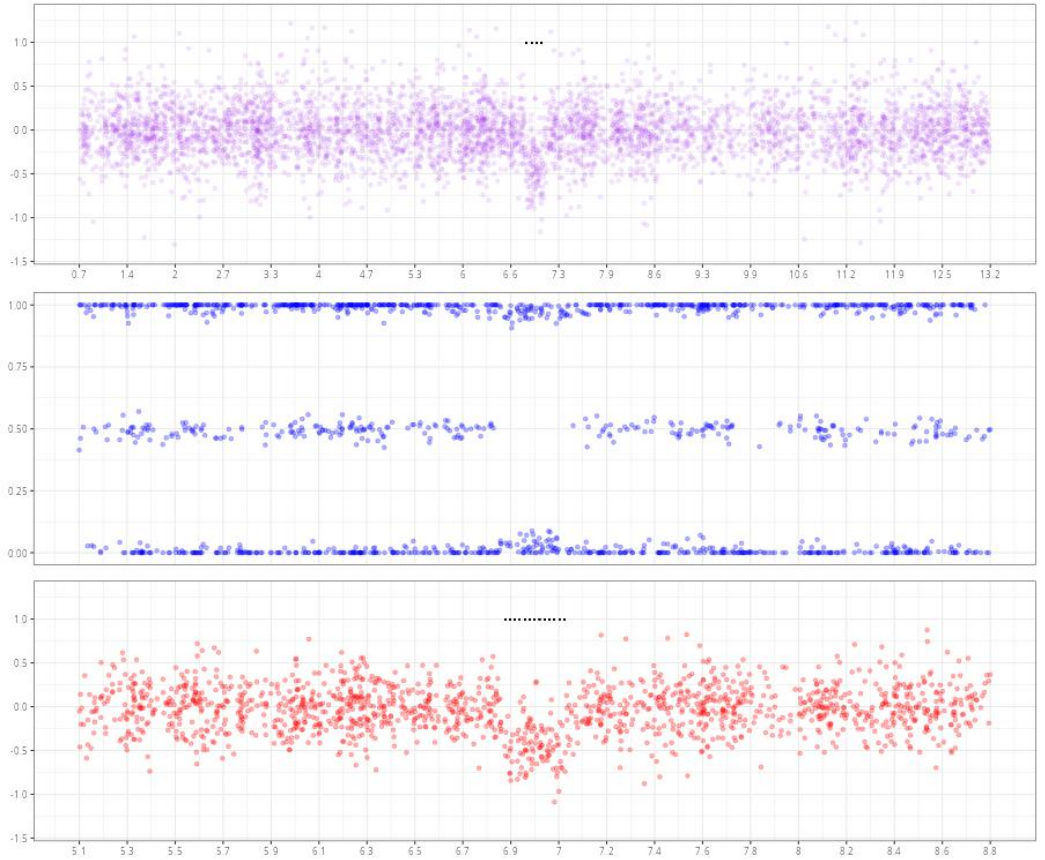
CNV calling limitations

- PennCNV uses an **HMM** to predict the state of each marker based of the previous one
- It builds larger CNVs from smaller sets of consecutive SNPs with the same predicted Copy Number
- This makes it **very sensitive to local noise**
- **LRR is not** always **stable** across the genome, even without CNVs, th… is partially related to **GC content**
- Two **main limitations** of CNV calling:
  - **Over segmentation** (easy to solve, CNV **stitching**)
  - **False positives** (not so easy to solve, QC filtering + **visual validation**)



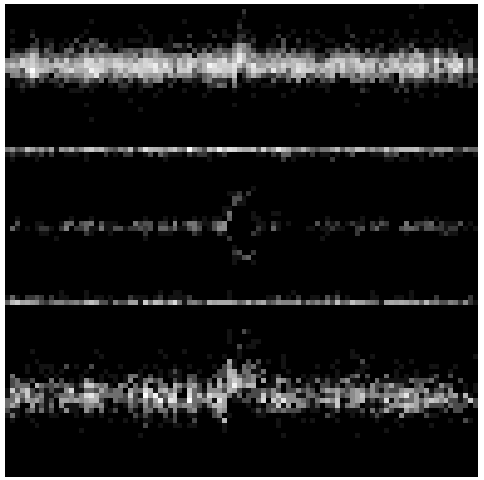
PennCNV false positives

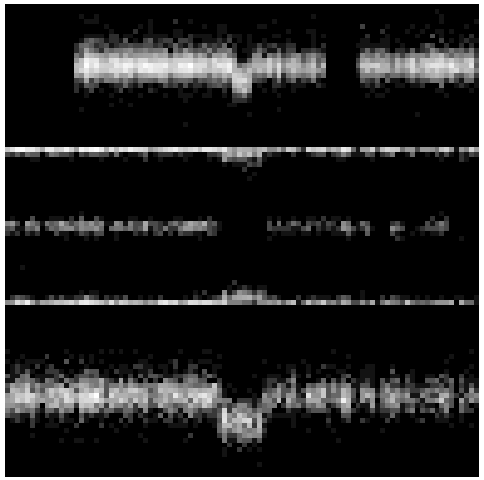
- **From 30% to 60%** of PennCNV calls are **false positives**
- This is only **partially** solved by standard filtering
- The only effective solution is **visual inspection** of the raw data trends for each CNVS call
- Extremely **time consuming** and not feasible genome wide

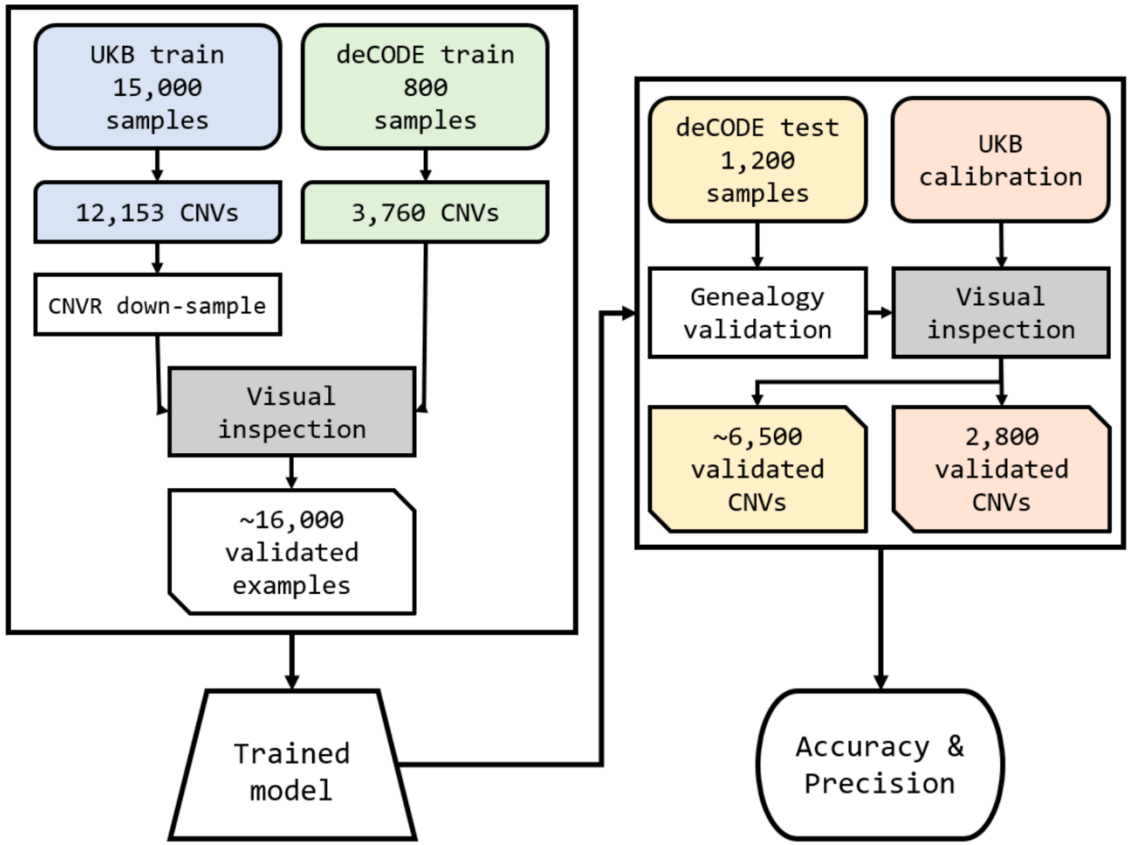


Our solution: automated visual validation

- **~16,000** human labelled **examples** of true and false CNV calls (the hardest task)
- **Simplified the image** for the computer
- Trained a relatively simple **Convolutional Neural Network** with **R torch**

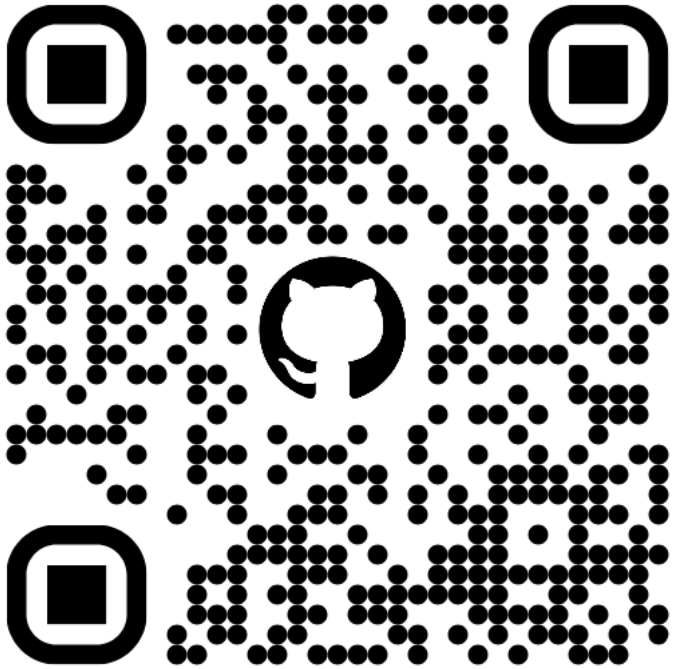






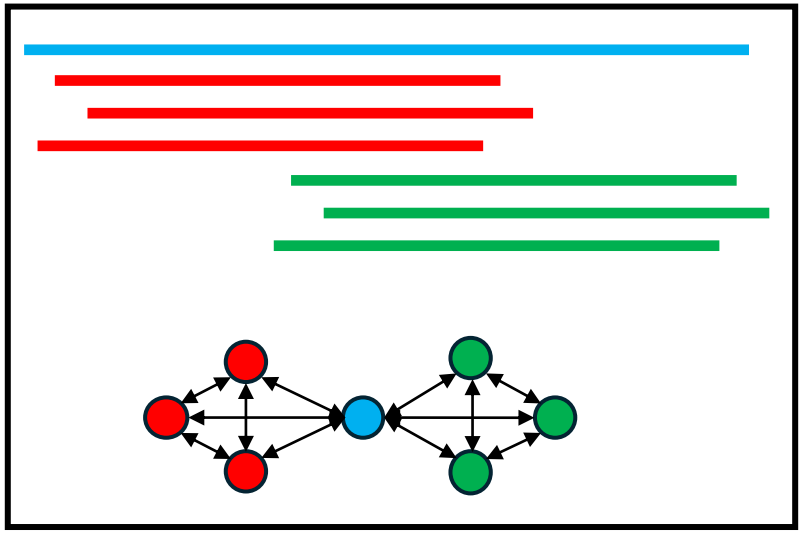
CNValidatron accuracy and R package

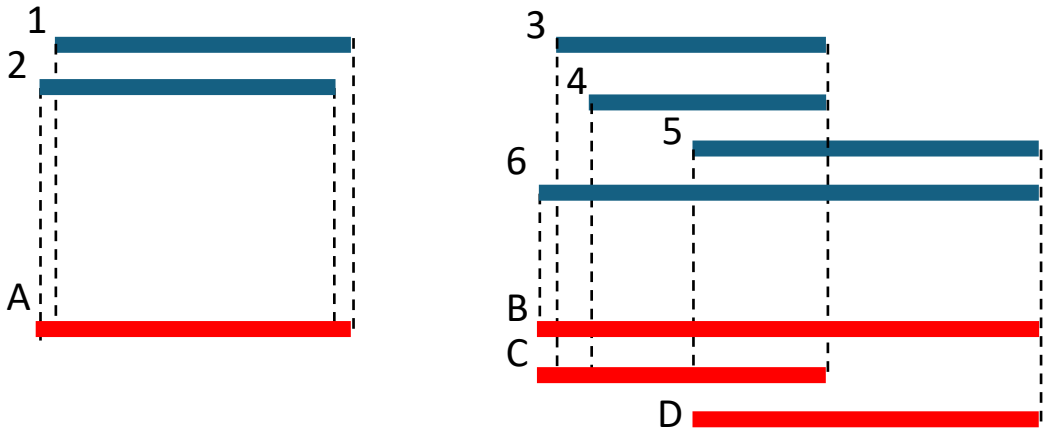
- Tested both in **UKB** and **deCODE** data
- **Accuracy** and precision are good, well **above 90%**
- Comparable to a trained human
- Fully achieved our primary goal of **reducing** the burden of **false positive** CNV calls
- The code is available on GitHub at https://github.com/SinomeM/CNValidatron_fl
- **BioRxiv** coming very soon



CNValidatron application, Genome Wide CNVs in the UKB

- **~525,000** validated CNVs from **~462,000** samples
- CNVs can be **complex** to analyse
- The easiest way to tackle the problem is to create **regular bins**/windows
- Each bin is a **marker**, each sample is a carrier if they have a CNV overlapping the bin

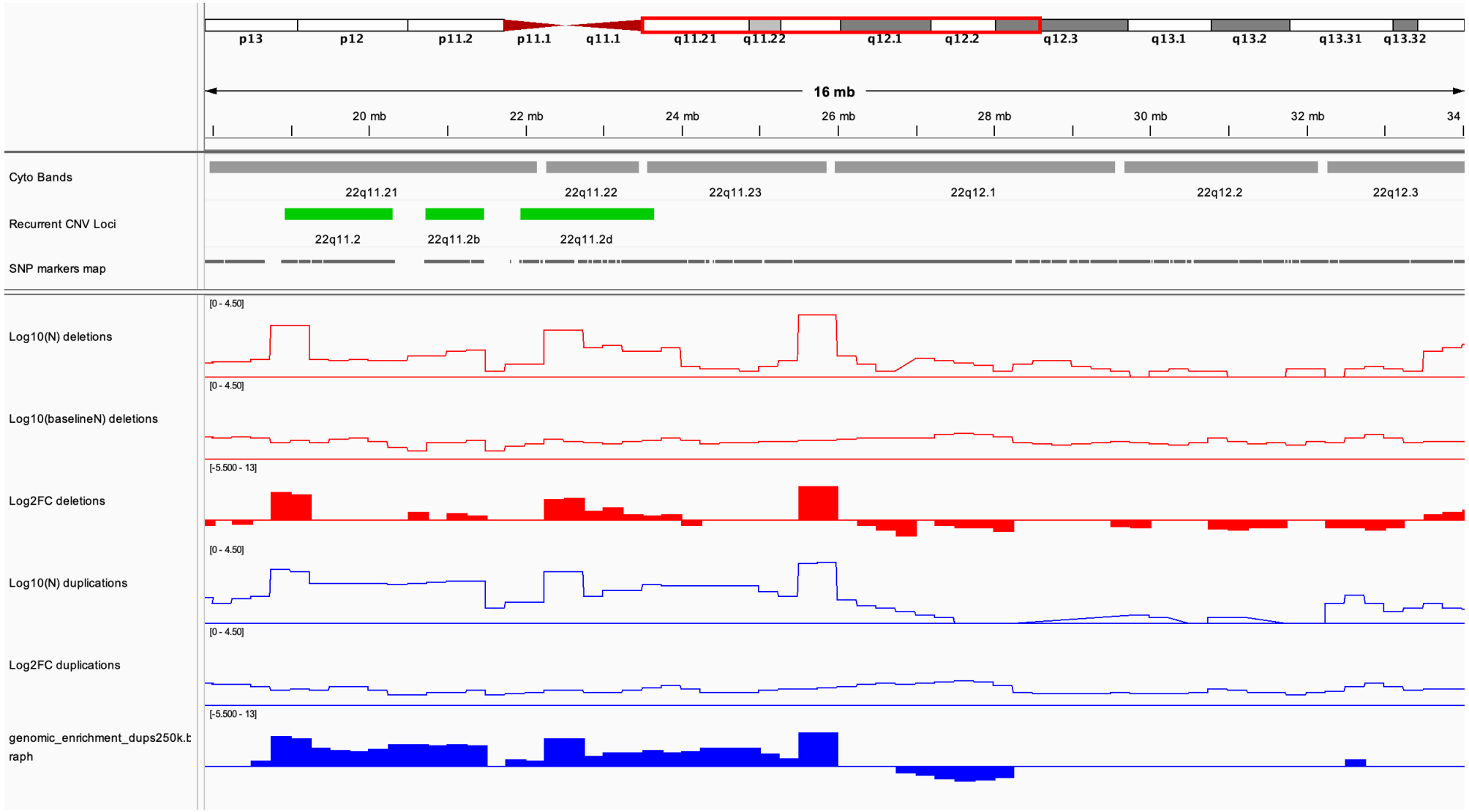




Given a set of **segments** (blue) which **grouping** (red) is better?

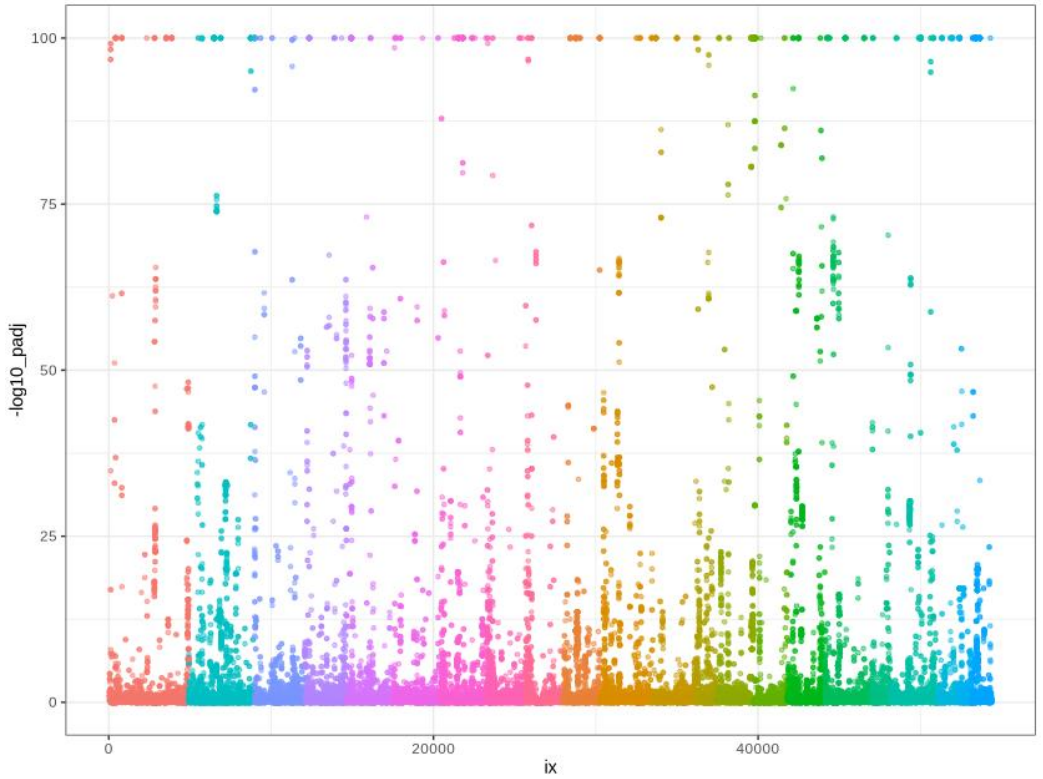
A,B | A,B,C | A,B,C,D ?

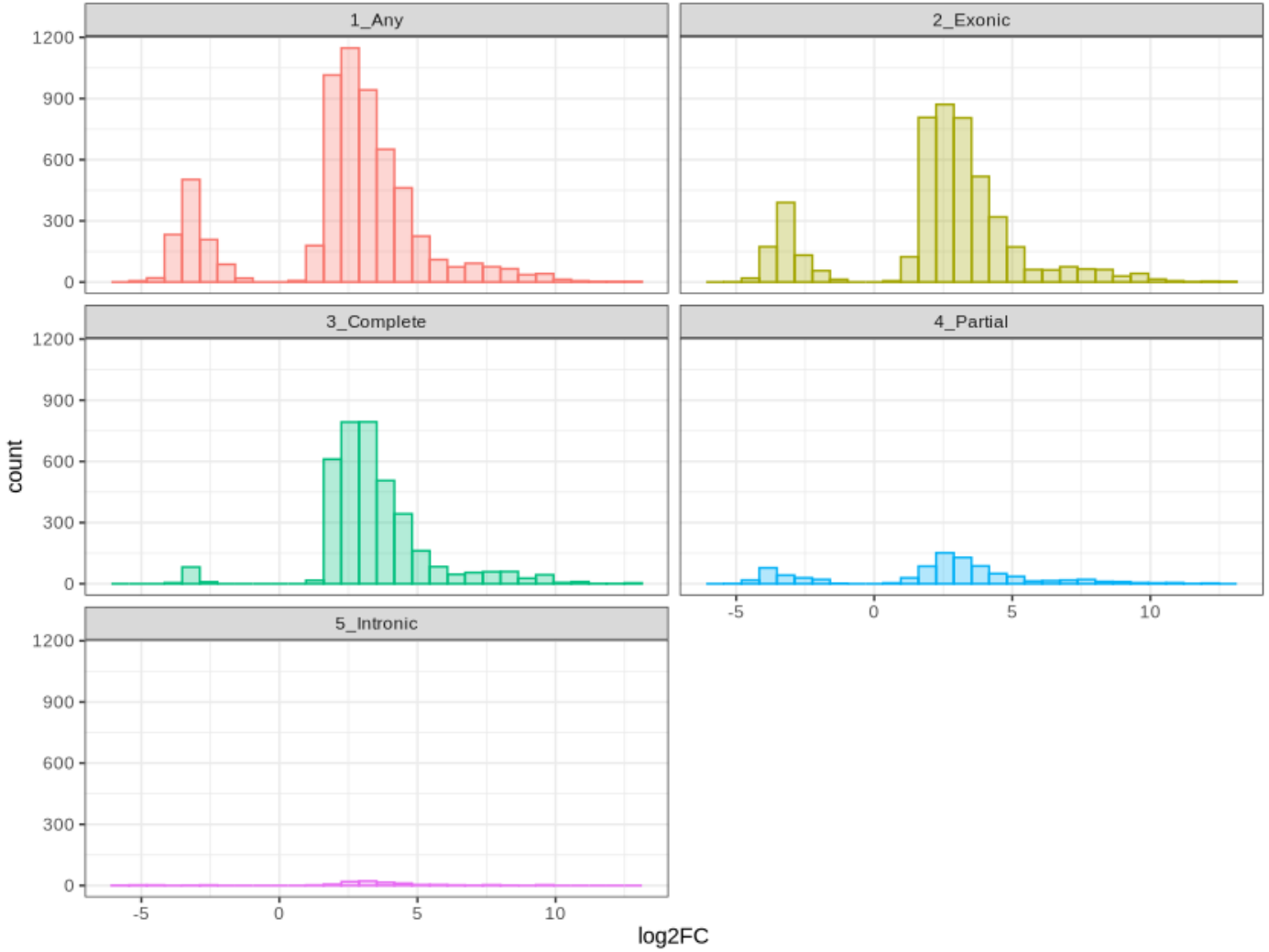
CNValidatron application, Genome Wide CNVs in the UKB



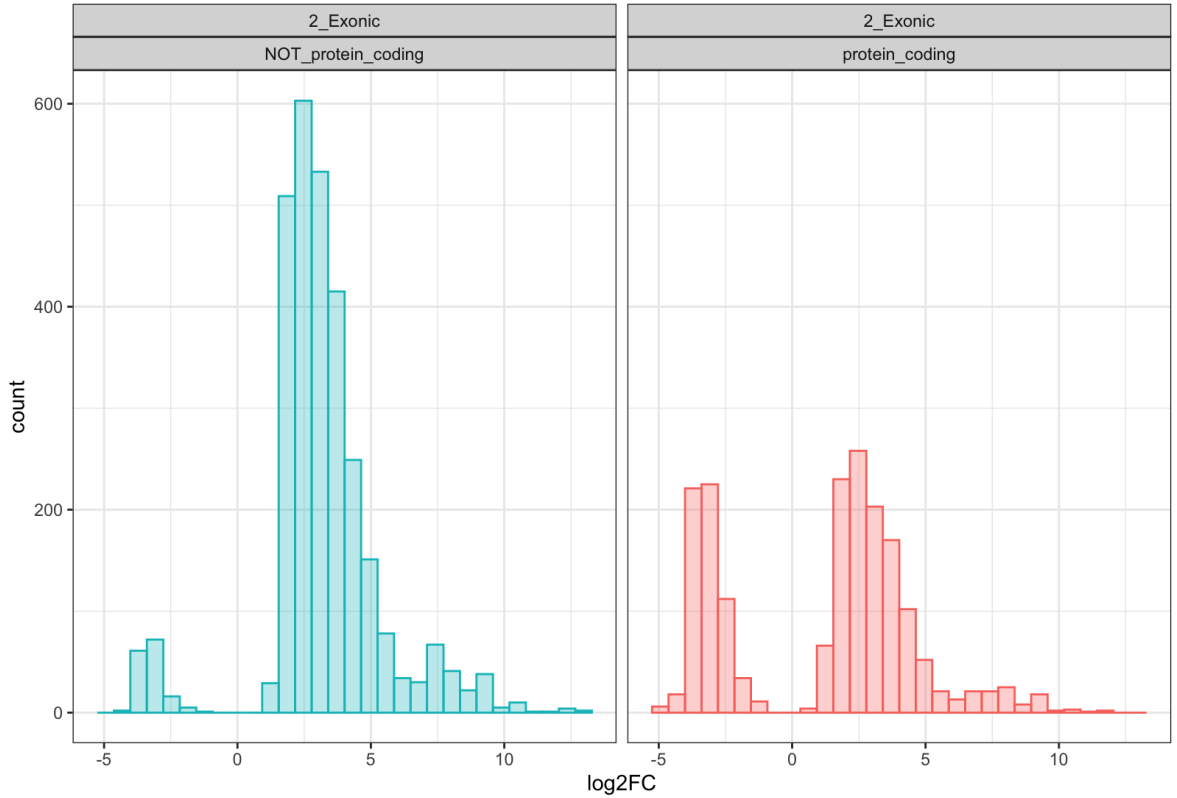
GW CNVs and Genes

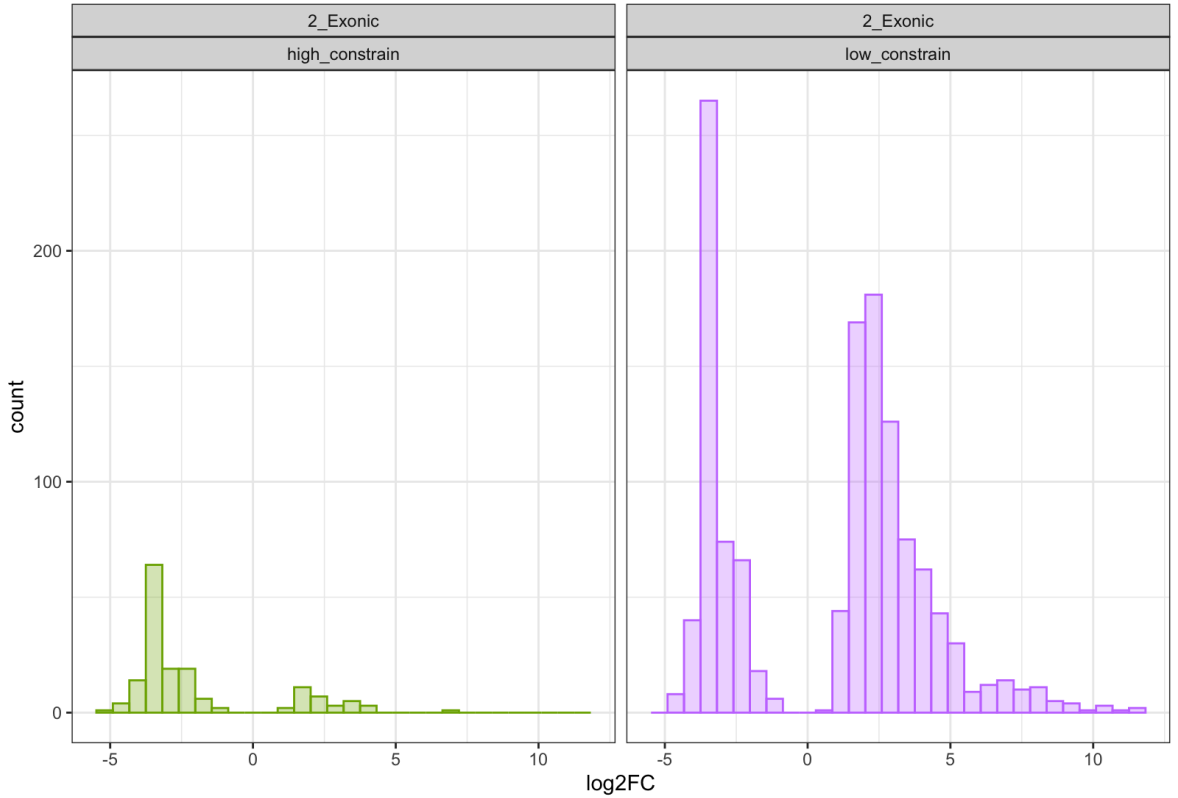
- What about **genes**?
- We can use them as bins and do the same





GW CNVs and Genes





Conclusions and Next Steps

Conclusion:

- CNV calling from **SNPs arrays** is still relevant (for large collections)
- The main limitation is **high false positives** rate
- We propose a software to **automate visual validation** of CNV calls
- Trained and tested on **UKB + deCODE** data (multiple arrays from the two main manufacturers)
- Accuracy is high and comparable to a trained human analysist

Next steps:

- **Human traits association**, already started in UKB
- Application in **other large biobanks**?

Thanks for your attention!

INSTITUTE OF BIOLOGICAL PSYCHIATRY

iPSYCH

uk biobank

Hreinn Stefansson & Bragi Walters

(deCODE genetics)

Thomas Werge & Andres Ingason (IBP)