*By Cecile Poulain and Florian Benitiere*
*01/09/2025*

# Recurrent CNV identification

Regions of Interest (ROIs) for recurrent CNVs (rCNVs) have been identified primarily using [Clinical Genome Resource](https://www.clinicalgenome.org/), including all documented recurrent CNVs, with the addition of specific genes of interest. Some regions were manually curated.

For each rCNV, we identified a geneset by looking at which transcripts fall within its boundary (>50 % overlap). We considered only transcripts and genes meeting the following conditions:

- **Canonical transcripts only**
- **Protein-coding genes**
- **Variants/transcripts overlapping >50%** with the CNV region (OverlapPC > 50)
- **Transcripts not overlapping problematic regions** as defined by UCSC (overlap <50%)

To validate our rCNV identification (labelled : *new_data)*, we used UKBB Cohort for which two others paper worked on computing rCNV frequency in the population :
- Kendall et al. 2019
- Crawford et al. 2019

Also, we compared it to a previous complex, and case specific, identification method used in the lab (labelled : *historic_lab*).

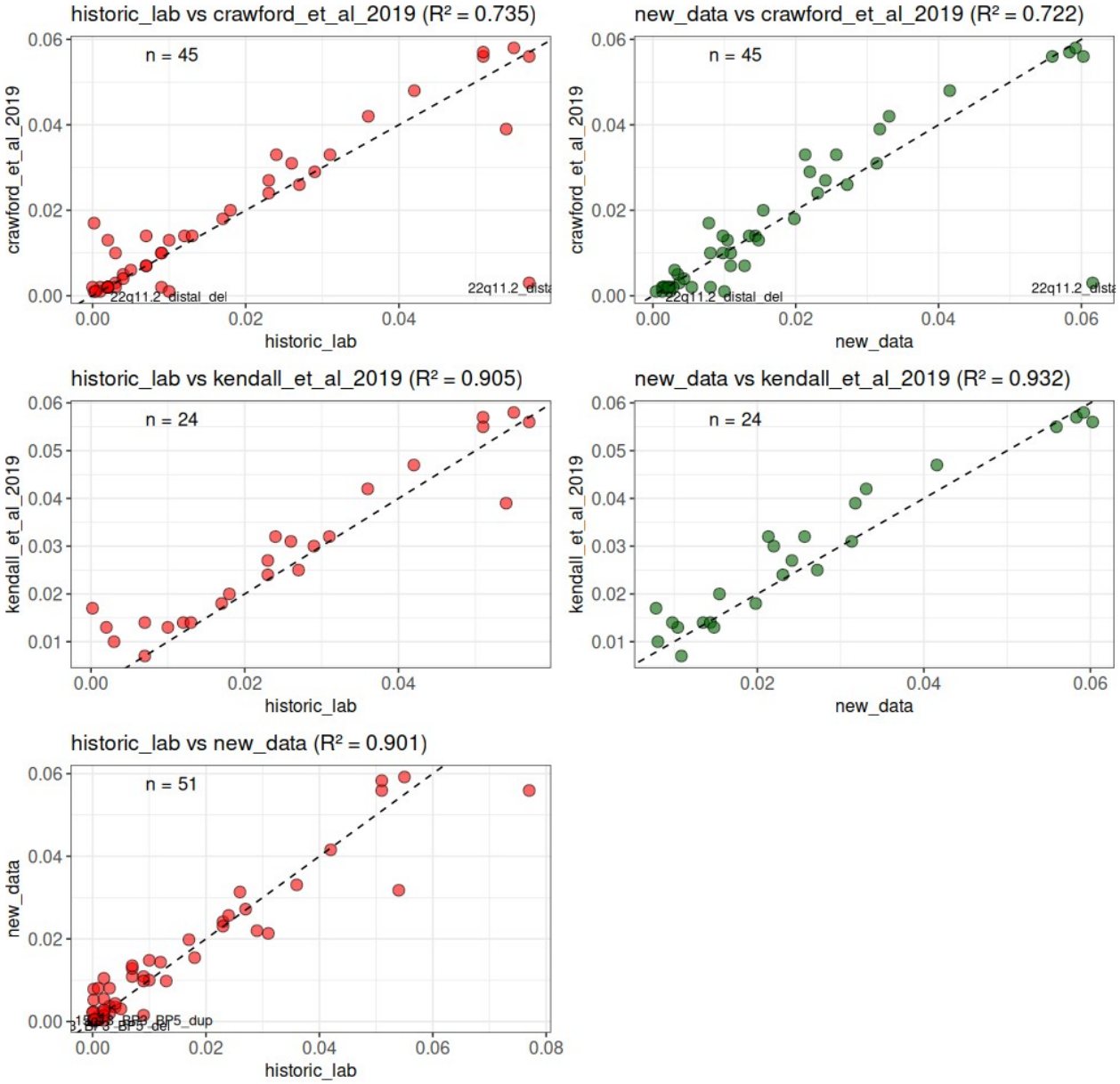
The results in the following figures show that the relation (R2) is better with the new method.

There is a specificity with the 22q11.2 when comparing with Crawford et al. explained by Cecile Poulain :
*Crawford presented only CNVs (fewer than 15) spanning from breakpoints D to F. « In this study, we aggregated all the CNV subtypes between breakpoints D and G (Mikhail et al. 2014). »*

*It's simply that the definition of 22q11.2 is not the same as the one given in the article by Kendall & Crawford.*

**Conclusion : The new method is simpler and seems to provide slightly better results.**

## Frequency < 0.06



## Frequency > 0.06