UC San Diego

Robust and accurate estimation of copy number for duplicated genes using WGS

Timofey Prodanov
Vikas Bansal

University of California San Diego

ISMB/ECCB July 2021

Segmental duplications

Long segments of duplicated DNA, usually low copy number.

Duplications longer than 3 kb with seq. similarity ≥ 97%:

- cover over 100 Mb
- cover ~ 1000 protein-coding genes, of them 120 disease-associated.

Problematic for short-read sequencing:

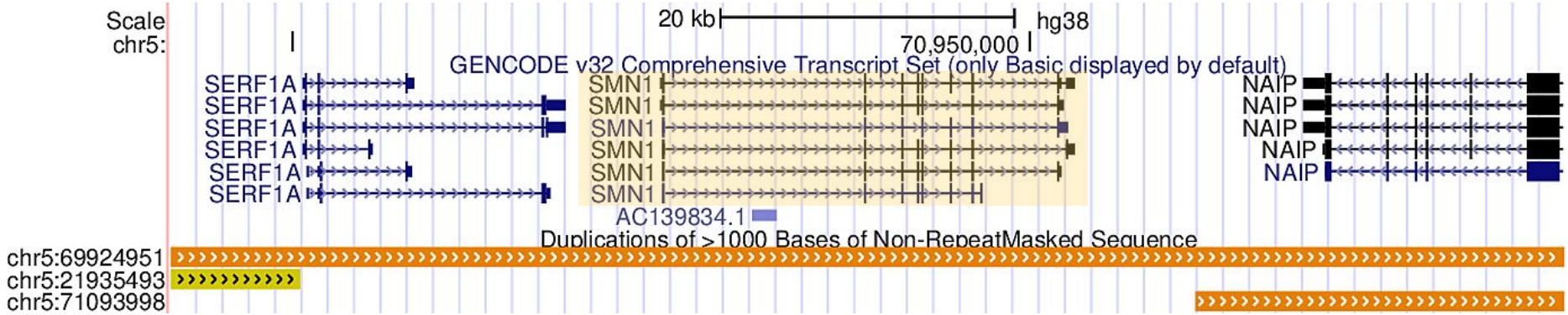
- Reads have ambiguity in alignment,
- Even high mapping quality alignments can be incorrect,
- CNVs and SNVs are difficult to identify.

SMN1/2 duplicated gene

Two copies: SMN1 and SMN2 lie within 205 kb duplication with 99.8% seq. similarity.

Point mutations and copy number changes cause **spinal muscular atrophy** (SMA).

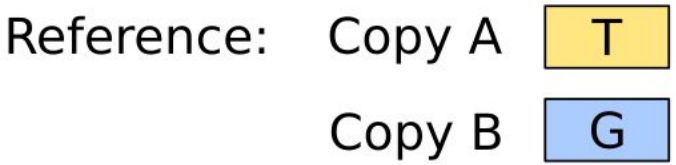
Paralog-specific copy number matters: psCN = 2,1 is healthy, psCN = 1,2 can be affected.



Paralogous Sequence Variants (PSVs)

PSV – small sequence difference between repeat copies.

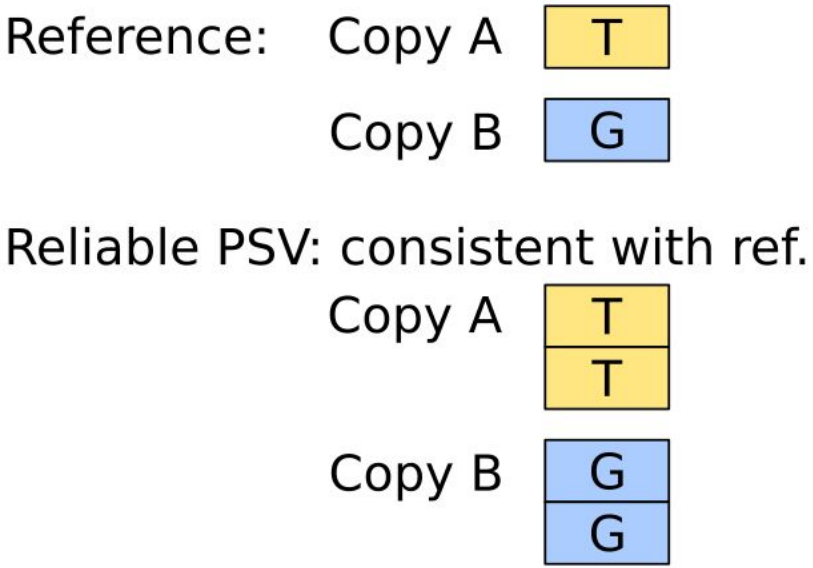
Often coincide with polymorphisms, and therefore can be unreliable:



Paralogous Sequence Variants (PSVs)

PSV – small sequence difference between repeat copies.

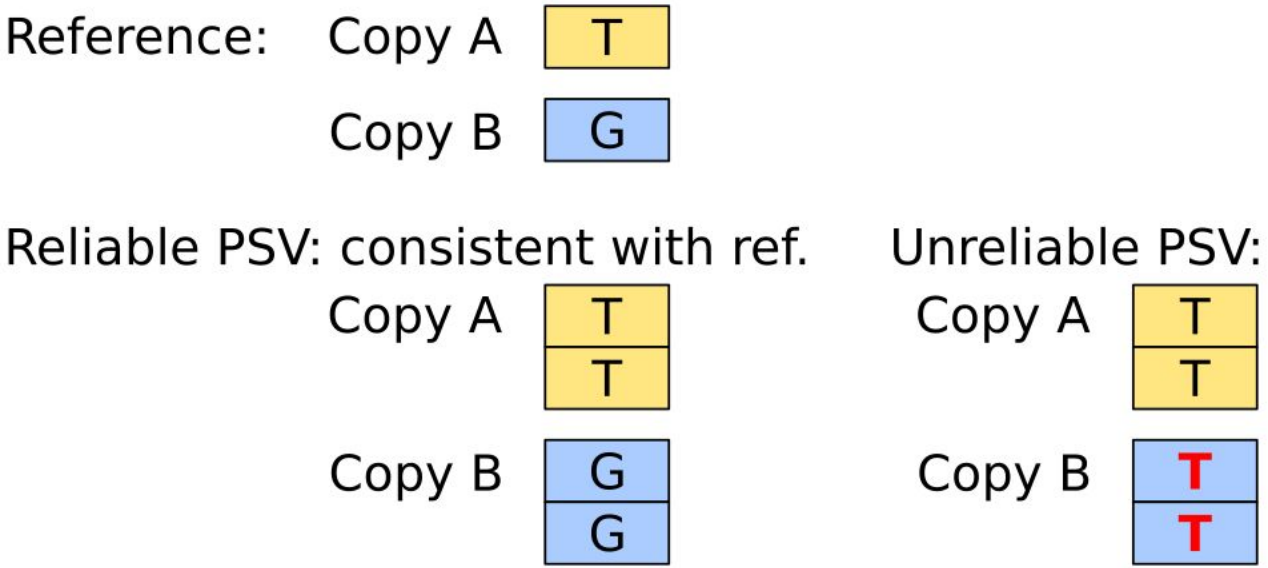
Often coincide with polymorphisms, and therefore can be unreliable:



Paralogous Sequence Variants (PSVs)

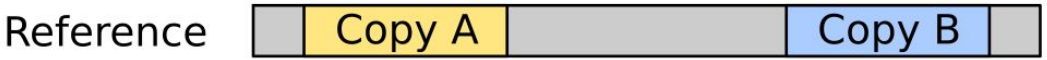
PSV – small sequence difference between repeat copies.

Often coincide with polymorphisms, and therefore can be unreliable:



Copy Number (CN)

N-copy duplication
(here 2 copies: A & B).

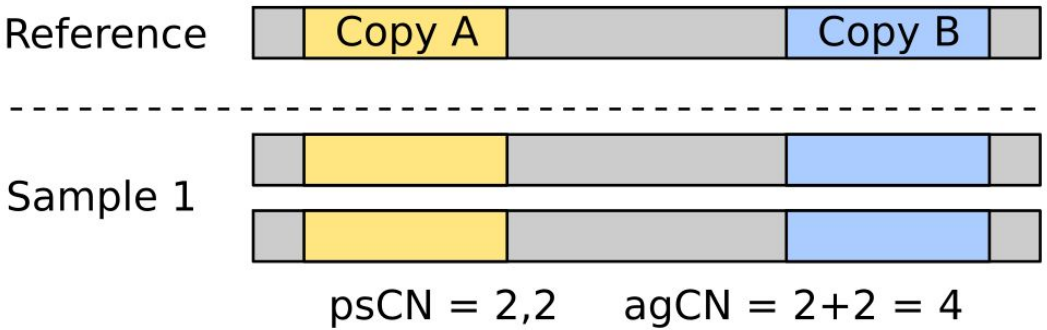


Copy Number (CN)

N-copy duplication
(here 2 copies: A & B).

Paralog-specific copy number
(psCN) – tuple of length N,
stores CN of each copy.

Aggregate copy number
(agCN) – sum copy number
of all copies.

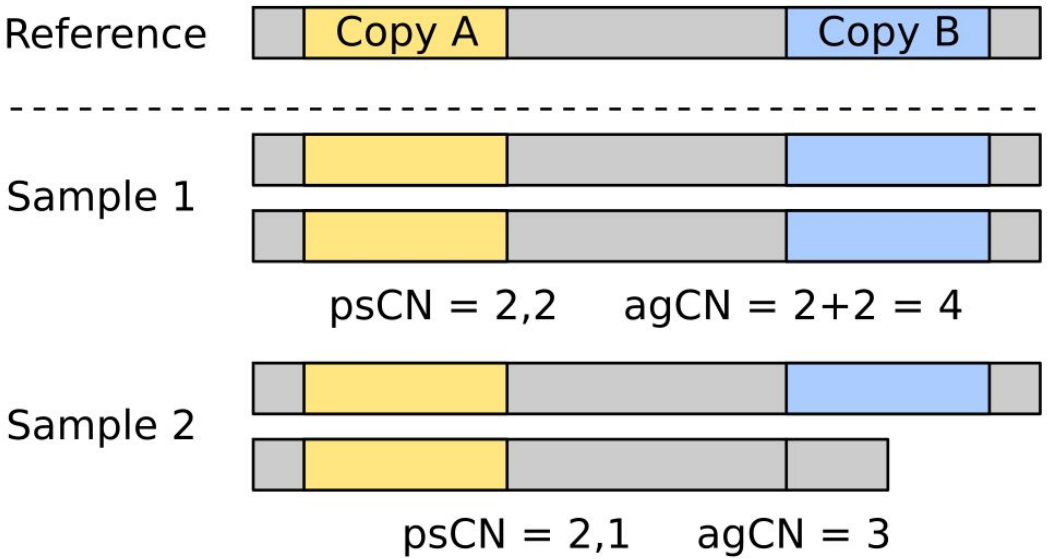


Copy Number (CN)

N-copy duplication
(here 2 copies: A & B).

Paralog-specific copy number (psCN) – tuple of length N, stores CN of each copy.

Aggregate copy number (agCN) – sum copy number of all copies.

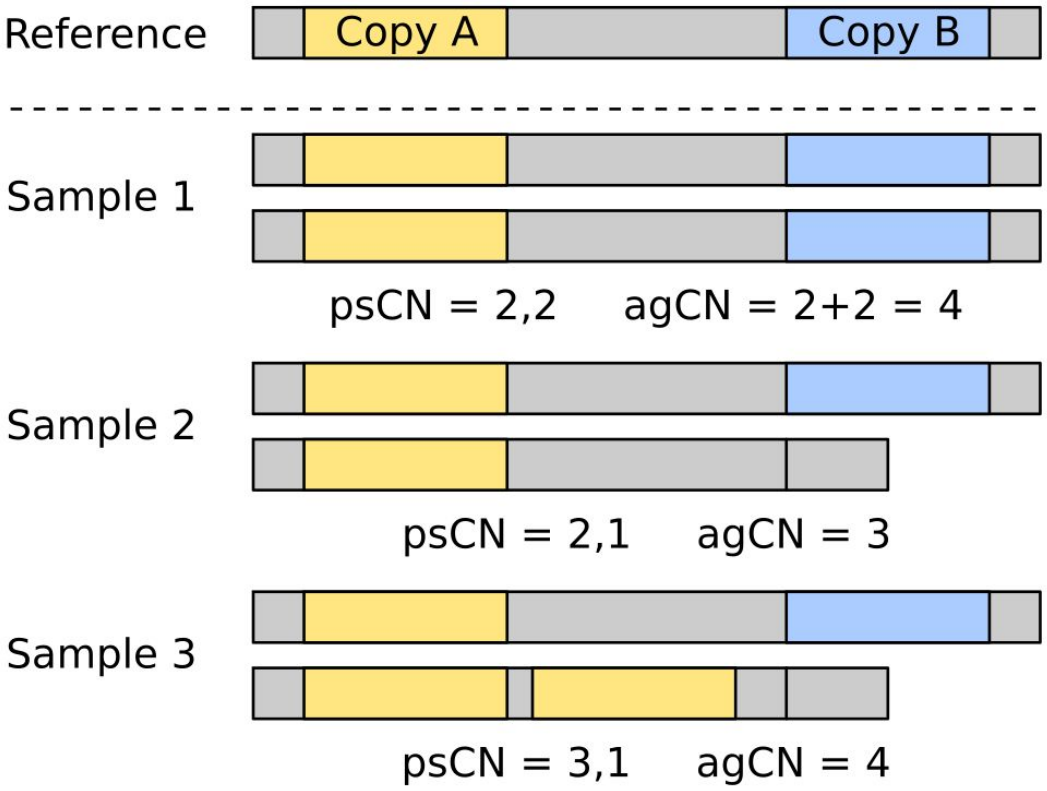


Copy Number (CN)

N-copy duplication
(here 2 copies: A & B).

Paralog-specific copy number (psCN) – tuple of length N, stores CN of each copy.

Aggregate copy number (agCN) – sum copy number of all copies.



Methods

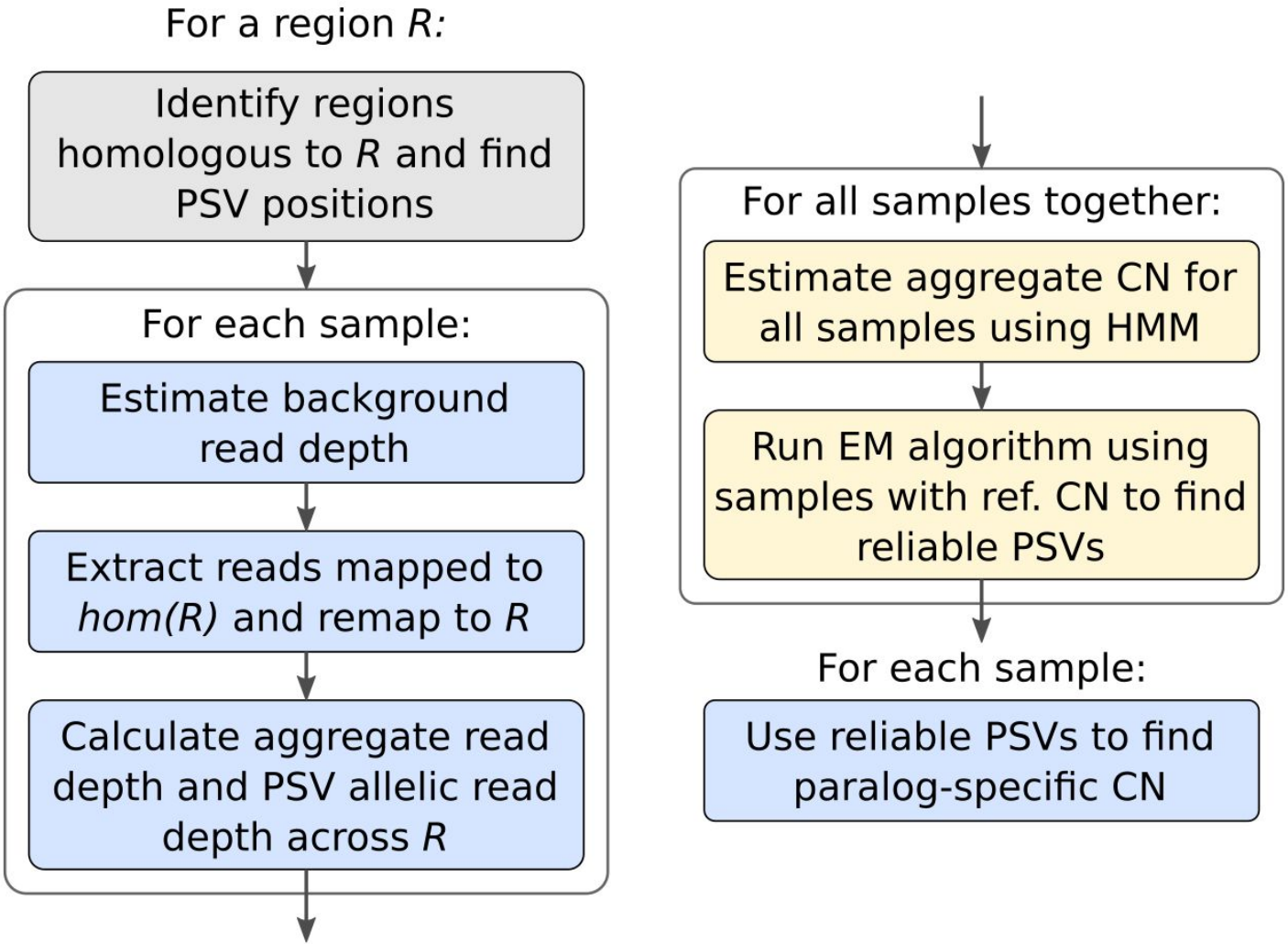
Parascopy

Input:

- Whole-genome sequencing (WGS) for multiple samples,
- List of duplicated regions (2-5 repeat copies).

Output:

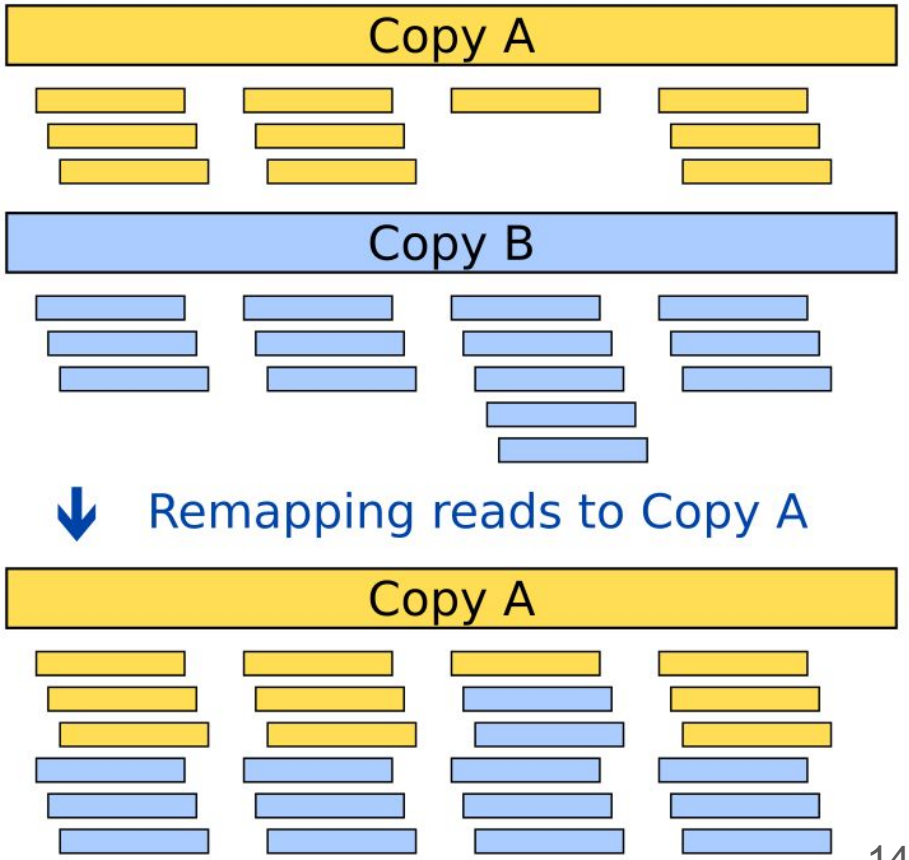
- **agCN** and **psCN** profiles for each region and each sample.


For a region R:
Identify regions homologous to R and find PSV positions
For each sample:
Estimate background read depth
Extract reads mapped to hom(R) and remap to R
Calculate aggregate read depth and PSV allelic read depth across R
For all samples together:
Estimate aggregate CN for all samples using HMM
Run EM algorithm using samples with ref. CN to find reliable PSVs
For each sample:
Use reliable PSVs to find paralog-specific CN

Pooling reads to a single repeat copy

Remap reads to a single repeat copy (here: from B to A).

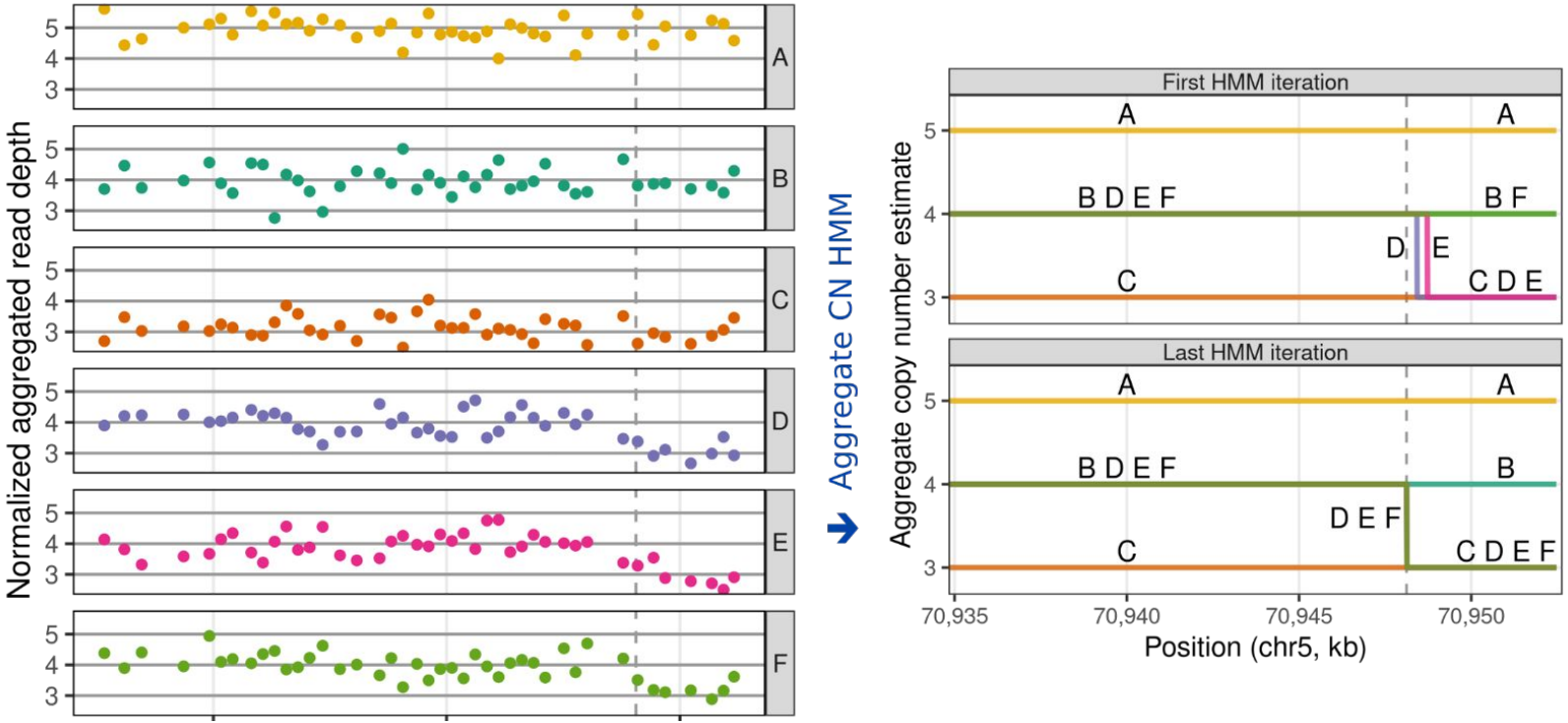
Allows to calculate aggregate read depth and PSV allelic read depth independent of mapping quality.



HMM for agCN estimation

Input: aggr. read depth in 100 bp windows Output: agCN profiles

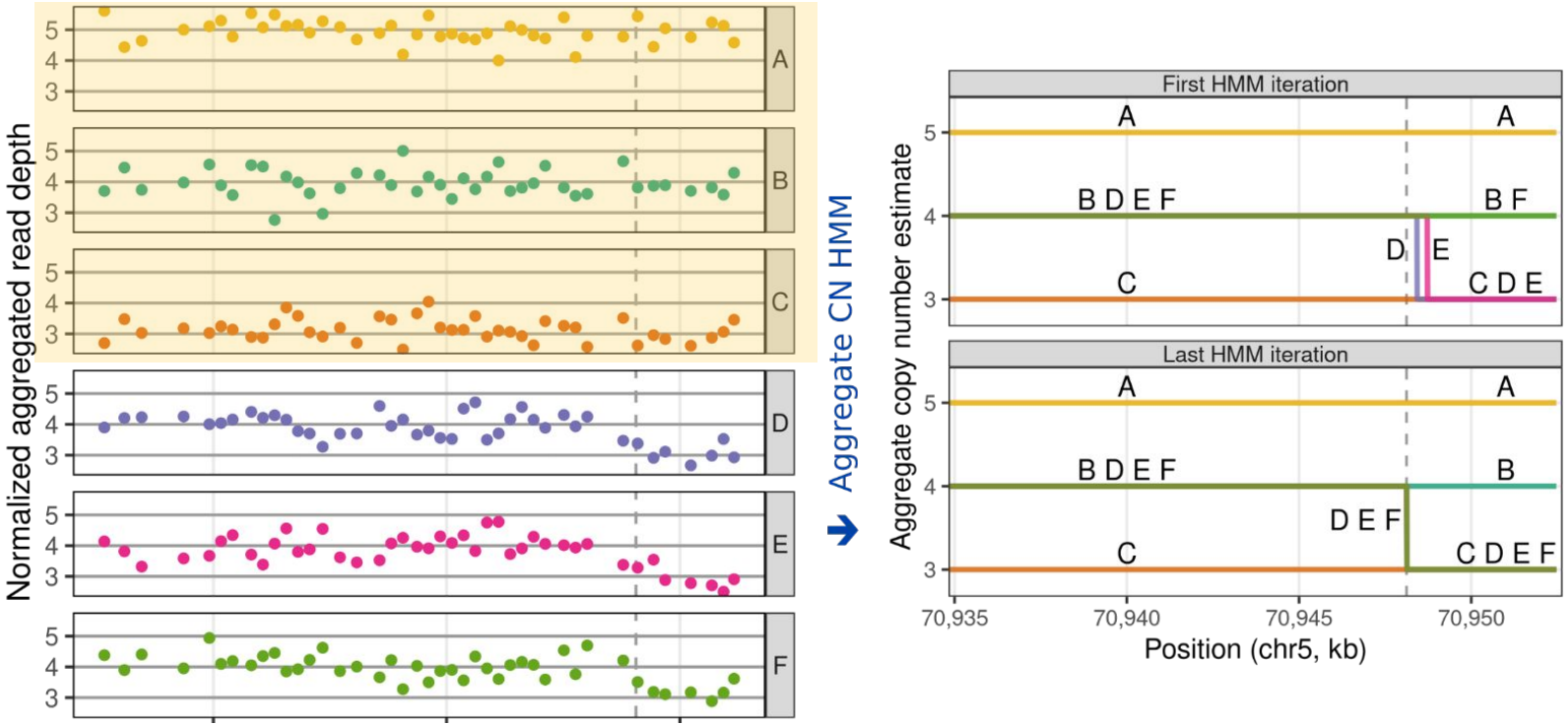
HMM parameters are refined based on multiple samples.



HMM for agCN estimation

Input: aggr. read depth in 100 bp windows Output: agCN profiles

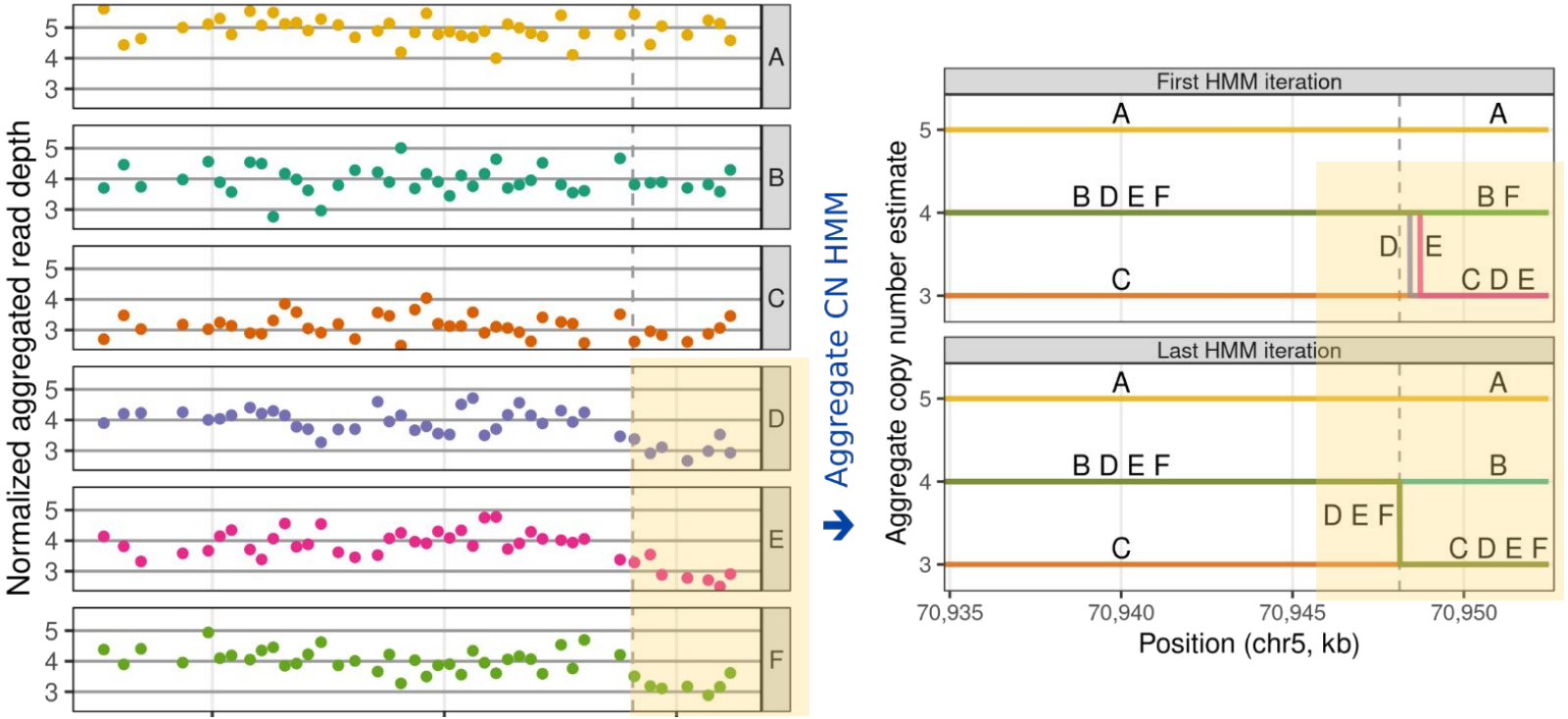
HMM parameters are refined based on multiple samples.



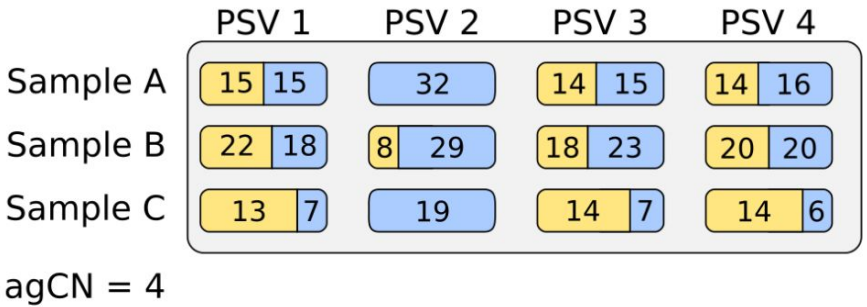
HMM for agCN estimation

Input: aggr. read depth in 100 bp windows Output: agCN profiles

HMM parameters are refined based on multiple samples.



EM algorithm for psCN estimation



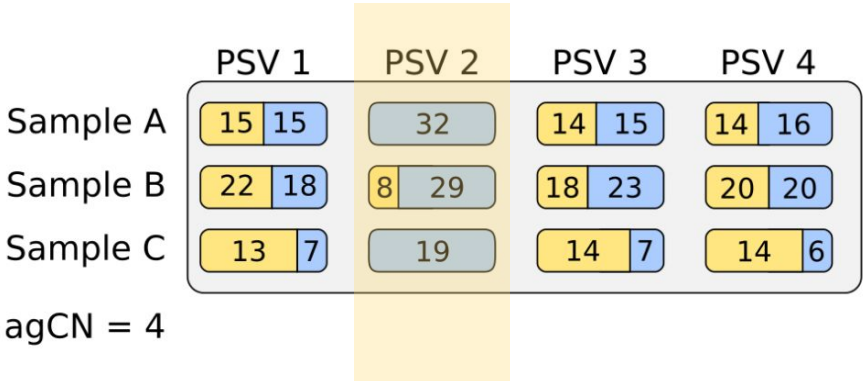
Input:

- agCN for each sample,
- PSV allelic read depth,

Output:

- psCN estimates.

EM algorithm for psCN estimation



Input:

- agCN for each sample,
- PSV allelic read depth,

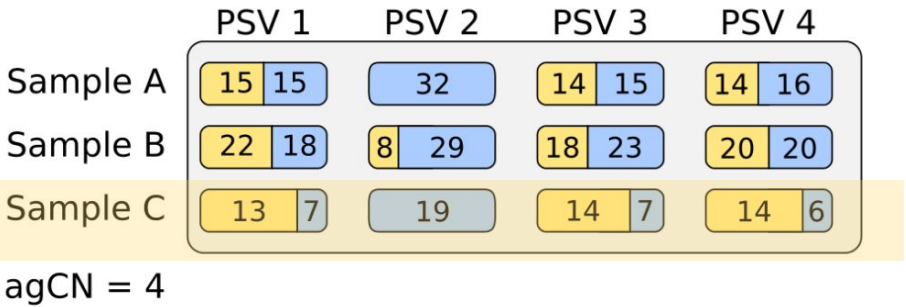
Output:

- psCN estimates.

Problem:

unreliable PSVs & non-ref. samples produce confusing results.

EM algorithm for psCN estimation



Input:

- agCN for each sample,
- PSV allelic read depth,

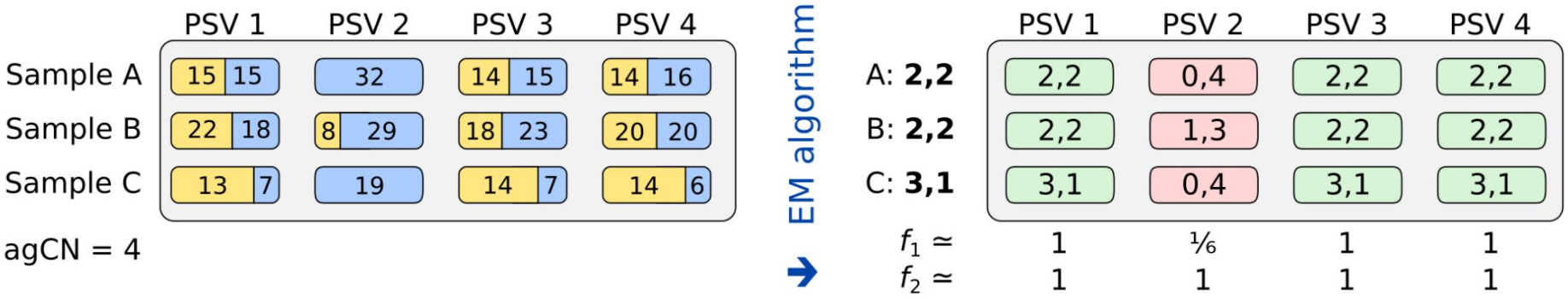
Output:

- psCN estimates.

Problem:

unreliable PSVs & non-ref. samples produce confusing results.

EM algorithm for psCN estimation



Run EM algorithm:

- Hidden variables: sample psCN,
- Parameters: PSV f-values: frequency of the ref. allele for each repeat copy.

PSV is reliable if all its f-values are close to 1.

Analyzing new samples

Parascopy allows to save HMM & EM parameters.

Use them to calculate agCN and psCN profiles of new samples.

Runtime

Different loci are analyzed independently, can be run in parallel.

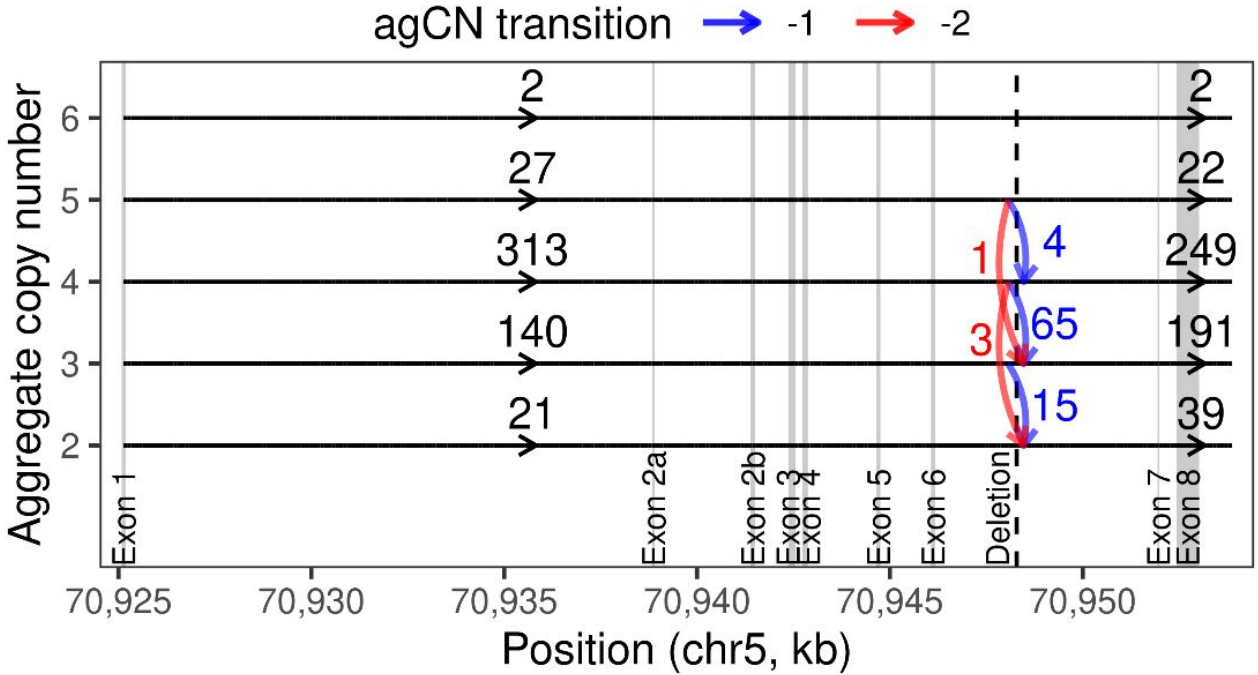
CN profiles for ~700 samples at 167 loci takes ~19 hours using 16 threads.

Analyzing a single sample with pre-computed model parameters takes ~25 min.

Results

SMN1/2 gene: agCN profiles

agCN HMM results: multiple samples allow to detect a known deletion event.

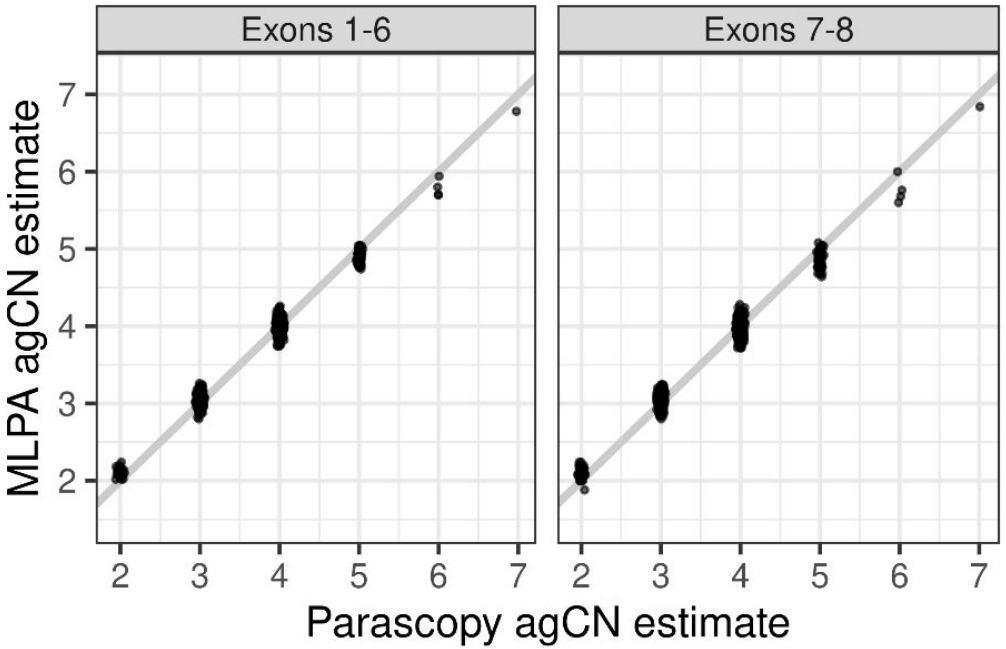


SMN1/2 gene: experimental validation

Validating agCN estimates using MLPA.

Two observations: exons 1-6 & 7-8 because of a common deletion in SMN2.

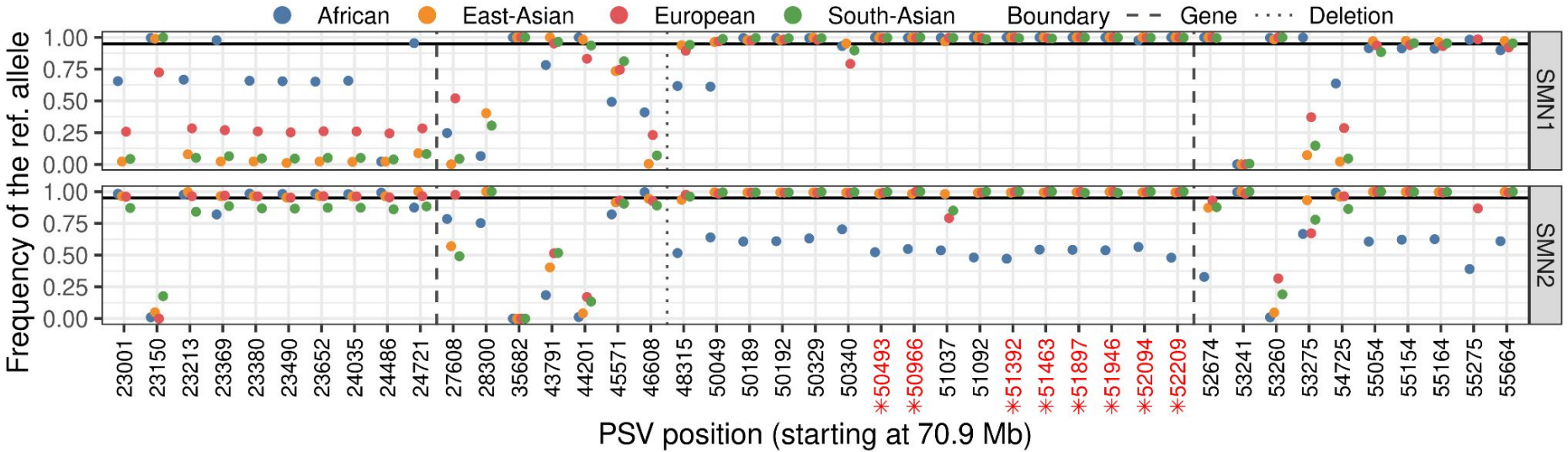
972 samples from
the 1000 genomes project:
100% concordance.



Experimental validation: Parascopy shows high accuracy

Gene	Data type (CN)	N	Custom method (%)	QuicK-mer2 (%)	Parascopy (%)
SMN1/2	Aggr. (exons 1-6)	972	100	78	100
	Aggr. (exons 7-8)		100	49	100
SRGAP2	Aggregate	40	98	63	100
	Paralog-specific		65	71	98
C4A	Aggregate	45		76	100
	Paralog-specific			49	67
AMY1C	Aggr. (corr. coef.)	225	0.881	0.886	0.911

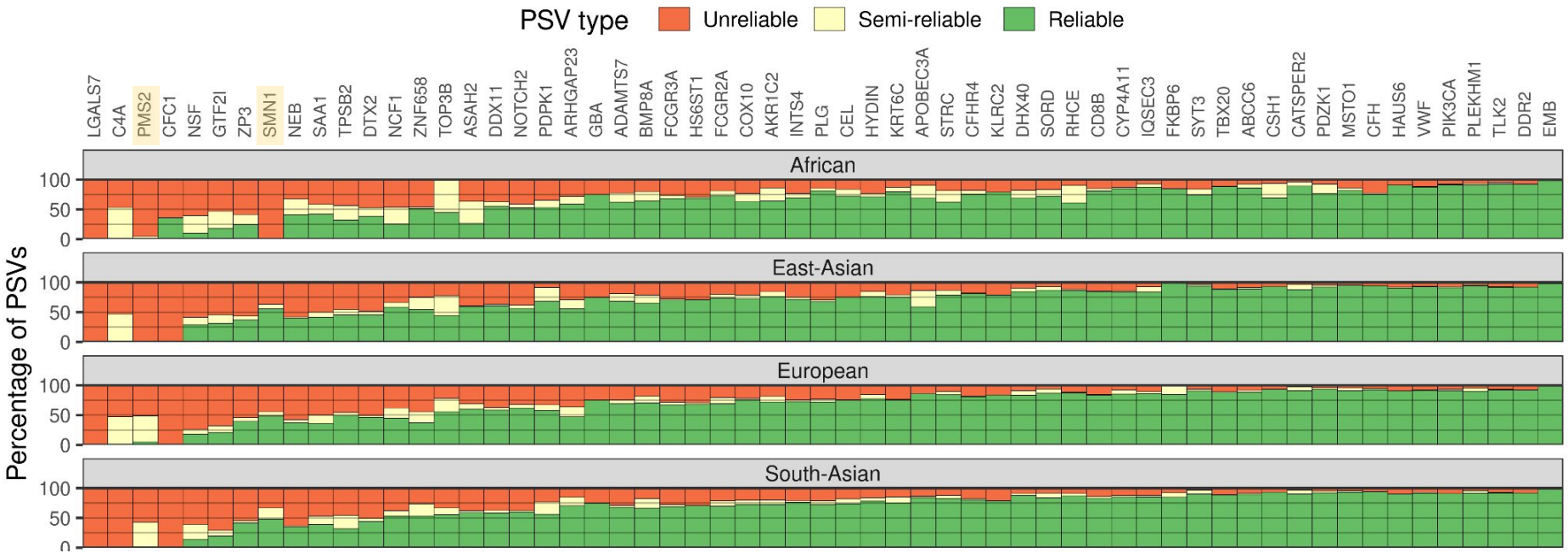
SMN1/2 gene: reliable PSV detection



Parascopy detects 10-19 reliable PSVs across 3 populations (0 in African pop.).

SMNCopyNumberCaller uses 8 reliable PSVs (in red).

Percentage of reliable PSVs



Varies across different genes from 0% to 100%.

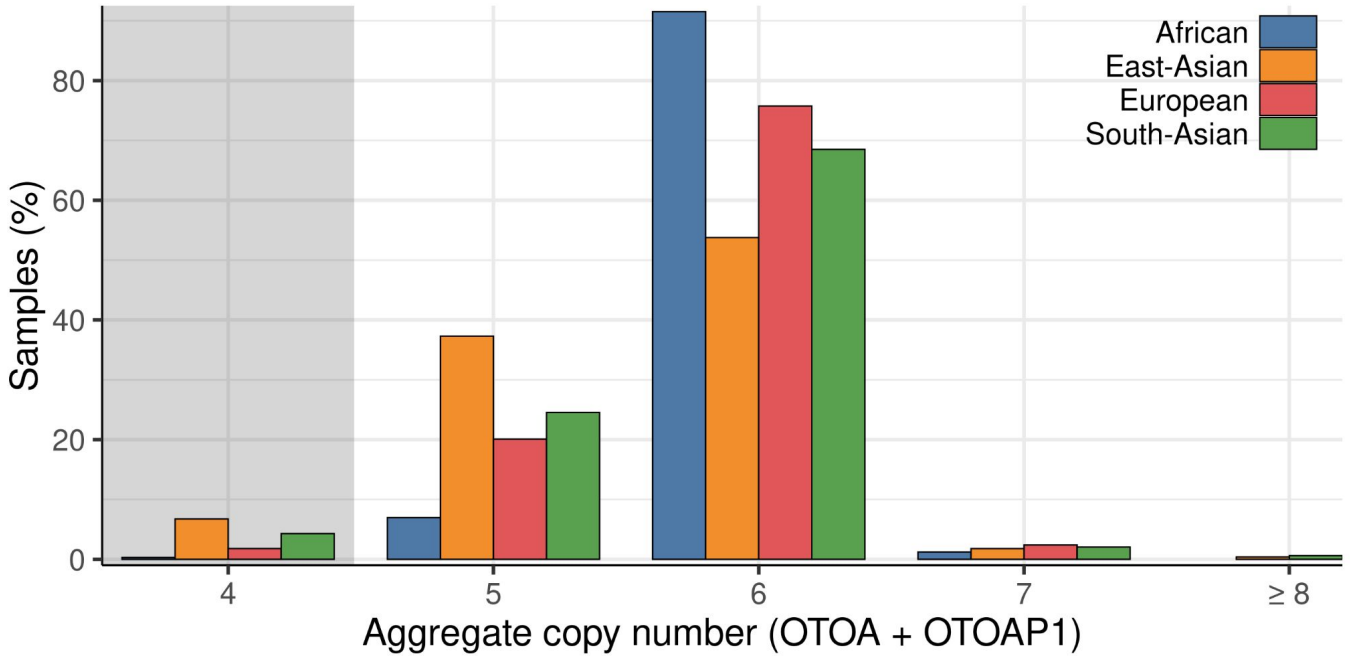
Often similar across different populations (calculated independently).

OTOA gene – missing copy in the reference

Two copies in the reference (OTOA & OTOAP1) => reference CN = 4.

Observed most common agCN = 6 (missing copy in the reference).

CHM13 assembly has 3 repeat copies.

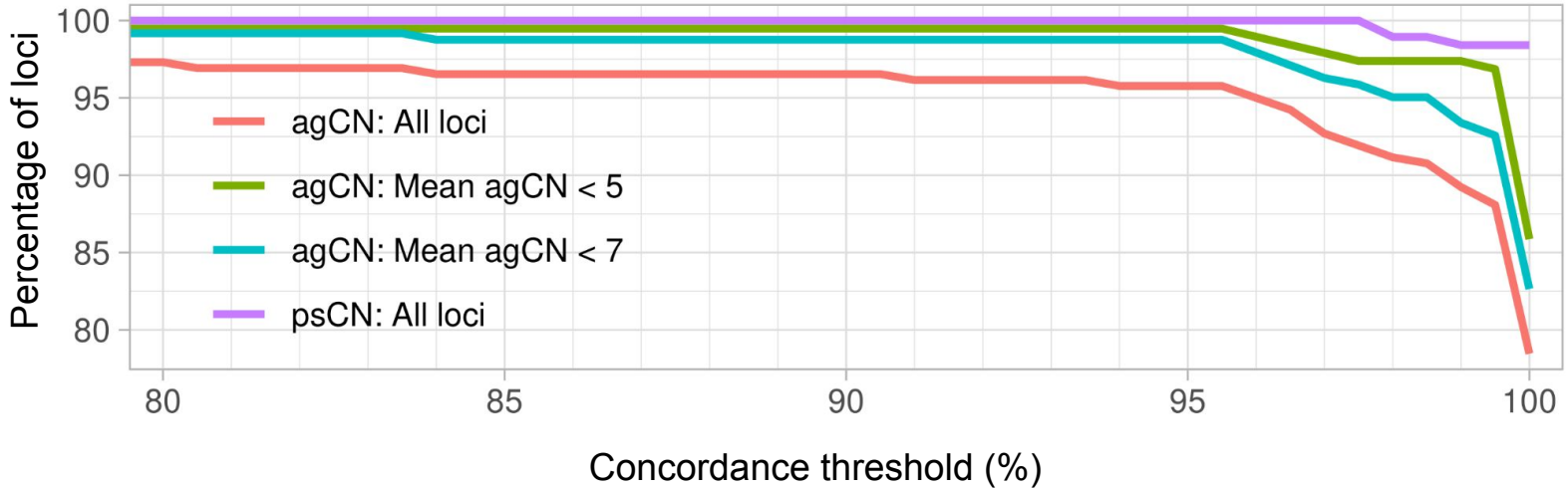


Parascopy robustness

We compared CN estimates across 167 duplicated loci.

Two independent sequencing datasets for 83 Han Chinese samples:

- PCR-free WGS, **1000 genomes,**
- PCR-based deep WGS, **BGI.**



Conclusions

Parascopy uses multiple samples to

- accurately estimate agCN,
- find reliable PSVs and use them to estimate psCN.

Parascopy has higher or equal accuracy compared to other sequencing-based methods.

Parascopy can analyze a large number of duplicated loci
with diverse repeat structures.

github.com/tprodanov/parascopy

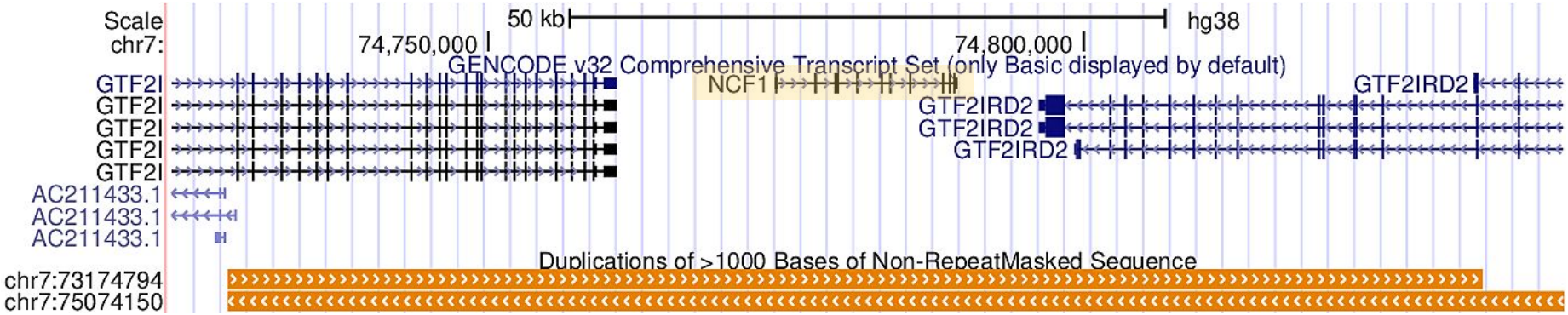
Q&A

NCF1 duplicated gene

Encodes neutrophil cytosolic factor 1 protein.

Three copies: NCF1, NCF1B and NCF1C. 106 kb duplication, 99.5% seq. similarity.

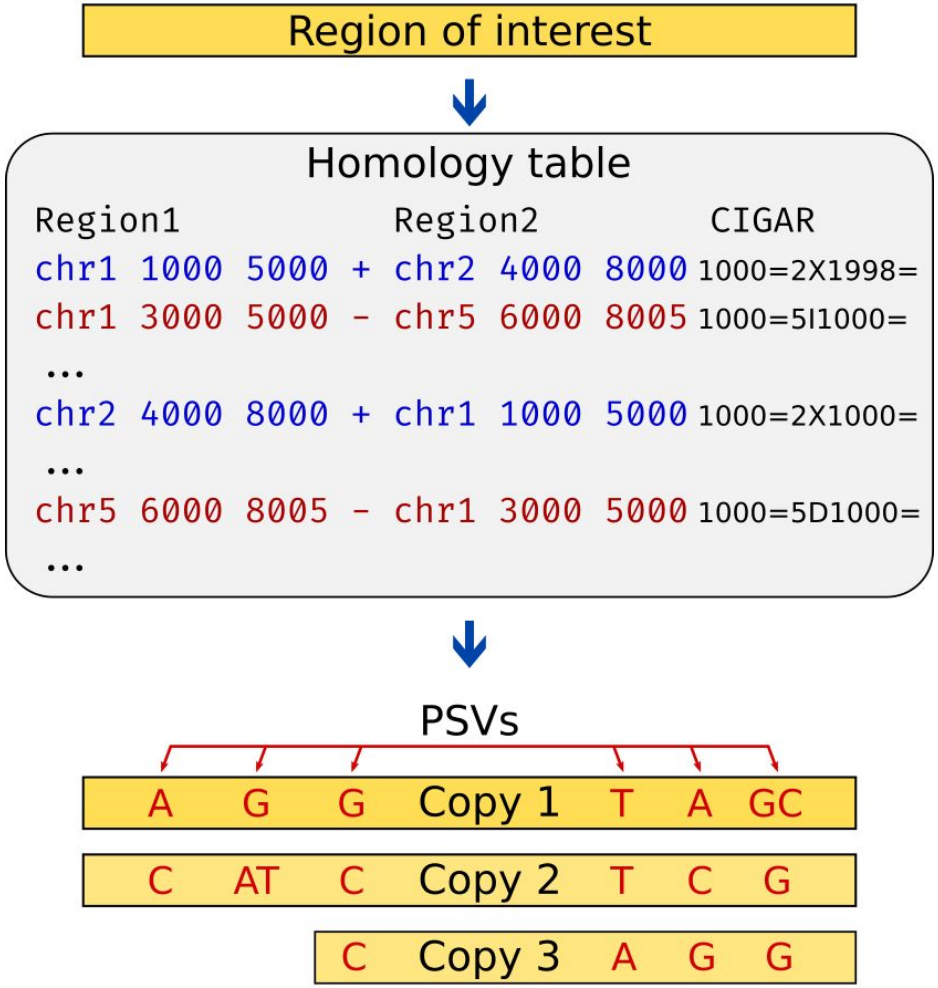
Mutations are associated with the **Chronic granulomatous disease**, and overall weaken immune system.



Homology table

Store alignments between repeat copies.

For each region of interest we reconstruct multi-copy duplications and extract PSV from pairwise alignments

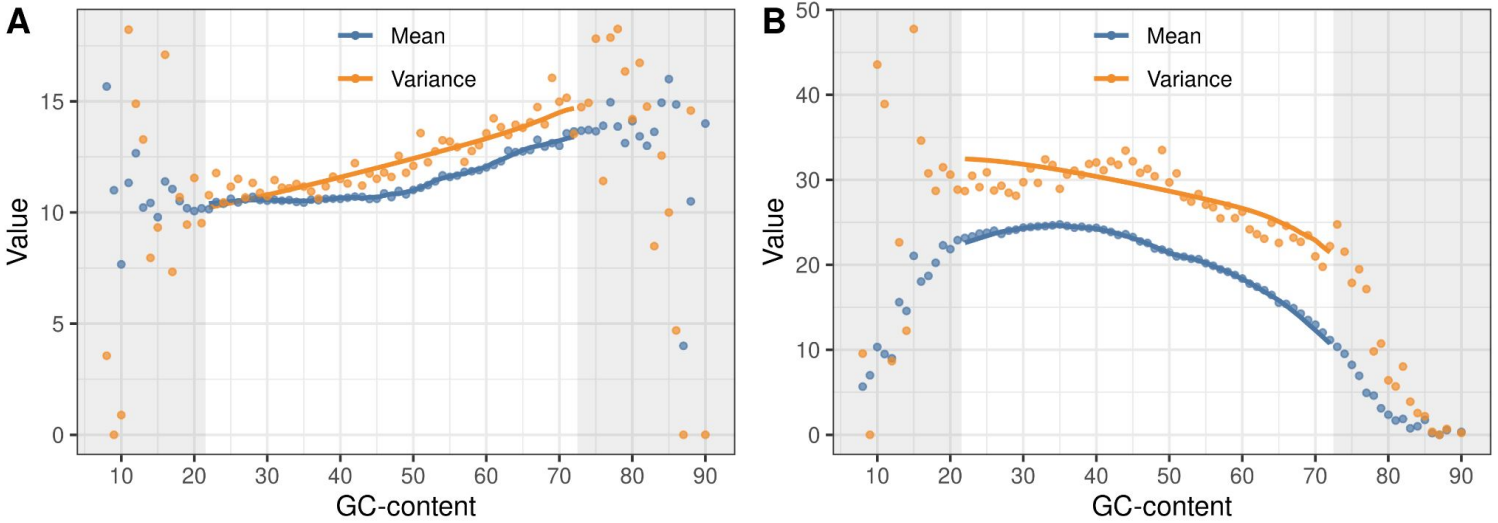


Background read depth

Use unique (non-duplicated) regions to estimate background read depth:

- For each sample,
- For each GC-content value.

Fit Negative-Binomial distribution.



Parascopy robustness

Robustness for various subsets of the 1000 genomes samples.

Use two independent HMM and EM parameters
(for example obtained using EUR or using EAS samples).