Institute of Medical Genetics and Applied Genomics

UNIVERSITÄTS KLINIKUM TÜBINGEN

Methods for detection of germline and somatic copy-number variants in next generation sequencing data

German Demidov

THESIS SUPERVISORS:
Prof. Dr. Stephan Ossowski
Prof. Dr. Tomas Marques-Bonet

EBERHARD KARLS UNIVERSITÄT TÜBINGEN

CRG Centre for Genomic Regulation

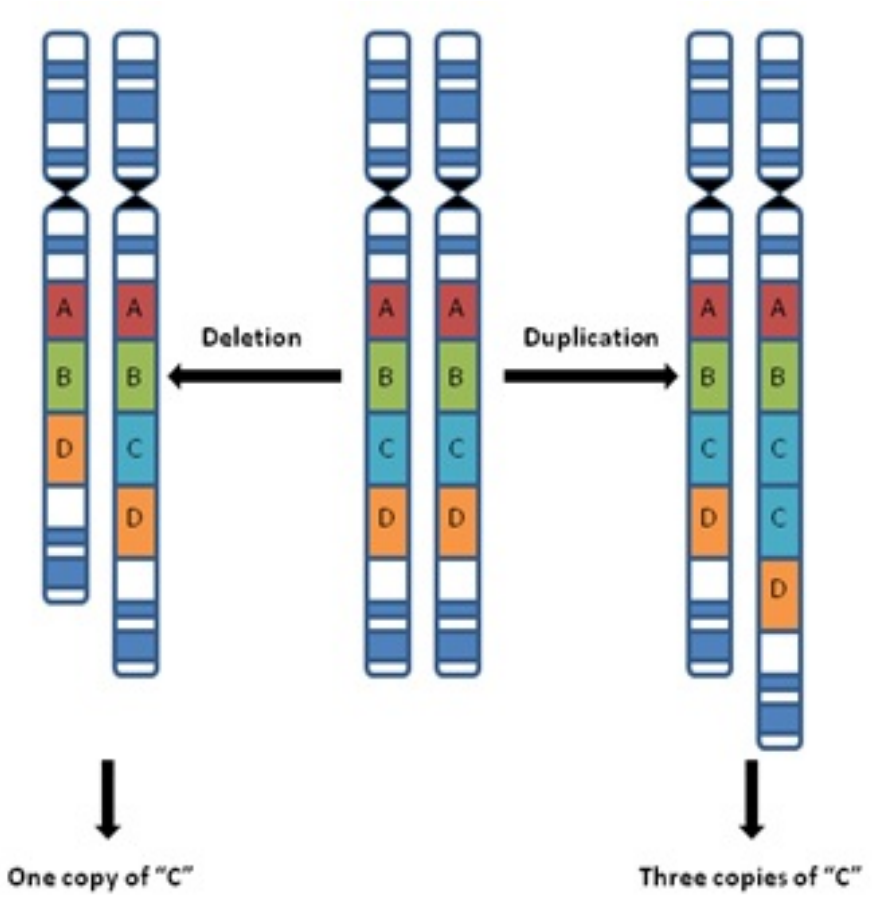
upf. Universitat Pompeu Fabra Barcelona

Overview

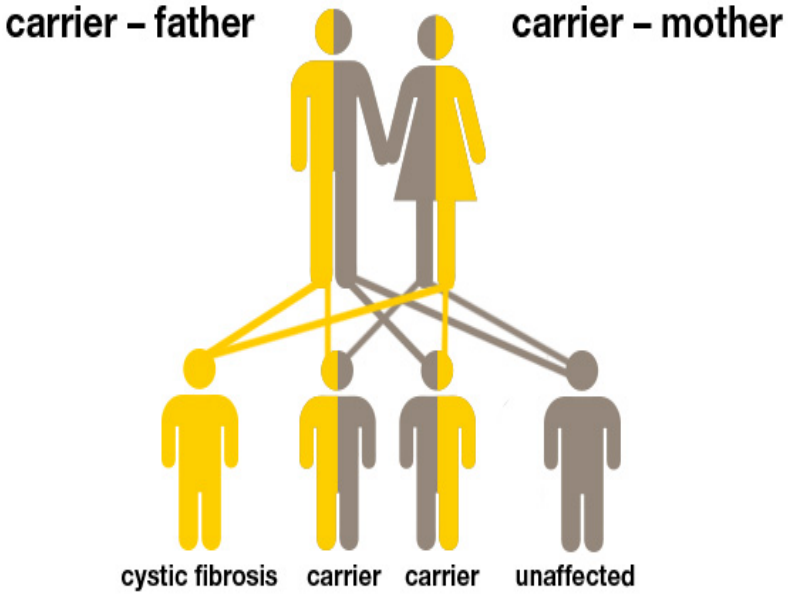
1. The **What**? − what is a copy number variant (CNV)? Copy number alteration (CNA)?
2. The **Why**? − why do we need to find them?
3. The **How**? − how people detect CNVs?
4. The **Novel**? − our method
5. The **Results**? − results in germline WES/shallow WGS/WGS and somatic WES/targeted panel sequencing (TPS)

What is a CNV/CNA and why they are important

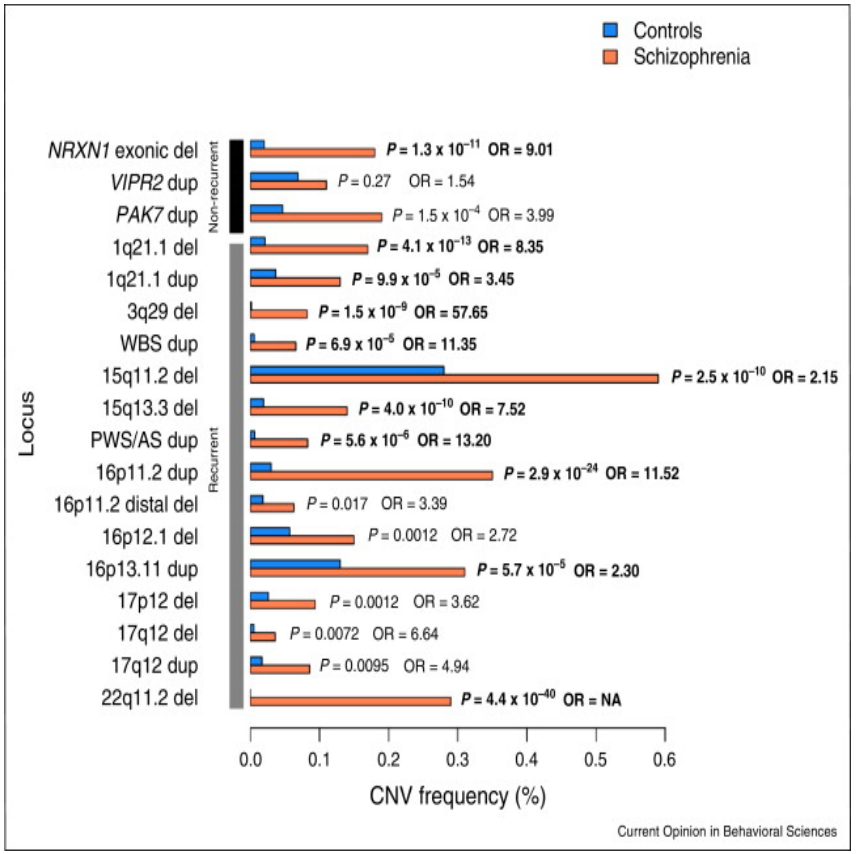
What is a copy number variant (CNV)?



Why it is important to detect CNVs?



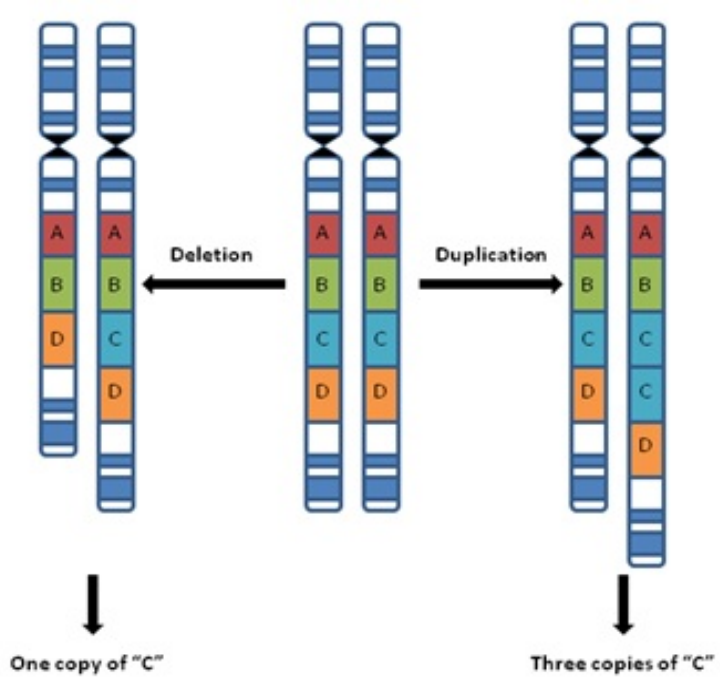
Deletions of CFTR in Cystic Fibrosis

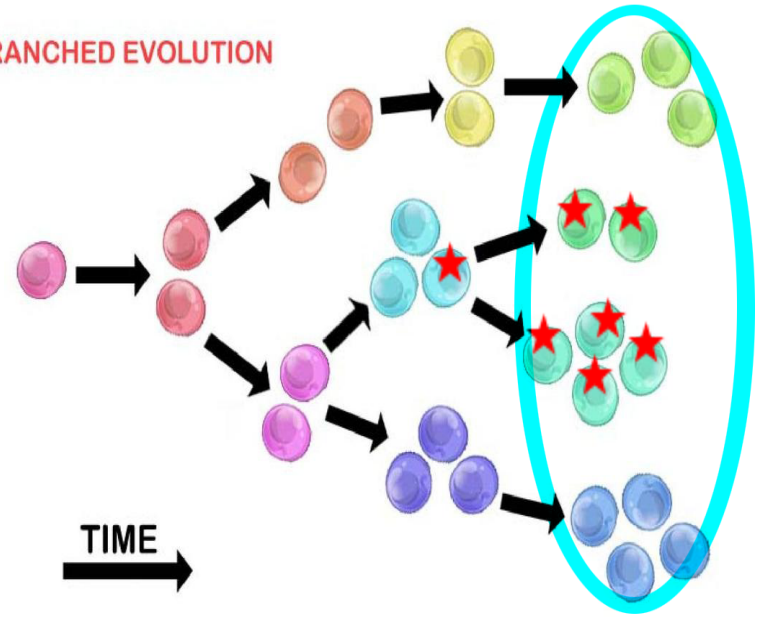


Copy Number Variation in Schizophrenia

(pictures from cysticfibrosis.org.uk website and Genetics of schizophrenia, Rees et al)

What is a copy number alteration (CNA)?





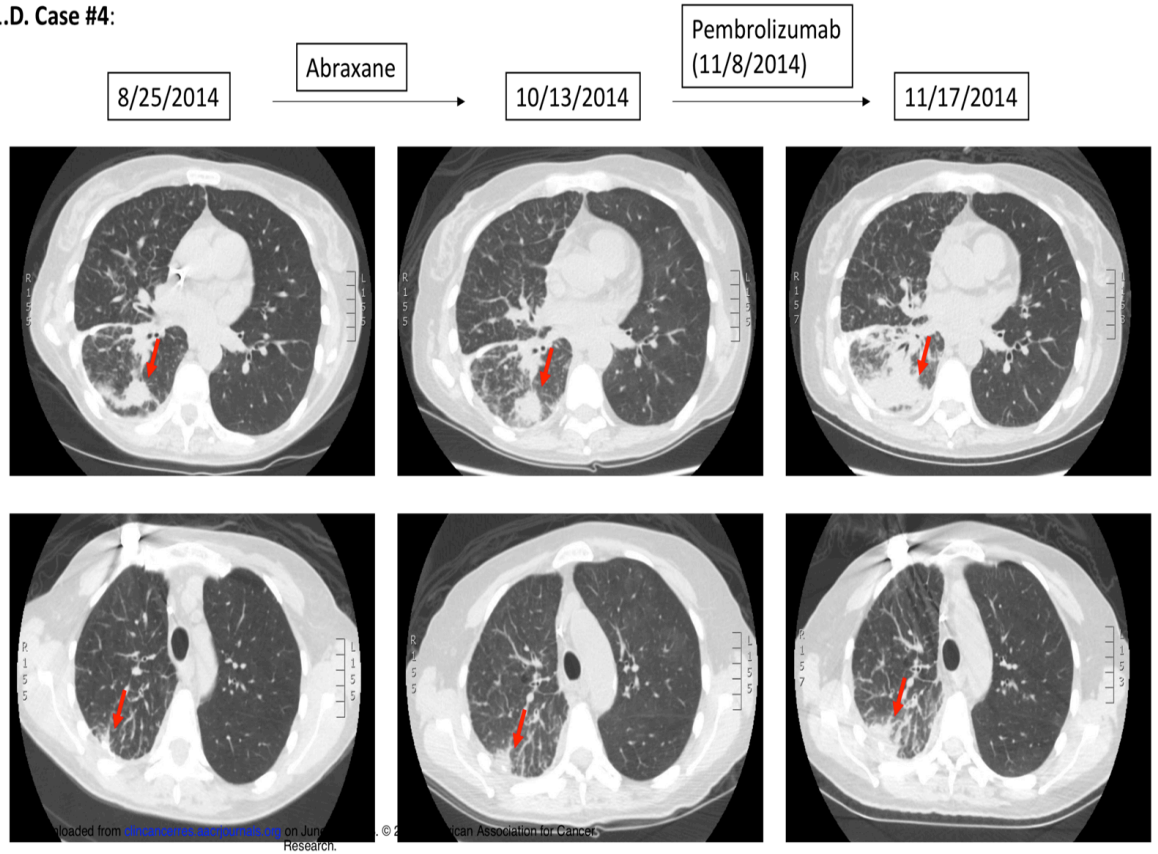
Tumor subclonal evolution. CNV event is shown as a red star. Cells that undergo sequencing are in blue ellipsoid.

(pictures from Autism Reading Room website and wikipedia)

Why it is important to detect CNAs?

A 50 year-old woman with lung adenocarcinoma harboring KIF5B-RET fusion and MDM2 amplification had gradual progression on Abraxane. This therapy was changed to pembrolizumab.

Figure 1.D. Case #4:



Picture from Hyper-progressors after Immunotherapy: Analysis of Genomic Alterations Associated with Accelerated Growth Rate, Kato et al, 2017, American Association for Cancer Research

Existing approaches

Tremendous amount of existing tools for Read Depth CNV detection

- **Tools for CNVs detection (WES/WGS/target panels), read depth:** `XHMM, ExomeDepth, ExomeCNV, ADTEx, CoNIFER, CNVnator, Control-FREEC, ONCOCNV, BIC-seq, cn.MOPS, CNV-seq, EXCAVATOR 1-2, ACEseq, CNVKit, CNVPanelizer, Genome STRiP 1-2, ExCNVSS, SegSeq, CNAseg, JointSLM, rSW-seq, CNVnorm, CNVeM, cnvHMM, CONTRA, CONDEX, PropSeq, VarScan2, ExoCNVtest, etc, etc…`

…¡madre mía!

Types of signal useful for CNVs detection

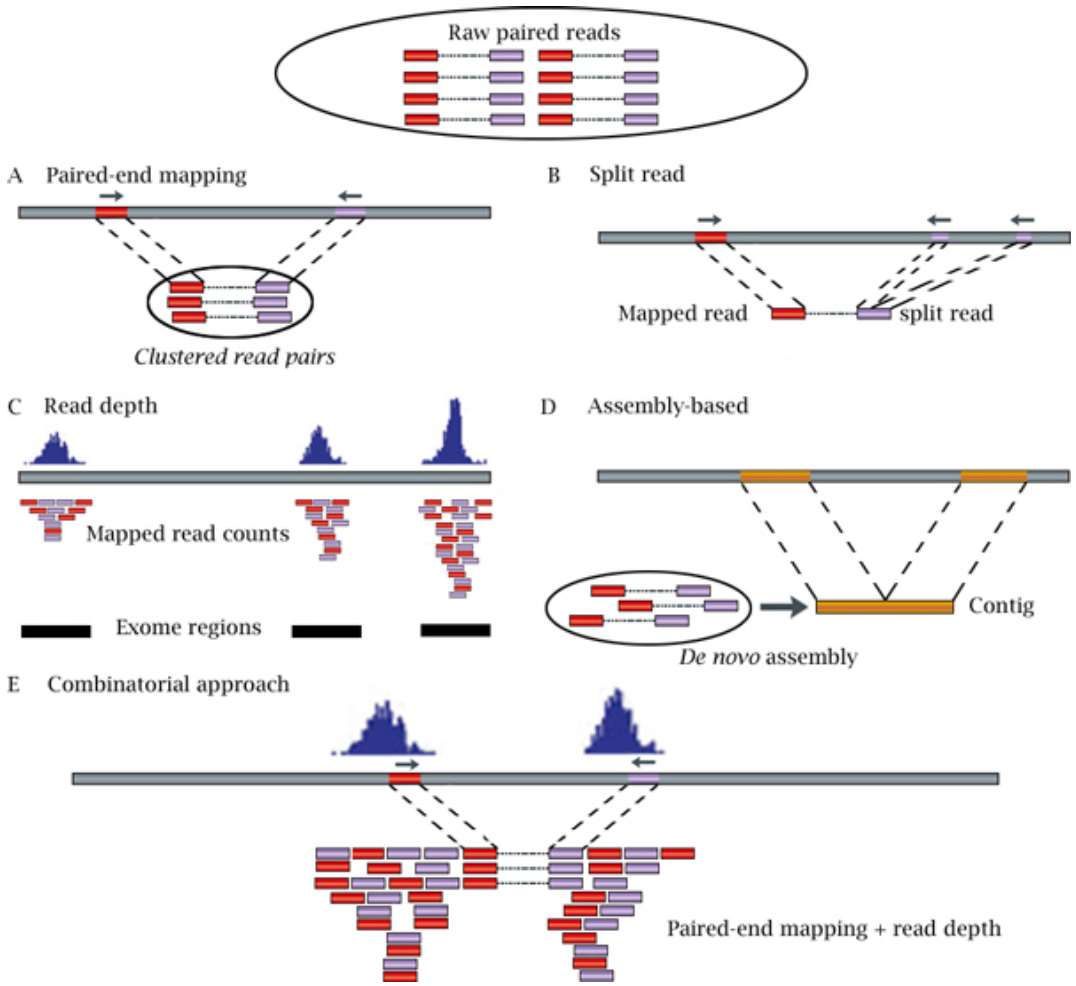


Image from: Computational tools for copy number variation (CNV) detection using next-generation sequencing data: Features and perspectives, Zhao et al, BMC Bioinformatics, 2013

Allele Frequency of Single Nucleotide Variants

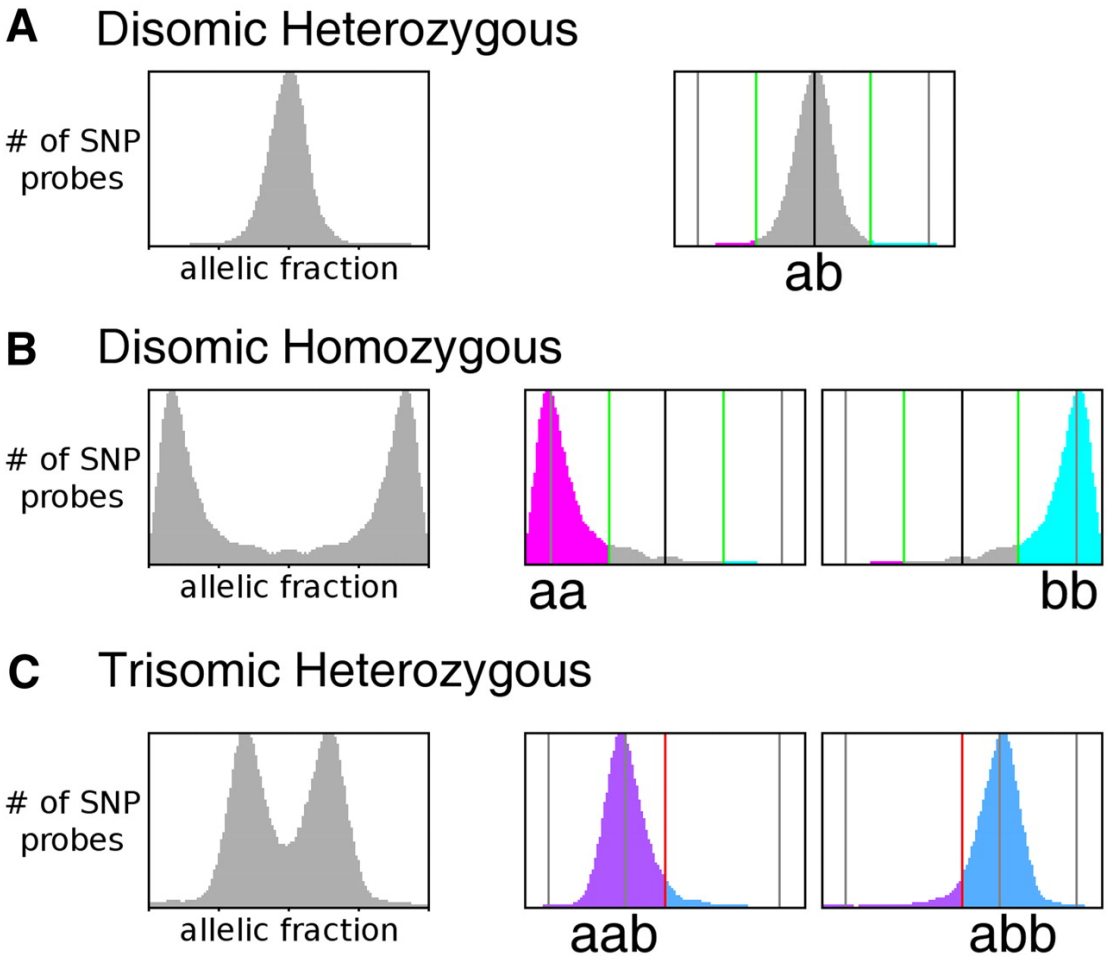
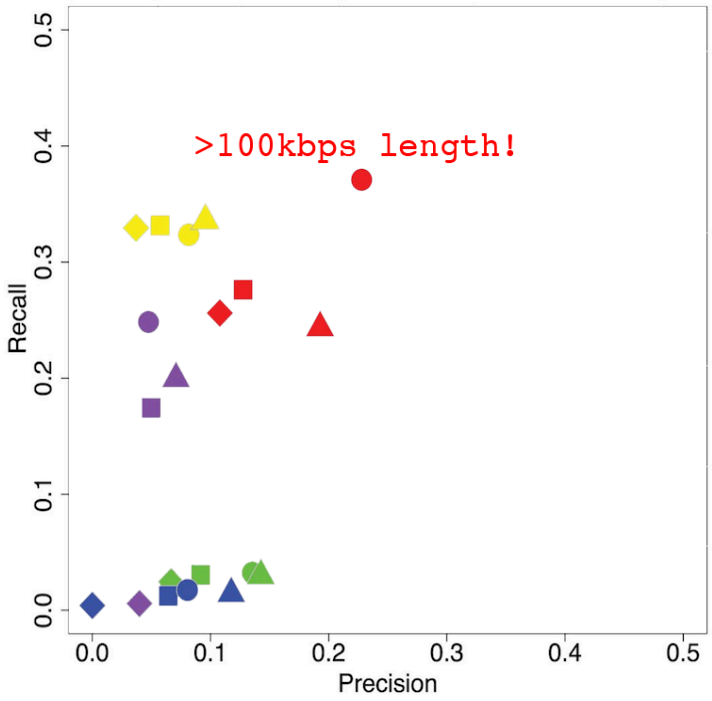


Image from: High-Resolution SNP/CGH Microarrays Reveal the Accumulation of Loss of Heterozygosity in Commonly Used Candida albicans Strains, Abbeyet al, G3: GENES, GENOMES, GENETICS December 1, 2011 vol. 1 no. 7 523-530; https://doi.org/10.1534/g3.111.000885

Why do we need one more?



- Accuracy
- Interpretability
- Limitations
- Guidelines

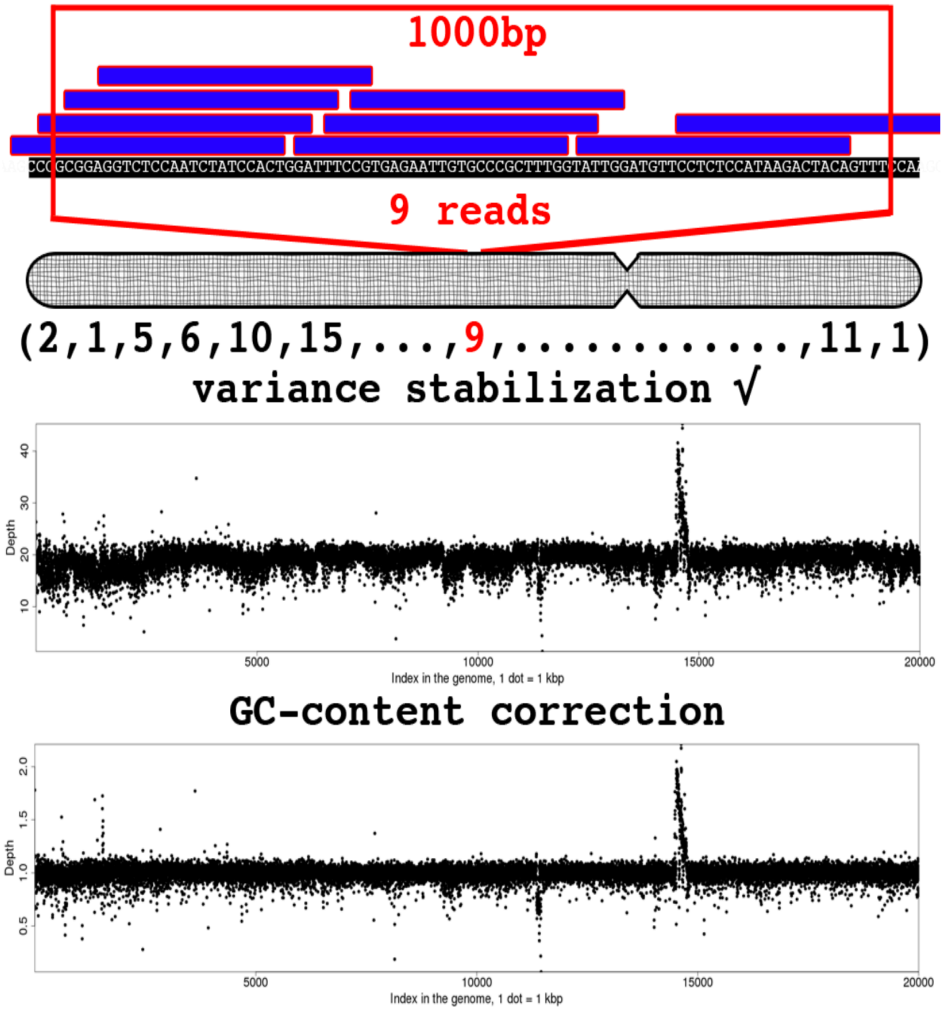
(pictures from D'Aurizio et al, EXCAVATOR2, Nucleic Acid Res., 2016.
I have to say the results to EXCAVATOR2 are not bad.)

Formulation of the problem

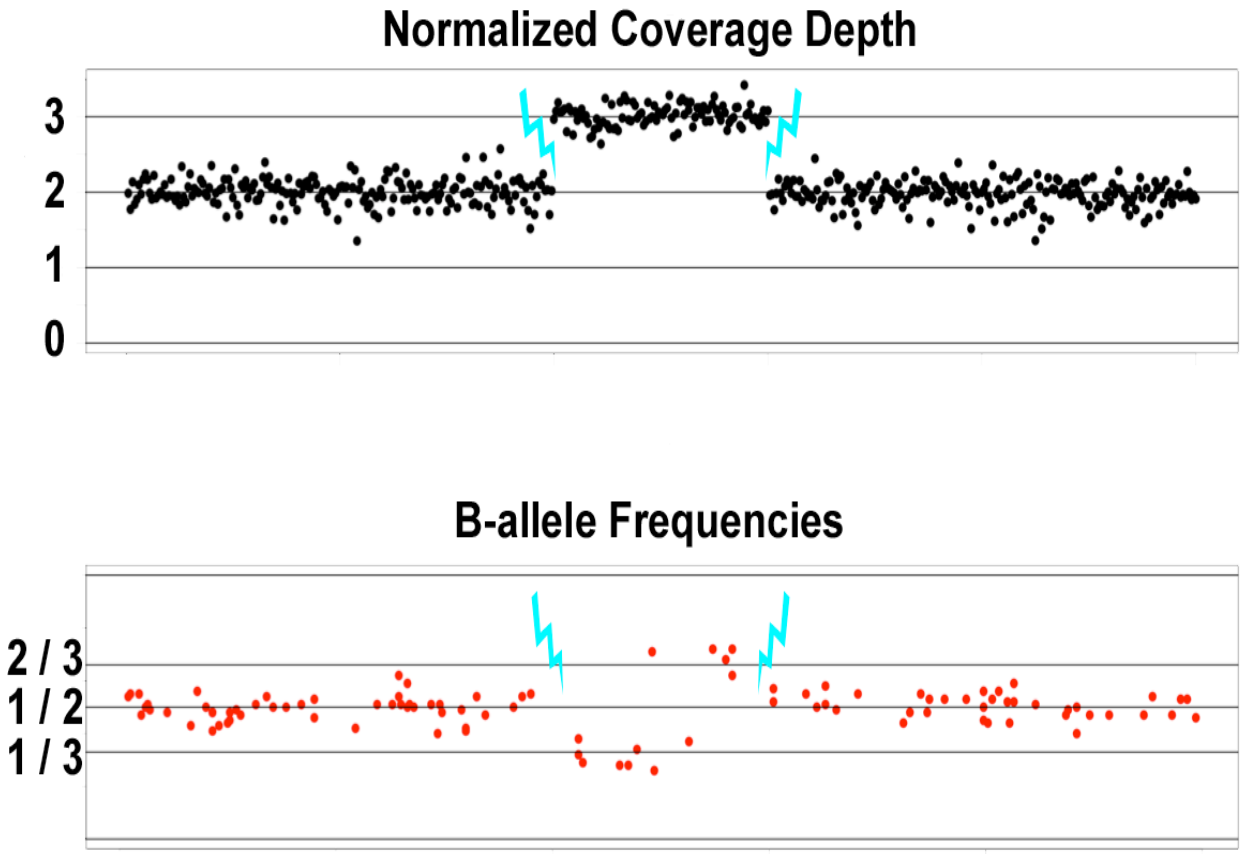
Having a vector of genomic regions {1,⋯,n}:

1. **Collect all the available data** corresponding to each genomic region
2. **De-noise** data
3. **Find segments** with high evidence of alternative copy number

Read Depth Signature for CNV detection: sequencing to numbers



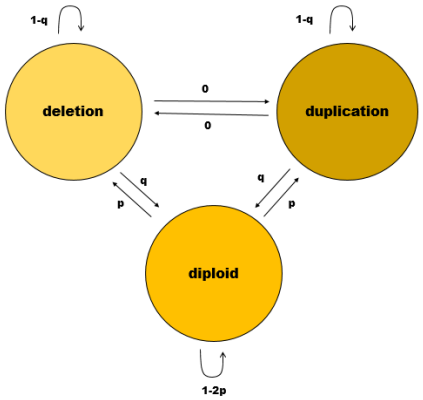
B-allele Frequency and Read Depth

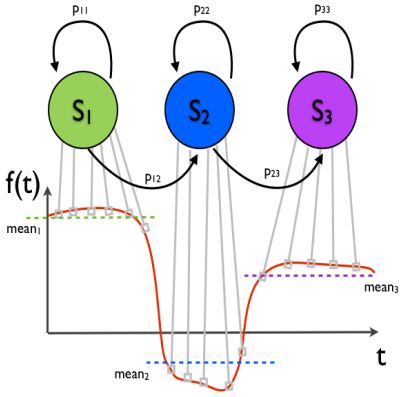


Bottom plot: B-allele Fraction: ratio between number of reads, supporting SNV's alternative allele / Total number of reads, covering position

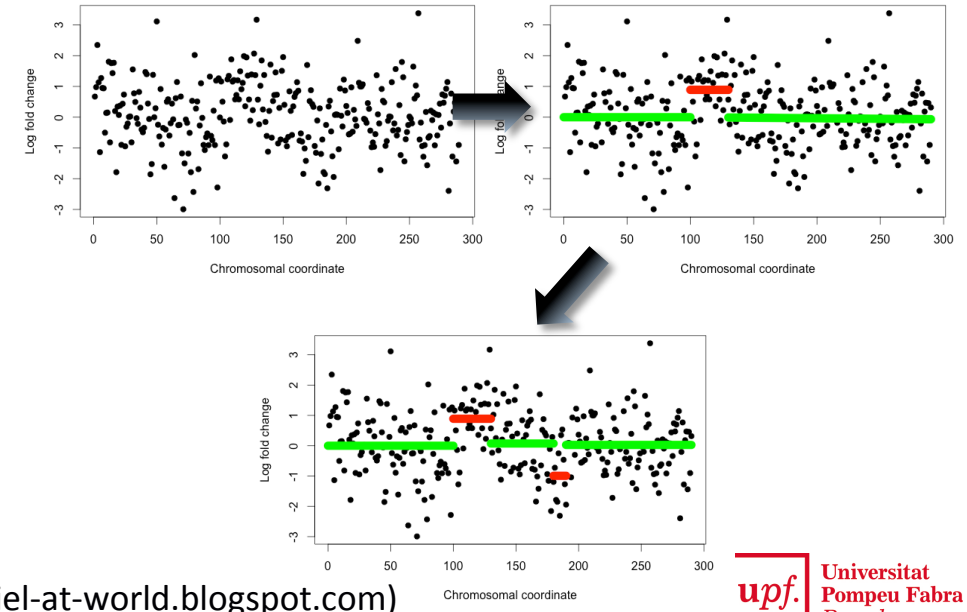
Popular Algorithms

- Hidden Markov Model (HMM)
- Makes assumptions on CNV states (i.e. heterozygous duplication => 1.5x increase of depth and SNVs ratio shifts to AA/B)





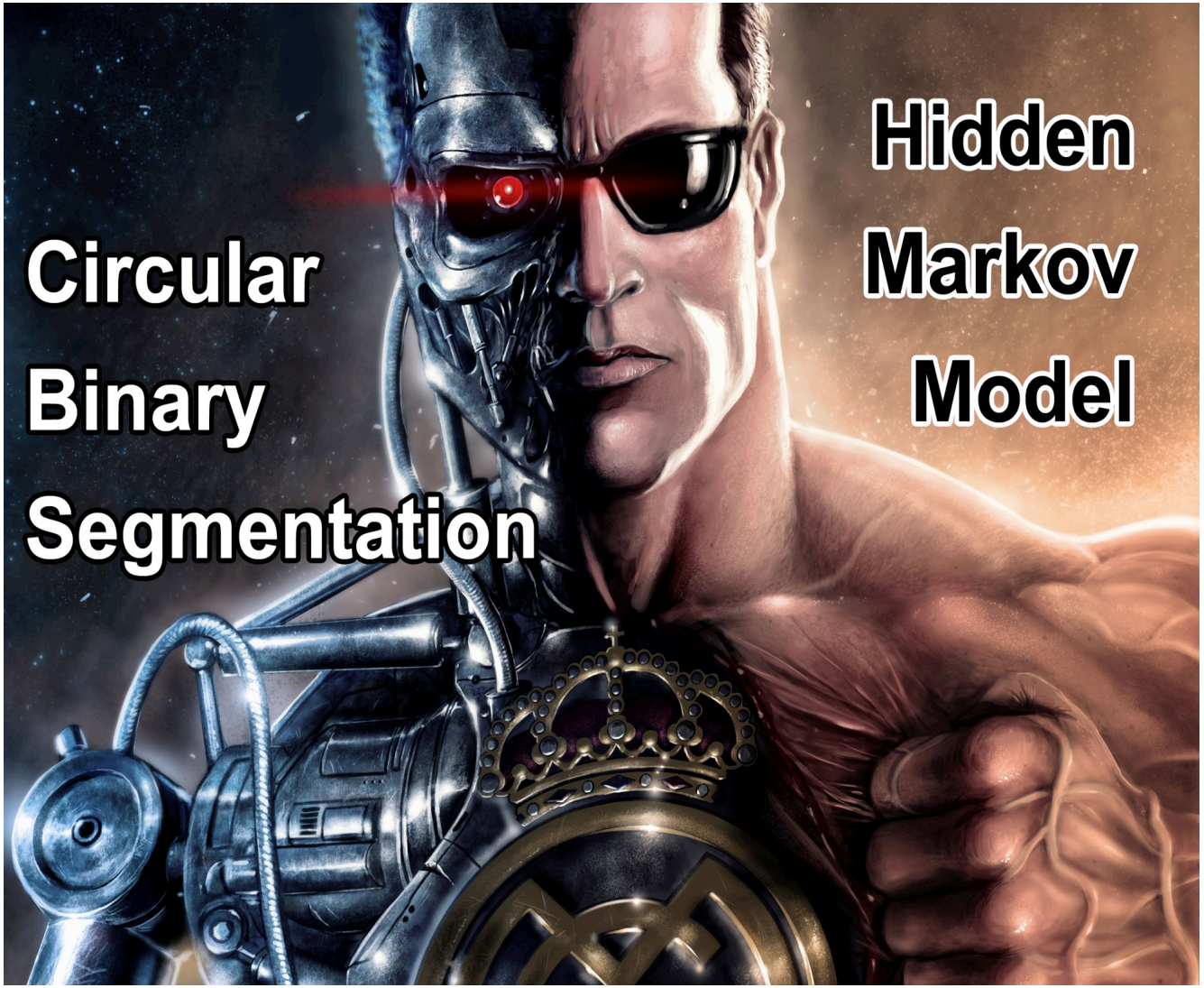
- Circular Binary Segmentation (CBS)
- Find a chromosomal segment with the most significant difference in means



(pictures from web site of XHMM and daniel-at-world.blogspot.com)

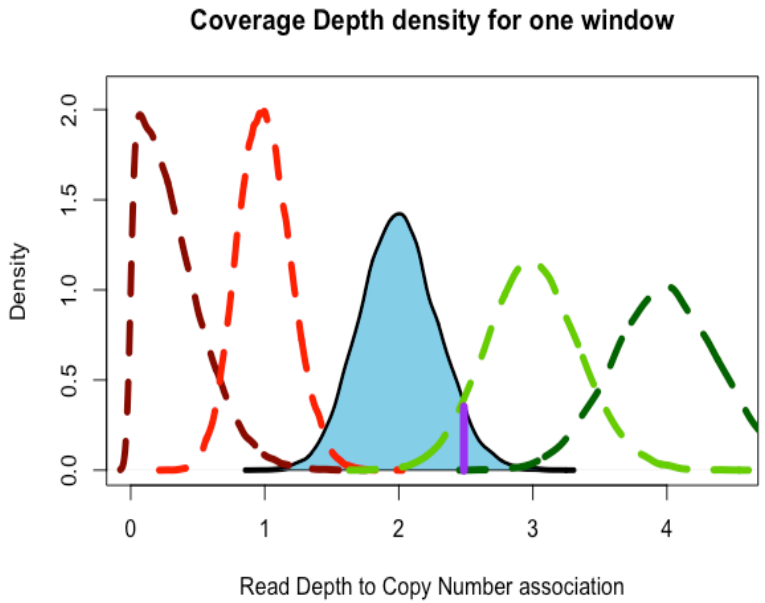
Our method

↓ ClinCNV ↓

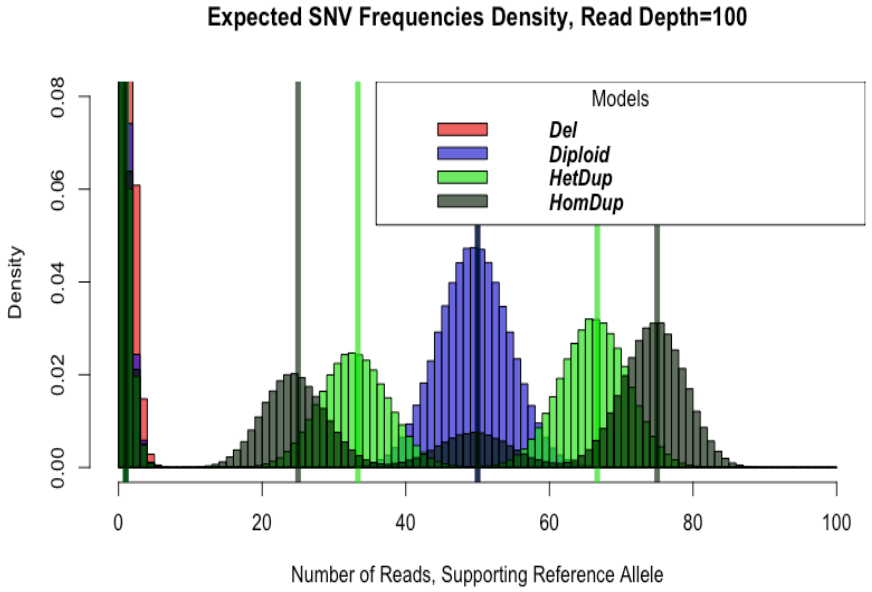


Fitting Statistical Models

Models for Read Depths



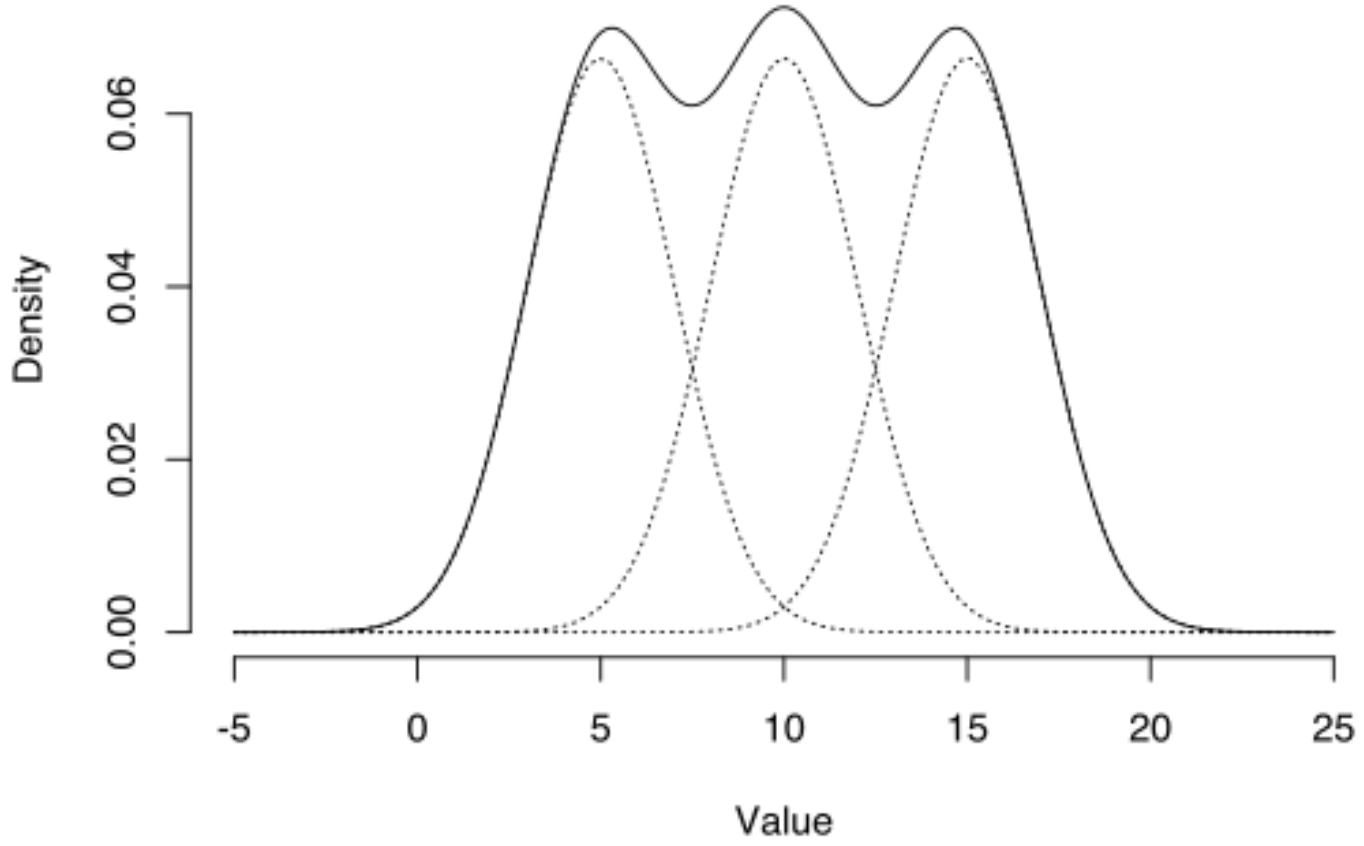
Models for B-allele Frequency (SNVs)



The height of the purple bar =
Likelihood of the data point 2.4
under null model or duplication model

Different type of models (states)

- As in HMM, we can put any states in our algorithm (image from https://www.statisticshowto.datasciencecentral.com/em-algorithm-expectation-maximization/)



Maximum Evidence Segment

LOGLIKELIHOODS OF THE MODELS

NORMAL	5	9	-2	-1	1	8	13	43	2	17	-9	10	45	-1
DUPLICATION	-3	-1	-5	-7	-8	-8	22	33	22	11	-1	61	72	-9
DELETION	-3	-1	33	17	5	41	7	-3	-2	1	-2	-1	19	1

LOGLIKELIHOOD DIFFERENCES OF THE MODELS

DUP-NORMAL: -8|-10|-3|-6|-9|-16|9 |-10|20| -6|8| 51| 27|-8

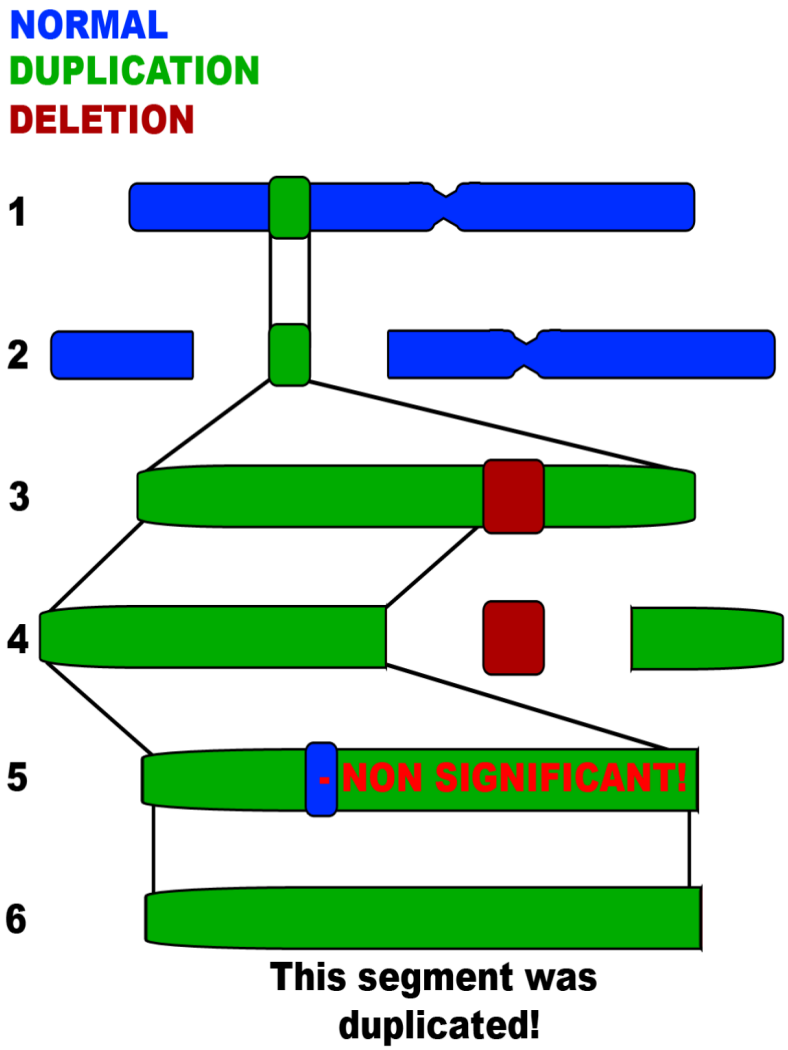
DEL-NORMAL: -8|-10|35|18| 4| 33|-6|-46|-4|-16|7|-11|-26| 2

1st Step: Find a Maximum Subarray Sum 100 in a first row, find a Maximum Subarray Sum 90 in the second

2nd Step: Detect CNV with the maximum Score (Duplication) at positions 8-12

3rd Step: Divide genomic region in 3 parts: 0-7, 8-12, 13 and repeat Step 1 for each sub-segment

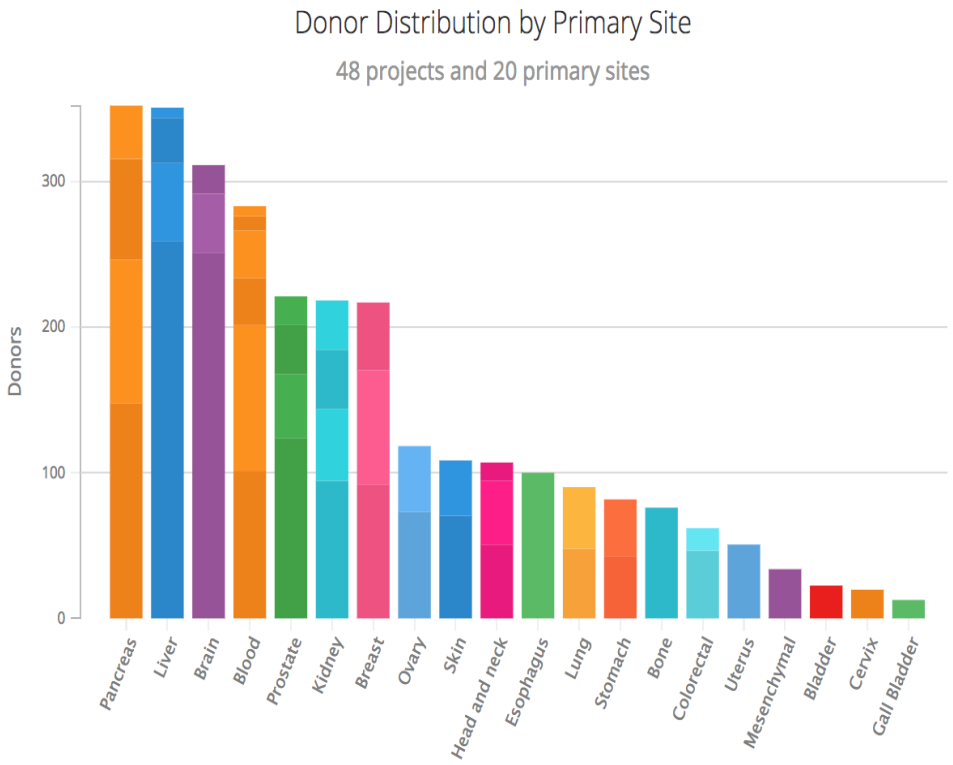
Recursive Maximum Evidence Segmentation



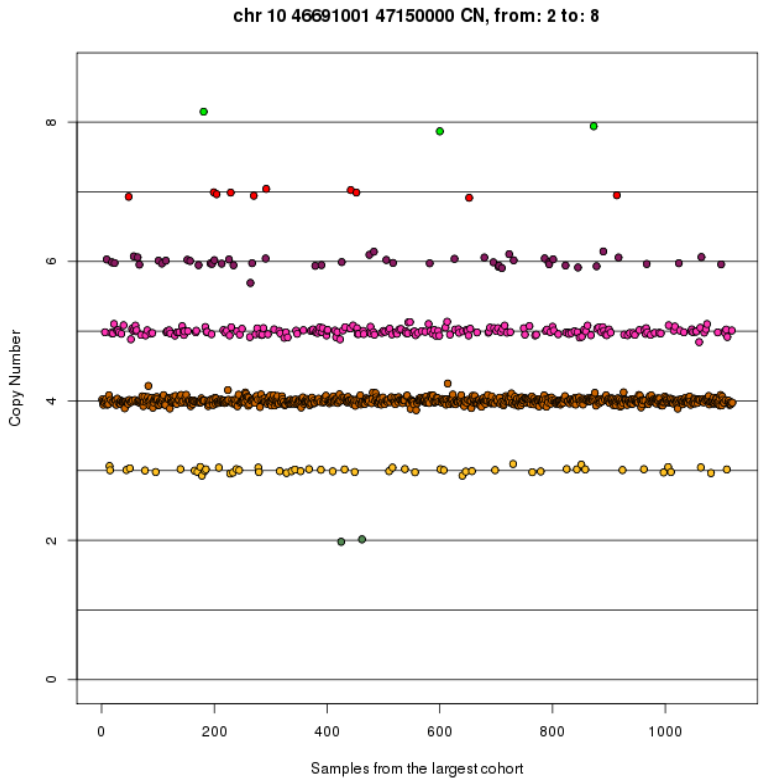
Results

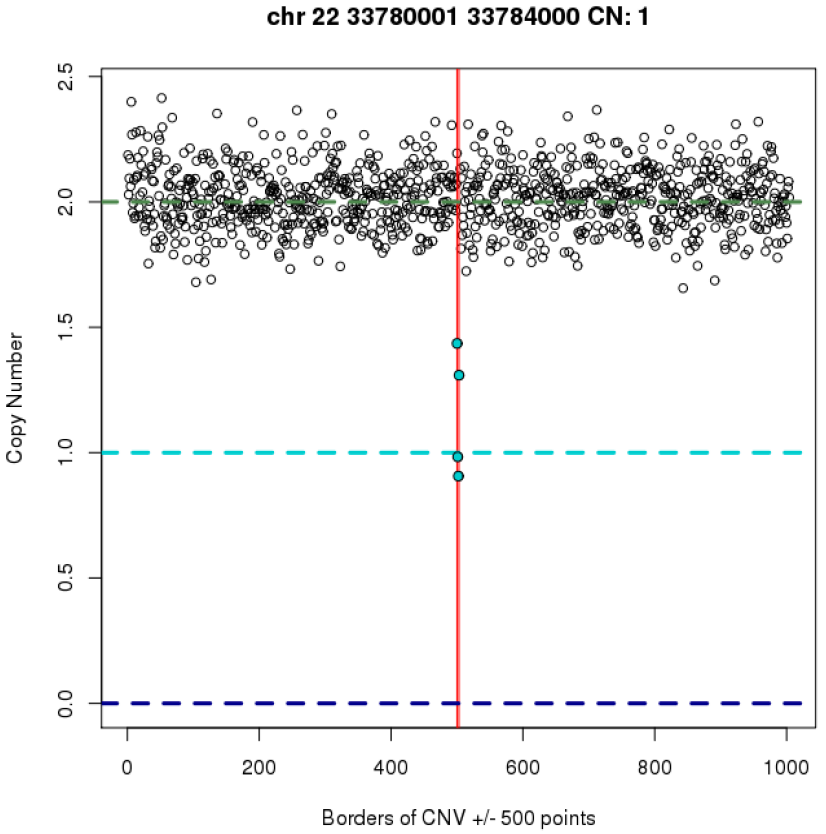
Case Study: Pan Cancer Analysis of Whole Genomes (PCAWG)

- Joint initiative by TCGA and ICGC
- >2800 donors, Whole Genome Sequenced
- Germline variants were studied (CNVs)
- Work Group 8: “Identification of cancer germline risk variants”

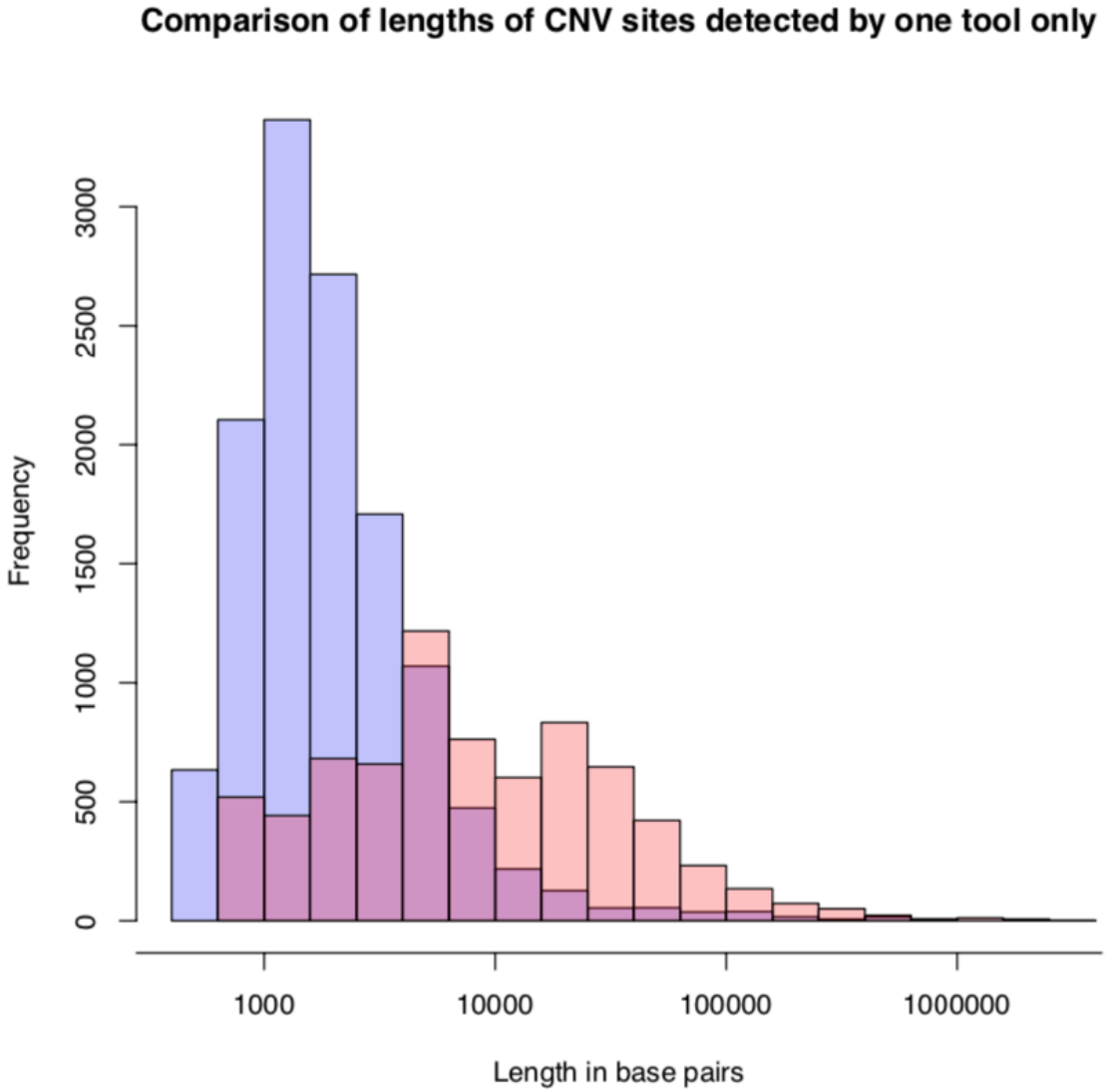


Common CNV vs Rare CNV

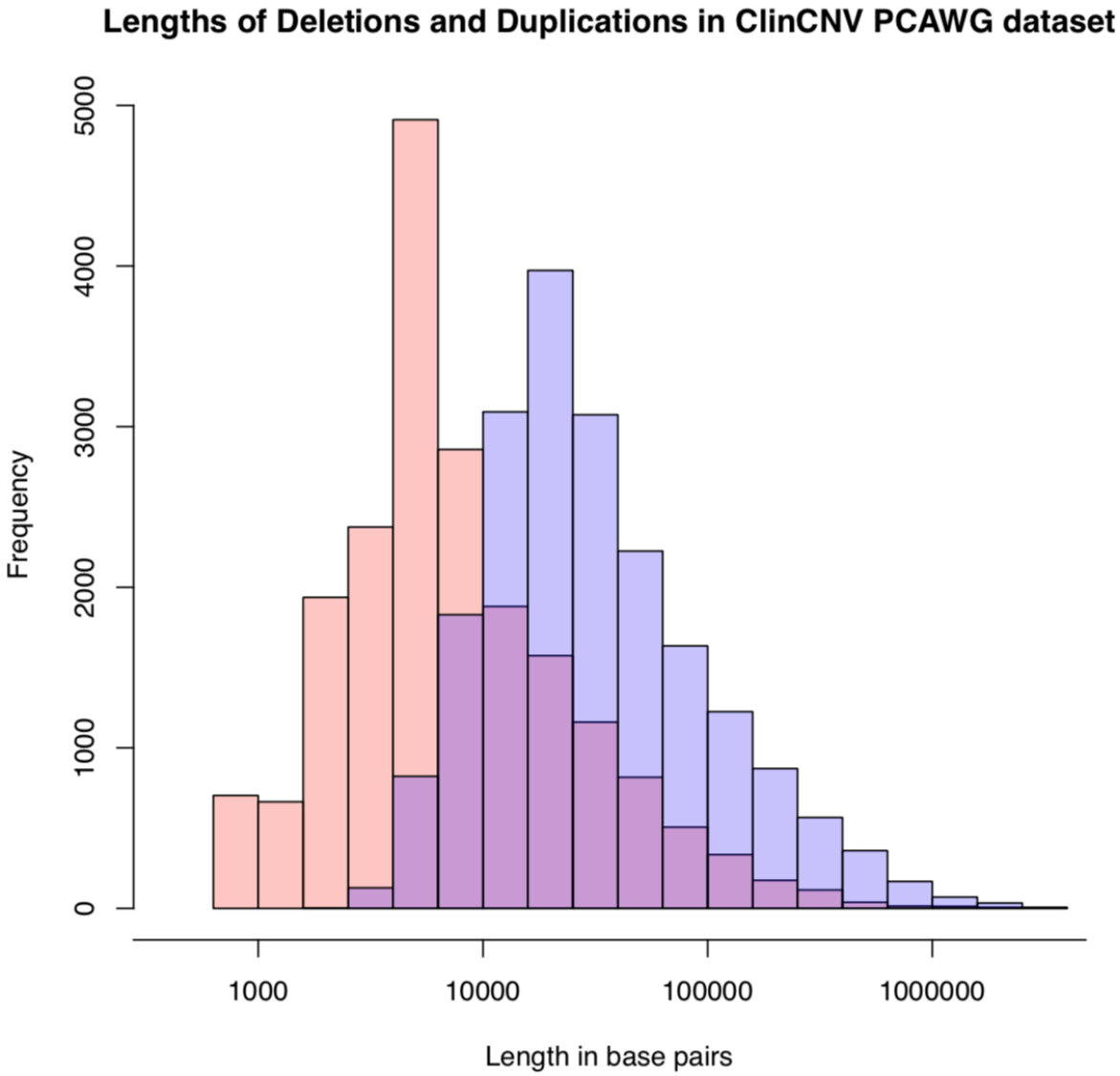




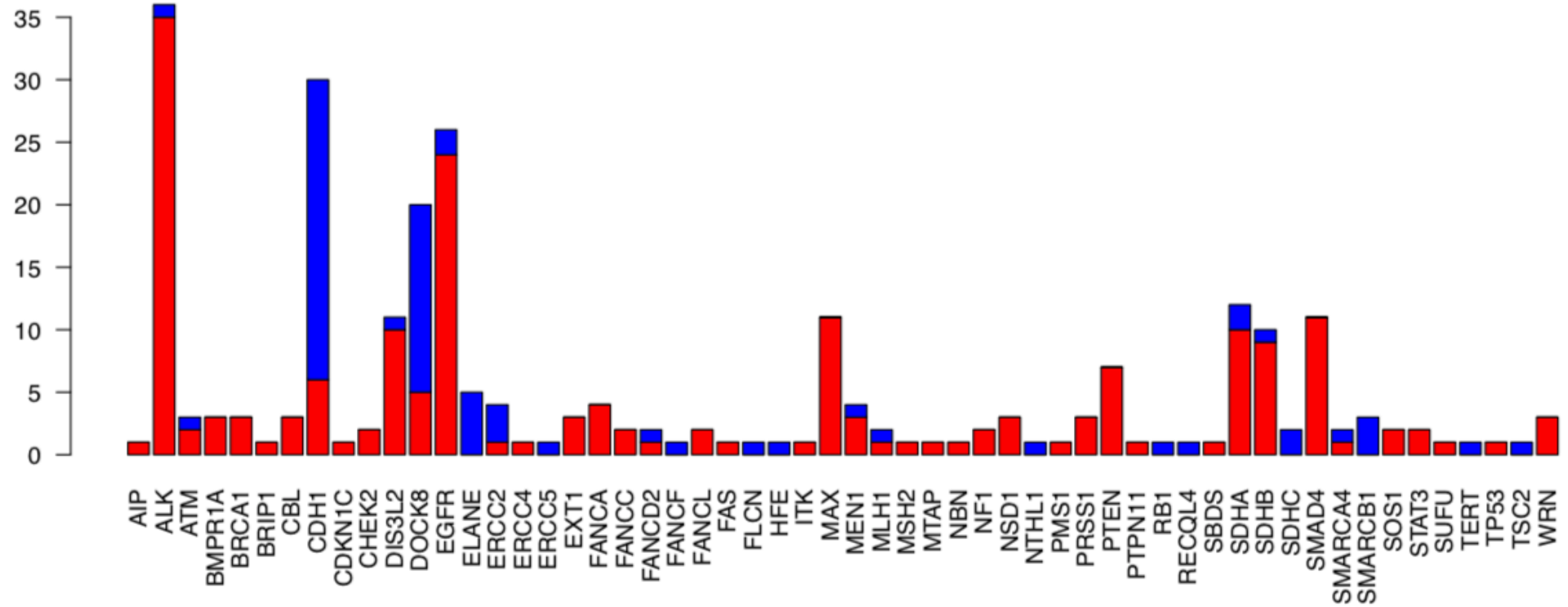
Paired end mapping method DELLY vs ClinCNV (5% FDR callsets)



Lengths of Deletions and Duplications



CNVs in Cancer Predisposition Genes

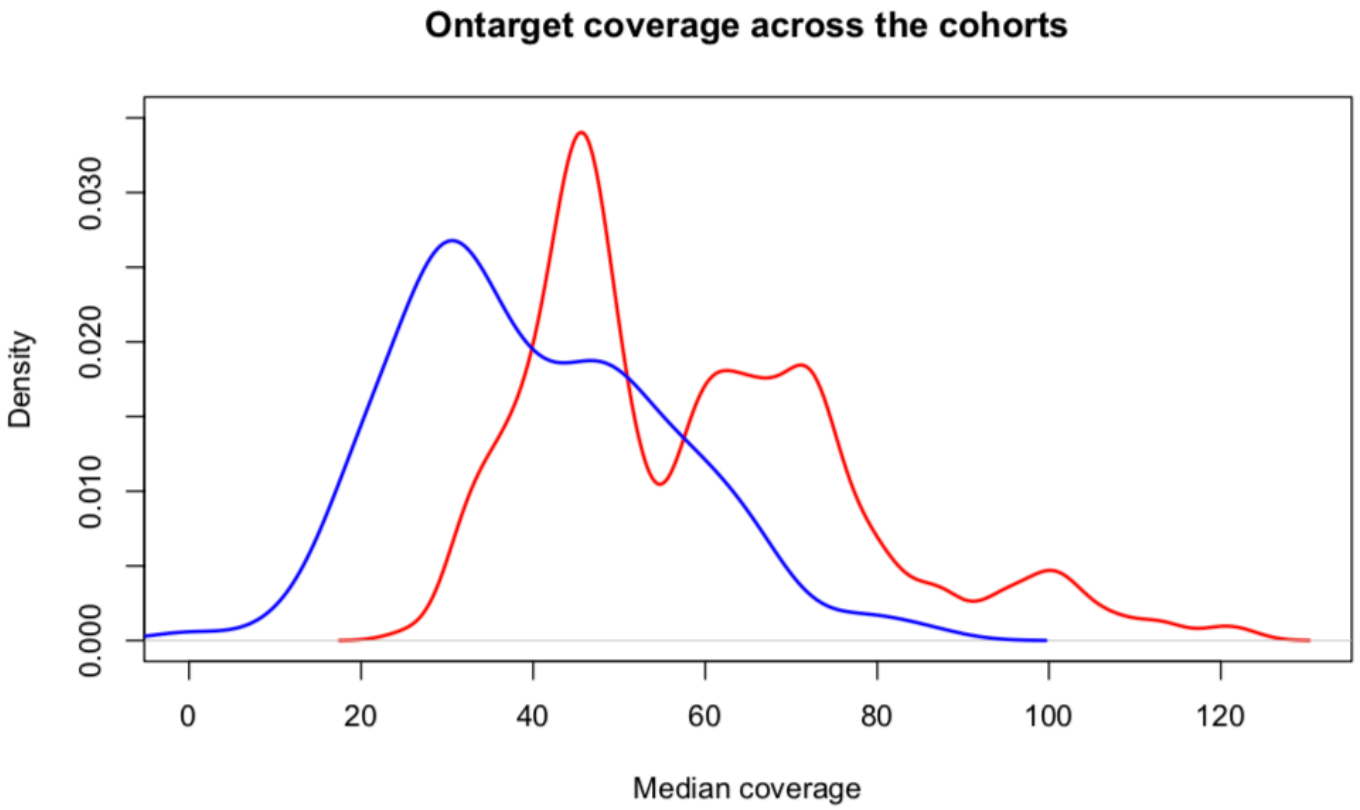


CRG Centre for Genomic Regulation

upf. Universitat Pompeu Fabra Barcelona

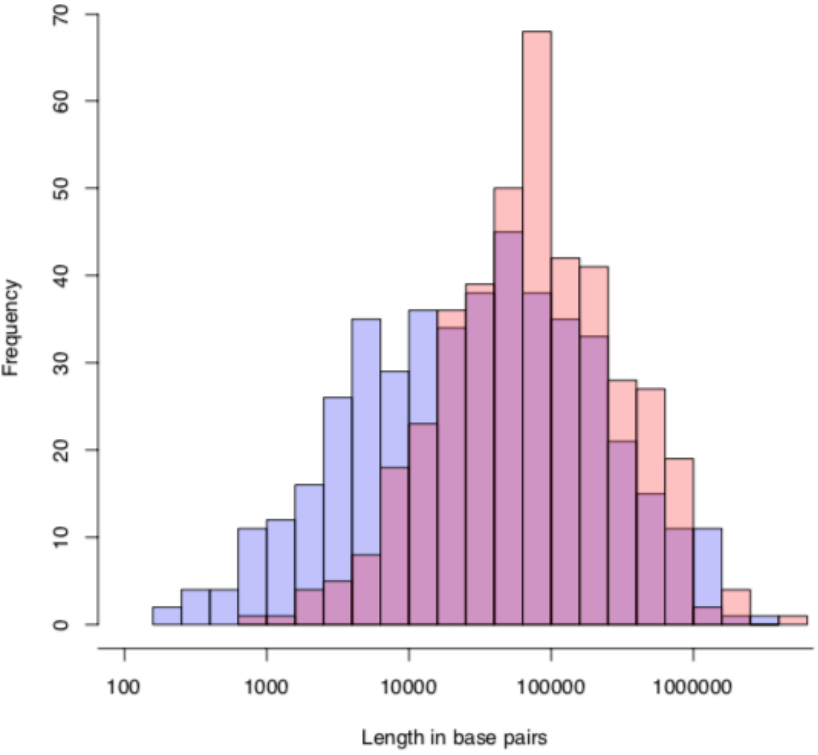
Germline CNVs in WES cohort (CLL)

- 435 samples sequenced with 2 different panels and different median depth
- Red shows deletion, blue: duplication

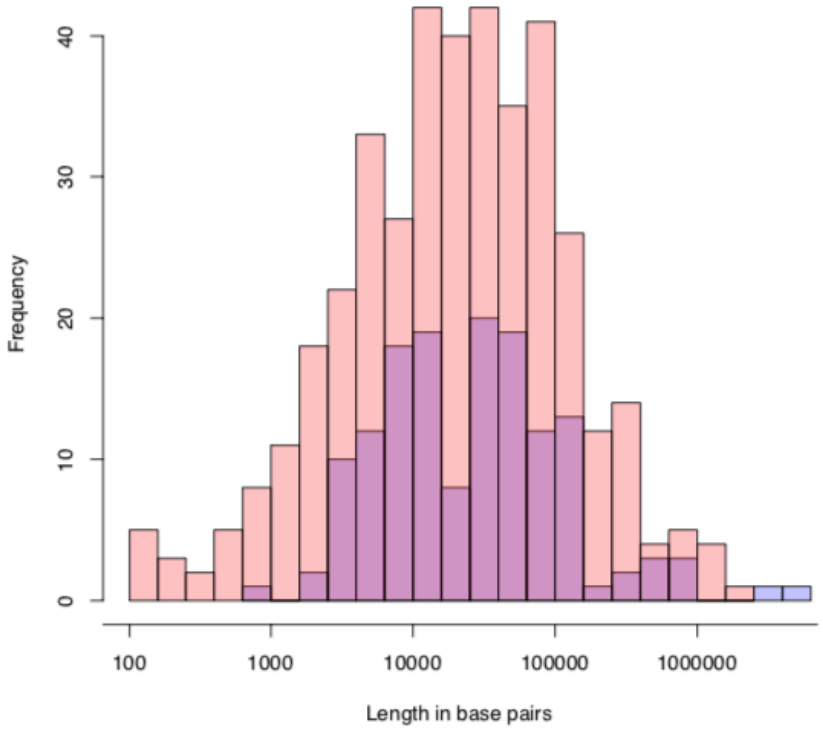


Germline CNVs in WES cohort (CLL)

- 435 samples sequenced with 2 different panels and different median depth
- Red shows deletion, blue: duplication

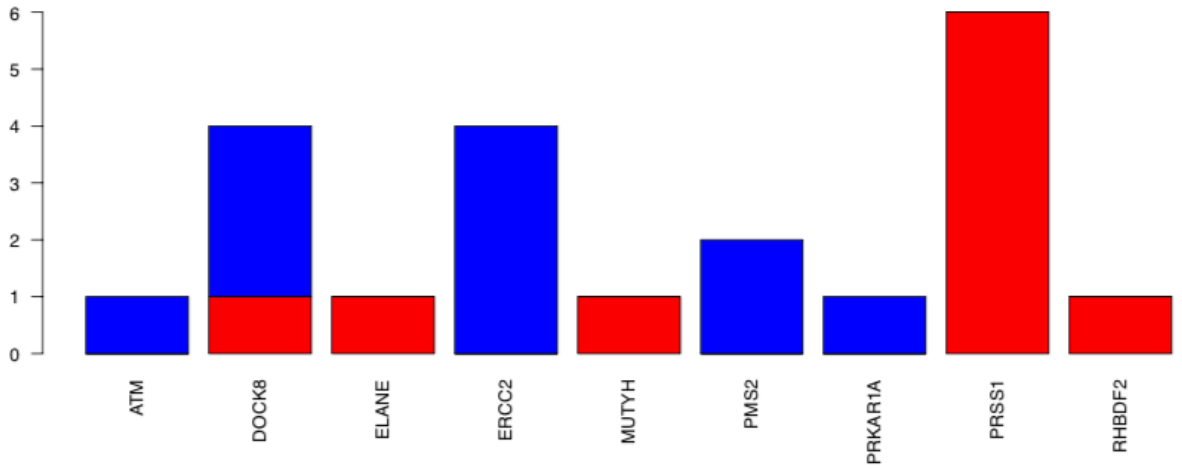


(a) Length of CNV sites (v4, 5% FDR).

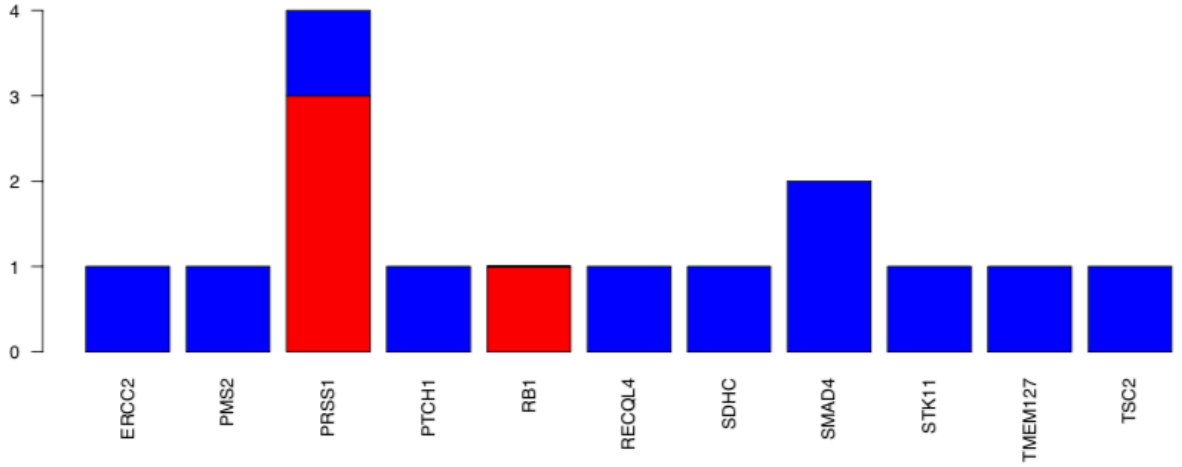


(b) Length of CNV sites (v5, 12% FDR).

CNVs in Cancer Predisposition Genes



(a) CNVs in cancer predisposition genes (v4 dataset, 5% FDR).



(b) CNVs in cancer predisposition genes (v5 dataset, 12% FDR).

ExomeDepth comparison

- 40 WES samples with available BAMs

	FDR for sites	FDR for variants	Overall # of sites
ClinCNV deletions (30)	0.143	0.004	238
ClinCNV deletions (34)	0.123	0.003	133
ExomeDepth deletions	0.149	0.005	130
ClinCNV duplications	0.25	0.02	179
ExomeDepth duplications	0.067	0.001	182

Comparison: Arrays vs NGS (50kbps)

- Can arrays be replaced with NGS? – which **arrays?**

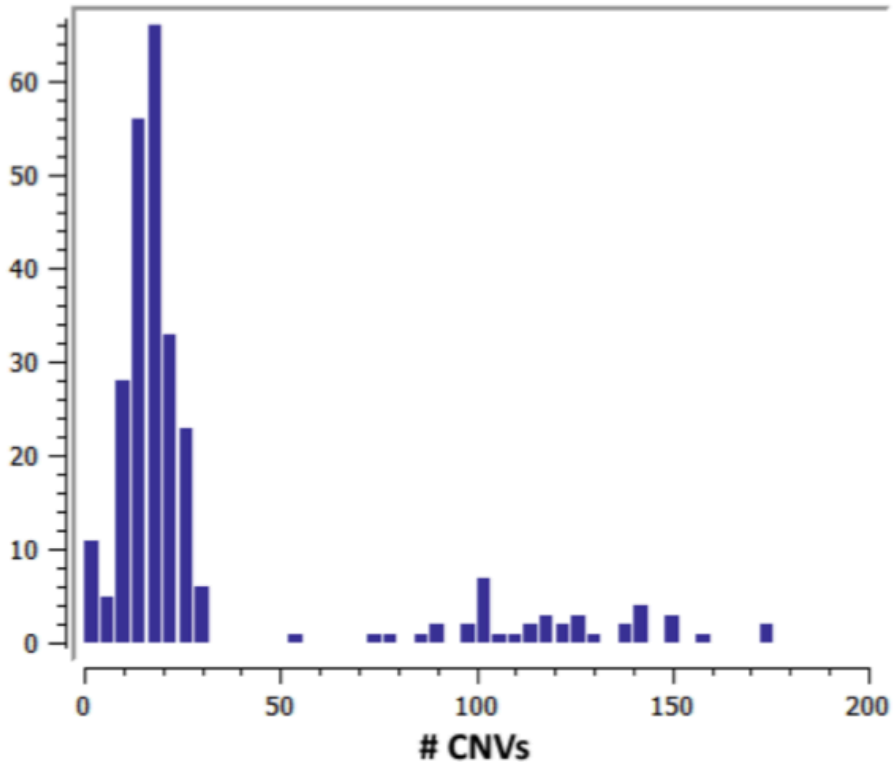
Type of microarray platform	Number of samples
CytoScan HD	39
CytoScan 750K	217
CytoScan Optima	12
Overall:	268

- and which **NGS?**

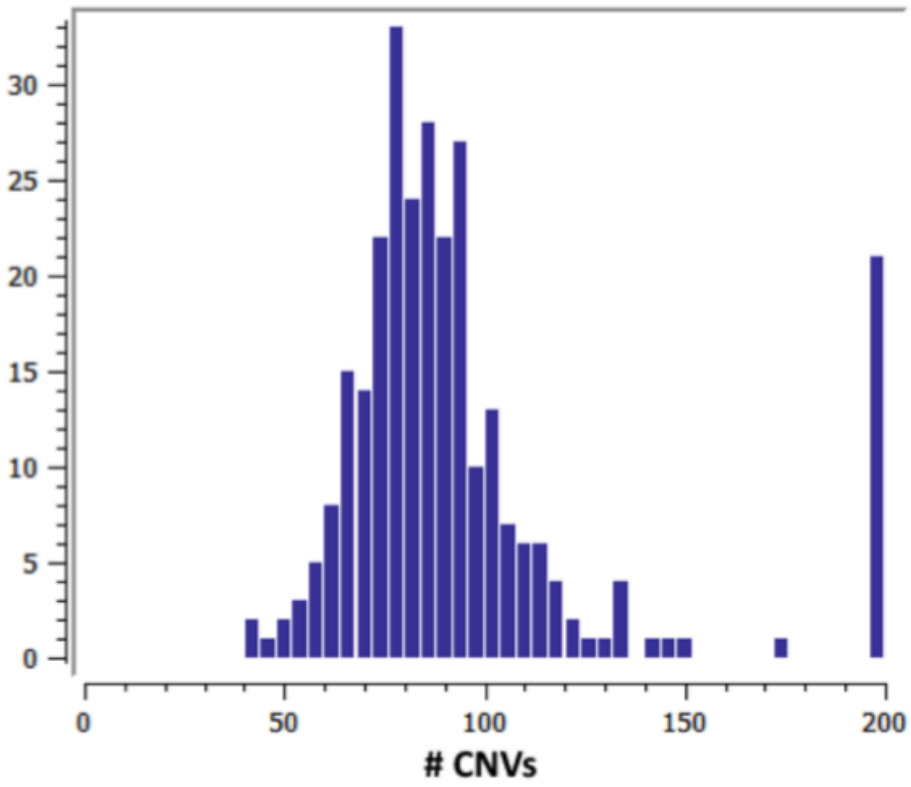
Type of NGS platform	Number of samples
WGS TruSeq PCR-free	9
WES ssHAEv6	197
WES ssHAEv7	79
Overall:	285

Comparison: Arrays vs NGS (50kbps)

- Number of CNVs for arrays



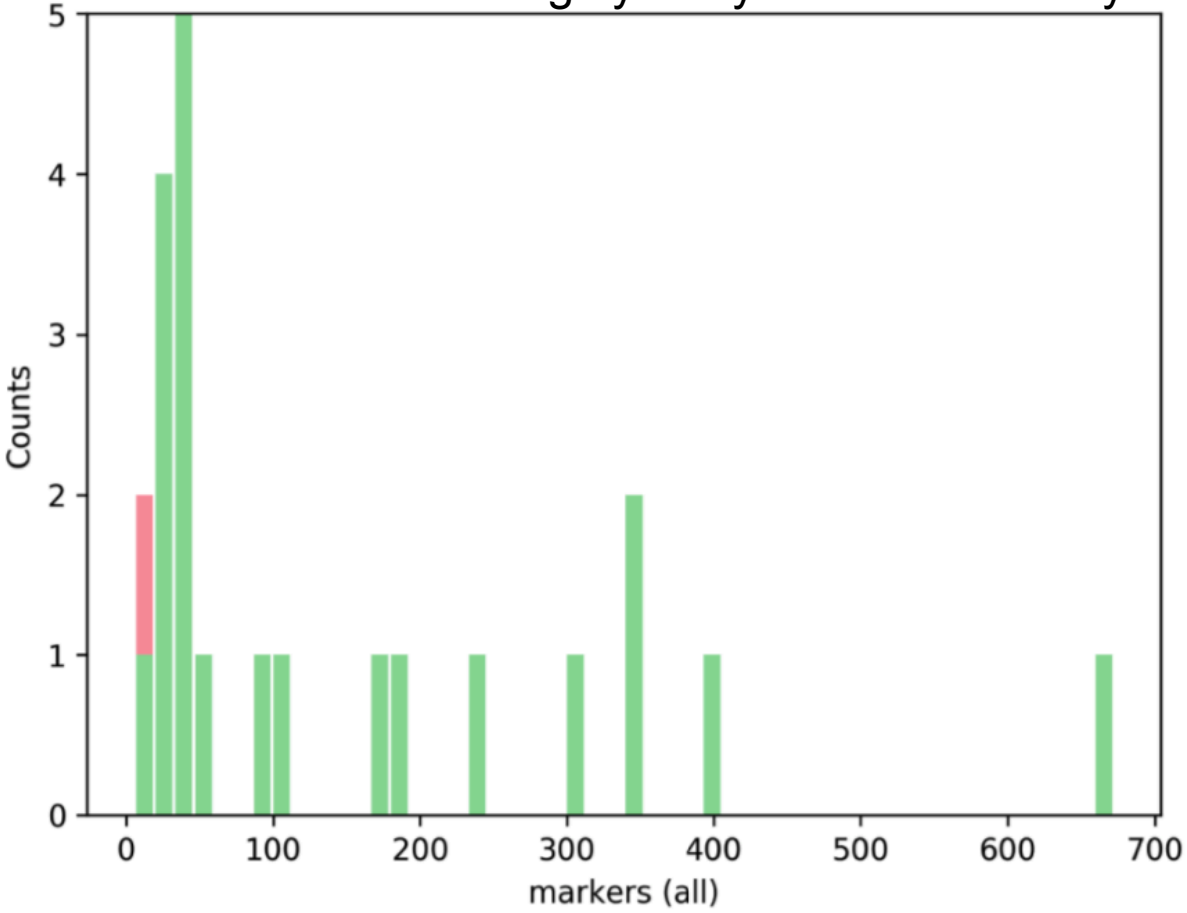
- Number of CNVs for NGS



Pictures from: Marc Sturm

Comparison: Arrays vs WGS (50kbps)

1 CNV is missed and it is highly likely to be FP of arrays



Author of histograms: Marc Sturm

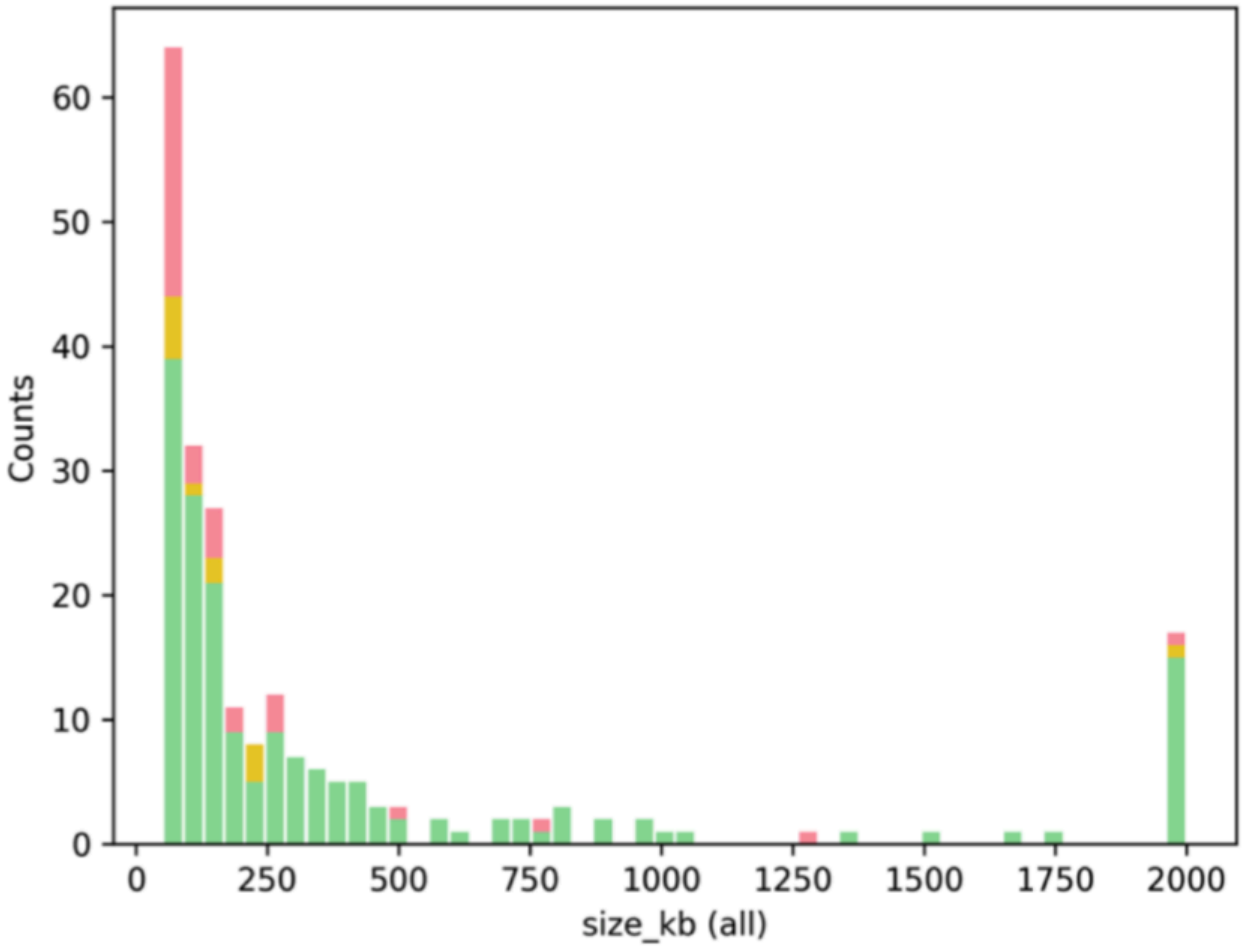
Are arrays perfect?

- After visual analysis, around **50%** of long False Negative CNVs were **“likely artifact”**

- Below 40 kbps around **85%** of array detected CNVs were not presented in WGS

Comparison: Arrays vs WES (at least 1 probe, 50kbps)

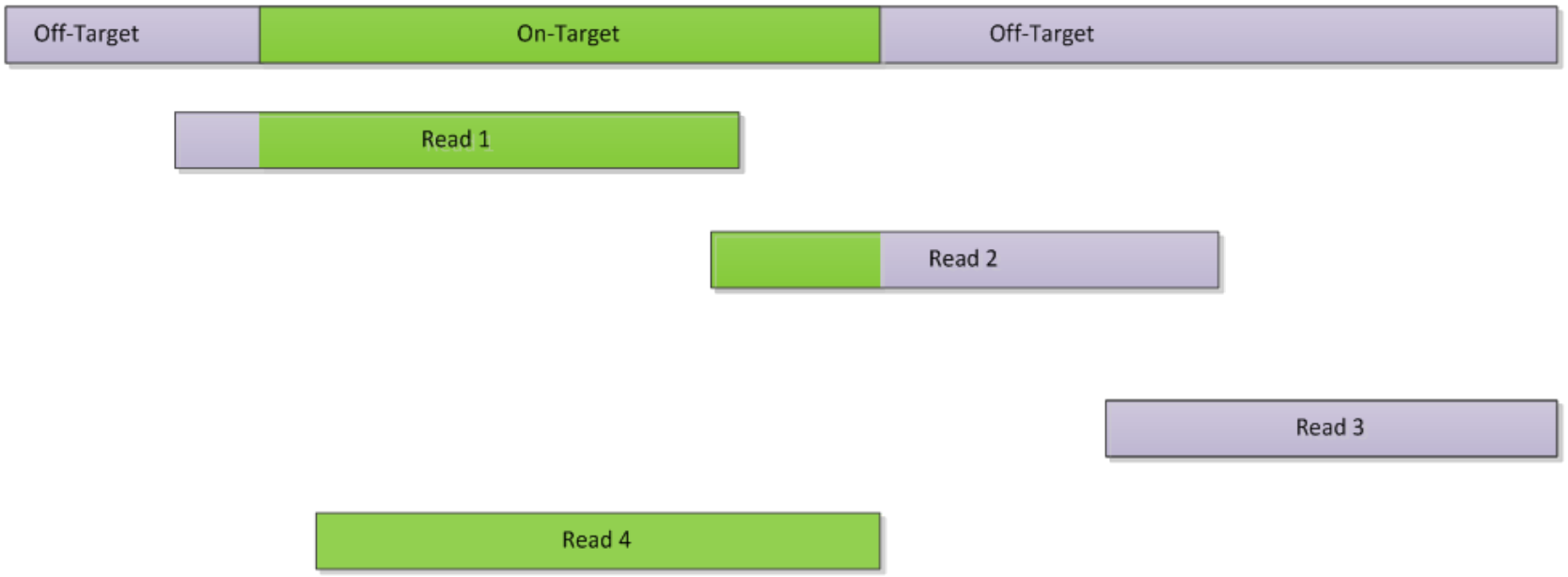
Only around 20% are missed!



Picture from: Marc Sturm

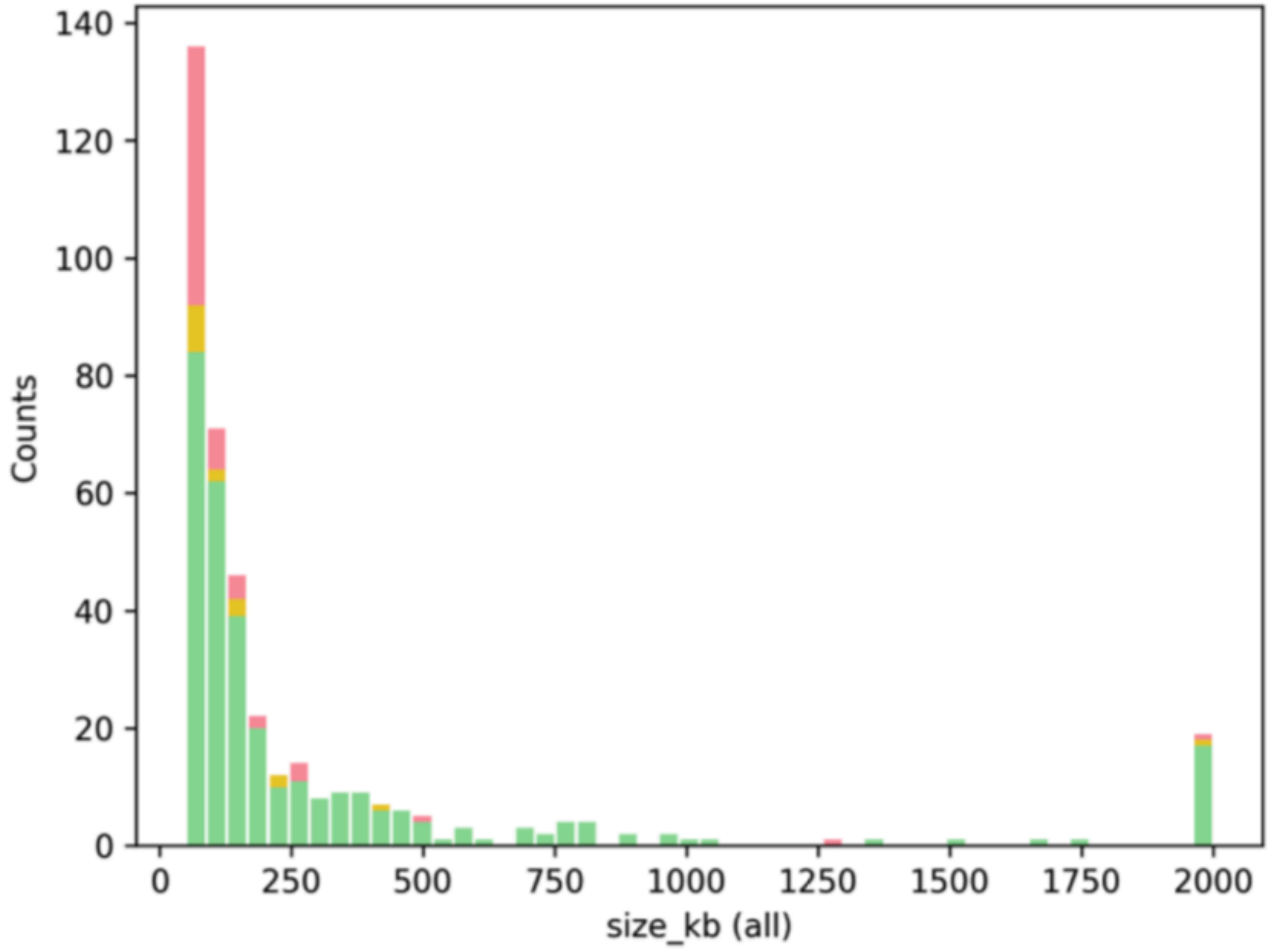
Off-target reads

Picture from: https://eu.idtdna.com/pages/education/decoded/article/how-important-are-those-ngs-metrics



Comparison: Arrays vs WES (off-target reads, 50kbps)

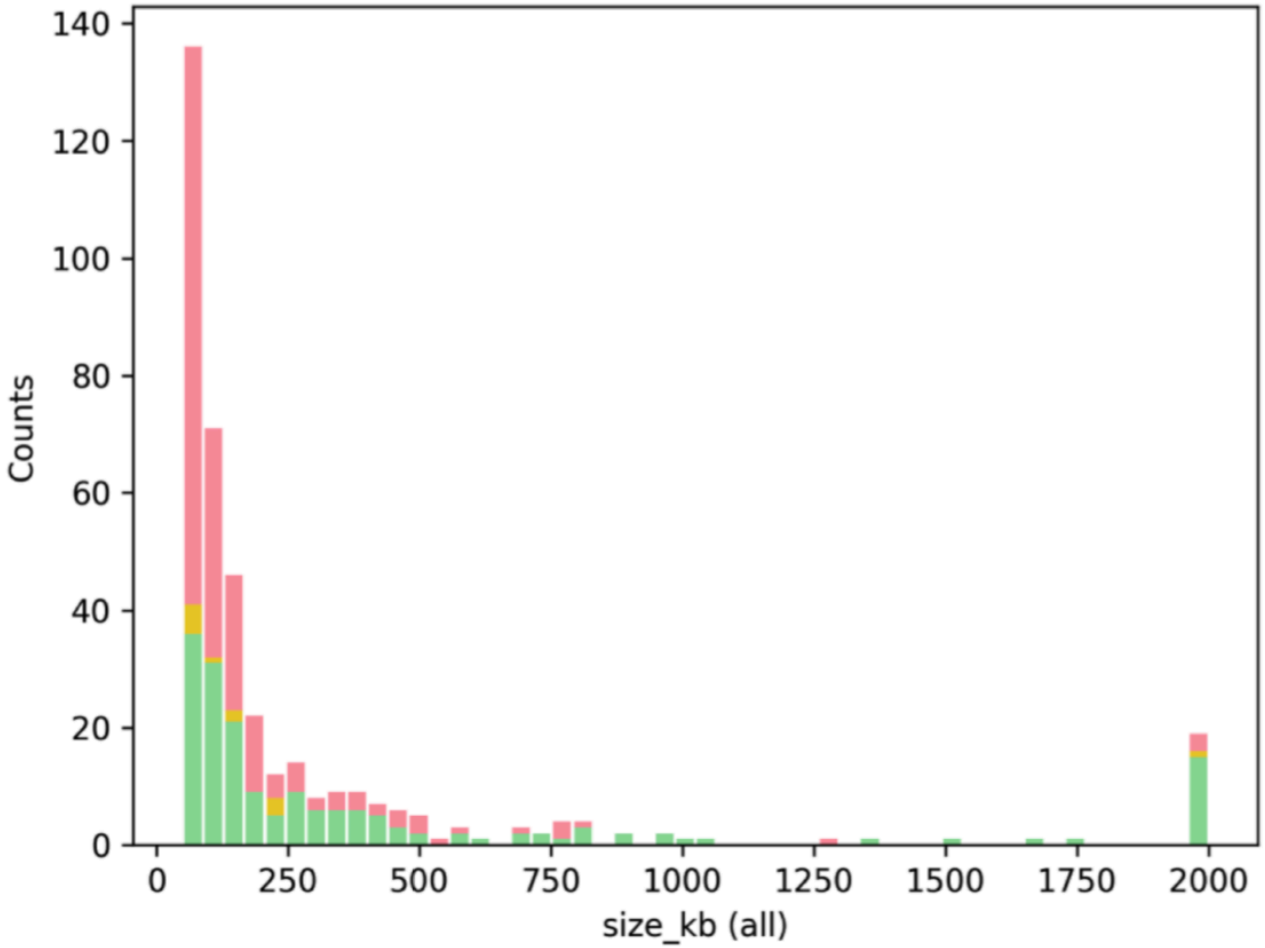
Only around 27% are missed!



Picture from: Marc Sturm

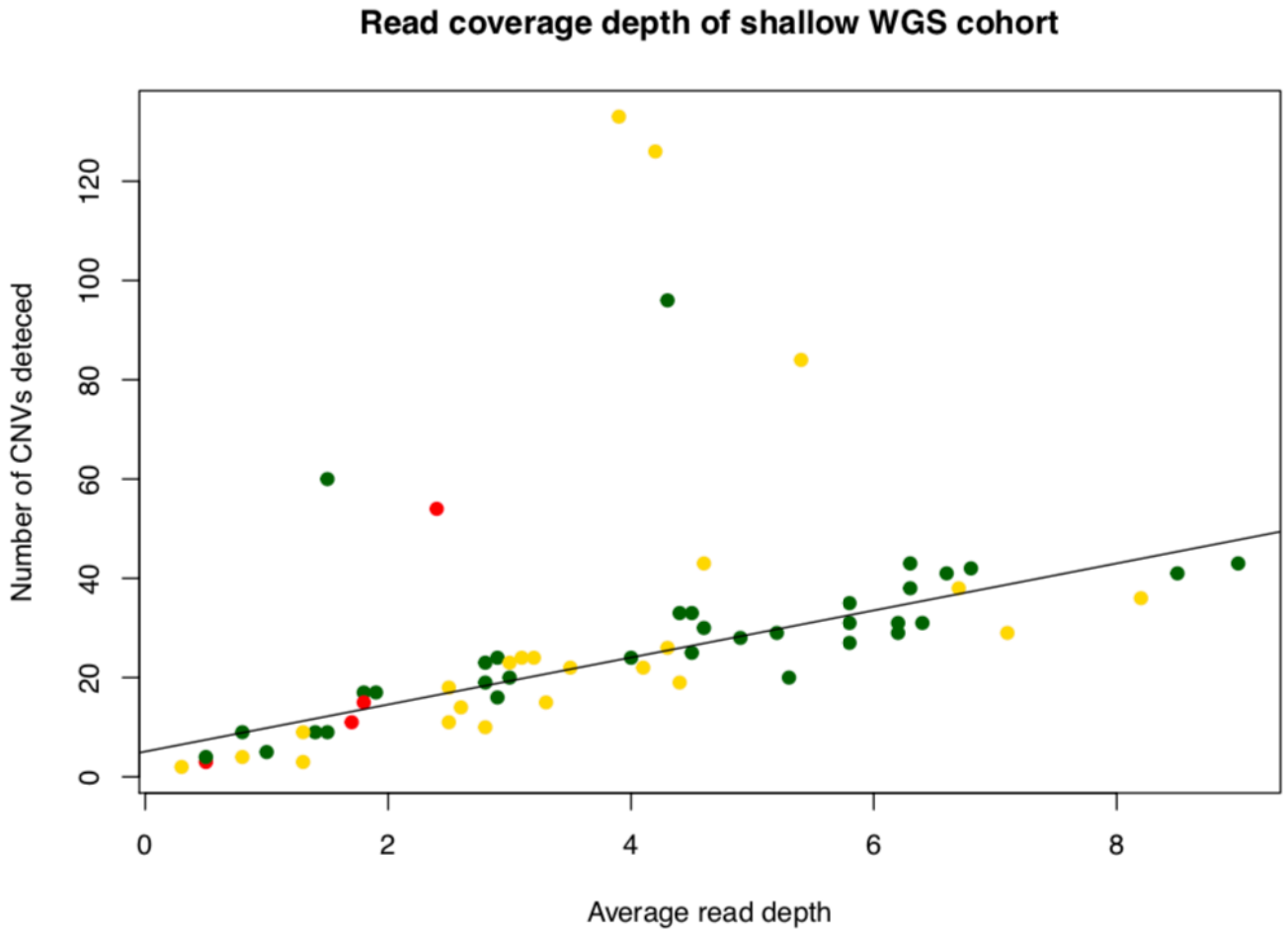
Comparison: Arrays vs WES (50kbps)

More than half of CNVs are missed! (mainly intronic/intergenic)



Picture from: Marc Sturm

Comparison: Arrays vs shallow WGS



Can ClinCNV replace arrays?

Arrays guarantee a resolution of 50kb. How about ClinCNV for NGS?

- **42x WGS**: yes, resolution is 3kbps

Can ClinCNV replace arrays?

Arrays guarantee a resolution of 50kb. How about ClinCNV for NGS?

- **42x WGS**: yes, resolution is 3kbps 
- Shallow **4x WGS**: yes, resolution is ~30kbps 

Can ClinCNV replace arrays?

Arrays guarantee a resolution of 50kb. How about ClinCNV for NGS?

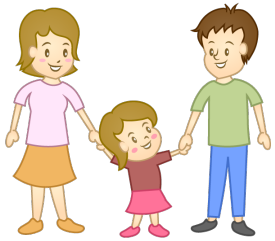
- **42x WGS**: yes, resolution is 3kbps 
- Shallow **4x WGS**: yes, resolution is ~30kbps 
- **WES** for **coding regions**: around 20% of CNVs are missed 

Can ClinCNV replace arrays?

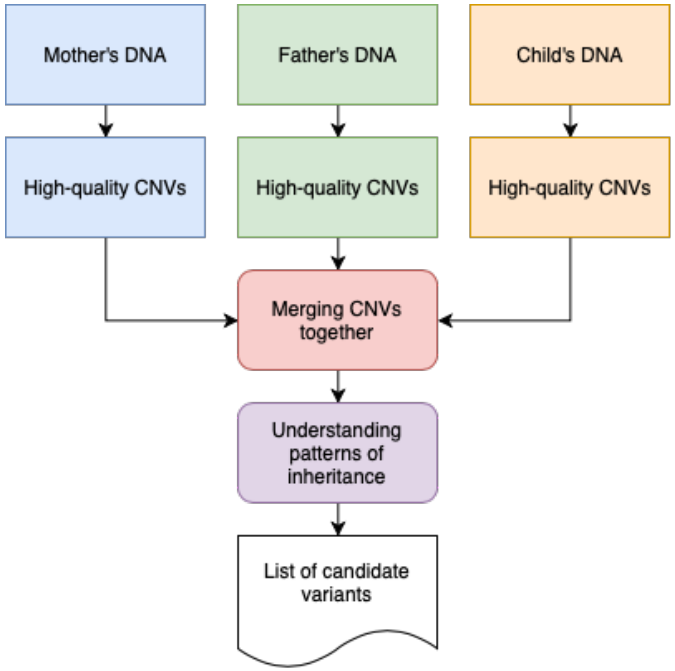
Arrays guarantee a resolution of 50kb. How about ClinCNV for NGS?

- **42x WGS**: yes, resolution is 3kbps 
- Shallow **4x WGS**: yes, resolution is ~30kbps 
- **WES** for **coding regions**: around 20% of CNVs are missed 
- **WES** with **off-target reads:** 200 kbps resolution, increased false positives and imprecise boundaries 

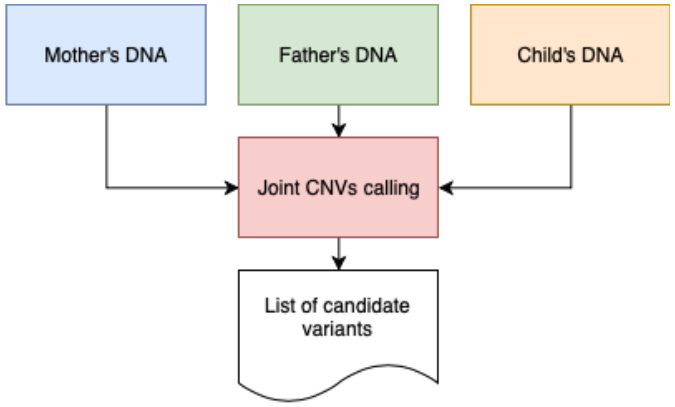
Detection of CNVs in Trios



- Single sample calling of CNVs



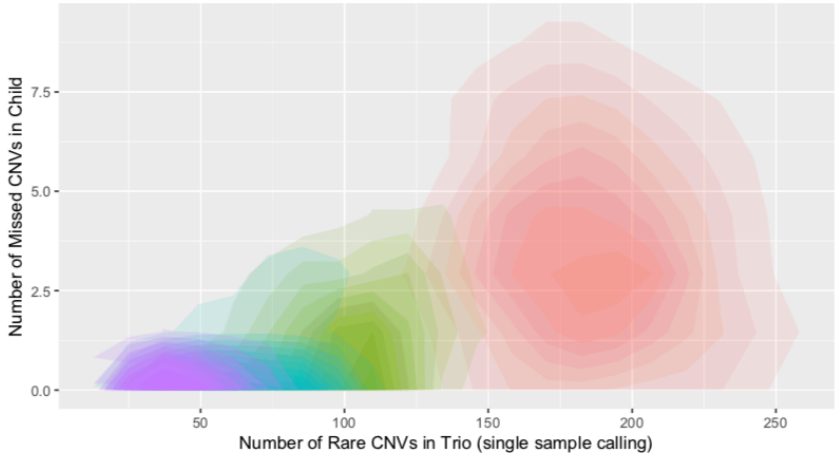
- Joint calling of CNVs



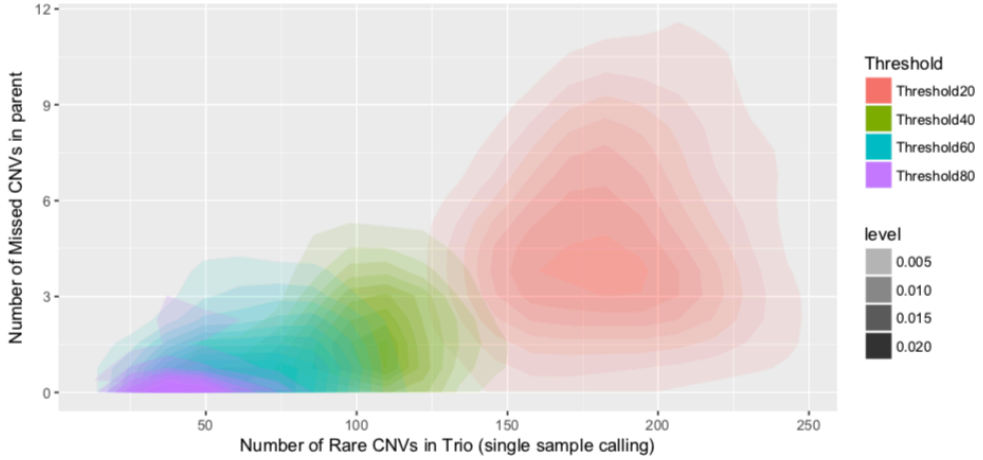
Advantage of Joint-Trio vs. Single-Sample Calling

Potential calling errors when using Single-Sample Calling:

False negative: CNV **in child missed** but **detected** in one of the parents

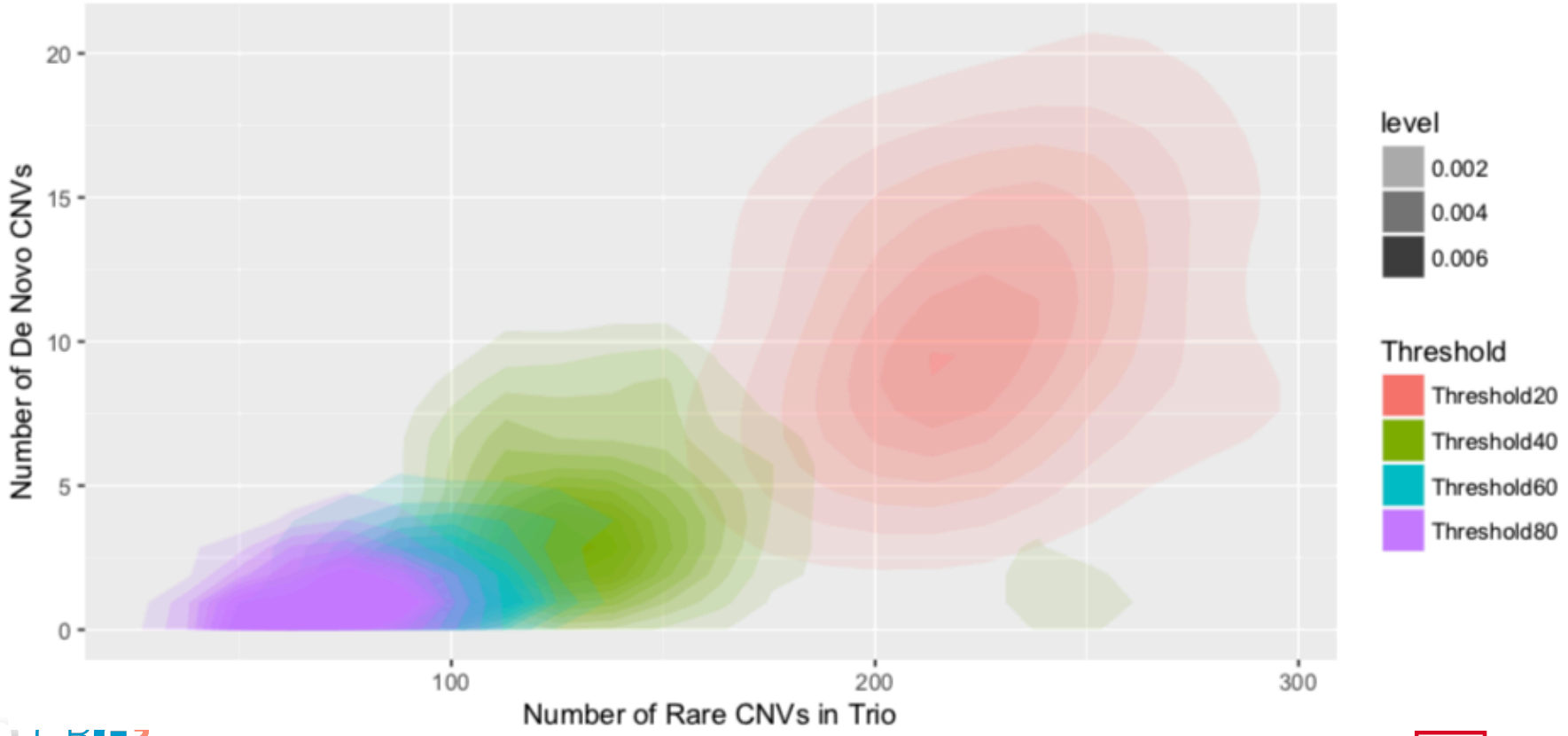


False positive 'de novo': CNV missed in one of the parents, but **detected in child**



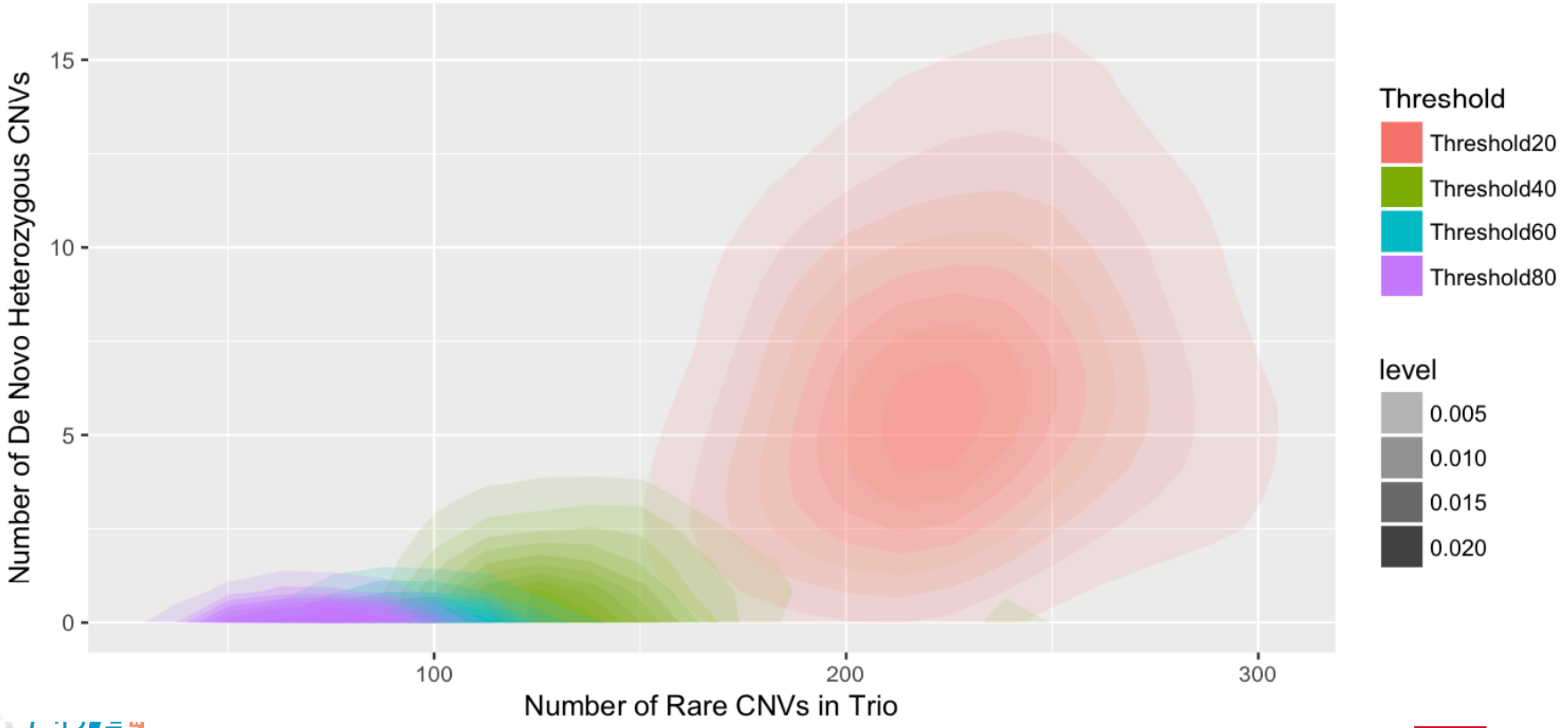
Detection of de novo CNVs in Joint-Trio Calling

- Mother copy-numbers: **A1, A2**
- Father copy-numbers **B1, B2**
- Child copy-numbers (inherited): **A1+B1, A2+B2, A1+B2, A2+B1**
- ***Novel CNV in child*** *('de novo mutation'): e.g. A1, C1*



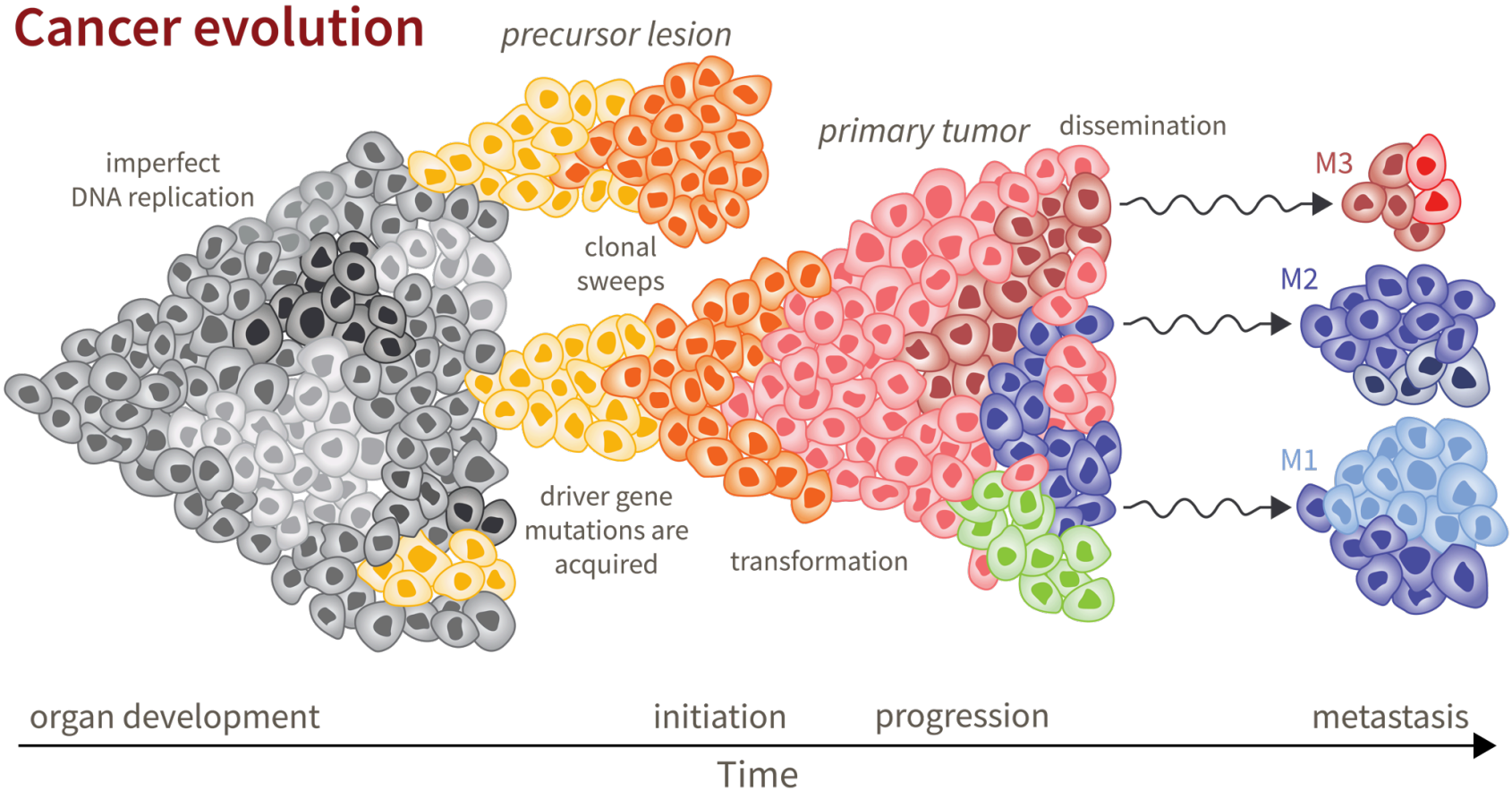
Detection of de novo CNVs in Joint-Trio Calling

- Mother copy-numbers: **A1, A2**
- Father copy-numbers **B1, B2**
- Child copy-numbers (inherited): **A1+B1, A2+B2, A1+B2, A2+B1**
- ***Novel CNV in child*** *('de novo mutation'): e.g. A1, C1, but C1 is likely 1 or 3 copies*



Somatic CNAs identification

Cancer

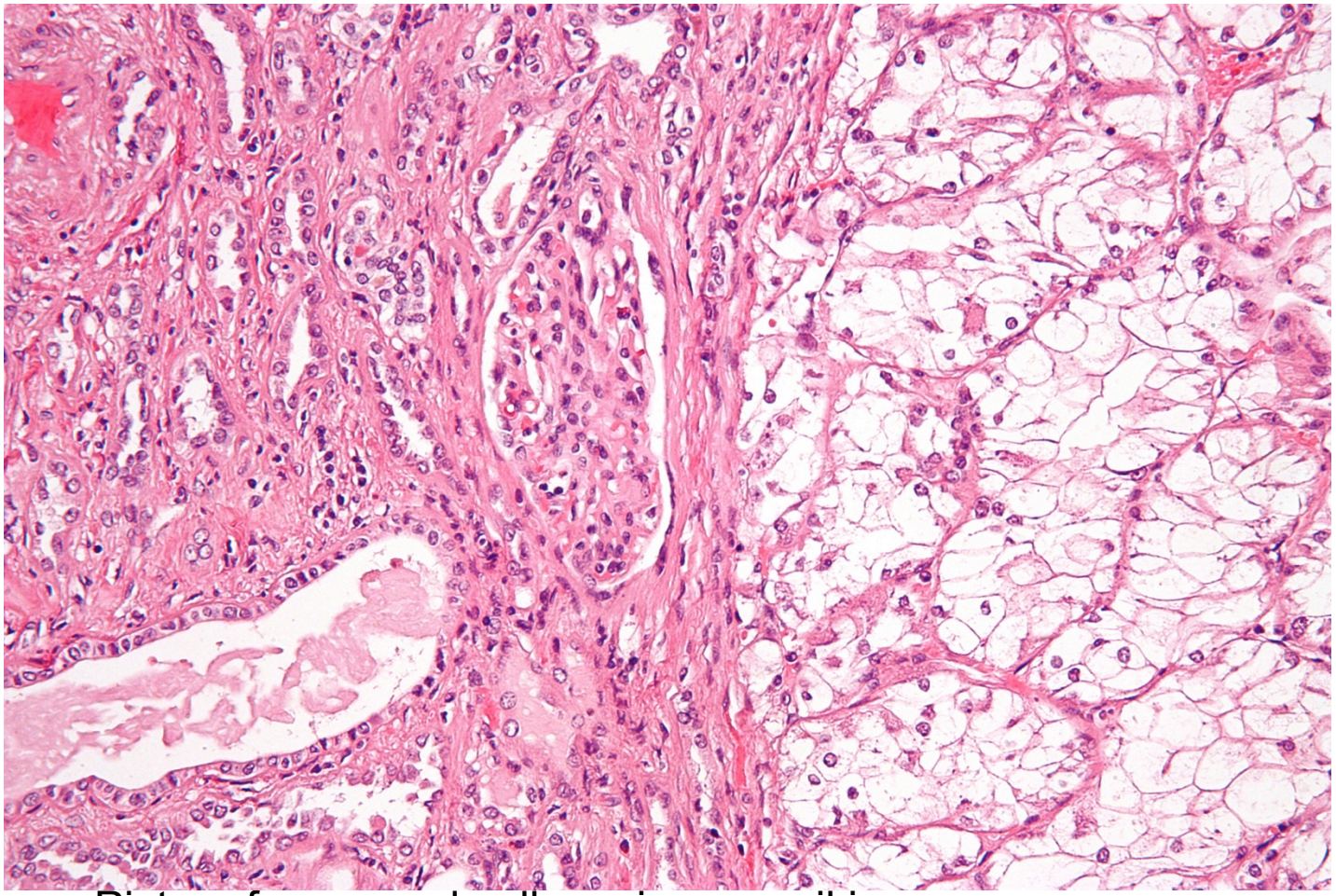


Picture from: https://reiterlab.stanford.edu

CNA states in Cancer

- We have 2 alleles in each genomic locus: **A1, A2,** each allele can have a different number of copies
- In tumor genomes these alleles may change copy-number: A1=>**T1,** A2=>**T2**
- Estimates of CNAs also depend on:
  - Cancer Cell Fraction (CCF) of CNAs
  - Purity of the tumor samples

Cancer Biopsy Purity



Picture from: renal cell carcinoma, wiki page

CNA states in Cancer

- T1 >= T2, we allow: T1 from **0** to **30** copies, T2 from **0** to **4** copies

- CCF can be from **5%** to **100%** with the step of **2.5%**

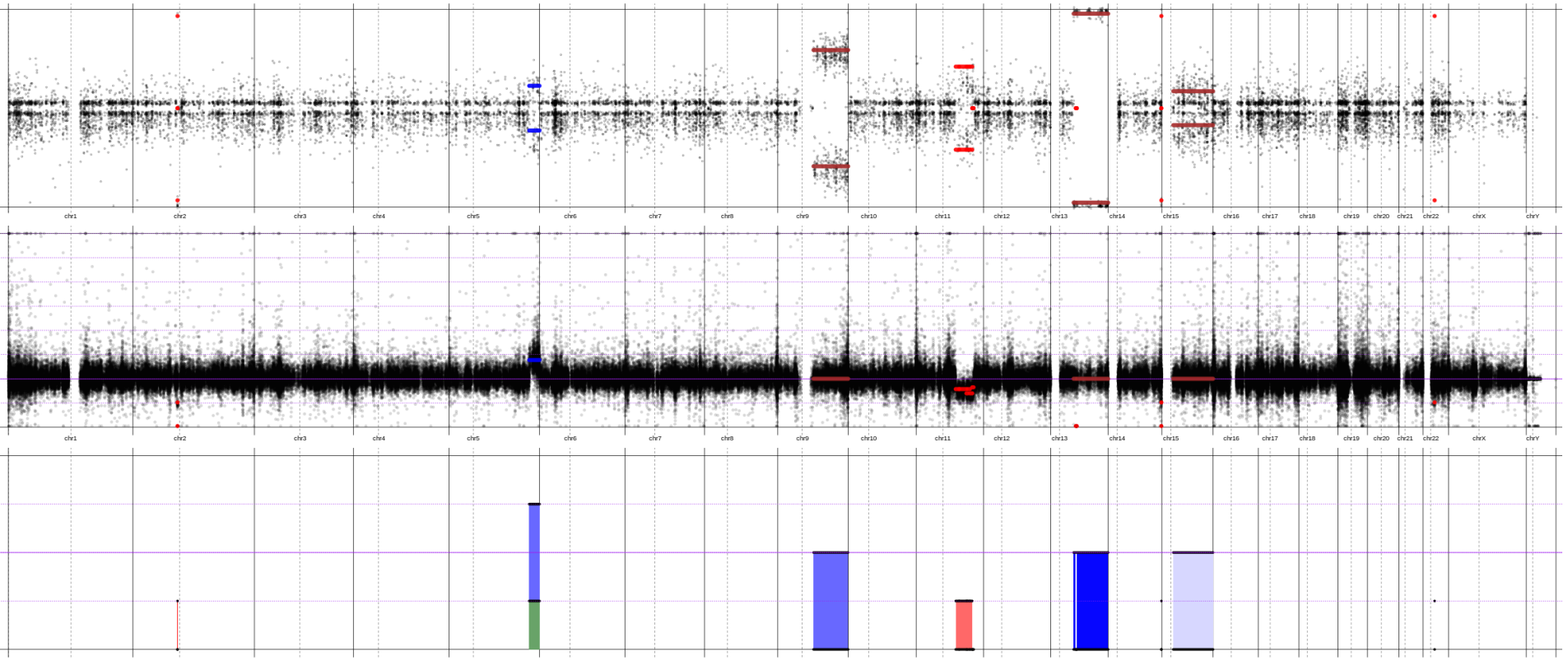
- This results in around **3000** states!

Recursive Maximum Evidence Segmentation: Complexity

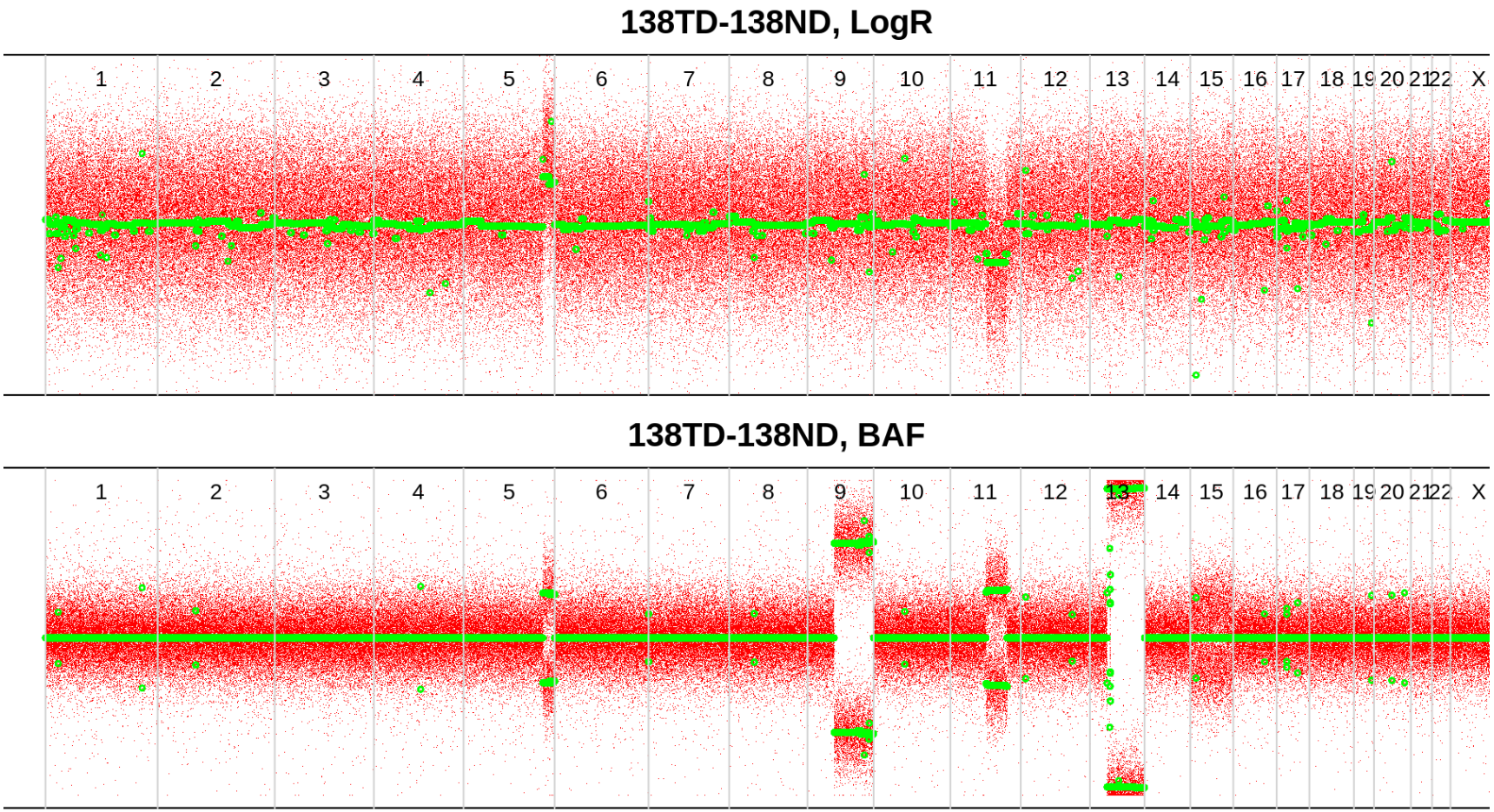
- HMM (Baum-Welch, Viterbi): **O(n * (s^2)) –** extremely slow when states **s** is large**!**
- CBS: **O(n * k) –** does not depend on number of states, but **requires post-filtering** and difficult to perform for 2 separate signals
- ClinCNV: efficient, linearly dependent on s! States are clearly defined as in HMM. Complexity of ClinCNV: **O(s * n * k)**

s − number of states, k − number of CNVs,
n − length of the genome

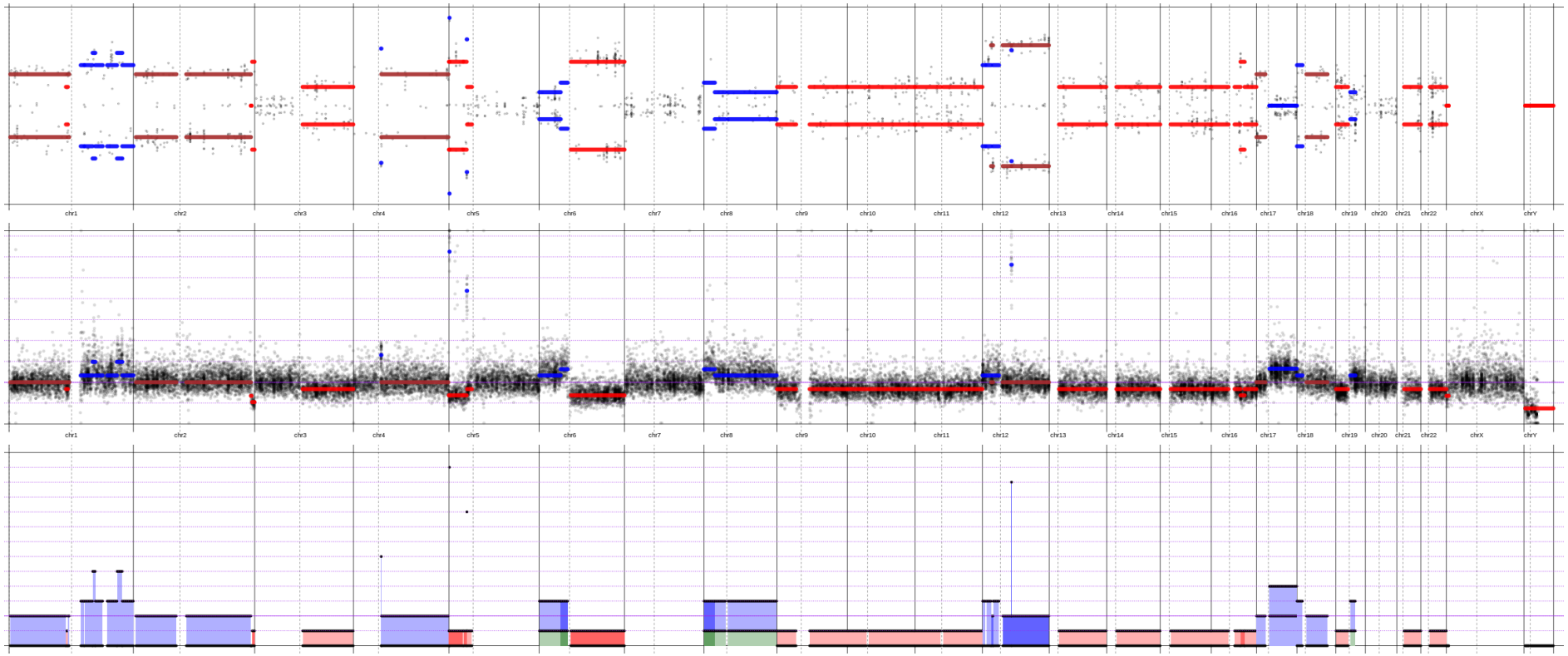
CLL sample, 3 clones, WES



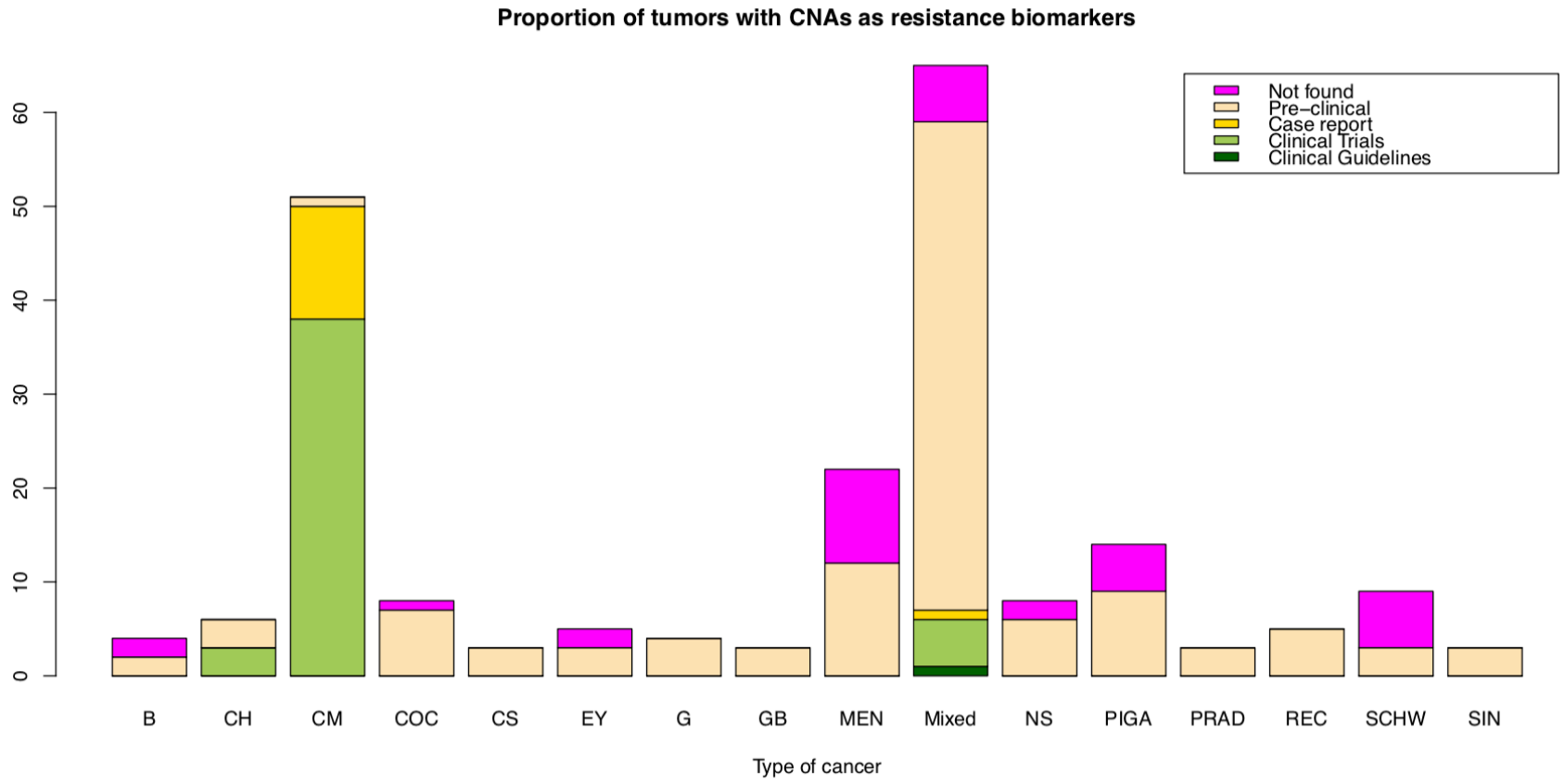
CLL sample, 3 clones, arrays



Tumor from diagnostics, panel sequencing



CNAs as biomarkers of response, resistance and toxicity



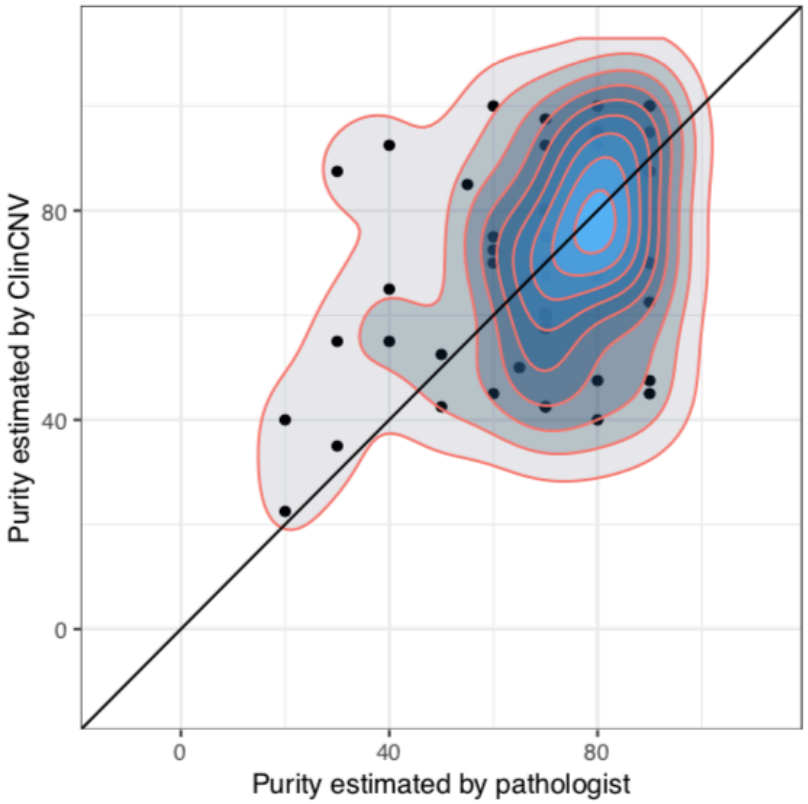
Comparison with other tools

- ClinCNV has the highest power of detection and provides sufficient number of metrics for quality control

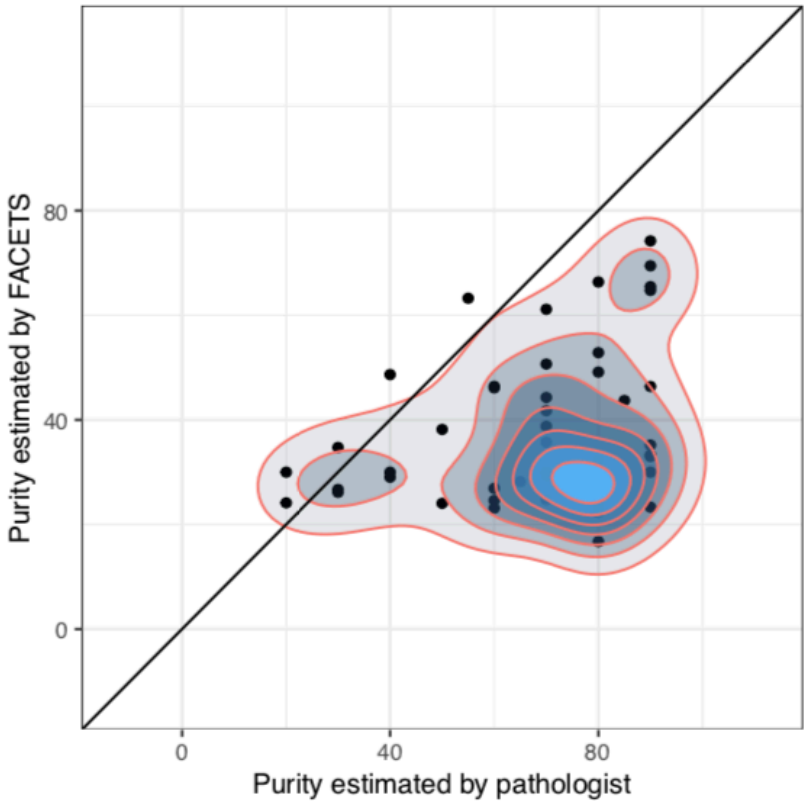
Results from \Mathced with	ClinCNV	CNV-Kit	FACETS
ClinCNV	178 / 0	50 / 128	122 / 56
CNV-Kit	39 / 13	52 / 0	30 / 22
FACETS	97 / 12	32 / 77	109 / 0

Purity estimation

ClinCNV

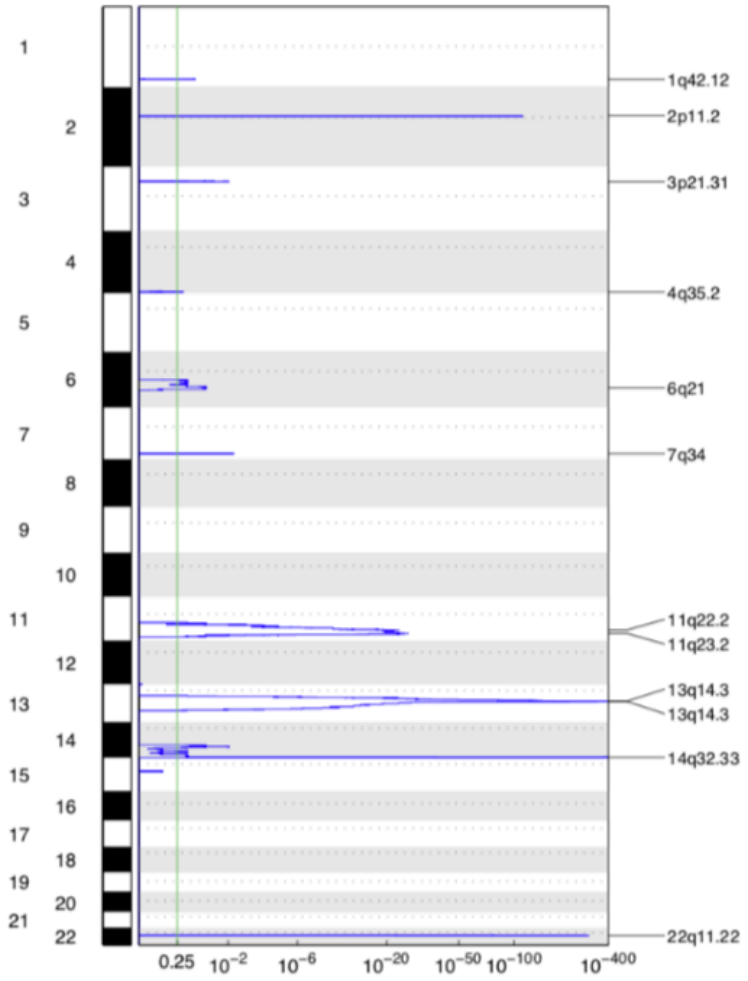


Facets

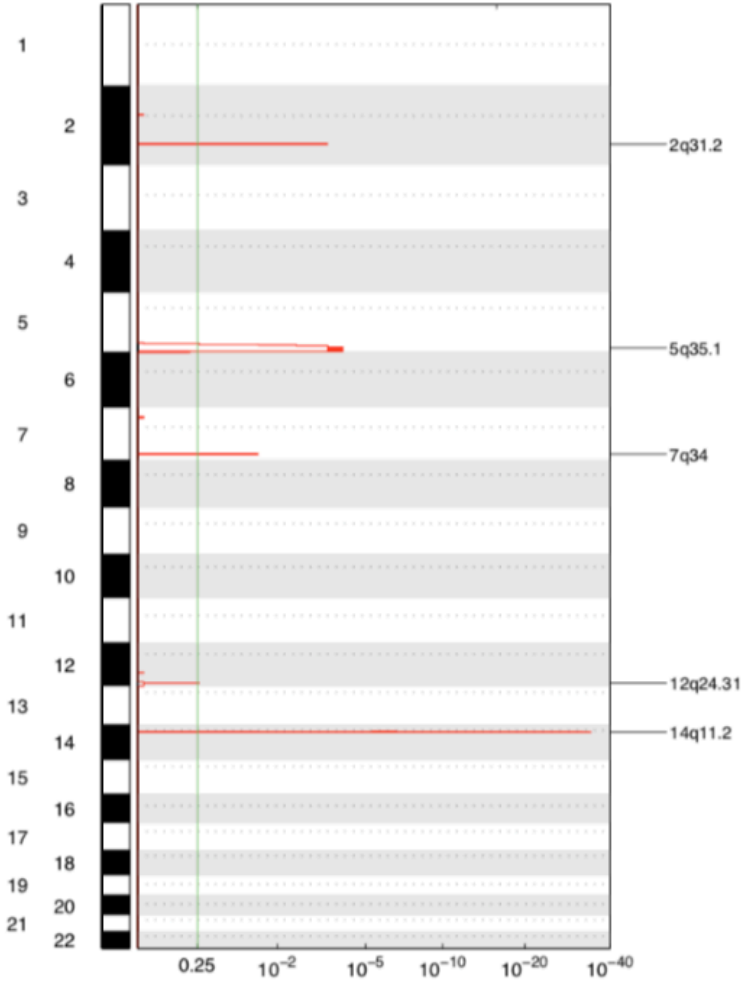


Recurrent CNAs regions in CLL

Recurrent deletions



Recurrent duplications



Plots produced with GISTIC2.0 software

Clinical Diagnostics with ClinCNV @ University Hospital Tübingen

ClinCNV was introduced into the clinical bioinformatics pipeline of UKT in mid-2018

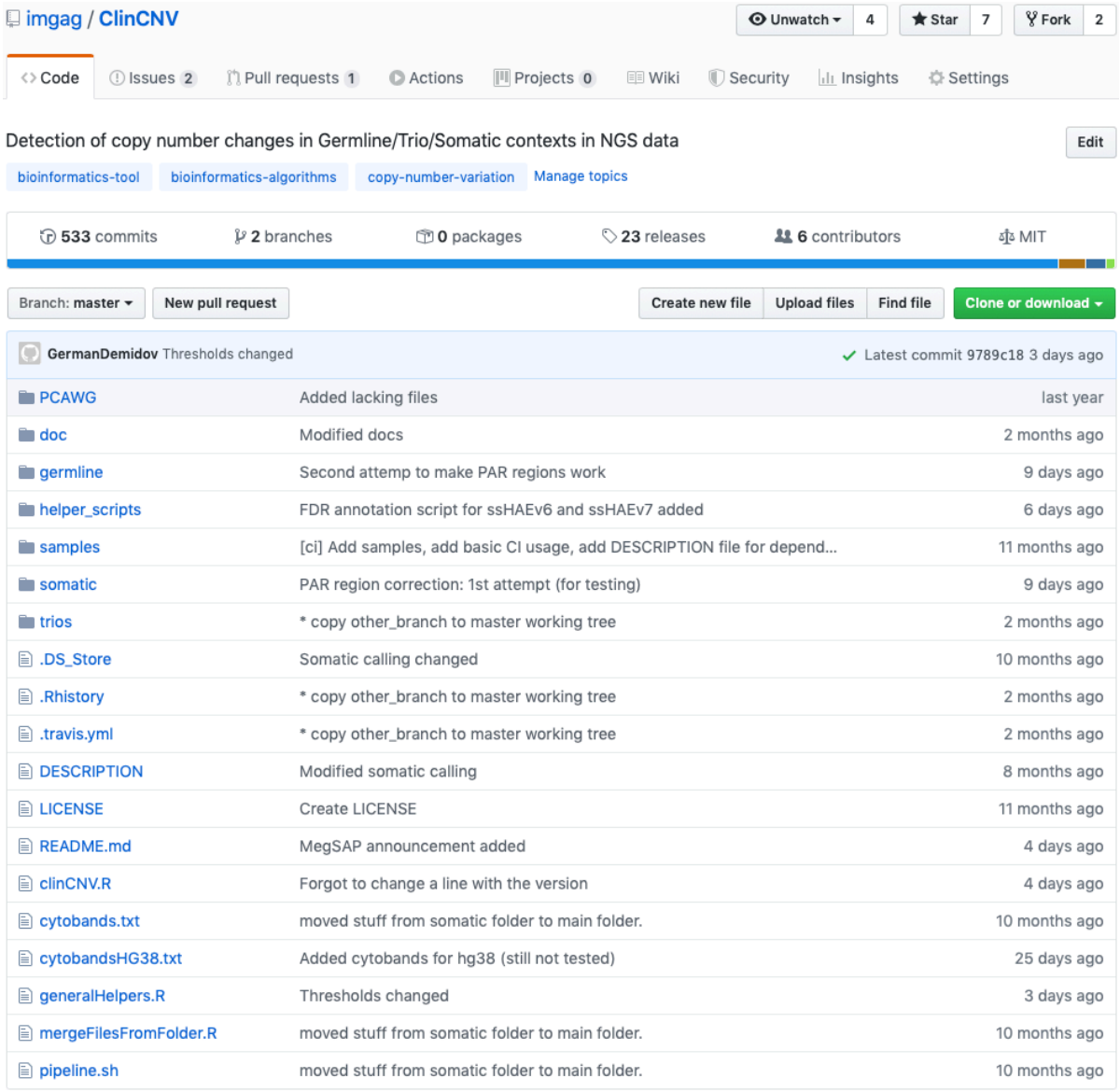
Germline, Rare Disease Diagnostics:

- 4.332 Exomes with Agilent SureSelect v7
- 2.900 Exomes with Agilent SureSelect v6
- 865 Whole Genome Sequencing
- 135 shallow WGS (around 4x coverage)

Somatic, Cancer diagnostics:

- 1.013 tumor samples (Panels with 500-710 important cancer genes)

https://github.com/imgag/ClinCNV



imgag / ClinCNV

Unwatch 4 | Star 7 | Fork 2

Code | Issues 2 | Pull requests 1 | Actions | Projects 0 | Wiki | Security | Insights | Settings

Detection of copy number changes in Germline/Trio/Somatic contexts in NGS data

Edit

bioinformatics-tool | bioinformatics-algorithms | copy-number-variation | Manage topics

533 commits | 2 branches | 0 packages | 23 releases | 6 contributors | MIT

Branch: master | New pull request | Create new file | Upload files | Find file | Clone or download

GermanDemidov Thresholds changed — Latest commit 9789c18 3 days ago

PCAWG	Added lacking files	last year
doc	Modified docs	2 months ago
germline	Second attemp to make PAR regions work	9 days ago
helper_scripts	FDR annotation script for ssHAEv6 and ssHAEv7 added	6 days ago
samples	[ci] Add samples, add basic CI usage, add DESCRIPTION file for depend...	11 months ago
somatic	PAR region correction: 1st attempt (for testing)	9 days ago
trios	* copy other_branch to master working tree	2 months ago
.DS_Store	Somatic calling changed	10 months ago
.Rhistory	* copy other_branch to master working tree	2 months ago
.travis.yml	* copy other_branch to master working tree	2 months ago
DESCRIPTION	Modified somatic calling	8 months ago
LICENSE	Create LICENSE	11 months ago
README.md	MegSAP announcement added	4 days ago
clinCNV.R	Forgot to change a line with the version	4 days ago
cytobands.txt	moved stuff from somatic folder to main folder.	10 months ago
cytobandsHG38.txt	Added cytobands for hg38 (still not tested)	25 days ago
generalHelpers.R	Thresholds changed	3 days ago
mergeFilesFromFolder.R	moved stuff from somatic folder to main folder.	10 months ago
pipeline.sh	moved stuff from somatic folder to main folder.	10 months ago

Publications connected with ClinCNV

- ClinCNV (somatic): bioRxiv, https://doi.org/10.1101/837971
- ClinCNV (germline):
- “Cancer immune control needs senescence induction by Stat1-dependent cell cycle regulator pathways in tumours” (accepted to Nature Communications)
- Novartis-sponsored study (>200 tumor-normal pairs, stage II melanoma, in preparation)
- Hyperprogression study (advanced melanoma + immunotherapy, 40 tumor-normal pairs, around half of them − hyperprogressors, in preparation)
- Several publications I lost track of

Acknowledgements

- Tobias Rausch, Jan Korbel (EMBL)
- Ossowski lab @ UT: Axel Gschwind, Marc Sturm, Jakob Admard, Leon Schütz
- Ossowski lab @ CRG: Francesc Muyas, Mattia Bosio, Hana Susak
- Diagnostics Teams: Sorin Armeanu-Ebinger, Franz Hilke, Christopher Schröder, Rebecca Buchert-Lo, Karin Schäferhoff (UKT)
- Artyomov lab: Max Artyomov, Kostya Zaitsev (WUSTL)
- NGS School squad: Anamaria Elek, Samantha Filipów, Alina Frolova, Kasia Kędzierska, Maja Kuzman, Maciej Łapiński, Leszek Pryszcz, Eugeniusz Tralle
- Supervisors: Stephan Ossowski (CRG, UT) & Tomas Marques (UPF, CRG)

CRG Centre for Genomic Regulation

upf. Universitat Pompeu Fabra Barcelona