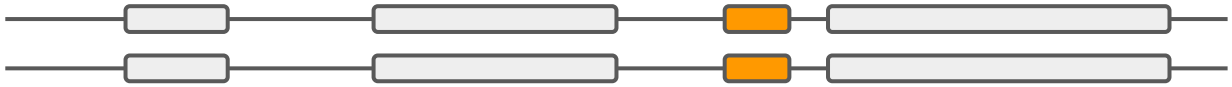
Improving detection of gene duplications in whole-genome sequencing data using allelic depth imbalance

Paweł Sztromwasser
BTM, Medical University of Łódź

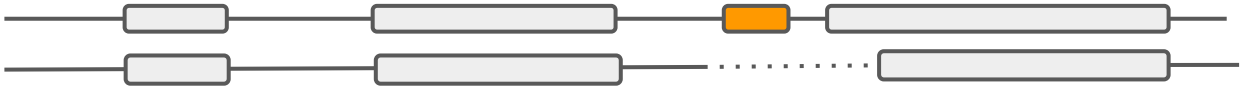
NATIONAL SCIENCE CENTRE
POLAND

Copy number variants

normal

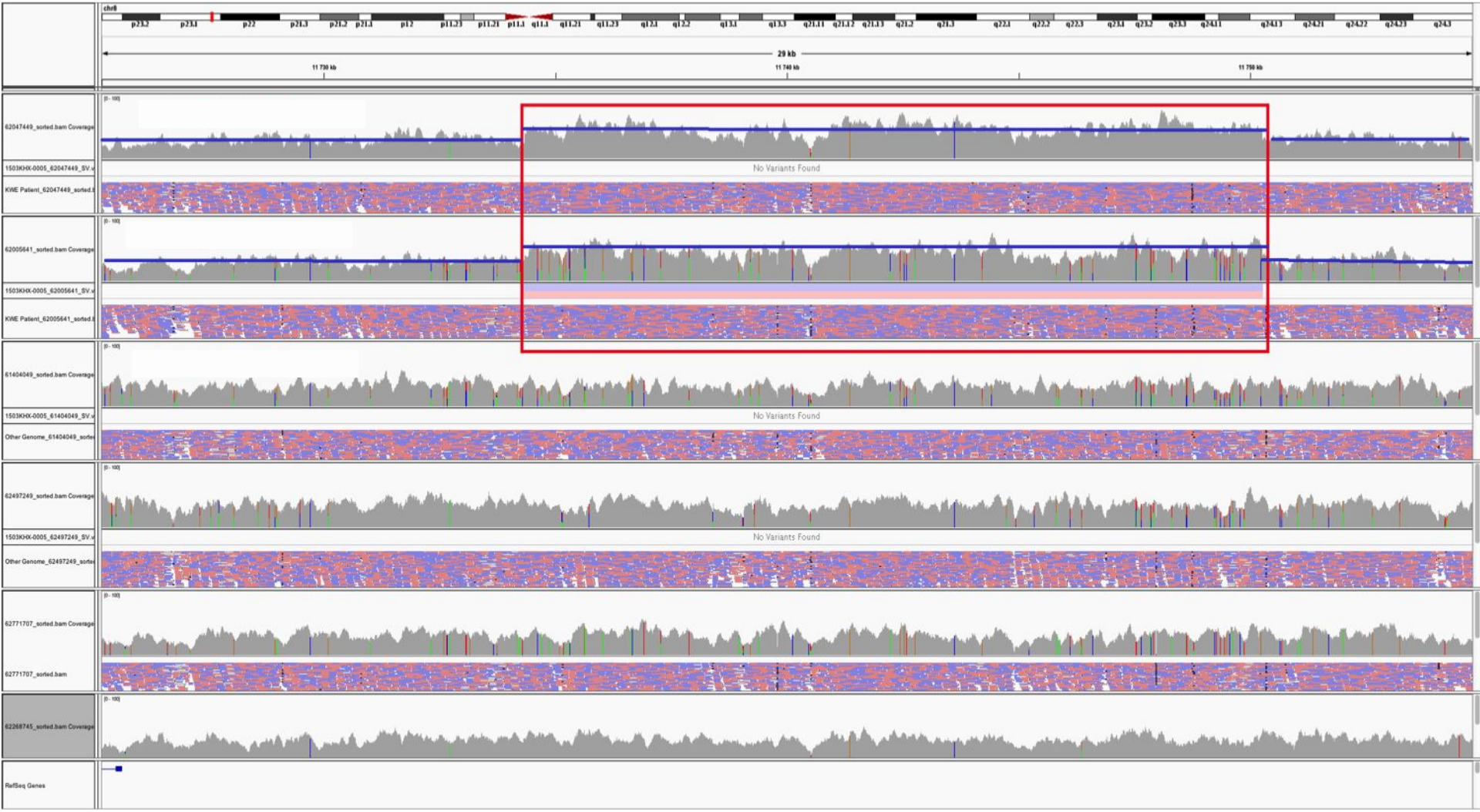


deletion

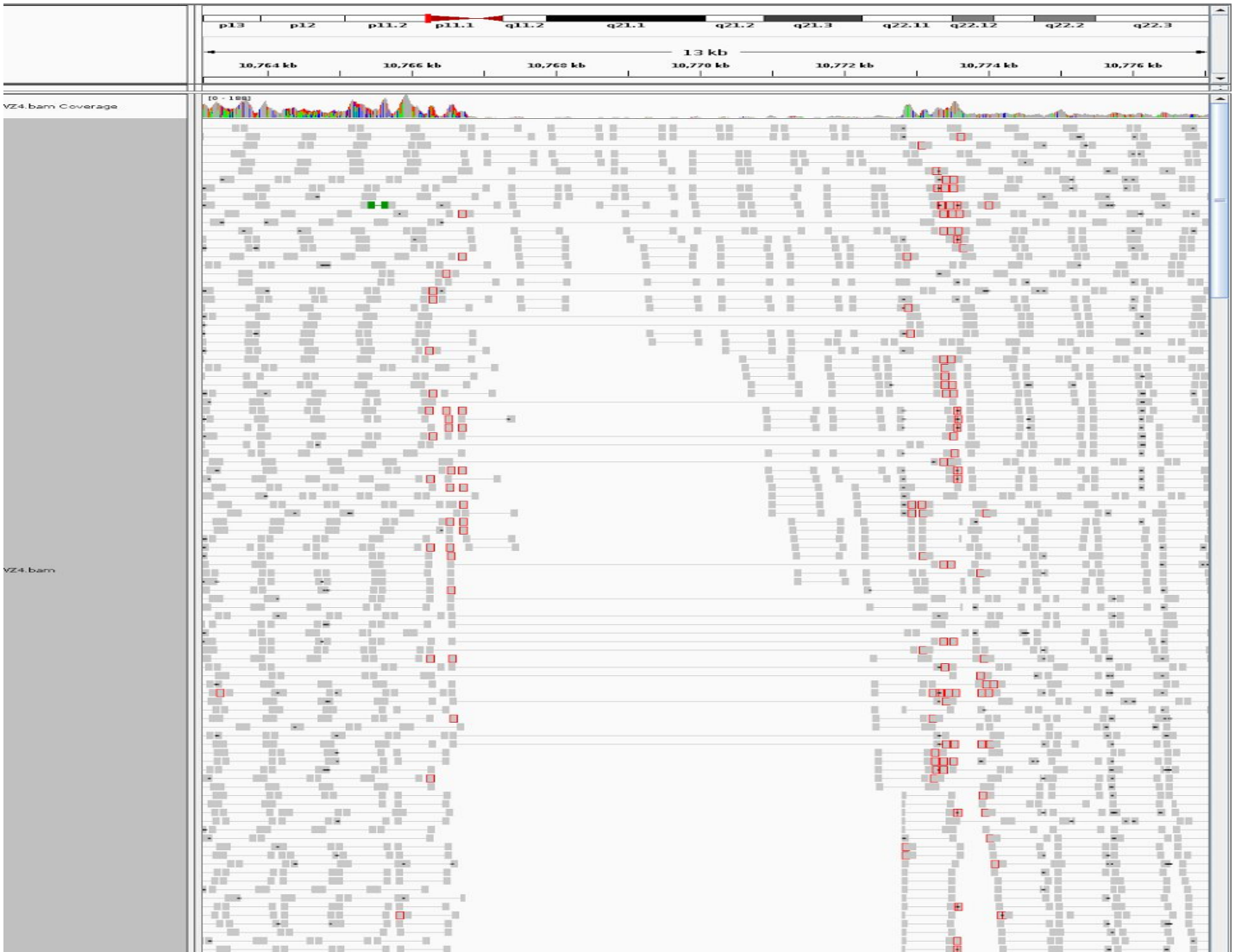


duplication

How to detect them in NGS data?



How to detect them in NGS data?

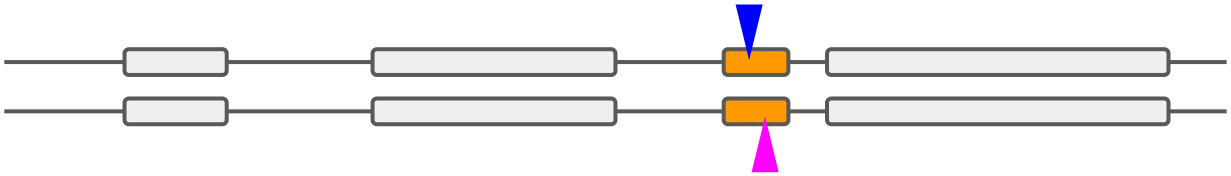


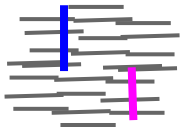
Benchmark results

	LUMPY, TrioCNV, ERDS	
	Precision (PPV)	sensitivity
1000 Genomes **deletions** (1947)	2.0 - 50.6 %	28.6 - 83.0%
1000 Genomes **duplications** (90)	1.4 - 6.5 %	7.8 - 43.3 %

Allelic depth imbalance (ADI)

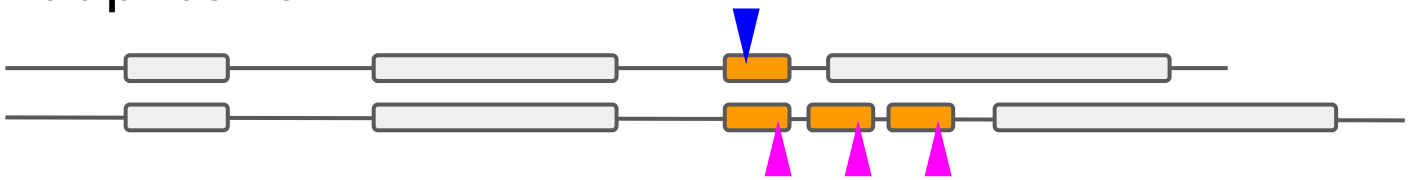
normal

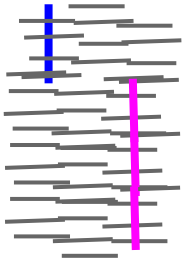




~ 50:50
~ 50:50

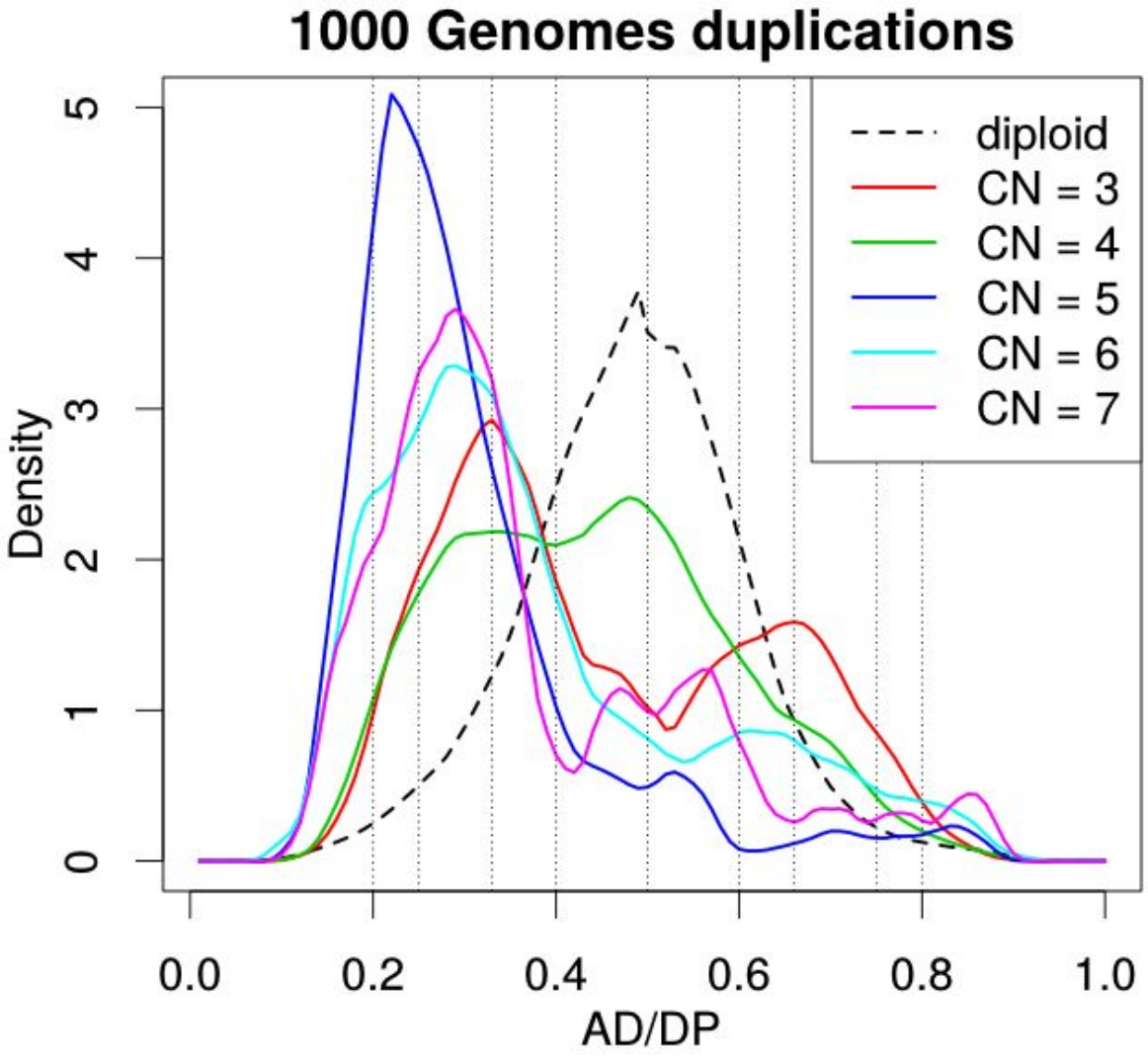
duplication





~ 25:75
~ 25:75

Proof of concept



Allelic Depth Imbalance (ADI) score

x - genomic interval, potential duplication

n - number of het variants overlapped by x

alternative allele depth

total depth

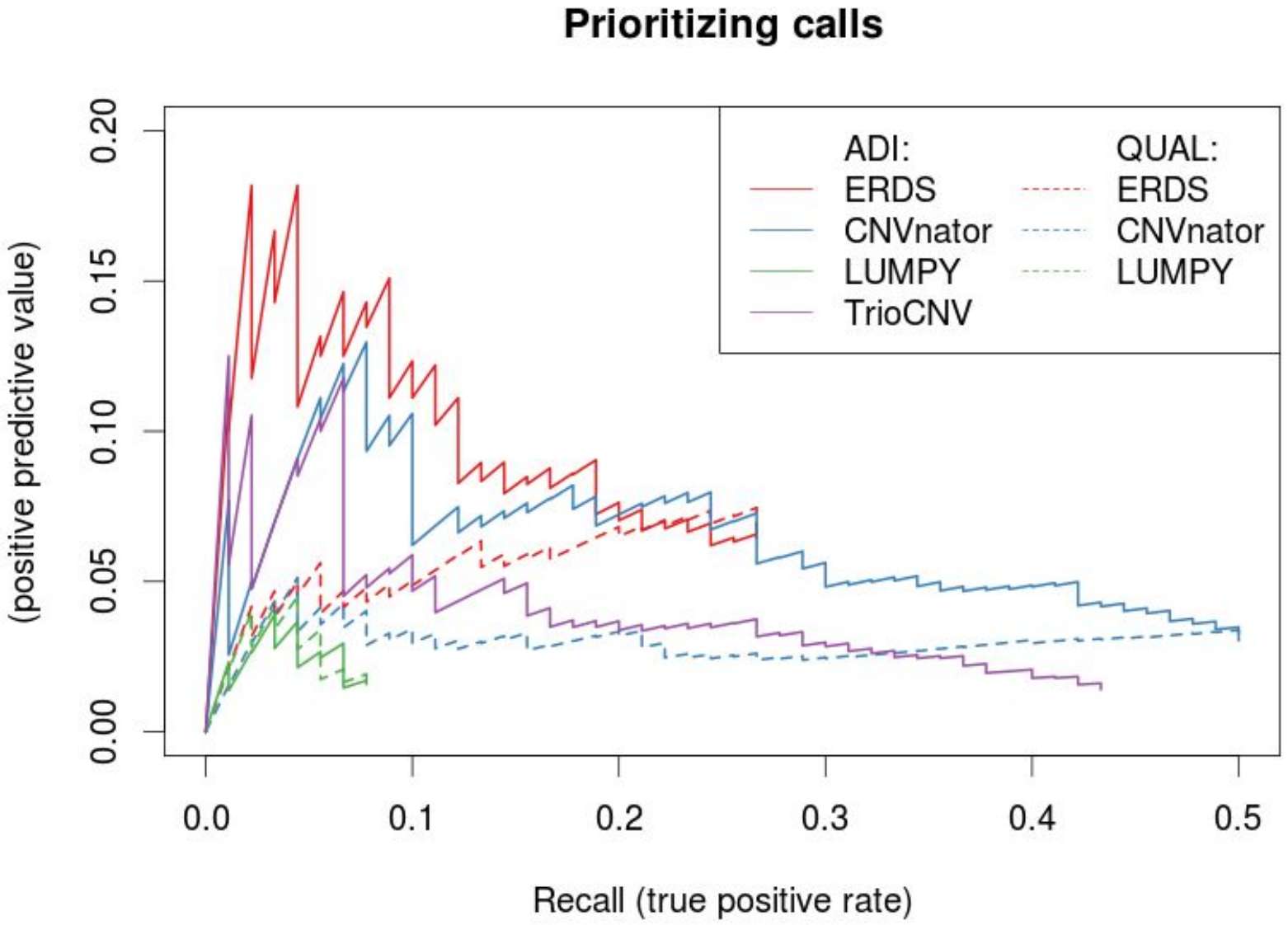
$$(1) \quad ADI_x = \sum_{i=1}^{n} |0.5 - \frac{AD_i}{DP_i}|$$

$$(2) \quad B_n = \{b_i : i = 1..1000 \; and \; b_i \; overlaps \; n \; heterozygous \; SNVs\}$$

$$ADI_{B_n} = \{ADI_{b_i} : b_i \in B_n\}$$

$$ADIscore(x, B_n) = rank(ADI_x, ADI_{B_n})$$

Results


Prioritizing calls
(positive predictive value)
0.00
0.05
0.10
0.15
0.20
0.0
0.1
0.2
0.3
0.4
0.5
Recall (true positive rate)
ADI:
ERDS
CNVnator
LUMPY
TrioCNV
QUAL:
ERDS
CNVnator
LUMPY

Acknowledgements

Inge Jonassen

Kjell Petersen

Vidar Steen

Stefan Johansson

Tomasz Stokowy

Wojciech Fendler

Project funded by Bergen Research Foundation and National Science Center in Poland
(POLONEZ 2016/23/P/NZ2/04251)

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 665778.

Materials

1. NA12878 genome

sequenced on Illumina Hiseq XTen, 150bp pair-end, 124GBps, 29x
(NA12878D library provided by DNAnexus)
https://kccg.garvan.org.au/confluence/pages/viewpage.action?pageId=31592745

2. CEU trio (NA12878, NA12891, and NA12892)

sequenced on Illumina HiSeq 2000, 101bp pair-end, 245-290GBps, 37-64x
(by 1000Genomes Consortium)
ftp://ftp-trace.ncbi.nih.gov/1000genomes/ftp/technical/working/20120117_ceu_trio_b37_decoy/

Reference CNVs

Copy number ⟶ ↓ CNV set	0	1	2	3	4	5	6	7
1000 Genomes	624	1411	135	**38**	**57**	**1**	**3**	**1**
Conrad (2010)	137	352	(3975)	**89**	**134**	**13**	**1**	**1**
Mills (2011) GS	617		-	**271***				

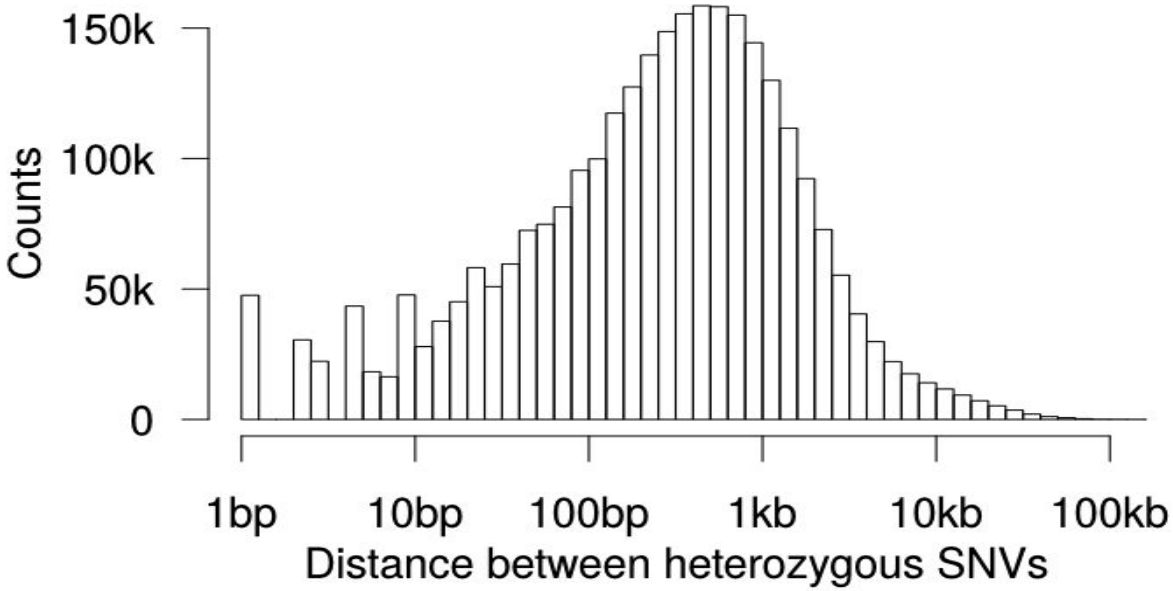
*) Conrad + 24 from McCarroll et al, 2008

10 out of 238 Conrad variants are overlapping a variant in the 1000 Genomes

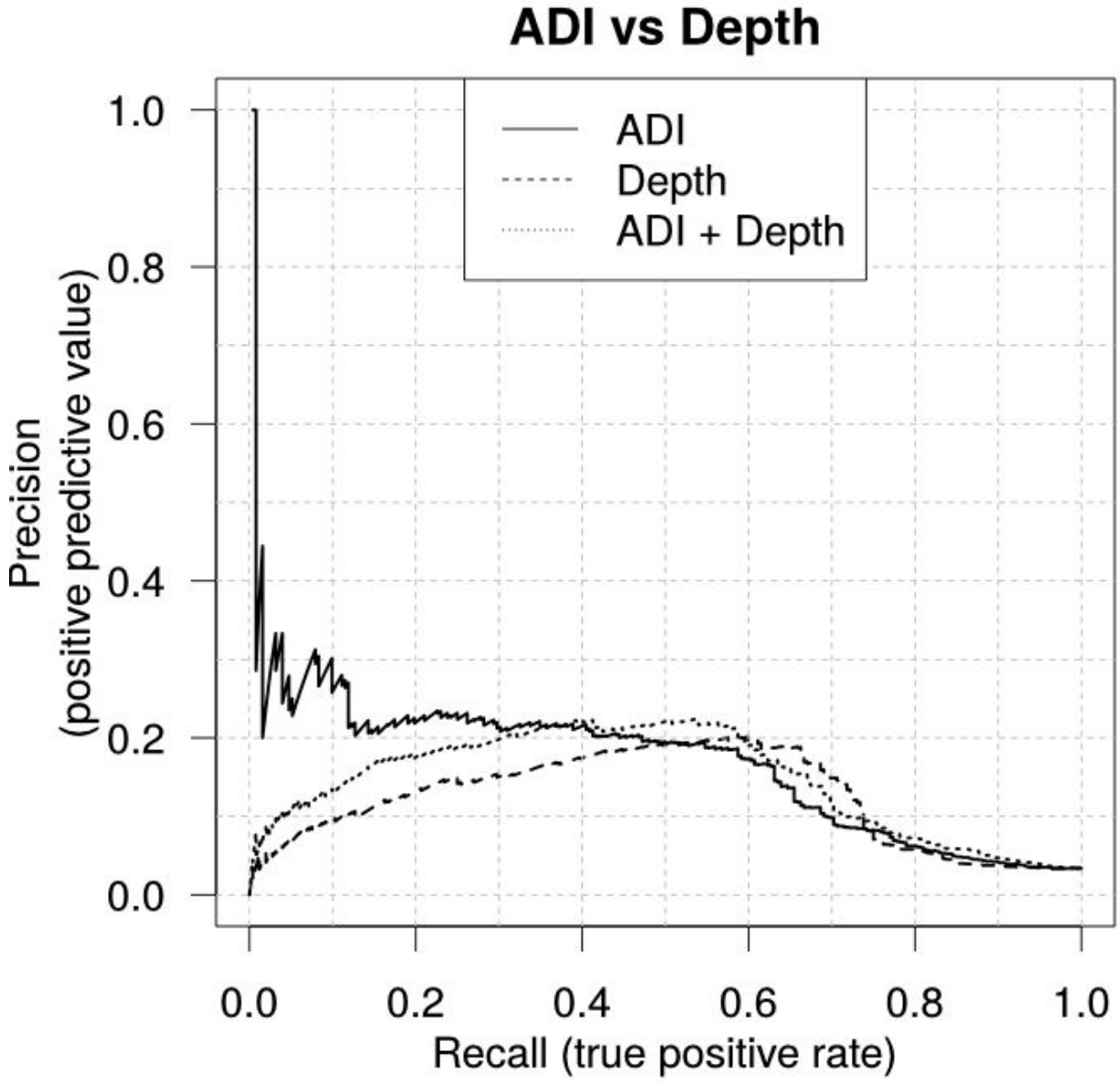
	1000 Genomes (100)	Conrad (238)	Mills (271)
1000 Genomes	-	10 (4%)	18 (7%)
Conrad	9 (9%)	-	243 (90%)
Mills	14 (14%)	238 (100%)	-

The 1000 Genomes Project Consortium (2015). Nature, 526(7571), 68–74.
Mills, et al. (2011). Nature, 470(7332), 59–65.
Conrad, et al. (2010). Nature, 464(7289), 704–712.
McCarroll et al. (2008). Nature Genetics 40, 1166–1174.

Can it work?


150k
100k
50k
0
Counts
1bp
10bp
100bp
1kb
10kb
100kb
Distance between heterozygous SNVs

Performance



Performance

	1000 Genomes	
	precision	sensitivity
LUMPY	1.6%	7.8%
TrioCNV	1.4%	43.3%
ERDS	6.5%	26.7%