MAPPING COPY NUMBER VARIATION BY POPULATION-SCALE GENOME SEQUENCING (1000 GENOME PROJECT)

Published in Nature, 3 February 2011

BioInformaatikute TeadusArtiklite Lugemise Klubi ettekanne

Maido Remm, 19.09.2011

ARTICLE

doi:10.1038/nature09708

Mapping copy number variation by population-scale genome sequencing

Ryan E. Mills[1]*, Klaudia Walter[2]*, Chip Stewart[3]*, Robert E. Handsaker[4]*, Ken Chen[5]*, Can Alkan[6,7]*, Alexej Abyzov[8]*, Seungtai Chris Yoon[9]*, Kai Ye[10]*, R. Keira Cheetham[11], Asif Chinwalla[5], Donald F. Conrad[2], Yutao Fu[12], Fabian Grubert[13], Iman Hajirasouliha[14], Fereydoun Hormozdiari[14], Lilia M. Iakoucheva[15], Zamin Iqbal[16], Shuli Kang[15], Jeffrey M. Kidd[6], Miriam K. Konkel[17], Joshua Korn[4], Ekta Khurana[8,18], Deniz Kural[3], Hugo Y. K. Lam[13], Jing Leng[8], Ruiqiang Li[19], Yingrui Li[19], Chang-Yun Lin[20], Ruibang Luo[19], Xinmeng Jasmine Mu[8], James Nemesh[4], Heather E. Peckham[12], Tobias Rausch[21], Aylwyn Scally[2], Xinghua Shi[1], Michael P. Stromberg[3], Adrian M. Stütz[21], Alexander Eckehart Urban[13,27], Jerilyn A. Walker[17], Jiantao Wu[3], Yujun Zhang[2], Zhengdong D. Zhang[8], Mark A. Batzer[17], Li Ding[5,22], Gabor T. Marth[3], Gil McVean[23], Jonathan Sebat[15], Michael Snyder[13], Jun Wang[19,24], Kenny Ye[20], Evan E. Eichler[6,7], Mark B. Gerstein[8,18,25], Matthew E. Hurles[2], Charles Lee[1], Steven A. McCarroll[4,26], Jan O. Korbel[21] & 1000 Genomes Project†

Genomic structural variants (SVs) are abundant in humans, differing from other forms of variation in extent, origin and functional impact. Despite progress in SV characterization, the nucleotide resolution architecture of most SVs remains unknown. We constructed a map of unbalanced SVs (that is, copy number variants) based on whole genome DNA sequencing data from 185 human genomes, integrating evidence from complementary SV discovery approaches with extensive experimental validations. Our map encompassed 22,025 deletions and 6,000 additional SVs, including insertions and tandem duplications. Most SVs (53%) were mapped to nucleotide resolution, which facilitated analysing their origin and functional impact. We examined numerous whole and partial gene deletions with a genotyping approach and observed a depletion of gene disruptions amongst high frequency deletions. Furthermore, we observed differences in the size spectra of SVs originating from distinct formation mechanisms, and constructed a map of SV hotspots formed by common mechanisms. Our analytical framework and SV map serves as a resource for sequencing-based association studies.

OBJECTIVES OF THE STUDY:

1. To compare performance of different methods and algorithms for discovery of structural variants (SV) from sequencing data.

2. To create a list of all SVs of 50 bp and larger in size within studied individuals for further reference.

Initial focus was on **deletions.** Less focus was placed on insertions and duplications. The balanced variations (inversions and chromosomal rearrangements) were not studied.

DATA:

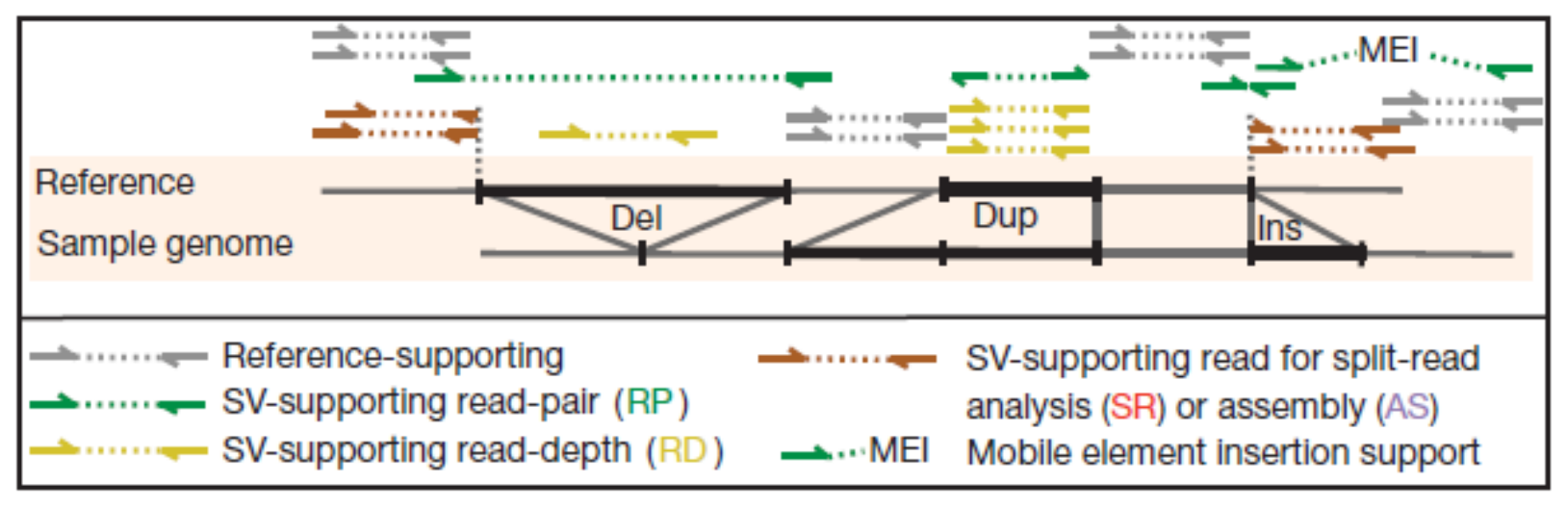
- High-coverage sequences (42x coverage)
  - 1 parent-offspring trio from CEU
  - 1 parent-offspring trio from YRI
- Low-coverage sequences (3.6x coverage)
  - 60 CEU
  - 60 JPT+CHB
  - 59 YRI

4 ALGORITHMS, 19 METHODS

- 6 methods using Read-Pair (RP)
- 4 methods using Read-Depth (RD)
- 4 methods using Split-Read (SR)
- 3 methods using local Sequence Assembly (AS)
- 2 methods using combination of RP and RD (PD)

Color-coding:

AS RP
SR RL Release set
PD RD



4 ALGORITHMS, 19 METHODS, 36 CALLSETS

- 19 methods were applied separately to low-coverage and high-coverage data and deletions and insertions were collected into separate datasets (callsets).

- Altogether 36 callsets: 15 callsets for low-coverage data and 21 callsets for high-coverage data (trios).

LOW-COVERAGE CALLSETS

Approach	Callset Origin	Discovery Algorithm Name and Reference*	Platform	Mapping Algorithm	Genomes Analyzed	SV Type	Algorithm Parameters Used
RD	AE	N/A[13]	Illumina	MAQ	8	DEL	window size (≥500bp); p-value ($P \leq 10^{-5}$)
	SD	Event-wise testing[15,#]	Illumina	MAQ	162	DEL	read mapping quality (≥Q30); window size (100bp); cluster size with merged events of same type(≤ 500bp); read depth (≤0.75 and ≥ 1.25 mean read depth); significance level ($P<10^{-6}$); event size (≥ 1kb); absolute difference between median read counts(>0.5)
	YL	CNVnator[10]	Illumina	MAQ	65	DEL	N/A
RP	BC	Spanner[13]	Illumina	MOSAIK	138	DEL	maximum mismatch threshold (4 for 36-43mer reads, 6 for 44-63mers, and 12 for 64mers and longer); hash size (15); Smith-Waterman bandwidth (17); alignment candidate threshold (25bp); local alignment search radius (100bp); hash position threshold (100); mapping distance (P-value≥0.99); minimum read-pairs (4, 2 from each side); map distance to annotated loci (≥400bp); gap between the F and R clusters (-30 bp < gap < 500 bp)
	BC	Spanner[13]	Illumina	MOSAIK	138	INS	maximum mismatch threshold (4 for 36-43mer reads, 6 for 44-63mers, and 12 for 64mers and longer); hash size (15); Smith-Waterman bandwidth (17); alignment candidate threshold (25bp); local alignment search radius (100bp); hash position threshold (100)
	SI	N/A[13]	Illumina	MAQ	144	DEL	MAQ mapping quality (≥20); read-pairs in a cluster (≥2); start/end distance (10 x median absolute deviation of the insert size distribution); event size (<1Mb)
	YL	PEMer[16,23]	SOLiD	CORONA	25	DEL	span-size (within 15% deviation from the median of span-size)
	WU	BreakDancer	Illumina	MAQ	138	DEL	RMAQ mapping quality (> 35); outer distance (> mean + 4stdev of the insert size)
SR	BC	Mosaik[13]	454	MOSAIK	22	INS	hash size (15bp); mismatch bases in alignments (≤5%); match bases aligned to one of the mobile element consensus sequences (40bp); gap length (≤6bp); alignment quality score (≥40);mobile element alignment length (>60bp); distance from annotated mobile elements (≥100bp)
	LN	Pindel[17]	Illumina	MAQ	145	DEL	MAQ mapping quality (>0); maximum deletion size (50kb); number of fragments for unmapped reads (2 for deletion and 3 for short insertions)
	LN	Pindel[17]	Illumina	MAQ	145	INS	MAQ mapping quality (>0); maximum deletion size (50kb); number of fragments for unmapped reads (2 for deletion and 3 for short insertions)
	YL	N/A[13]	454	BLAT	5	DEL	N/A
	YL	N/A[13]	454	BLAT	5	INS	N/A
PD	BC	Spanner[13]	Illumina	MOSAIK	138	TDUP	mapping quality values of read pairs (≥30); mapping distance between the pairs (p-value<0.02%); number of supporting read pairs (≥3); minimum deletion size (50bp); "Alignability" in the clustered regions (> 0.01); Net read c overage over all samples (< 2.5 x the expected coverage); event length (≤250bp); copy number (≤2.2)
	BI	Genome STRiP[18]	Illumina	MAQ	168	DEL	clusters of paired-ends (N >= 2); apparent insert size (> the median of the insert size distribution + 10 x the median absolute deviation of insert size from the median for that lane/library);

HIGH-COVERAGE CALLSETS

Approach	Callset Origin	Discovery Algorithm Name and Reference*	Platform	Mapping Algorithm	Genomes analyzed	SV Type	Algorithm Parameters Used
RD	SD	Event-wise testing[15]	Illumina	MAQ	6	DEL	read mapping quality (≥Q30); window size (100bp); cluster size with merged events of same type(≤ 500bp); read depth (≤0.75 and ≥ 1.25 mean read depth); significance level (P<10-6); event size (≥ 1kb); absolute difference between median read counts (>0.5); median read-depth (<1.25); common deletion regions (>4 occurrencies)
	UW	mrFAST[19]	Illumina	mrFAST	6	DEL	RepeatMasker (on human reference genome build 35, with the sensitivity option "-s" enabled); Tandem Repeats Finder (mask tandem repeats ≤500bp); edit distance (≤ 2); unique PDervals (5 kb of unmasked sequence); windows (6/7 consecutive 5 kb windows with read depth >average-2stdev)
	YL	CNVnator[10]	Illumina	MAQ	6	DEL	N/A
RP	AB	AB large indel tool[13]	SOLiD	MAPREADS	1	DEL	read-pairs in a cluster (≥2)
	BC	Spanner[13]	Illumina	MOSAIK	6	DEL	maximum mismatch threshold (4 for 36-43mer reads, 6 for 44-63mers, and 12 for 64mers and longer); hash size (15); Smith-Waterman bandwidth (17); alignment candidate threshold (25bp); local alignment search radius (100bp); hash position threshold (100); mapping distance (P-value≥0.99); minimum read-pairs (4, 2 from each side); map distance to annotated loci (≥400bp); gap between the F and R clusters (-30 bp < gap < 500 bp)
	BC	Spanner[13]	Illumina	MOSAIK	6	INS	maximum mismatch threshold (4 for 36-43mer reads, 6 for 44-63mers, and 12 for 64mers and longer); hash size (15); Smith-Waterman bandwidth (17); alignment candidate threshold (25bp); local alignment search radius (100bp); hash position threshold (100)
	SI	N/A[13]	Illumina	MAQ	6	DEL	MAQ mapping quality (≥20); read-pairs in a cluster (≥2); start/end distance (10 x median absolute deviation of the insert size distribution); event size (<1Mb)
	UW	Variation Hunter[20]	Illumina	mrFAST	6	DEL	high-quality reads (average phred score ≥ 20); edit distance (≤ 2 with the mrFAST); size threshold (average± 4×stdev)
	WU	BreakDancer	Illumina	MAQ	6	DEL	MAQ mapping quality (> 35); outer distance (> mean + 4stdev of the insert size)
	YL	PEMer[16,23]	454	PEM	1	DEL	p-value cutoff of 0.05
	YL	PEMer[16,23]	454	PEM	1	INS	p-value cutoff of 0.05
SR	BC	Mosaik[13]	454	MOSAIK	2	INS	hash size (15bp); mismatch bases in alignments (≤5%); match bases aligned to one of the mobile element consensus sequences (40bp); gap length (≤6bp); alignment quality score (≥40);mobile element alignment length (>60bp); distance from annotated mobile elements (≥100bp)
	LN	Pindel[17]	Illumina	MAQ	6	DEL	MAQ mapping quality (>0); maximum deletion size (50kb); number of fragments for unmapped reads (2 for deletion and 3 for short insertions)
	YL	N/A[13]	454	BLAT	1	DEL	N/A
	YL	N/A[13]	454	BLAT	1	INS	N/A
AS	BG	SOAPdenovo[21]	Illumina	SOAP	6	DEL	prealignment (BLAT v. 30 with –fastMap and –maxPDron=50); scaffold set alignment (LASTZ V1.01.50 with high-scoring segment pairs (HSP) chaining option, ambiguous 'N' treatment, and gap-free extension tolerance up to 50kb); Best hits were further confirmed using "axtBest"
	BG	SOAPdenovo[21]	Illumina	SOAP	6	INS	prealignment (BLAT v. 30 with –fastMap and –maxPDron=50); scaffold set alignment (LASTZ V1.01.50 with high-scoring segment pairs (HSP) chaining option, ambiguous 'N' treatment, and gap-free extension tolerance up to 50kb); Best hits were further confirmed using "axtBest"
	OX	Cortex[13]	Illumina	CORTEX	1	DEL	event size (≤1kb for "bubble calling" algorithm and ≤40kb for "reference assisted" algorithm)
	OX	Cortex[13]	Illumina	CORTEX	1	INS	event size (≤1kb for "bubble calling" algorithm and ≤40kb for "reference assisted" algorithm)
	UW	NovelSeq[22]	Illumina	mrFAST	6	INS	event size (≥200bp)
PD	BC	Spanner[13]	Illumina	MOSAIK	6	TDUP	mapping quality values of read pairs (≥30); mapping distance between the pairs (p-value<0.04%); number of supporting read pairs (≥3); minimum deletion size (50bp); "Alignability" in the clustered regions (> 0.01); Net read coverage over all samples (< 2.5 x the expected coverage); event lenght (≤250bp); copy number (≤2.2)

SENSITIVITY AND FDR

		Real situation: (can be tested by PCR or microarray)		
		Positive (P)	Negative (N)	
Software prediction results:	Positive	True Positive (TP)	False Positive (FP)	**False Discovery Rate** = FP / (TP + FP)
	Negative	False Negative (FN)	True Negative (TN)	
		Sensitivity = TP / (TP + FN)	**Specificity** = TN / (FP + TN)	

- **Sensitivity:** Sn = TP / (TP + FN)
- **Specificity**: Sp = TN / (FP + TN)
- **Accuracy**: ACC = (TP + TN) / (P + N)
- **False Discovery Rate:** FDR = FP / (TP + FP)

VALIDATION OF METHODS (SENSITIVITY GOLD STANDARD)

- **Sensitivity in detecting deletions** estimated for three gold standard sources, i.e., sets of published deletions (Conrad, 2010; McCarroll, 2008; Kidd, 2008;Mills, 2006).
  SVs in these publications were identified with capillary sequencing (median=0.2kb), tiling CGH microarrays (median=2kb), and fosmid sequencing (median=6kb).

- Only 1bp overlap required for recording positive prediction!

- Individual methods show sensitivity between 0% and 80%.

- In final "release set" sensitivity was 69% (low-coverage set) to 82% (high-coverage set).
- With more stringent sensitivity criterion (>50% overlap) the sensitivity was 51% (low-coverage) to 70% (high-coverage).

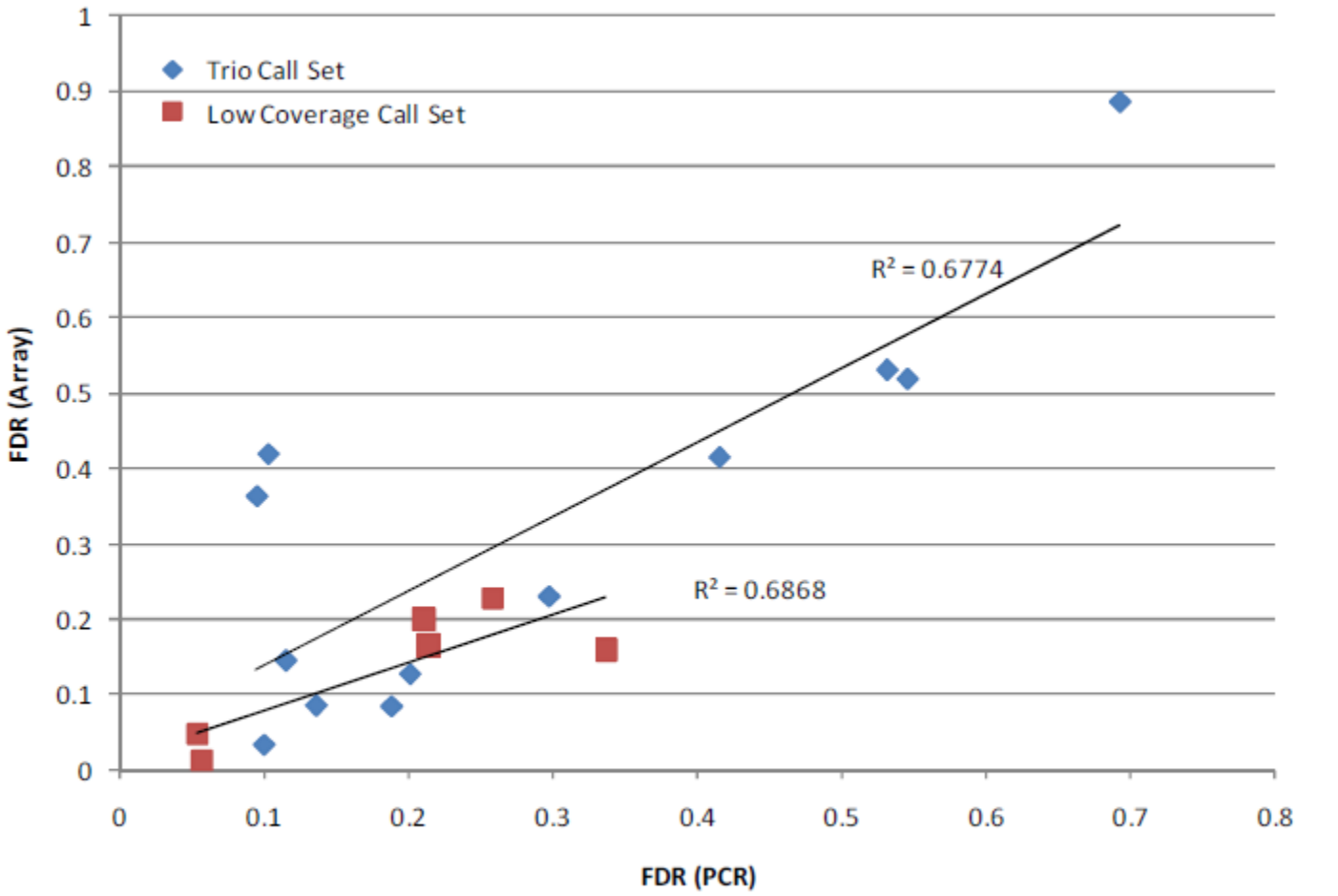
VALIDATION OF METHODS (FDR RATES)

Findings in each callset were validated using PCR and CGH.
PCR primers were designed for randomly chosen SV predictions from each callset.
Custom **array-CGH DNA Microarrays** were used to validate deletions and duplications in the high coverage trios. Affy 6.0, Illumina 1.0 and NimbleGen 2.1M arrays were also used for some individuals.

$$FDR = \frac{CGH_{invalidated}}{CGH_{validated} + CGH_{invalidated}} * \frac{CGH_{validated} + CGH_{invalidated}}{N} + \frac{PCR_{invalidated}}{PCR_{validated} + PCR_{invalidated}} * (1 - \frac{(CGH_{validated} + CGH_{invalidated})}{N})$$

Final FDR is weighted average from both experiments

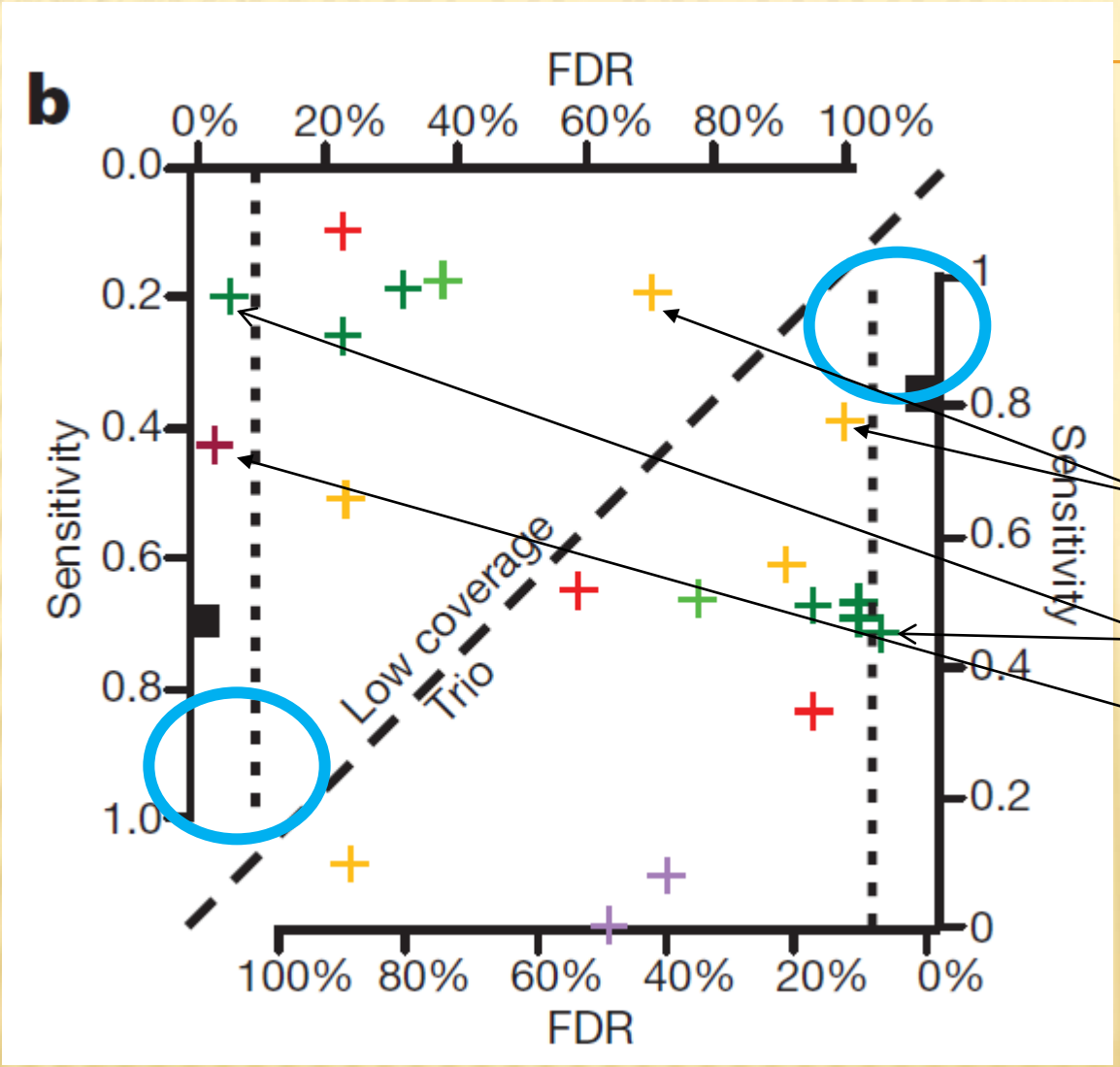
VALIDATION OF METHODS (FDR RATES)



Supplementary Figure 3. Correlation of PCR and array based FDR estimates. FDR estimates based on PCR and arrays are displayed both for trio (blue) and low coverage (red) callsets.

- **PCR and microarrays have only moderate agreement with each other on presence of structural variants**

VALIDATION OF METHODS (FDR RATES)



CNVnator (Gerstein, YaleU)

Spanner (Marth, BostonC)

GenomeSTRiP
(McCarroll, Broad Institute)

- **Sensitivity and FDR on 2 individuals (low- and high coverage).**

CONCLUSIONS

- None of the sequence-based methods is reliable for individual SV calling in inheritance studies or in medical diagnostics. For GWAS studies ??

- For example, one of the best methods **Spanner** has:
  - FDR ca 9% and
  - sensitivity ca 40%
  in high-coverage deletion callset.

RELEASE SET

- For final release only methods with overall FDR<10% were used + some experimentally validated SVs.

- These methods were:
  Spanner (from Marth group, Boston College)
  Mosaik (from Marth group, Boston College)
  GenomeSTRiP (from McCarroll group, Broad Institute)

RELEASE SET OF STRUCTURAL VARIATIONS:

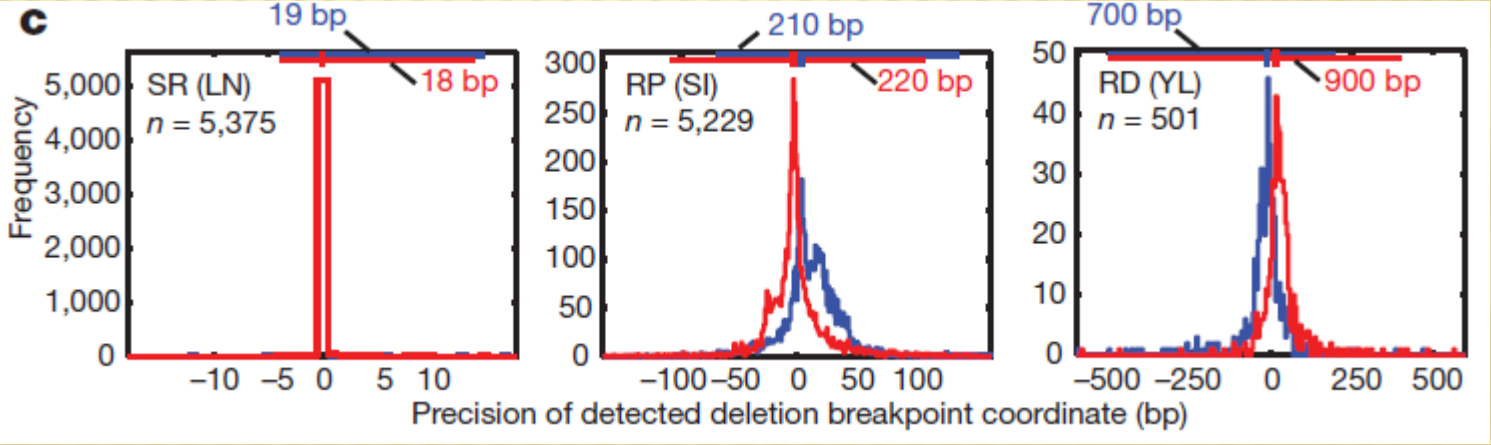
- 28 000 structural variations described from given individuals (cell lines)
  - 22 000 deletions,
  - 5 400 mobile element insertions,
  - 500 duplications,
  - 100 insertions

Half of these were "novel" SVs, missing from dbVAR, DGV and from other sequenced genomes.

MAPPING OF BREAKPOINTS:

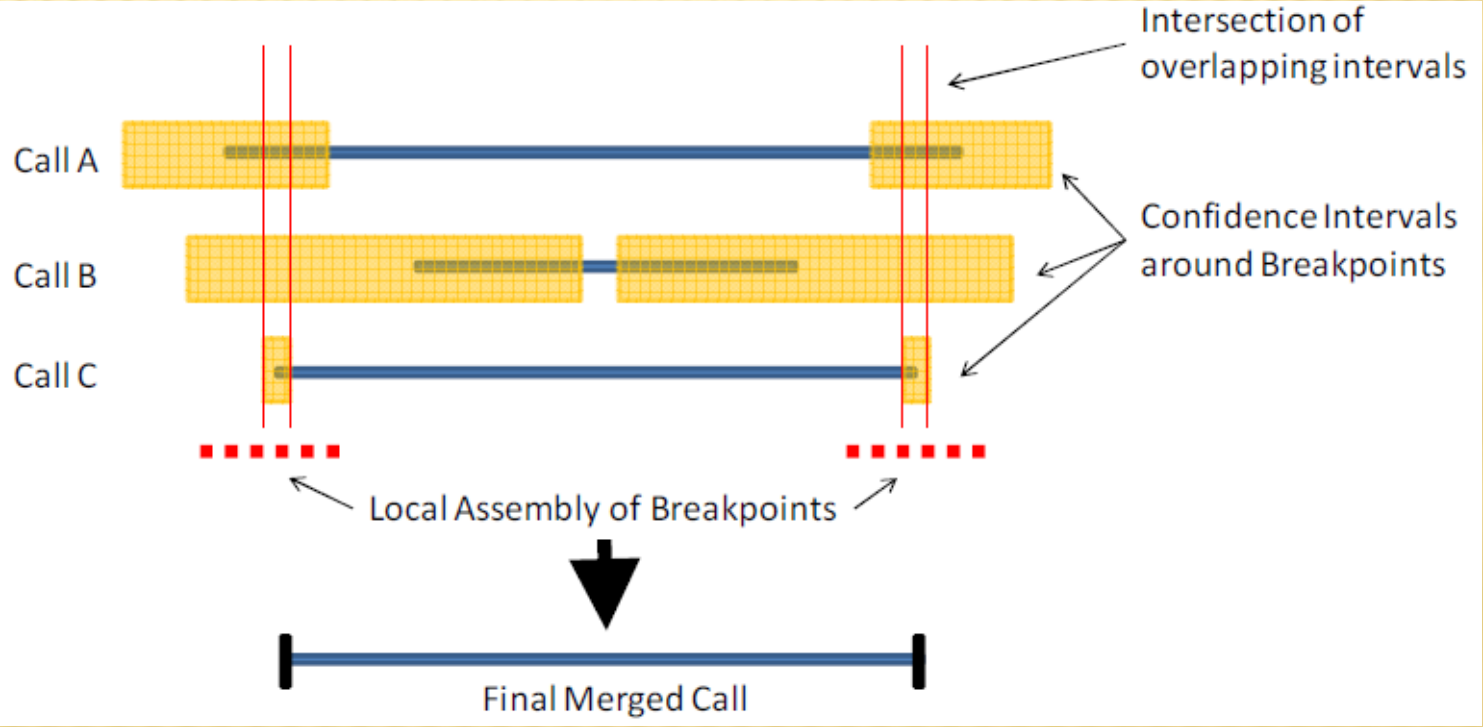
- Sequence data allows mapping of breakpoints with single nucleotide precision. This was done for ca 15000 SVs.

- Different methods have different precision



MAPPING OF BREAKPOINTS:

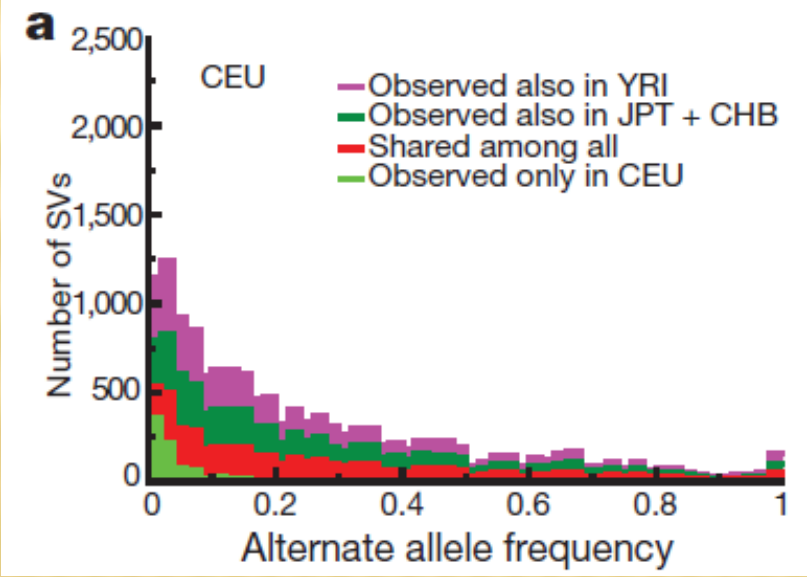
- Sequence data allows mapping of breakpoints in single nucleotide precision. This was done for ca 15000 SVs.

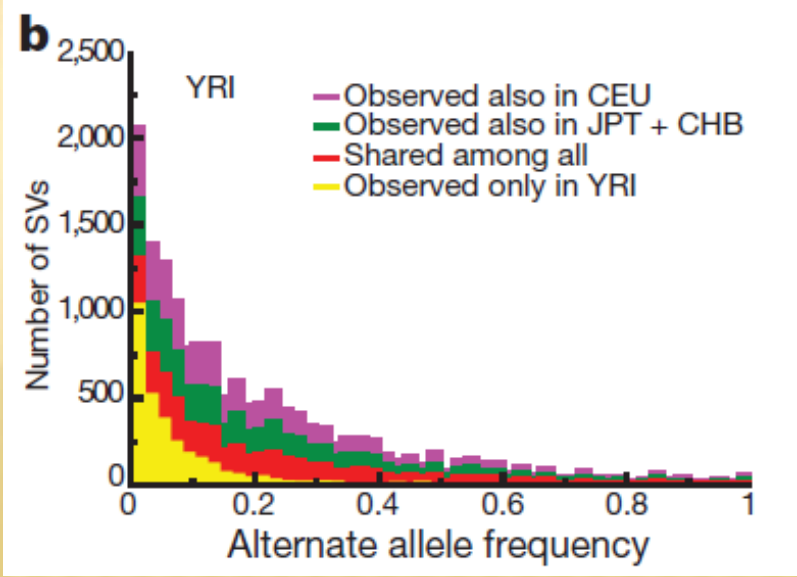


POPULATION GENETICS

- Common SVs (MAF > 5%) were typically shared across populations, whereas rare alleles were frequently observed in only one population.

- 81% of deletions display linkage disequilibrium (LD) with SNPs at level $r^2>0.8$





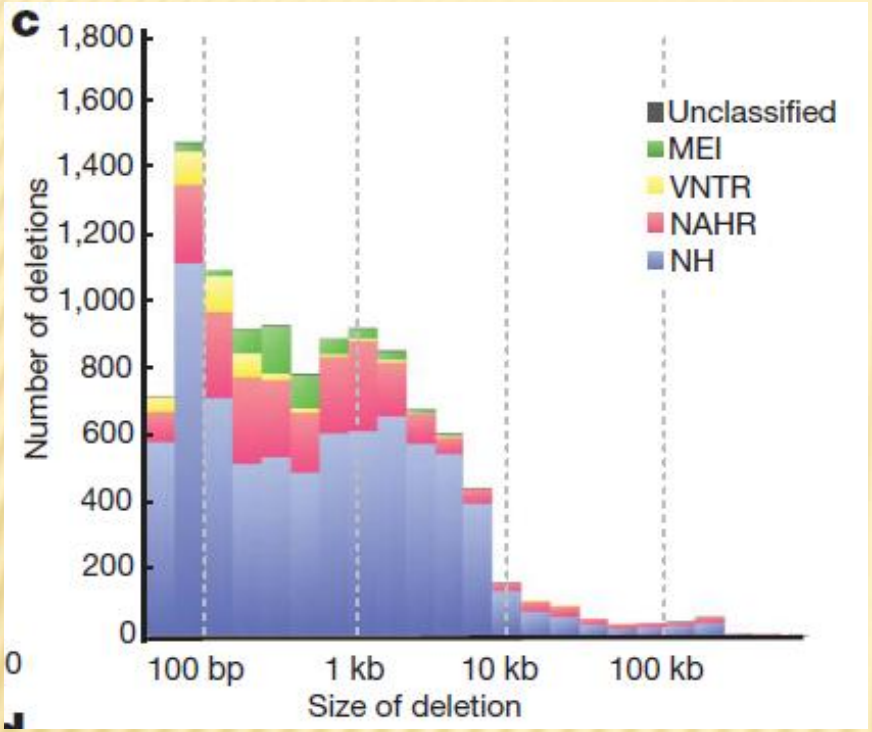
RE-CLASSIFICATION OF DELETIONS

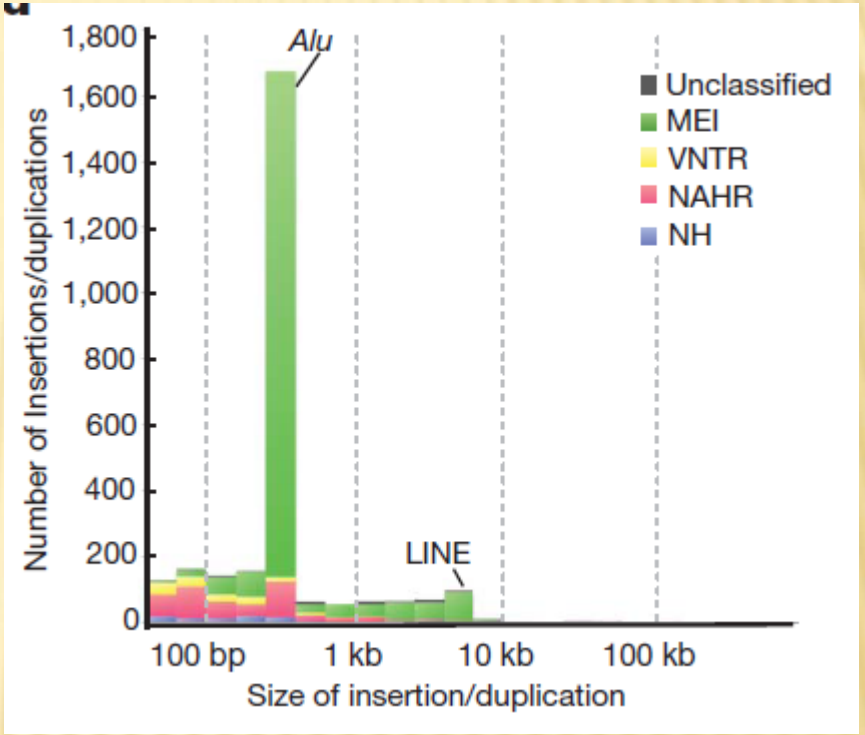
- 11 000 nucleotide-level deletions were compared to **primate genomes** using BreakSeq classification approach (Nat. Biotechnology, 2010).

- Only 60% confirmed as deletions
- 23% are actually duplications
- 17% undetermined

60/ (60+23) = **28% of determined deletions are NOT deletions.** They only look like deletions because all comparisons are done wrt reference genome (single individual).

MECHANISM OF DELETION AND INSERTION





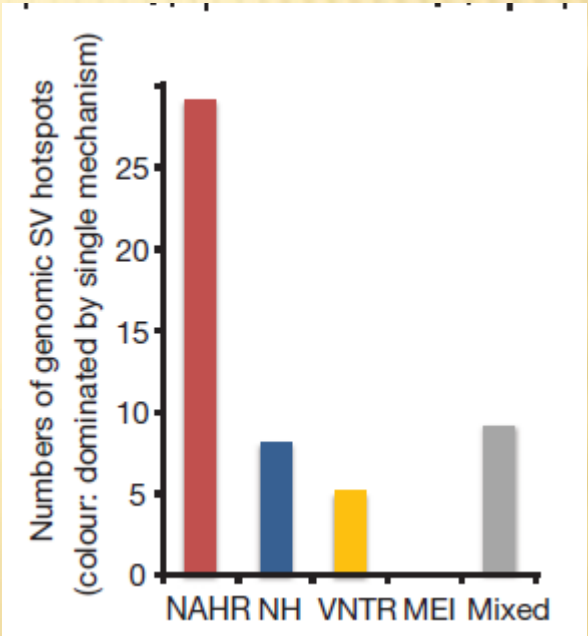
Deletion types

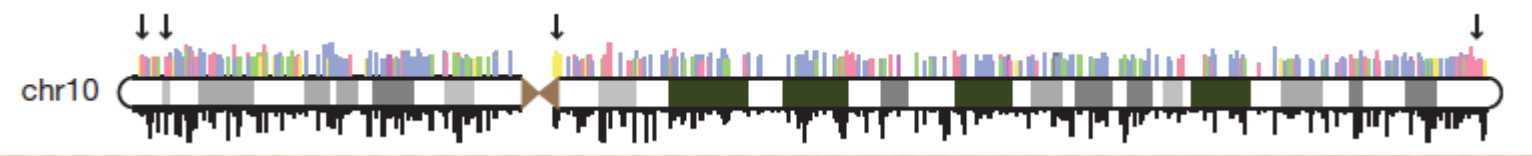
Insertion types

MEI: mobile element insertion
VNTR: variable number of tandem repeats (polymerase slippage)
NAHR: non-allelic homologour recombination (error of recombination)
NH: non-homologous end joining (DNA repair mechanisms)

MECHANISM OF DELETION AND INSERTION

51 hotspots of SVs over the entire genome were detected, 6 of them are in regions of known genetic disorders previously associated with recurrent *de novo* deletions, including Miller-Dieker syndrome and Leri-Weill dyschondrosteosis.





CONCLUSIONS

- **Sequencing-based methods are not yet reliable for most types of SV analyses. Even GWAS might be problematic.**
- **Reference genome is not representing ancestral state. Better to compare with ancestral genome.**
- **28 000 SVs available from 1000GP webpage, majority of them are mapped to single nucleotide precision.**