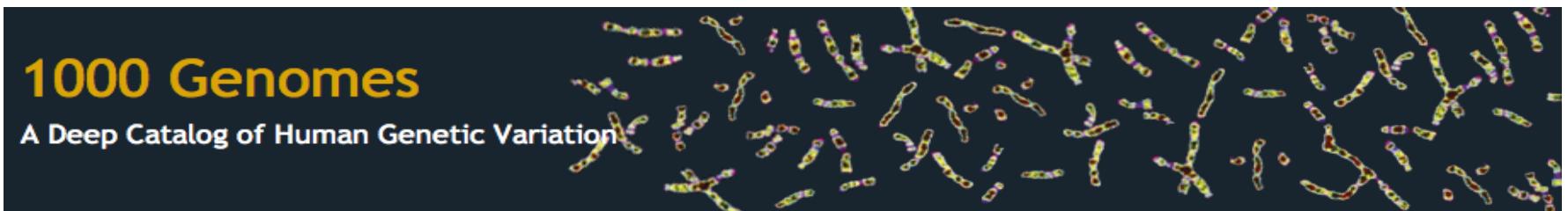


1000 Genomes data tutorial at ASHG

Structural variants

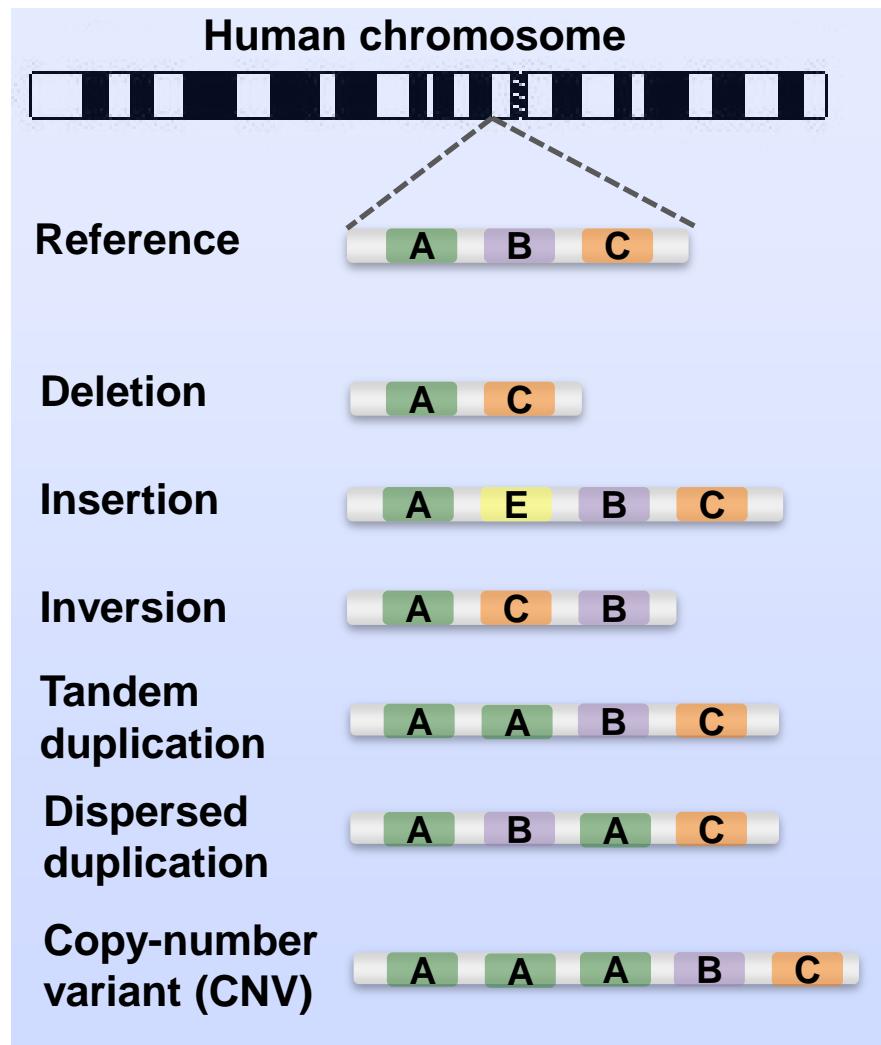
Jan Korbel

European Molecular Biology Laboratory (EMBL) Heidelberg
Genome Biology Research Unit



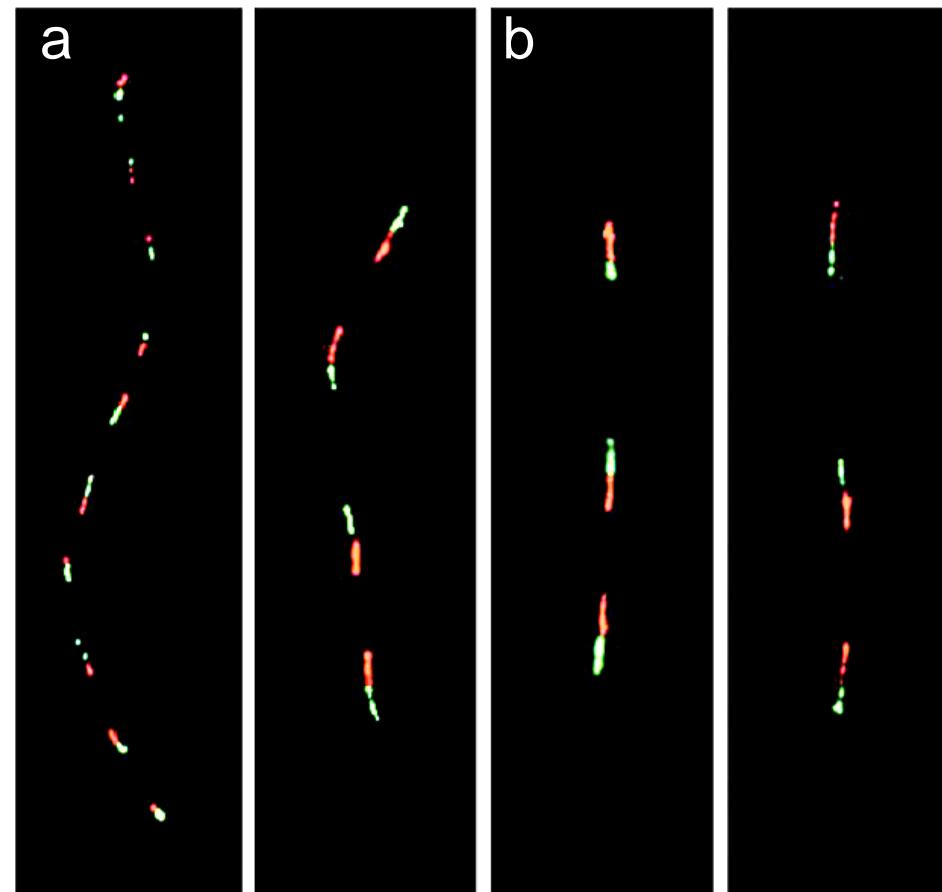
Structural variants (SVs) in the genome

[polymorphic rearrangements of the genome of 50bp up to hundreds of kb in size]



~0.5% of the genome affected by SVs
in each individual

Example: *AMY1* copy-number variation



a, Japanese; **b**, African (Biaka) individual
[Perry *et al.*, *Nat. Genet.* 2007]

SV discovery considering evidence from multiple sources

Read Pairs (RP)

No SV

sample

Deletion

Mobile element (MEI) insertion

Tandem duplication

reference

MEI

reference

reference

Read Depth (RD)

Duplication

reference

Deletion

reference

Split Reads (SR)

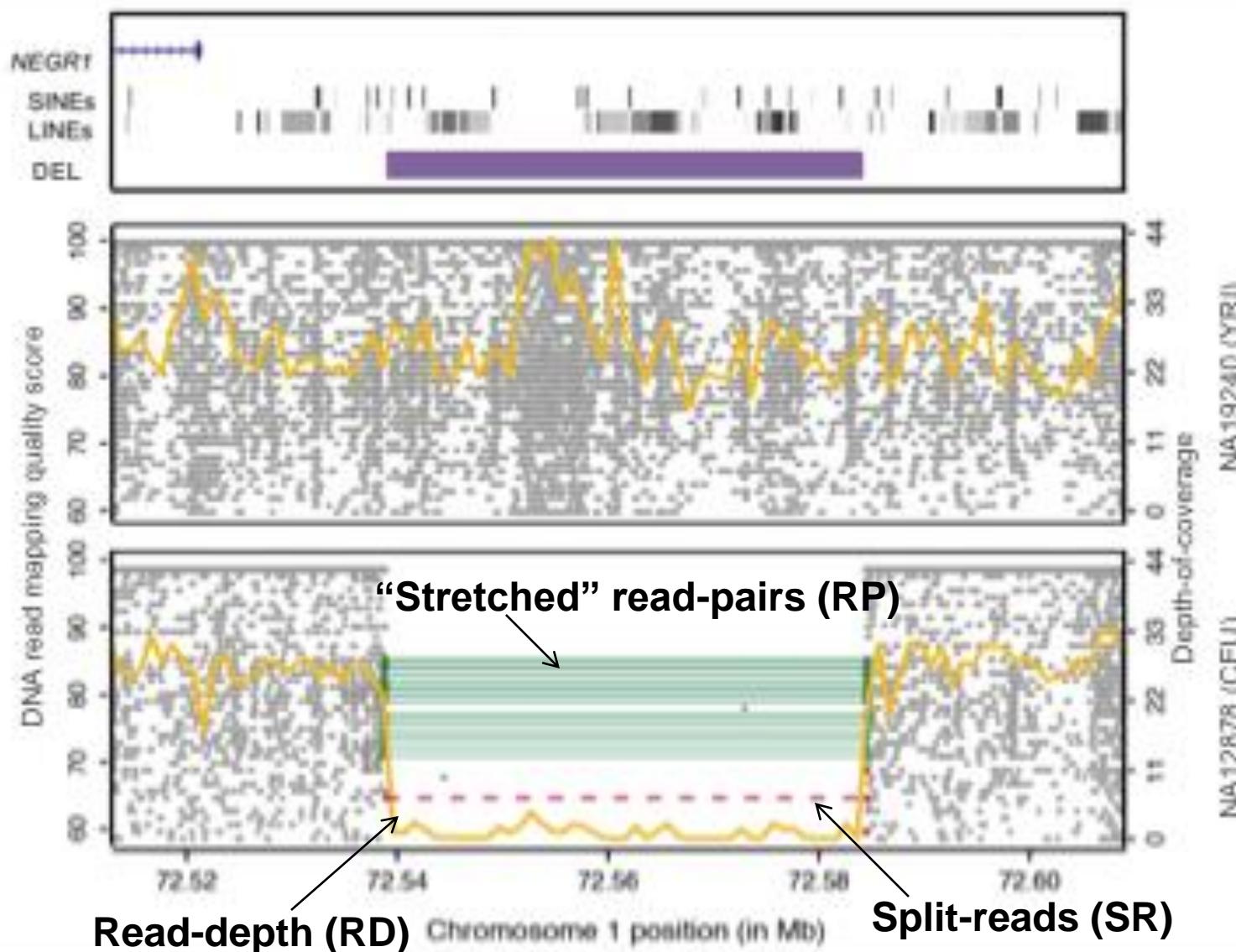
Deletion

Assembly (AS)

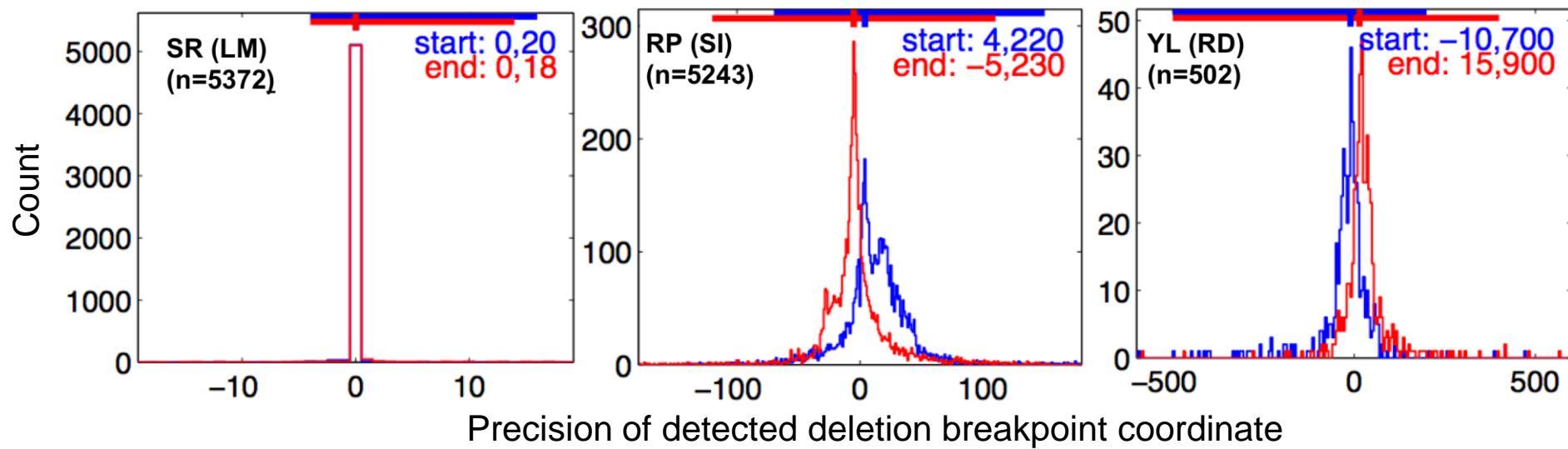
Novel sequence insertion

reference

A deletion simultaneously detected by paired-end mapping (PEM), read-depth analysis, and split reads



Ascertainment differences among deletion discovery methods: SV breakpoint precision



Blue and red histograms: breakpoint residuals for predicted start/end coordinates relative to assembled coordinates. Horizontal lines at the top of each plot mark the 98% (2.3 sigma) confidence intervals.

Individuals analyzed in the pilot 1 (low-coverage) and pilot 2 (trio) studies of the 1000 Genomes Project

	Trios	Low coverage
Samples	6	179
Raw data	1.08Tbp	2.22Tbp
Deletions	11,248	15,893
Mobile element insertions	2,531	4,775
Tandem Duplications	256	407
Novel sequence insertions	174	-
SV breakpts	6,169	9,092

Deletion genotypes from the 1000 Genomes Project

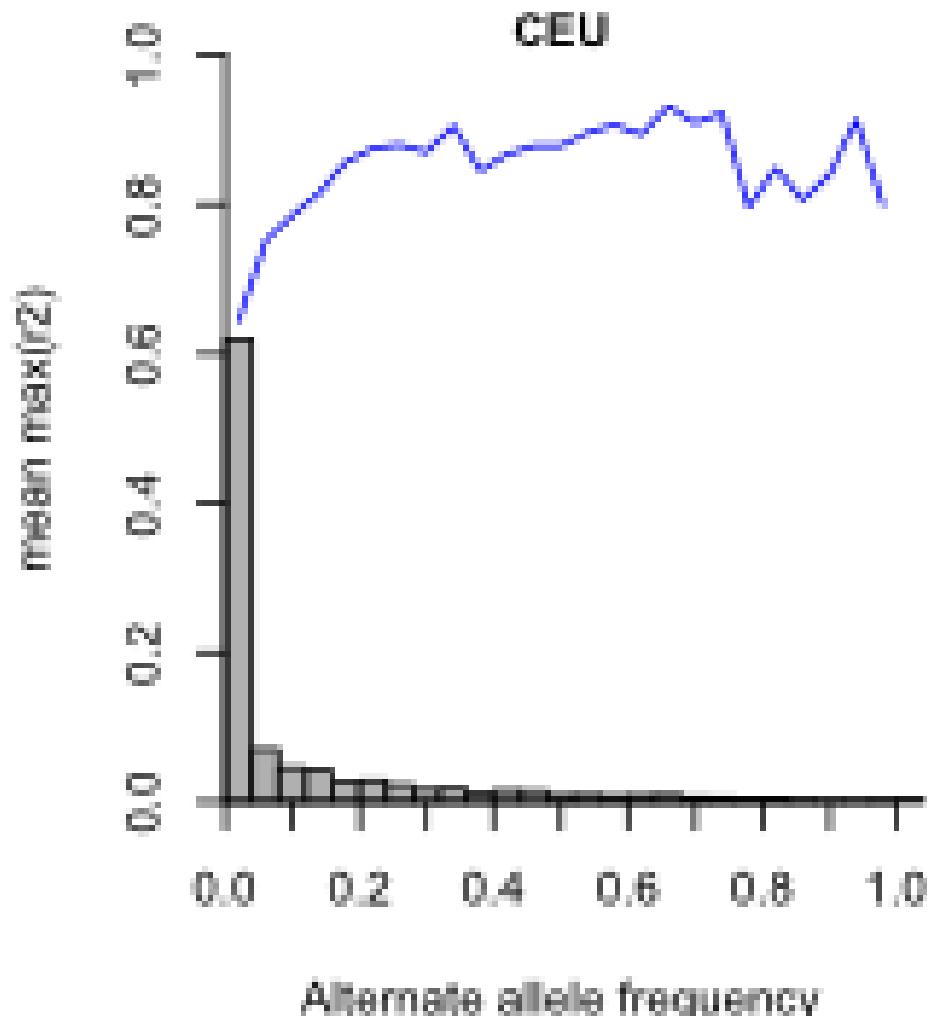
- 13,826 deletion polymorphisms (48 bp – 960 kb) genotyped in 156 genomes using Genome STRiP (Handsaker et al., manuscript submitted)
- Concordance with array-based genotypes: **99.1%** (for 1,970 deletions from Conrad et al., 2009)

Genome STRiP integrates multiple features of sequencing data

- Read depth
- Read pairs
- Split reads

Deletions and SNPs on shared haplotypes

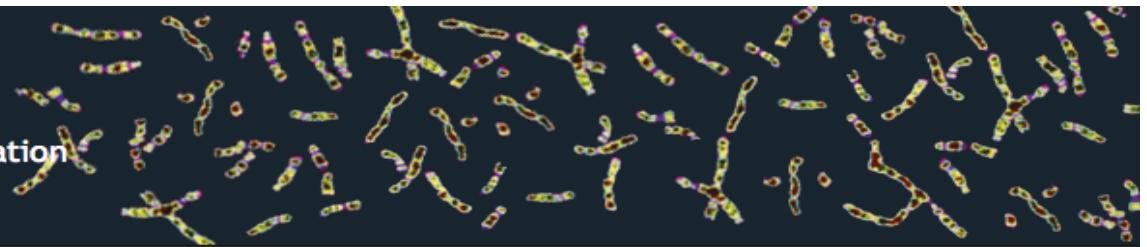
LD between 1000 Genomes deletions and HapMap3 SNPs



81% of common
deletions are tagged
by one or more
HapMap SNPs ($r^2 >$
0.8)

1000 Genomes

A Deep Catalog of Human Genetic Variation



Data formats

SV Pilot Paper Data Release

(for 1000 Genomes Project pilot 1 [low-coverage] and pilot 2 [trios])

- **SV data is available as different formats, providing different levels of detail**
 - (1) Variant Call Format (VCF) – Primary
 - Contains SV discovery (release) set and deletion genotypes
 - Standardized format (version 4.0)
 - (2) Master Validation Format (MVT) – Auxiliary
 - Raw data from individual SV discovery methods
 - Includes additional information regarding validation and original SV coordinate predictions
 - (3) SV breakpoint information available as textfiles – Auxiliary
 - SV breakpoint junctions generated by the TIGRA targeted assembly algorithm (FASTA format)
 - BreakSeq annotations available (mechanism & ancestral state for SVs with assembled breakpoints) in GFF format

SV discovery set as VCF format

- Accessible as tab-delimited files
 - These can be converted into **Excel** spreadsheets
 - They can also be processed with **vcftools**: <http://vcftools.sourceforge.net/>
 - **PERL** module (Vcf.pm), also available through vcftools
- Format
 - #CHROM POS ID REF ALT QUAL FILTER INFO
 - [POS] is the position **before** the variant
 - [ID] links the variant to the original SV discovery method and callset (SV master validation tables)
 - [REF] and [ALT] show exact sequence if breakpoints are known, otherwise a variant-specific tag is used: (, <DUP:TANDEM>, <INS:ME:ALU>, <INS:ME:L1>, <INS:ME:SVA>)
 - [INFO] contains various information including [END] as the SV end coordinate

Example VCF Records for SVs

[POS]: Position before variant

```
#CHROM POS ID REF ALT QUAL FILTER INFO
1 1152535 P1_M_061510_1_86 GGCAGGGAAAGGCGAGCTCGTGGCCAGGCCCTGCAGGGAAAGGCCAGGCTCGTGGCCAGGCCCT
TCGTGGCCAGGCCGGCGGGAAAGGCGAGCTCGTGGCCAGGCCGGAGGCCAGGCTCGTGGCCAGGCCCT G
BKPTID=BC_Pilot1_del_6;END=1152680;HOMLEN=38;HOMSEQ=GCAGGGAAAGGCCAGGCTCGTGGCCAGGCCCTGCAGGGAAAGG;SVLEN=-145;SVTYPE=DEL;
VALIDATED;NOVEL;VALMETHOD=ASM;SVMETHOD=RP
```

Endpoint of SV

Reference Allele Sequence
(if breakpoint resolution)

Alternative Allele (with deletion)

Alternative Allele:
(With no breakpoint resolution)

```
1 1404466 P1_M_061510_1_3_G <DEL> .
. . CIEND=-200,1300;CIPOS=-991,309;
END=1405825;IMPRECISE;SVLEN=-1359;SVTYPE=DEL;VALIDATED;DBVARID=esv11756;VALMETHOD=CAV;SVMETHOD=RD
```

Estimated length
(negative for deletions)

Confidence Intervals around
Imprecise breakpoints

Master Validation Tables

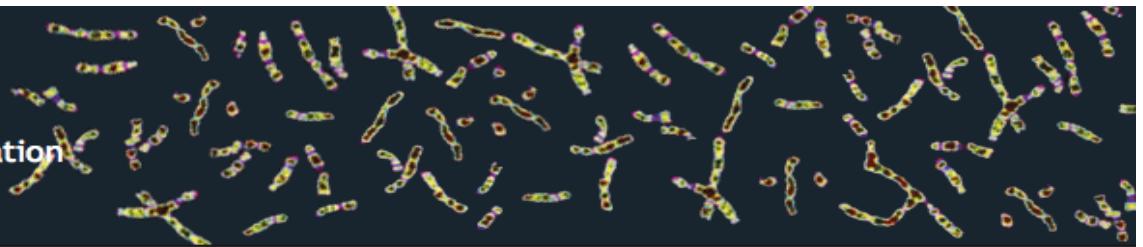
- Contains all reported SVs in standardized format for each individual algorithm
 - May find SVs in particular regions of interest which may be real but did not meet our stringent criteria (FDR <10%) for inclusion in release set.
- Reports specific validation results for each call
 - e.g., whether a call was validated by PCR, arrays, or sequence assembly
- Contains other meta information not found in VCF files
 - e.g. sequence technology and mapping algorithm used, assembled breakpoint sequences
- Particular fields of interest (see readme for more information)
 - [SAMPLES]: Which samples SV was originally discovered in
 - [SEQUENCE_TECHNOLOGY]: Sequencing platform used to make call
 - [MAPPING_ALGORITHM]: Mapping algorithm used to map reads to reference
 - [*_VALIDATION_*]: Results from various validation experiments

Master Validation Table Format

- Complete call sets with validation information
 - Tab-delimited Files
 - MasterValidation.Pilot1.all.leftmost.061510.txt
 - MasterValidation.Pilot2.all.leftmost.061510.txt
 - Assembled Breakpoints (if available)
 - MasterValidation.Pilot1.deletion.061510a.assembly.fasta
 - MasterValidation.Pilot2.deletion.061510a.assembly.fasta
- Merged call sets with refined breakpoint information
 - Similar format as complete call set files
 - [MERGED_ID] consistent with VCF [ID] field
 - [ID] column links back to complete call set files
 - MasterValidation.Pilot1.deletion.leftmost.061510a_mergedValPlus.txt
 - MasterValidation.Pilot2.deletion.leftmost.061510a_mergedValPlus.txt

1000 Genomes

A Deep Catalog of Human Genetic Variation



Information on ancestral state of SVs and of formation mechanism involved

[inferred with the BreakSeq algorithm; Lam *et al.*, *Nat. Biotechnol.*, 2010]

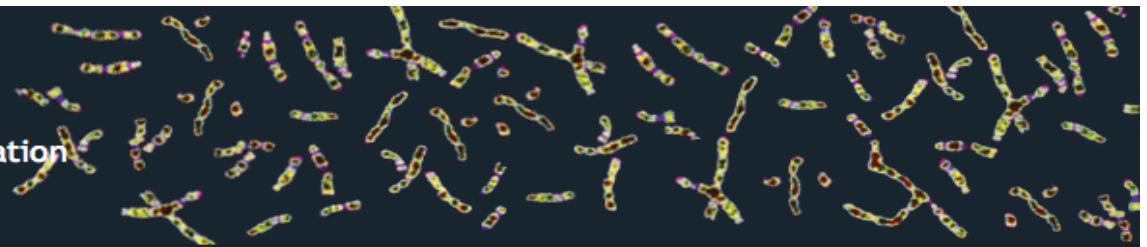
BreakSeq's GFF Format

#	Column	Description	Example
1	seqname	Chromosome	chrX
2	source	Source name	Yale
3	feature	Event type (Insertion/Deletion/Inversion)	Insertion
4	start	Start coordinate	13330
5	end	End coordinate	13331
6	score	<EMPTY>	.
7	strand	<EMPTY>	.
8	frame	<EMPTY>	.
9	additional attributes	Inserted sequence* ID Mechanism (e.g., non-allelic homologous recombination) Ancestral State of SVs (discriminates deletions from insertions)	Iseq "AATTGGGGCCTATAGTCCA"; Id "LIB000001"; Mech "NAHR"; Ancestral "Deletion"; etc

* for insertion; inserted sequences can be stored in a separate FASTA file

1000 Genomes

A Deep Catalog of Human Genetic Variation



Displaying SVs in the 1000 Genomes Browser

[presently available for deletions]

Example: deletion displayed on the 1000 Genomes Browser

v: Homo_sapiens - Region in detail - Chromosome 1: 72,510,000-72,610,000
http://test.1000genomes.org/Homo_sapiens/Location/View?r=1:72510000-72610000;time=1288798027596.596

Gmail PubMed Spiegel Google Scholar Calendar Maps LEO UCSC bl2seq DNA liftover BLAT isPcr EMBL Wikipedia Most Visited Thesaurus Oracle Mozilla Firefox Start ...
Google Mail - SVs in 1000 gen... v: Homo_sapiens - Region in d... +

1000 Genomes

A Deep Catalog of Human Genetic Variation

Home > Human

Location: 1:72,510,000-72,610,000

Location-based displays Chromosome 1: 72,510,000-72,610,000

- Whole genome
- Chromosome summary
- Region overview
- Region in detail**
- Comparative Genomics
- Genomic alignments (0)
- Synteny (0)
- Genetic Variation
- Resequencing (6)
- Linkage Data
- Markers

- Configure this page
- Add custom data to page
- Export data
- Bookmark this page

Chromosome bands Contigs Ensembl/Havana g... Chromosome 1 p31.1 q12 q41 q43 100.00 Kb Forward strand 72.51 Mb 72.52 Mb 72.53 Mb 72.54 Mb 72.55 Mb 72.56 Mb 72.57 Mb 72.58 Mb 72.59 Mb 72.60 Mb p31.1 AL512292.5.1.135389> RP11-1017019.1 RP11-292017.1 NEGR1 72.51 Mb 72.52 Mb 72.53 Mb 72.54 Mb 72.55 Mb 72.56 Mb 72.57 Mb 72.58 Mb 72.59 Mb 72.60 Mb Ensembl Homo sapiens version 54.36 (NCBI36) Chromosome 1: 72,510,000 - 72,610,000

Location: 1 : 72510000 - 72610000 Go << < > >>

Resembli settings Selected genome: Not chosen

Chromosome bands Ensembl/Havana g... Reference 1KG Low coverage... Ensembl/Havana g... Gene Legend

RP5-1017019.1-001 > Known pseudogene Vega gene RP11-292017.1-001 > Known pseudogene Vega gene

< NEGR1-001 Known protein coding Ensembl/Havana merge transcript Reverse strand Known protein coding Known pseudogene

Reference only displayed for less than 0.2 Kb.

100.00 Kb p31.1 Forward strand 100.00 Kb p31.1 Known pseudogene

Configuring the display You currently have 1 tracks in the overview panel and 103 tracks in the main panel turned off. To change the tracks you are displaying, use the "Configure this page" link on the left.

1000 Genomes release 3 - June 2010 © EBI Empowered About 1000Genomes | Contact Us | Help

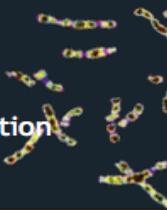
Large deletion

Imputing deletions into GWAS

- These deletions can be imputed into GWAS using existing tools (Beagle, MACH, etc.)
- Data availability

<http://www.1000genomes.org>

- Genotypes in VCF file format (Danacek *et al.*, manuscript submitted)
- Genotype calls at 95% confidence
- Genotype likelihoods for imputation

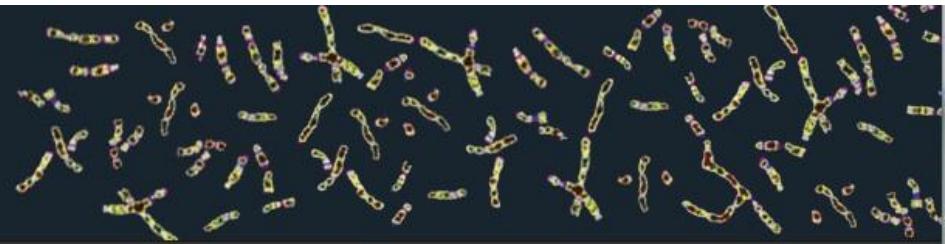


Further Information & Data Links

- 1000 Genomes Pilot Project SV Release Data & Readme files (links)
 - ftp://ftp-trace.ncbi.nih.gov/1000genomes/ftp/pilot_data/paper_data_sets/a_map_of_human_variation/low_coverage/sv/
 - ftp://ftp-trace.ncbi.nih.gov/1000genomes/ftp/pilot_data/paper_data_sets/a_map_of_human_variation/trio/sv/
- Link to auxiliary master validation tables & breakpoint assembly/analysis tables
 - ftp://ftp-trace.ncbi.nih.gov/1000genomes/ftp/pilot_data/technical/working/20100916_paper_data/companion_papers/mapping_structural_variation/
- More information on the description of SVs in the VCF format
 - <http://vcftools.sourceforge.net/specs.html>
- More information on bgzip and tabix (compression and coordinate indexing)
 - <http://samtools.sourceforge.net/tabix.shtml>

Acknowledgements

1000 Genomes
A Deep Catalog of Human Genetic Variation



1000 Genomes Project Structural Variation Group

WashU - Ken Chen, Asif Chinwalla, Li Ding

WT Sanger Inst - Klaudia Walter, Yujun Zhang, Aylwyn Scally, Don Conrad

Yale/Stanford - Mark Gerstein, Mike Snyder, Zhengdong Zhang, Jasmine Mu, Alex Eckeart Urban, Fabian Grubert, Alexej Abyzov, Jing Leng, Hugo Lam

EMBL - Jan Korbel, Adrian Stütz, Tobias Rausch

Univ of Washington - Jeff Kidd, Can Alkan

EBI - Daniel Zerbino, Mario Caccamo, Ewan Birney

Oxford - Zamin Iqbal, Gil McVean

LSU - Miriam Konkel, Jerilyn Walker, Mark Batzer

Simon Fraser - Iman Hajirasouliha, Fereydoun Hormozdiari

CSHL/AECOM/UCSD - Jonathan Sebat, Kenny Ye, Seungtai Yoon, Lilia Iakoucheva, Shuli Kang, Chang-Yun Lin

Illumina - Kiera Cheetham

AB - Heather Peckham, Yutao Fu

BC - Chip Stewart, Gabor Marth, Deniz Kural, Michael Stromberg, Jiantao Wu

Broad Inst - Josh Korn, Jim Nemesh, Steve McCarroll, Bob Handsaker

HMS - Ryan Mills, Mindy Shi

BGI - Ruiqiang Li, Ruibang Luo, Yingrui Li, Jun Wang

Leiden Univ – Kai Ye

Co-chairs: Matthew Hurles, Evan Eichler, Charles Lee

1000 Genomes Project

Structural Variation Glossary

- **Structural variant (SV)**: deletion, duplication, or insertion ($\geq 50\text{bp}$) relative to the reference genome (NCBI build36).
- **Ancestral state**: inferred SV class (deletion, duplication, insertion) relative to likely ancestral genome.
- **SV genotype**: allelic state determination of SVs in each genome (e.g., homozygous reference allele, homozygous SV allele, heterozygous SV allele).
- **SV breakpoint**: boundary (start- and end-coordinate) of SV, in case of breakpoint assembly and/or split-read analysis available at nucleotide resolution.