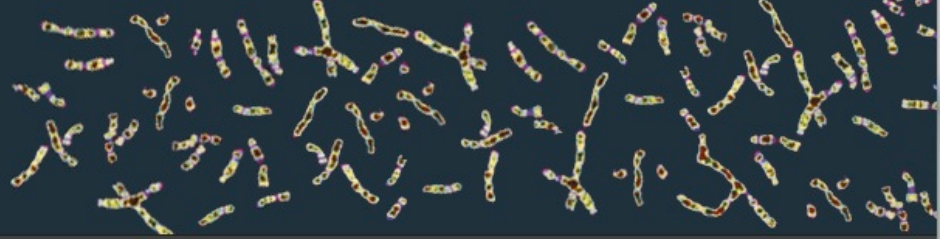
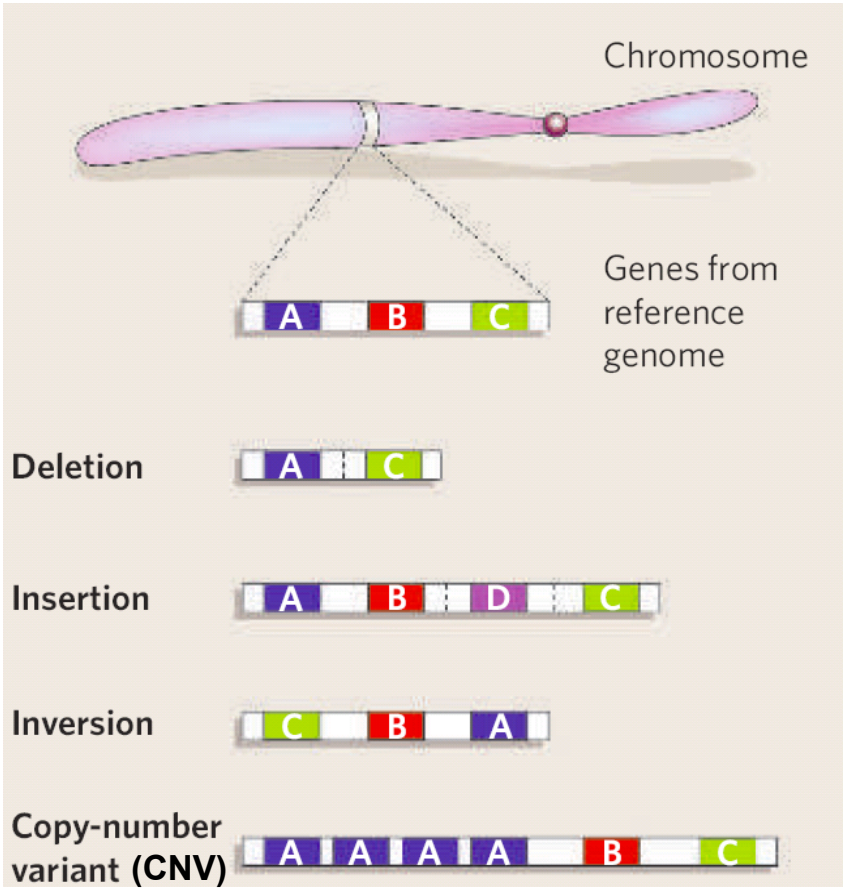


1000 Genomes Project Data Tutorial

Structural Variants

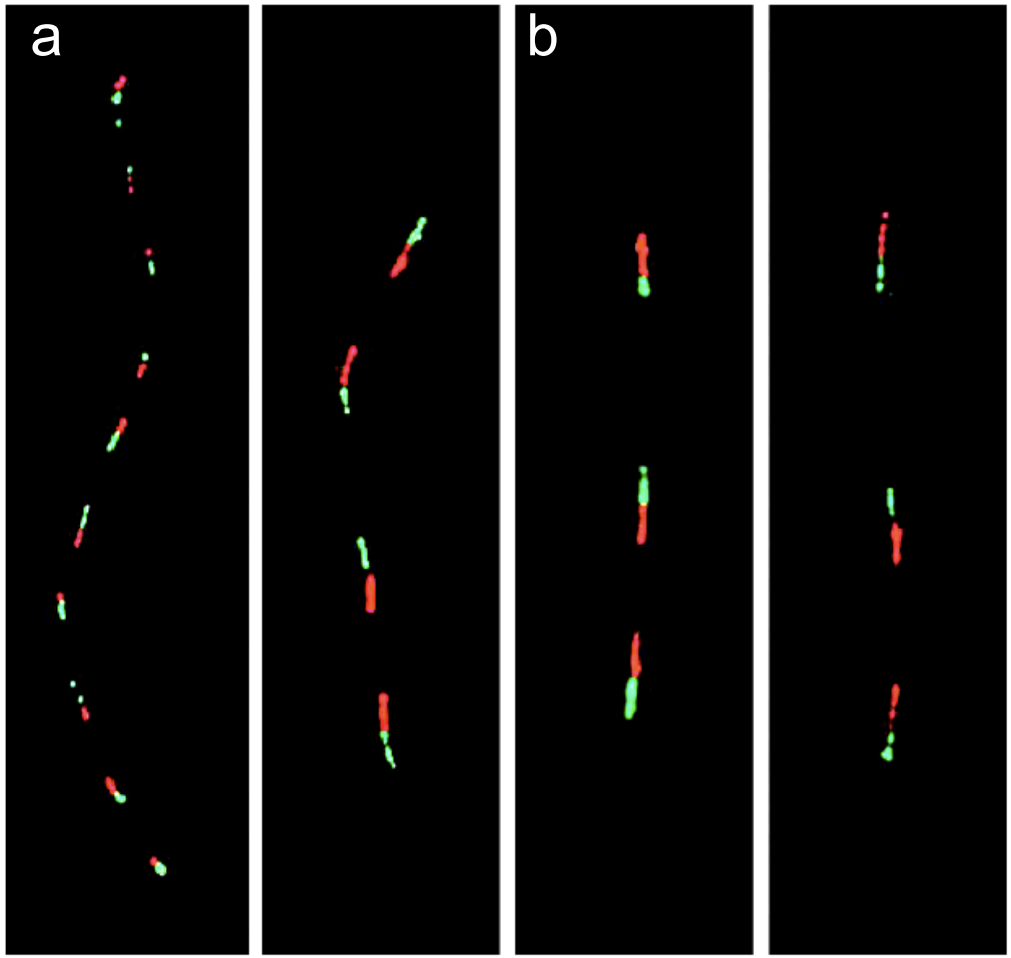
Ryan Mills, Ph.D.
Brigham and Women's Hospital
Harvard Medical School
Boston, MA

Structural Variants (SVs) in the Genome



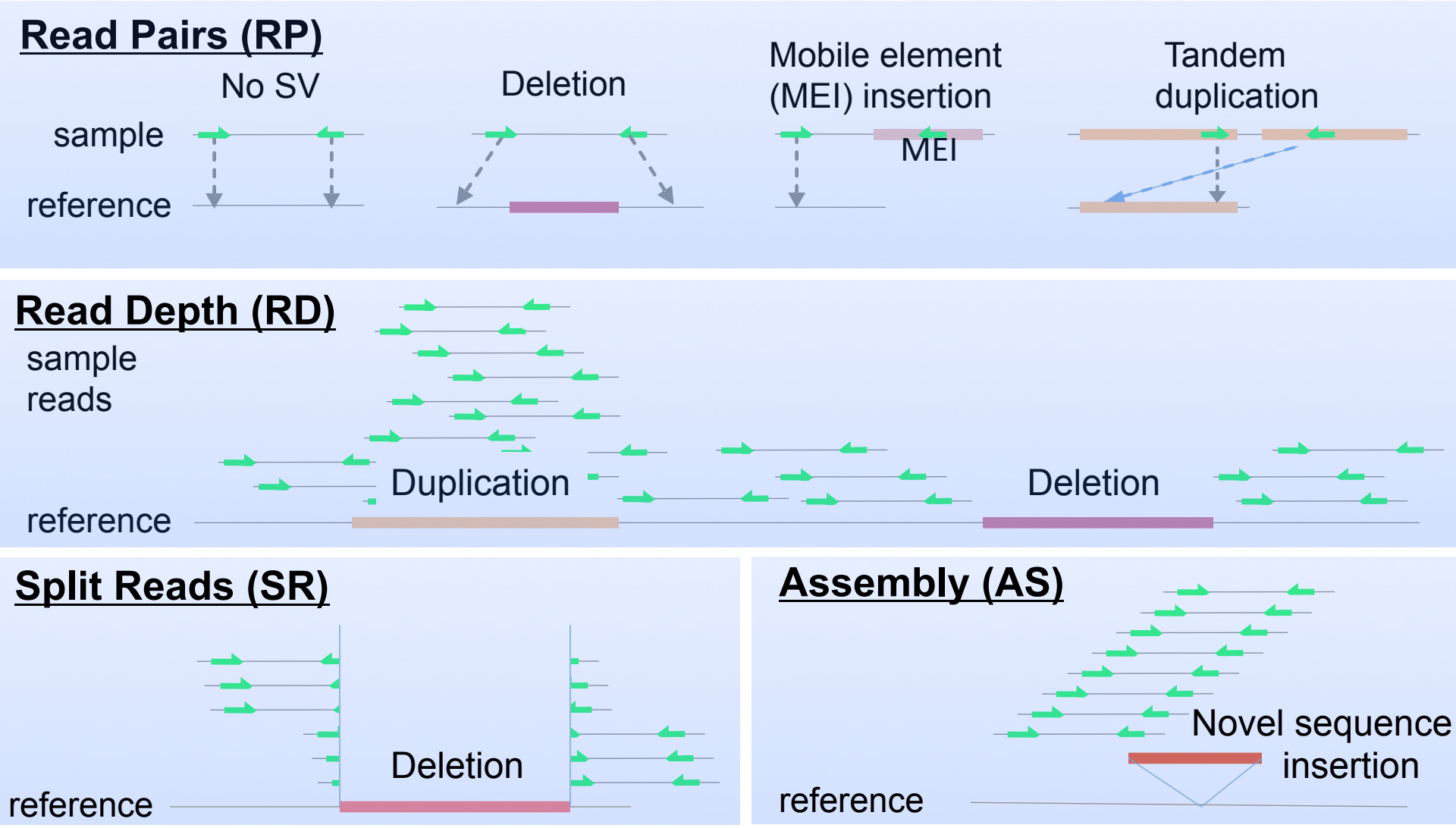
~0.5% of the genome according to current estimates

Striking *AMY1* gene copy-number variation

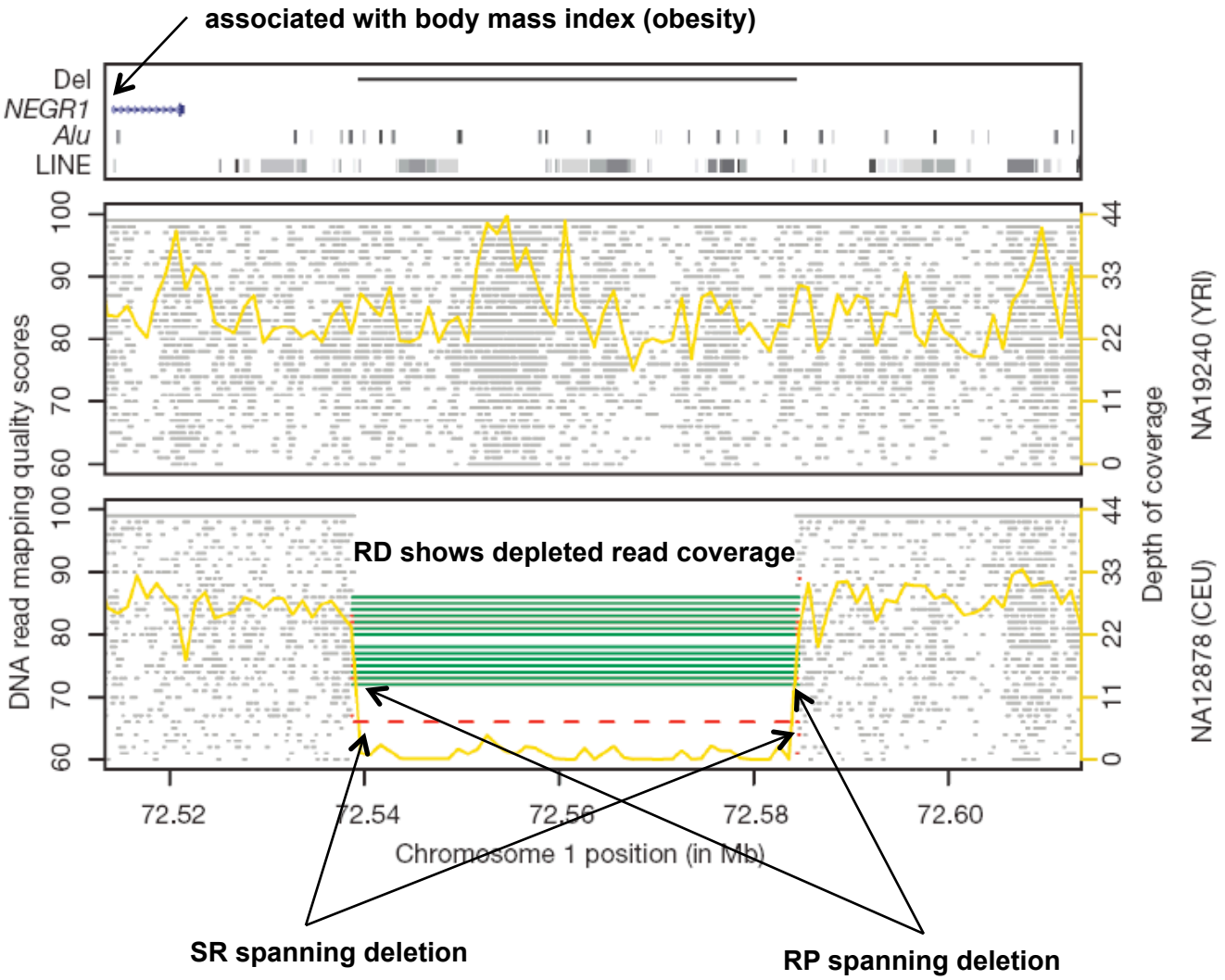


a, Japanese; **b**, African (Biaka) individual
[Perry *et al.*, *Nat. Genet.* 2007]

SV discovery considering evidence from multiple sources



Example SV with diverse support



Klaudia Walter

SV Discovery Algorithms

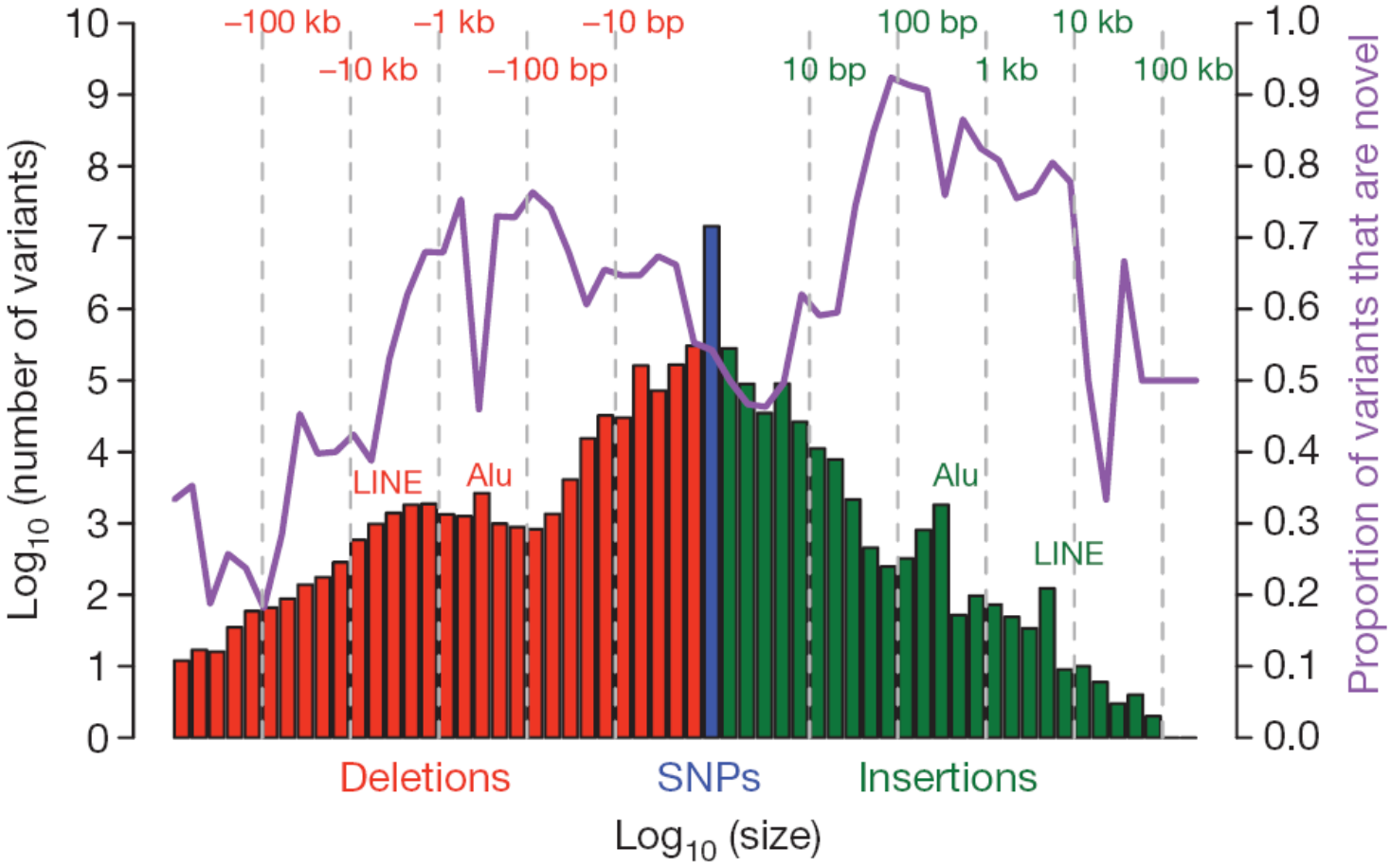
- Event-wise testing
- CNVnator
- Spanner
- PEMer
- BreakDancer
- Mosaik
- Pindel
- GenomeSTRiP
- mrFast
- AB large indel tool
- VariationHunter
- SOAPdenovo
- Cortex
- NovelSeq
- Various others

For full list of algorithms and parameters, please see:
http://www.nature.com/nature/journal/v470/n7332/extref/nature09708-s1.pdf

Tools for SV Discovery Assessment

- 1000 Genomes Project data provides a rich data set for developing and assessing detected structural variants
- The SuperArray annotator has been created to measure the efficacy of SV discovery algorithms
  - http://www.broadinstitute.org/gsa/wiki/index.php/SuperArray
- Tigra_SV and AGE algorithms allow for the identification and assessment of precise breakpoint locations for some discovered SVs
  - Tigra_SV: http://genome.wustl.edu/software/tigra_sv
  - AGE: http://sv.gersteinlab.org/age/
- GenomeSTRiP allows for the genotyping of discovered variants across multiple genomes
  - http://www.broadinstitute.org/gsa/wiki/index.php/Genome_STRiP

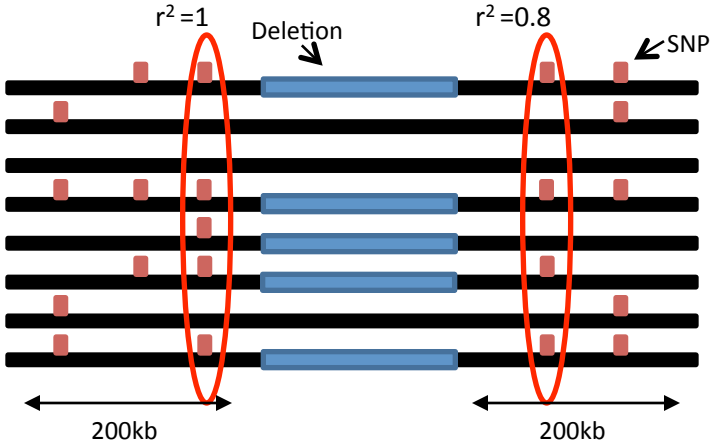
Length and Novelty of Discovered Variants

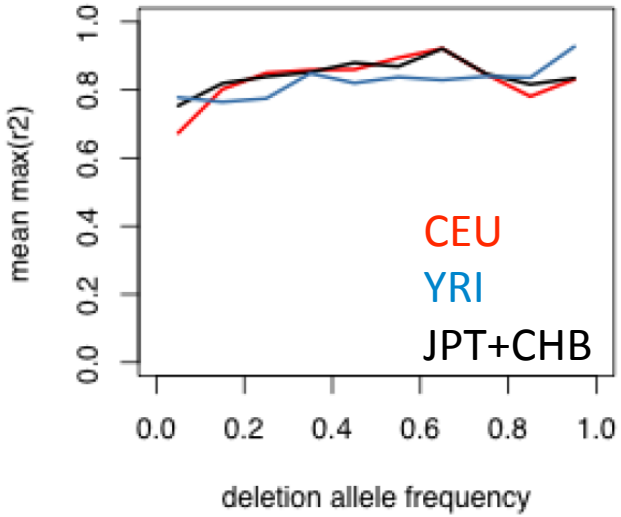


1000 Genomes Project Consortium, 2010

Deletions and SNPs on shared haplotypes

- Pearson's correlation coefficient (r^2) was calculated between genotyped deletions and HapMap3 SNPs to assess linkage disequilibrium (LD)
- For each deletion, the maximum r^2 among SNPs flanking the breakpoints (within 200kb) was determined
- 79% of common (MAF > 0.05) deletions were observed to be strongly correlated with a SNP (r^2>0.8)





SVs in high LD with tagged SNPs can be found here:
http://www.nature.com/ng/journal/v43/n3/extref/ng.768-S3.xlsx

(figure adapted from Handsaker et al., 2011)

Data formats and access

Location of Data Files

- Pilot phase SV discovery and genotyping data release
  - 185 samples in total
  - <10% false discovery rate, validated calls labeled
  - Includes call sets (.vcf) and breakpoint assembly sequences (.fasta)
  - ftp://ftp-trace.ncbi.nih.gov/1000genomes/ftp/pilot_data/paper_data_sets/companion_papers/mapping_structural_variation/
- Phase 1 released integrated variants and phased genotypes
  - 1092 individuals
  - Highly accurate but conservative deletion data set
  - Includes SNPs, Indels and SVs
  - ftp://ftp-trace.ncbi.nih.gov/1000genomes/ftp/release/20110521
- Further SV data releases forthcoming (Winter 2011)
  - Will be announced at project website
  - Will include larger, more sensitive set of deletions as well as duplications, insertions, and inversions
  - www.1000genomes.org

Can also be accessed from 1000 Genomes Project Browser:

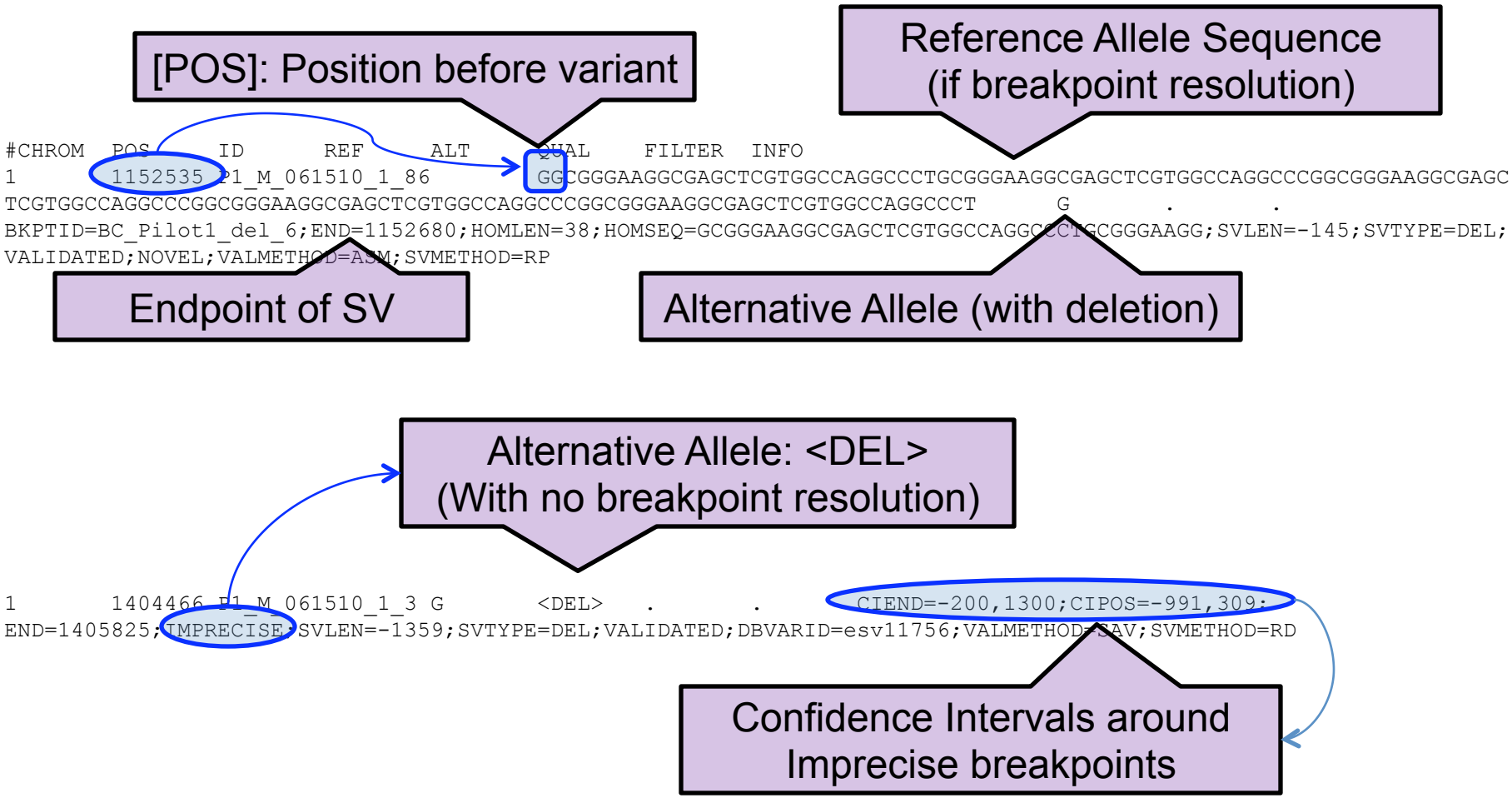
http://browser.1000genomes.org/

SV discovery set in VCF format

- Compressed with bgzip (.gz), indexed with tabix (.gz.tbi) [e.g., to enable quick retrieval of data lines overlapping specific genome regions].
- 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 usd: (<DEL>, <DUP:TANDEM>, <INS:ME:ALU>, <INS:ME:L1>, <INS:ME:SVA>)
  - [INFO] contains various information including [END] as the SV end coordinate

- Detailed specifications are available at http://vcftools.sourceforge.net/specs.html

Example VCF Records for SVs



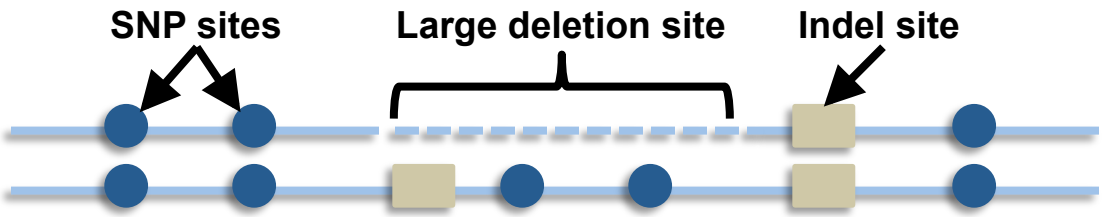
Processing VCF genotypes with *vcftools*

- *--012* converts vcf file into large matrix with samples as columns and genotypes as 0,1,2 representing the number of non-reference alleles
- *--IMPUTE* converts vcf file into IMPUTE reference-panel format
- *--BEAGLE-GL* converts vcf into input file for the BEAGLE program
- *--plink* converts vcf into PLINK PED format

Full list of commands can be found here:
http://vcftools.sourceforge.net/options.html

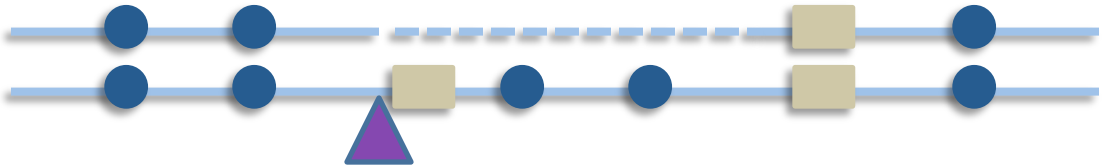
Integrated SV Genotypes

Strategies for integrating deletions with other types of variation



Previous Approach
Remove SNPs under SVs for imputation
(1000G pilot, Handsaker et al., 2010)

Current Approach
Treat SVs as point events
(1000 Genomes phase 1)

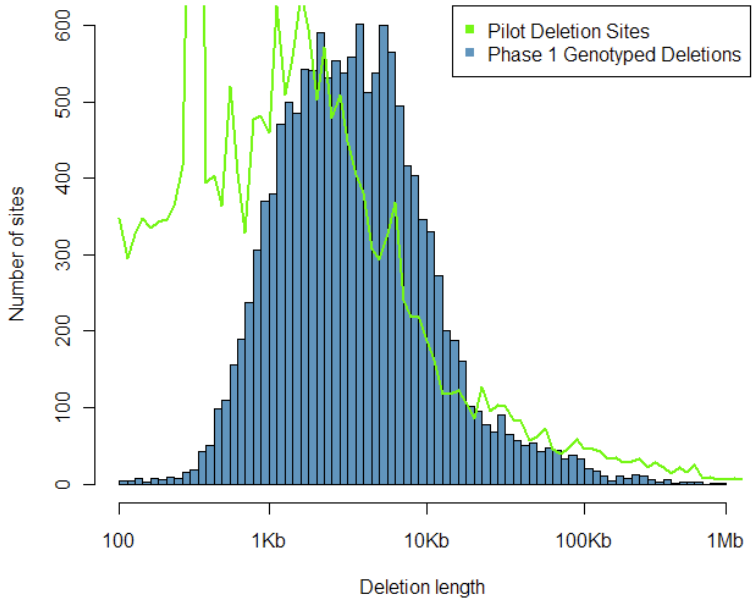


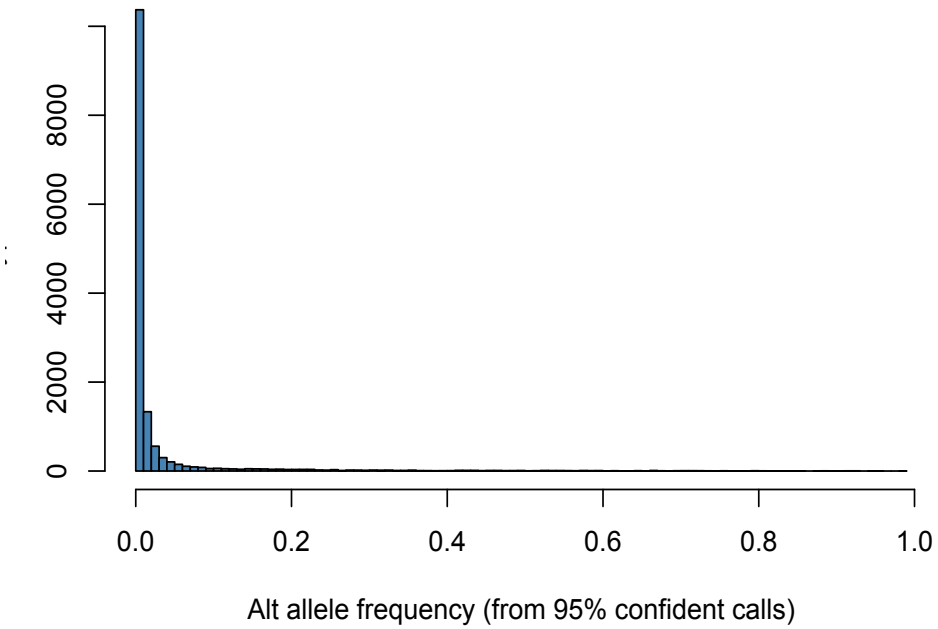
Matt Hurles

Site selection for integrated call set

- Candidate sites
  - Deletions called by Genome STRiP
  - Deletions from other callers with SAV validation $p < 0.01$
  - Autosome and chrX
- Genotyped using read depth + read pairs
  - Read depth cluster separation
  - Normalized read depth within 50% of genome-wide average
  - Length of unique sequence > 100bp
- Redundant call removal using genotype likelihoods
- Additional filters
  - Remove sites with inbreeding coefficient < -0.15
  - Remove sites where all samples > 95% confident homref
- Final set
  - 14,422 sites
  - Median length of 2.9 Kbp

Length and frequency spectrum for genotyped sites

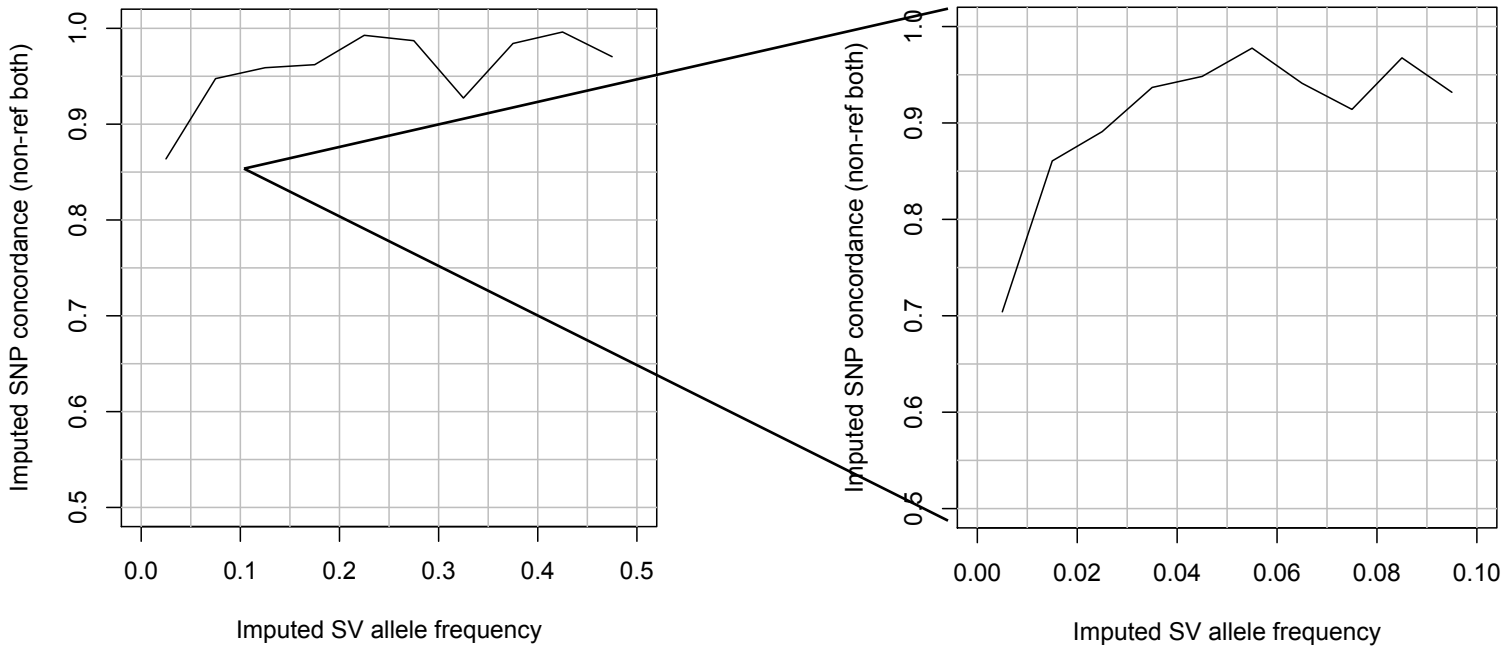




Bob Handsaker

Concordance with SNP genotypes

- Calculated using purely imputed genotypes compared to Conrad et al, 2010
- Follows similar trends as imputed SNP genotypes
- High imputation accuracy across frequency spectrum, with the exception of less concordance at lower frequencies



Bob Handsaker

Acknowledgements

1000 Genomes Project
Structural Variation Analysis Group

WashU - Ken Chen, Asif Chinwalla, Li Ding

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

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

EMBL - Jan Korbel, Adrian Stütz

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, Jayon Lihm

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, Marcin von Grotthuss

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

Leiden Univ – Kai Ye

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