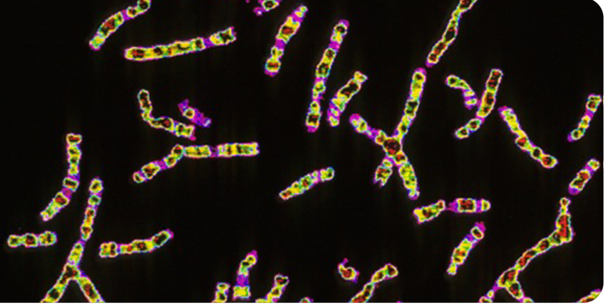
1000 Genomes

A Deep Catalog of Human Genetic Variation



1000 Genomes Project Phase III Tutorial

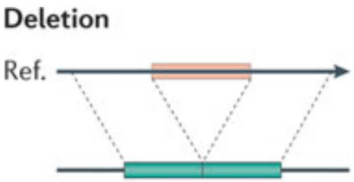
Structural Variants (SVs)

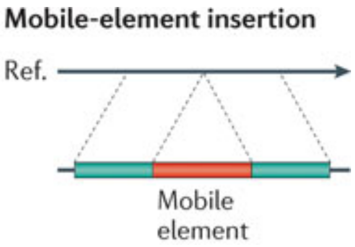
Eugene J. Gardner
University of Maryland
Institute for Genome Sciences
Baltimore, MD
On Behalf of the 1000 Genomes Structural Variation Analysis Group

SV Discovery Over the Three Phases of the 1000 Genomes

Pilot:

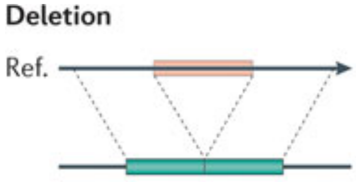
- Deletions (DEL)
- Mobile Element Insertions (MEI)





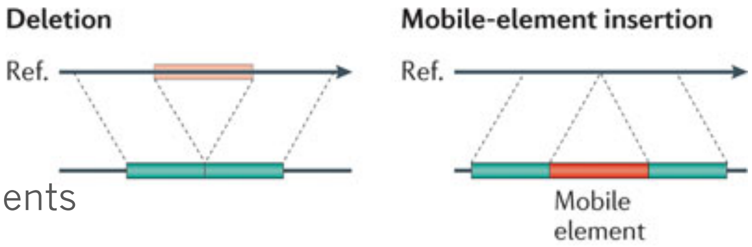
Phase I:

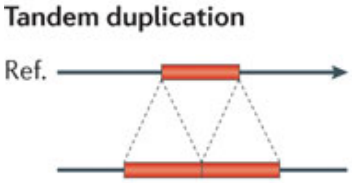
- Only *LARGE* DEL

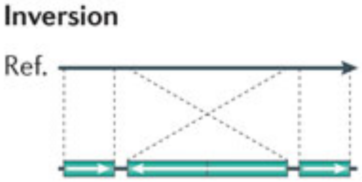


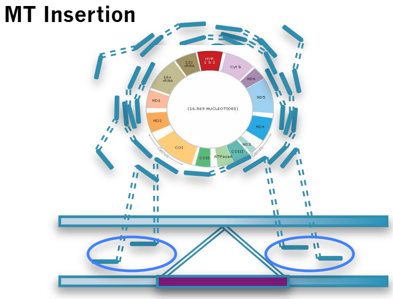
Phase III:

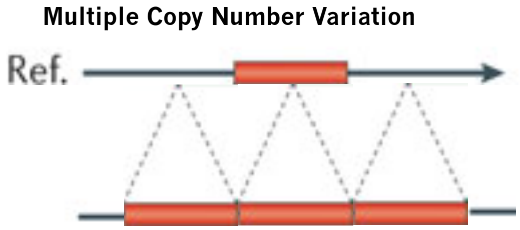
- DEL – Many more deletions including more 50-500bp
  - 7 Total callers
- MEI – Redesigned from ground up to use BWA Alignments
- New types of variation!!!
  - Duplications (DUP) – Copy Number = 2
  - multiple Copy Number Variation (mCNV) – Copy Number = 3+
  - Inversions (INV)
  - Nuclear Mitochondrial Insertions (NUMT)











Cartoons from Alkan, Coe, and Eichler. Nature Reviews Genetics. 2011.

SV Callers and What They Discover

Deletion (DEL)

GenomeStrip
Breakdancer
CNVnator
Delly
Variation Hunter
UWash RD
Pindel (*Short Deletions*)

multiple Copy Number Variation (mCNV)

UWash SSL
GenomeStrip

Duplications (DUP)

Delly
UWash RD
GenomeStrip

Inversions (INV)

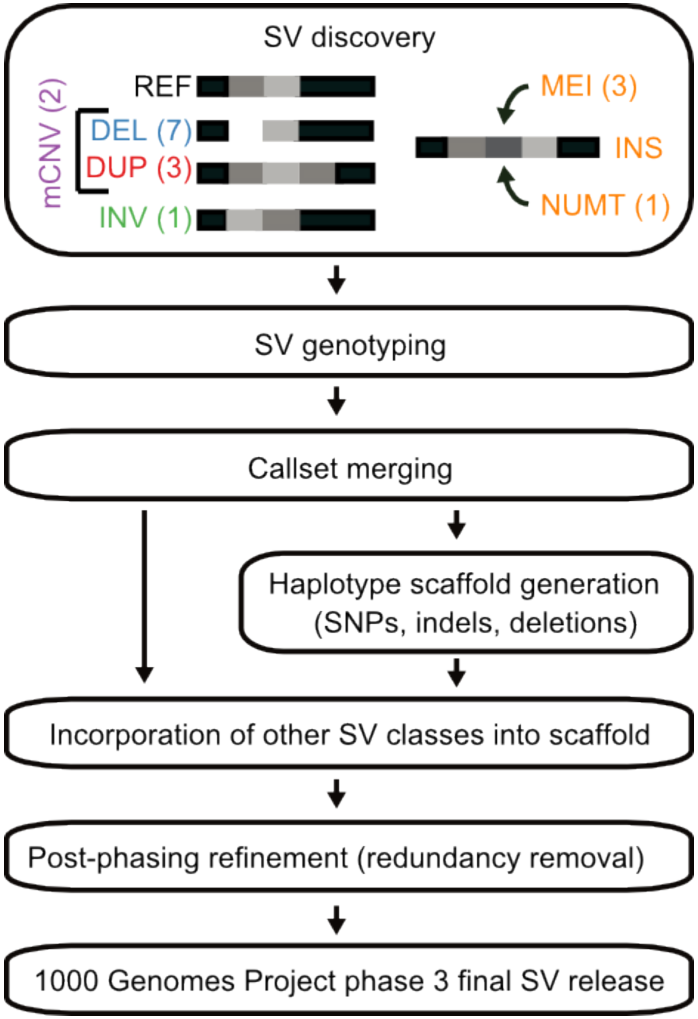
Delly

Mobile Element Insertions (MEI)

MELT

Mitocondrial Insertions (NUMT)

Dinumt



SV genotying

Callset merging

Haplotype scaffold generation (SNPs, indels, deletions)

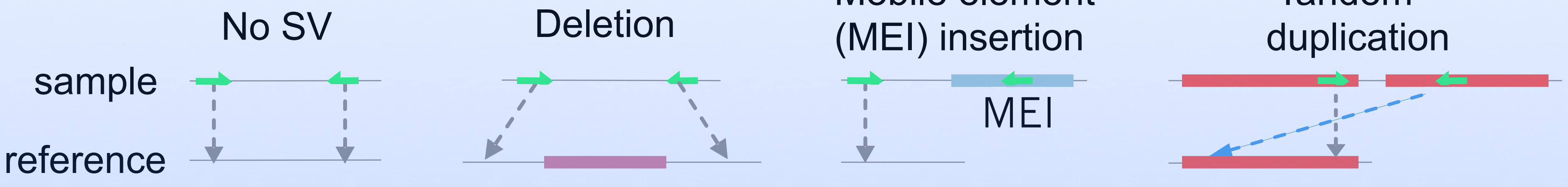
Incorporation of other SV classes into scaffold

Post-phasing refinement (redundancy removal)

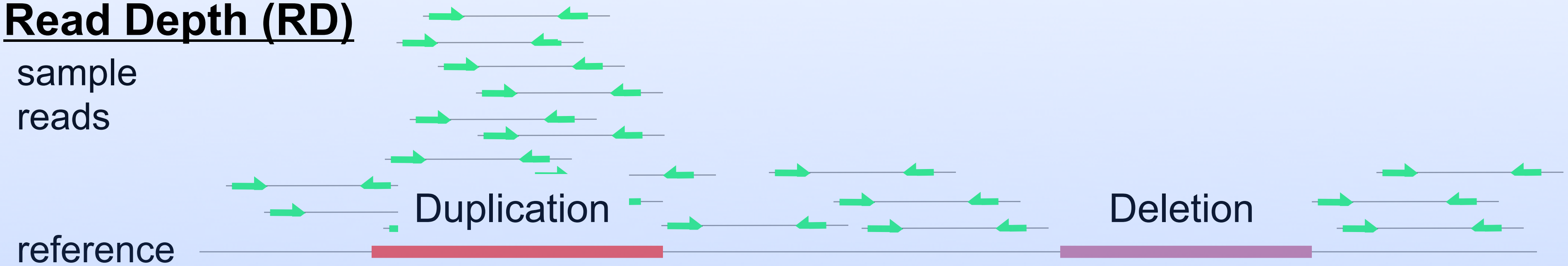
1000 Genomes Project phase 3 final SV release

Detection of SVs Using Multiple Pieces of Evidence

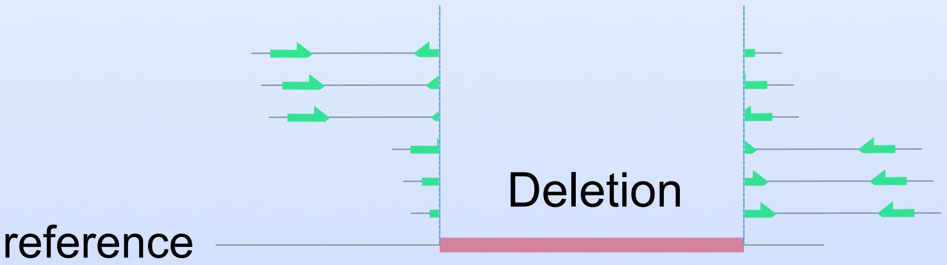
Read Pairs (RP)



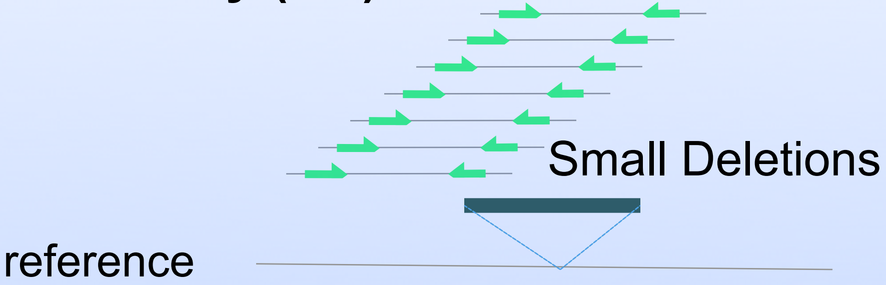
Read Depth (RD)



Split Reads (SR)

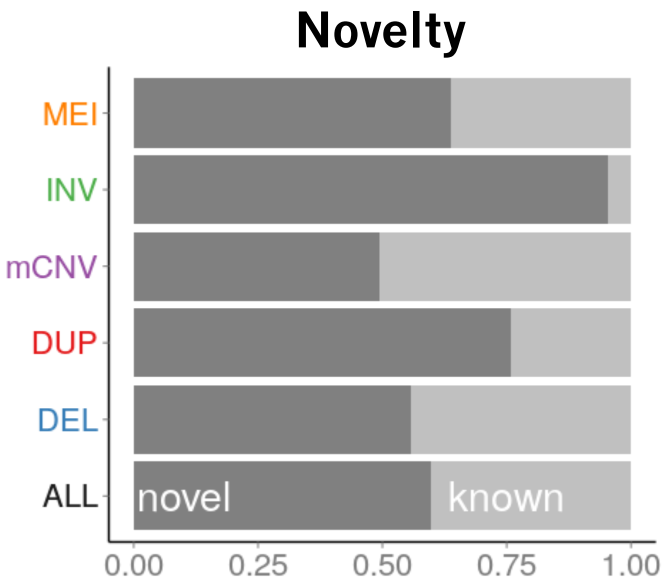


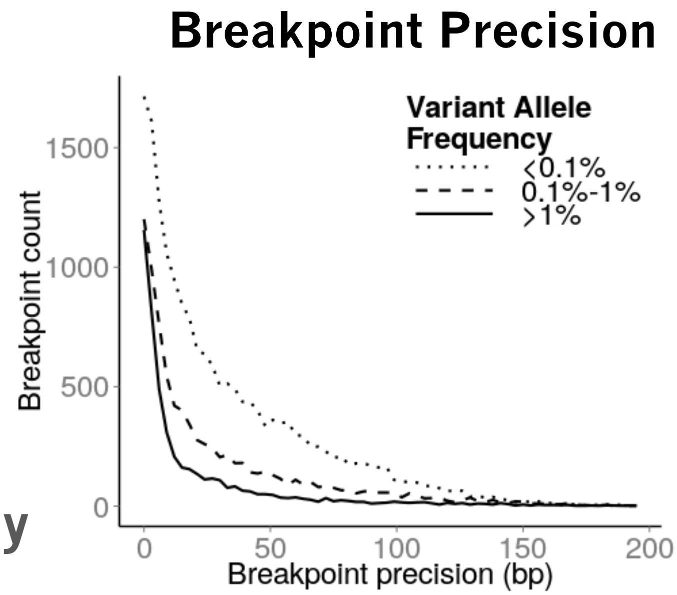
Assembly (AS)



Phase III Improvements

- **Expanded dataset**
  - Total number of SVs has vastly increased (69,353 vs. 14,422)
  - 26 populations from 5 super populations
  - ~60% novel variants
- **Technology Development**
  - Many new callers (algorithm development as a tool for the community)
    - Improved technology and efficiency
  - Genotypes for ALL call-sets
    - Improves use for GWAS studies in all 26 populations
  - Genotyped deletions from 50bp – 500bp
    - Only had >500bp before
  - Enhanced breakpoint precision



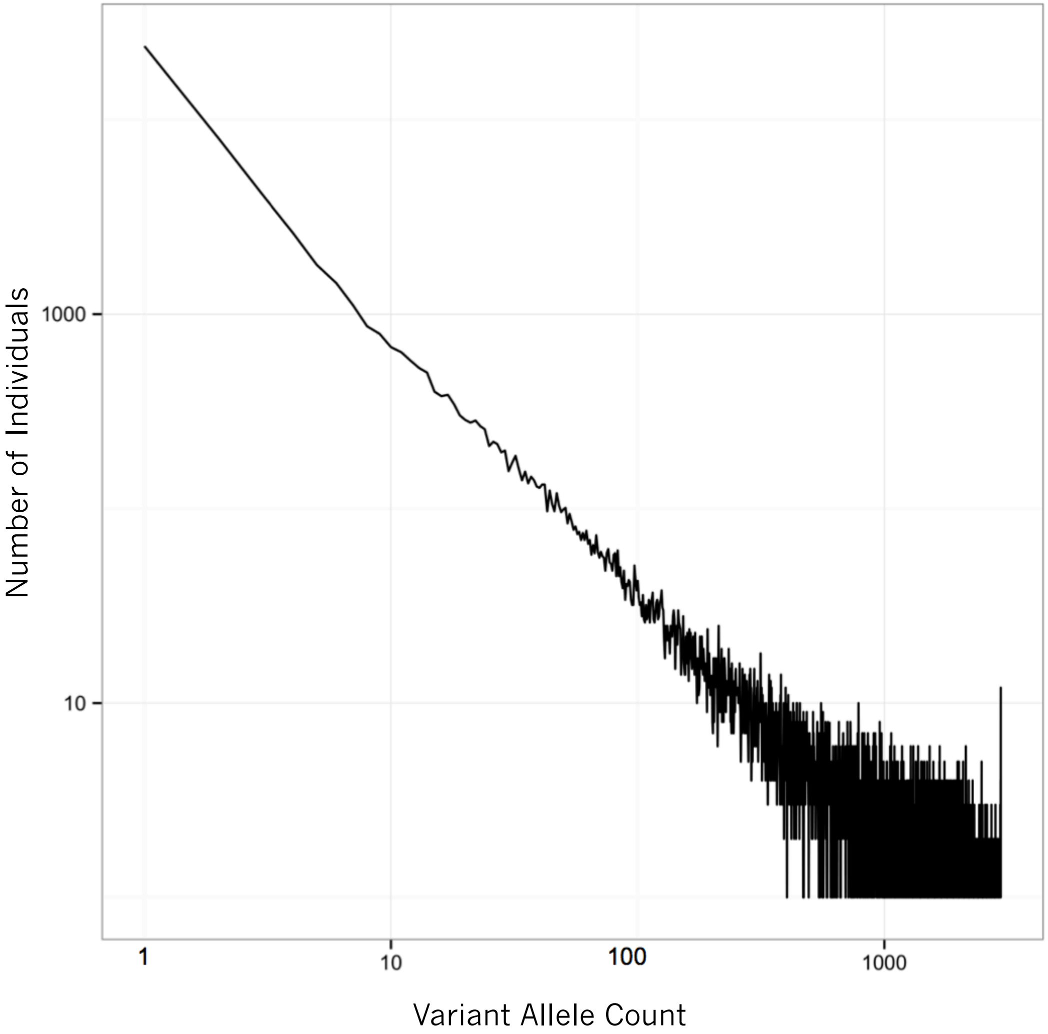


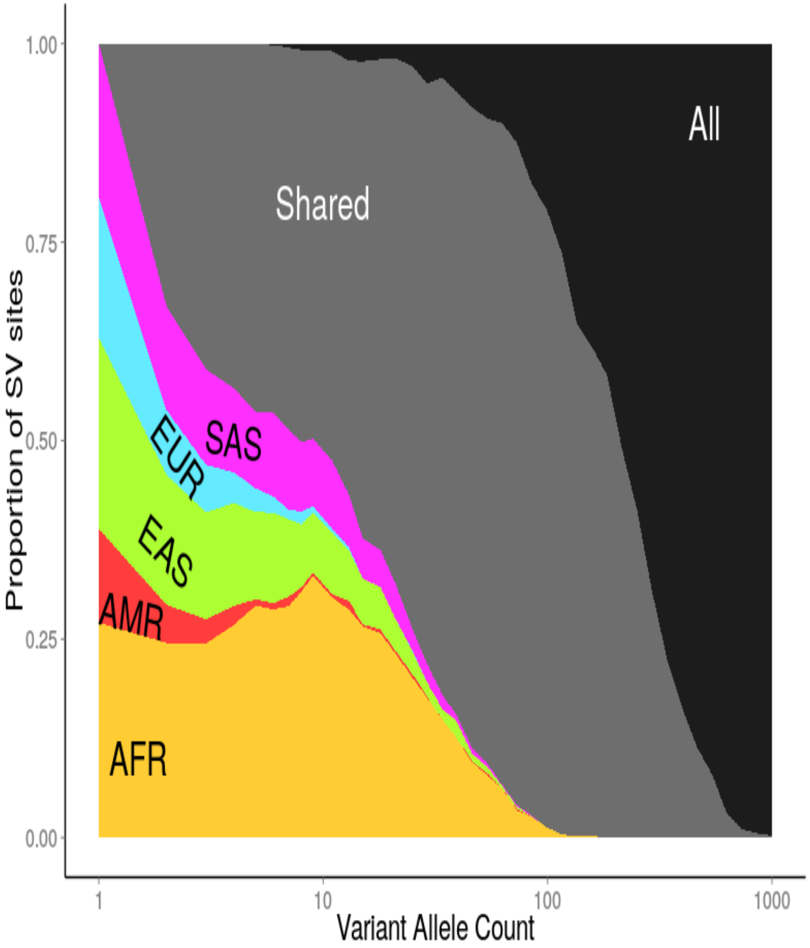
Utilizing the power that comes with more advanced sequencing technology (high coverage, read pairs, long reads)

Summary of 1KGP Phase III SV Calls

SV type	No. sites	Pilot/Phase I	Mean size of SV sites	Alleles per individual (mean)	Affected portion of genome	Site FDR	Genotype concordance (non-ref.)	Sensitivity estimates	Tools
DEL	42,491	23,594 (Union Set), 14,422 (Genotyped)	9,633 bp	1,879	0.14%	2%^{A} - 4%^{O}	97.5%^{C}	84.4%^{C}	BreakDancer, Delly, CNVnator, GenomeSTRiP, Pindel, UWash-RD Variation-Hunter
DUP	6,136	501 (Tandem Only)	66,105 bp	17	0.044%	1%^{A} - 4%^{O}	91.4%^{C}	53.2%^{C}	Delly, GenomeSTRiP, UWash-RD
mCNV	3,014	N/A	37,115 bp	195	0.48%	1%^{A} - 2%^{O}		NA	GenomeSTRiP, UWash-RD
INV	858 (100 simple; 758 complex)	N/A	33,767 bp	24	0.0024%	11%^{L} (20%^{L} for one-sided)	Markus contacted – based on Pang et al	24% (67% for Inv. <5kb; 32% for <50kb)	Delly
MEI (*Alu*, L1, SVA)	16,684 (12,786, 3060, 838)	3,276 (2882, 345, 49)	(268 bp, 3,063 bp) 957 bp)	706 (583, 91, 32)	(0.0054%, 0.0078%, 0.0012%)	3.7%^{P} (3.1%, 3.6%, 11.9%)	98%^{D} (97%, 98%, 98%)	82.9$^{\#}$ - 96.0%^{V}	MELT
NUMT	170	N/A	815bp	3	0.00010 %	10%^{P}	86.1%^{P}	*NA**	Dinumt

Most Variants are Rare, and Those Variants are Only in One Population





Figures Courtesy of Tobias Rausch and Jan Korbel

Addition and Improvement of Multiple SV Types

- Addition of multiple new SV types with expected length distributions

- Improvement over previous releases in terms of length distribution (DEL calls)

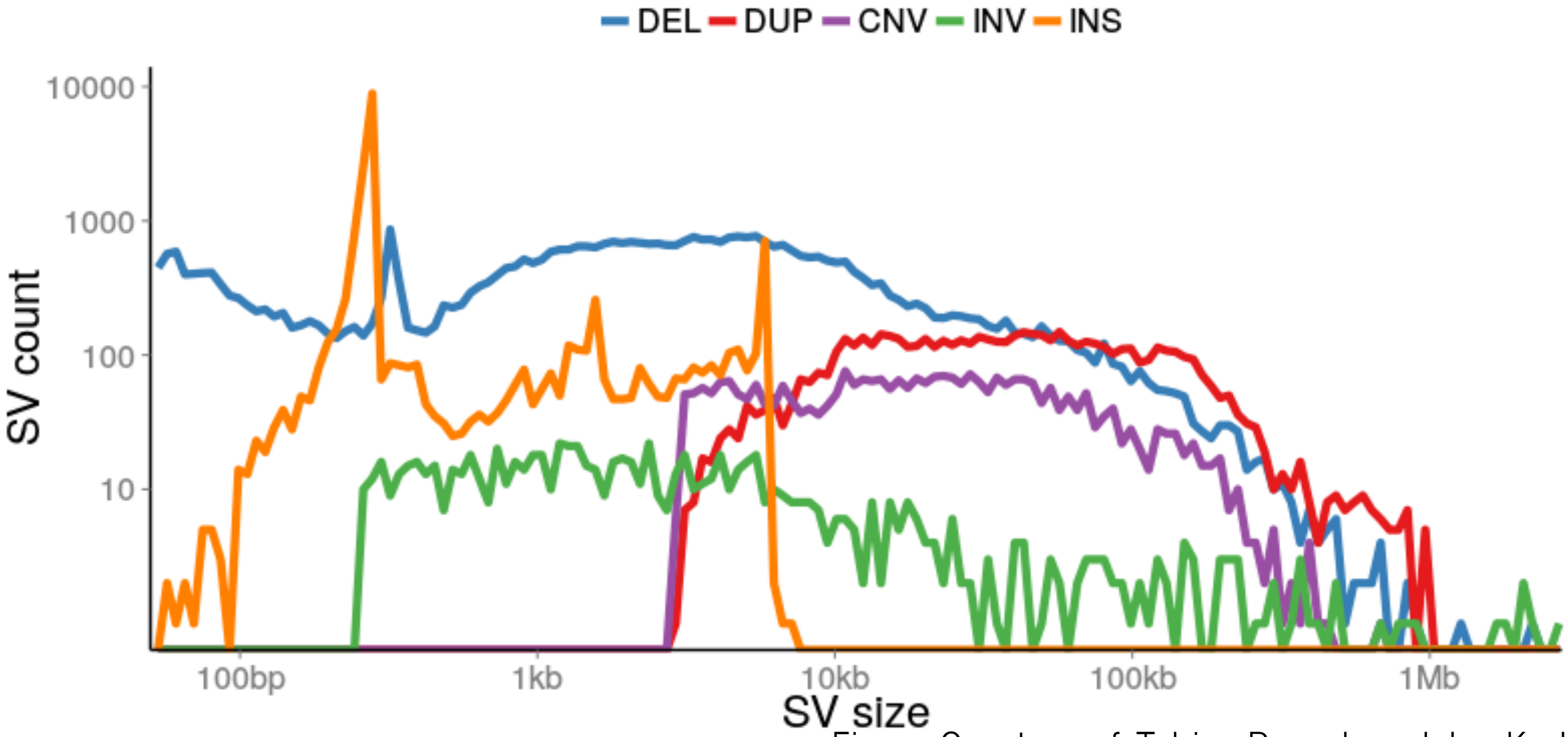


Figure Courtesy of Tobias Rausch and Jan Korbel

SVs are in LD with Nearby SNPs

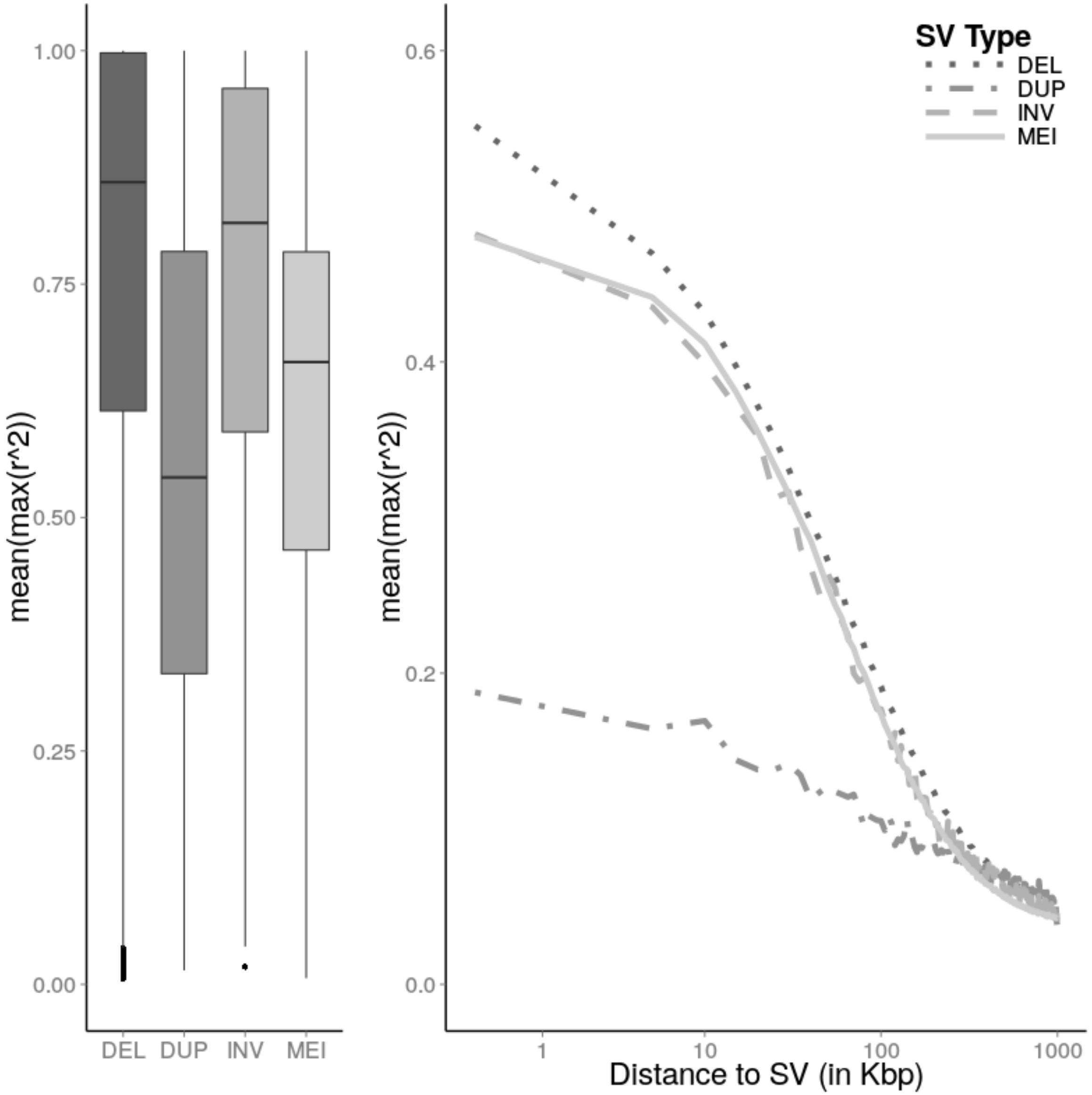


Figure Courtesy of Tobias Rausch and Jan Korbel

How to Access and Use SV Data

All individual calls talked about in this presentation can be found here:
ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/supporting/input_callsets/

The final SV calls can be found here:
ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/working/20130723_phase3_wg/merged_sv_genotypes/ALL.wgs.mergedSV.v3.20130502.svs.genotypes.vcf.gz

A final list of algorithms will be released with the 1KGP Phase III/SV Companion!

- All calls are in Variant Call Format (VCF) v4.1
  - Specification at: http://samtools.github.io/hts-specs/VCFv4.1.pdf
- VCF Record for SV is slightly different that for a SNP or InDel
  - Header is same as standard VCF format
  - Often no specific sequence included in ALT/REF field
  - INFO field includes several different conventions
  - These vary based on the SV-Type being called

VCF Format with an MEI Specific Example

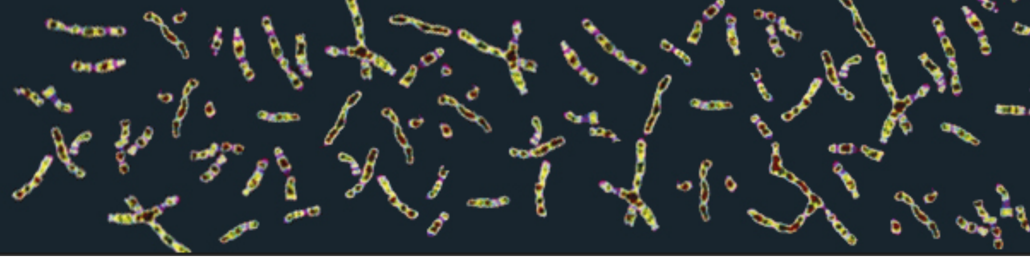
```
CHROM POS       ID                REF ALT            QUAL  FILTER   INFO

1     3011887   ALU_umary_ALU_3   A   <INS:ME:ALU>   .     .        TSD=AGAAAGTGGAGTA;SVTYPE=ALU;MEINFO=AluYa5,1,281,-;SVLEN=280;CS=ALU_umary
```

- First 6 Columns:
  - Fairly similar in all SV types!
  - CHROM – Chromosome Variant is on
  - POS – Position on Locus
  - ID – Specific ID from the specific Callset (no real convention here, up to the group that contributed the calls)
  - REF – All *EXCEPT PINDEL* (Identify with CS Tag in INFO field) simply lists the 1st base before POS
    - Pindel lists the entire sequence of the deletion
  - ALT – Shows the variant type (i.e. LINE1, INV, etc.)
  - QUAL – May or may not be included based on callset, typically is Phred-based score for call quality when included
  - FILTER – Only PASS variants are included in final VCF, even if they do not say pass

- INFO Field – Column 7
  - Common to ALL SV Types:
    - SVTYPE – One of the types listed on the table on slide 8 (INV, MEI, DUP, etc.)
    - CS – Callset this variant was derived from. Allows end-users to go back to raw calls
  - SV Specific – MEI
    - TSD – Described Target Site Duplication
    - MEINFO – Specific MEI info such as subspecies and start/stop positions
  - A Note on END/SVLEN:
    - Use SVLEN if it is available. END is not necessarily the exact length of the event

- GENOTYPES – Column 8+:
  - ALL Reported SV variants have GTs for every reported individual in the 1KGP
  - Have either been phased onto Haplotypes (DEL Calls) or Imputed onto the Scaffolds generated by this phasing
  - Fully supported for GWAS/Pop Gen studies

1000 Genomes
A Deep Catalog of Human Genetic Variation



1000 Genomes Structural Variation Analysis Group

WashU – Li Ding, Kai Ye

WT Sanger Institute – Klaudia Walter

Yale – Ekta Khurana, Yan Zhang, Jing Zhang, Mark Gerstein

EMBL – Adrian Stütz, Tobias Rausch, Markus Fritz

EMBL-EBI – Oliver Stegle

Univ of Washington – Peter Sudmant, Fereydoun Hormozdiari, John Huddleston, Elif Dal, Fatma Balci, Ben Nelson

Mayo Clinic – Nick Parrish, Alexaj Abyzov

UCSD – Amina Noor, Danny Antaki, Madhu Gujral, Jonathan Sebat

Bilkent University – Can Alkan

LSU – Miriam Konkel, Jerilyn Walker, Sarah Brantley, Mark Batzer

Univ of Maryland – Eugene Gardner, Scott Devine

Univ of Michigan – Gargi Dayama, Ryan Mills

Univ of Michigan – Sarah Emery, Jeff Kidd

Univ of Michigan – Goo Jun, Goncalo Abecasis

UT / MD Anderson – Wanding Zhou, Zechen Chong, Xian Fan, Ken Chen

Broad Institute – Bob Handsaker, Steve McCarroll

Boston College –Eric Garrison, Wanping Lee, Alistair Ward, Gabor Marth

Baylor CoM – Fuli Yu

UNC Charlotte – Mindy Shi

Jax Genomic Medicine – Ankit Malhotra, Dariusz Plewczy, Kamen Radew

MSSM – Ali Bashir

Co-chairs: Jan Korbel (EMBL)
Evan Eichler (U. Washington)
Charles Lee (Jax)