Structural Variation in the 1000 Genomes Project

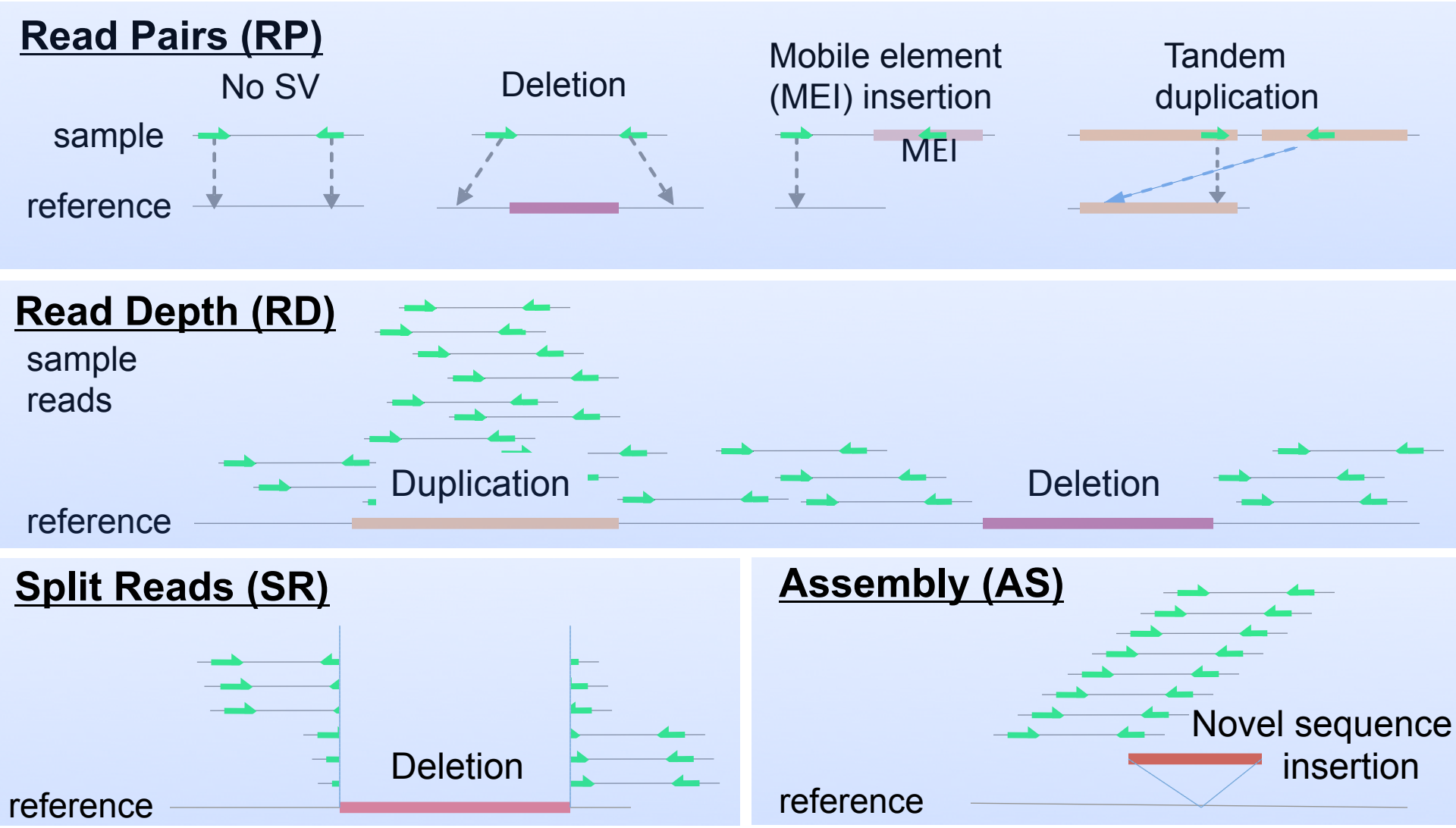
Ryan Mills
on behalf of the 1000 Genomes
Structural Variation Analysis Group

Department of Computational Medicine & Bioinformatics
Department of Human Genetics
University of Michigan

November 7, 2012

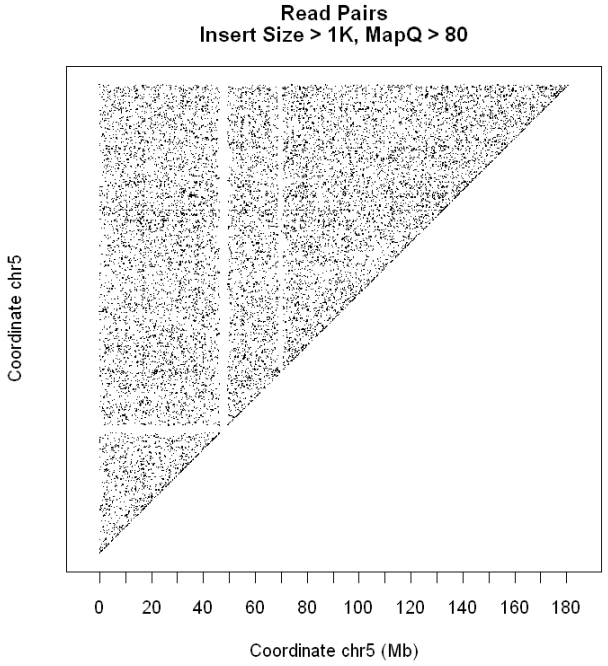
Slides courtesy of Bob Handsaker

Ascertaining large variants from short reads



Why is structural variation calling challenging?

- Artifacts abound
  - Millions of chimeric molecules generated during library construction
  - Read depth varies across the genome and across libraries
  - Alignment algorithms are misled by the genome's repeats
- Low-coverage sequencing
  - Data is not definitive in each genome
  - False discoveries can accumulate across genomes
- Deep genomes
  - Increased depth can help, but methodology is more important



Structural Variation in 1000 Genomes

What are the goals?

- Create a comprehensive catalog of human structural polymorphism
- Create a reference panel for imputing structural polymorphisms

Structural Variation in 1000 Genomes

What are the goals?

Create a comprehensive catalog
of human structural polymorphism

Create a reference panel
for imputing structural polymorphisms

1000 Genomes Project Phases

Pilot Phase (2010/2011)

179 lowcov genomes
2 deep trios
Plus exome sequencing

Phase 1 (2012)

1092 lowcov genomes

Phases 2 & 3 (2013)

2500 lowcov genomes
500 deep genomes from Complete Genomics

Structural Variation Goals

- Variant catalog of multiple variant types
- Genotypes for some variants (deletions, mobile element insertions)

- Expanded catalog of deletions
- Integrated haplotypes combining deletions with SNPs / indels

- Expanded variant catalog covering many variant types
- Integrated haplotypes, perhaps combining multiple variant types

1000 Genomes Project Phases

	Structural Variation Goals
Pilot Phase (2010/2011) 179 lowcov genomes 2 deep trios Plus exome sequencing	• Variant catalog of multiple variant types • Genotypes for some variants (deletions, mobile element insertions)
Phase 1 (2012) 1092 lowcov genomes	• Expanded catalog of deletions • Integrated haplotypes combining deletions with SNPs / indels
Phases 2 & 3 (2013) 2500 lowcov genomes 500 deep genomes from Complete Genomics	• Expanded variant catalog covering many variant types • Integrated haplotypes, perhaps combining multiple variant types

Pilot Phase SV Discovery Algorithms

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

A full list of algorithms and parameters can be found here:
http://www.nature.com/nature/journal/v470/n7332/extref/nature09708-s1.pdf

Variant Catalog – Pilot Phase

1000 Genomes Pilot

ftp://ftp-trace.ncbi.nih.gov/1000genomes/ftp/pilot_data/paper_data_sets /companion_papers/mapping_structural_variation/

Deletions

22,025 sites with genotype information for 156 pilot samples

Mobile element insertions (ALU, LINE, SVA)

5,371 sites with genotype information for 156 pilot samples

Tandem duplications

500 sites, with (sparse) genotype information

Novel insertions (other than transposable elements)

128 sites

Mapping copy number variation by population-scale genome sequencing. Mills, et al. *Nature* 2011 Feb 3;470(7332):59-65

1000 Genomes Project Phases

Structural Variation Goals

Pilot Phase (2010/2011)

179 lowcov genomes
2 deep trios
Plus exome sequencing

- Variant catalog of multiple variant types
- Genotypes for some variants (deletions, mobile element insertions)

Phase 1 (2012)

1092 lowcov genomes

- Expanded catalog of deletions
- Integrated haplotypes combining deletions with SNPs / indels

Phases 2 & 3 (2013)

2500 lowcov genomes
500 deep genomes from Complete Genomics

- Expanded variant catalog covering many variant types
- Integrated haplotypes, perhaps combining multiple variant types

1000G Phase 1 – Deletion discovery

Five deletion discovery algorithms were used

BreakDancer	*Read pairs, Washington University*
CNVnator	*Read depth, Yale*
Delly	*Read pairs/depth, EMBL*
Genome STRiP	*Read pairs/depth, Broad Institute*
Pindel	*Split reads, Leiden University*

Three validation methods

Omni 2.5 SNP arrays, probe intensity rank-sum test

Array CGH, 2 x 1M arrays, run on 25 samples

Attempted PCR on 100 sites from each algorithm

Deletion call sets and validation results

Call Set	Sites (post-merging)	Estimated FDR	Estimated Redundancy
Union call set	113,694	unknown	high
Filtered call set	23,594	1.4 – 3.7%	12%
Genotyped call set	14,422	1.4 – 3.7%	low

The union call set is a merge (90% reciprocal overlap) of calls from the 5 discovery methods plus sites with assembled breakpoints from the 1000 Genomes pilot.

The filtered call set was constructed to have a low false discovery rate (< 5%) based on results from the three validation methods.

The genotyped subset was selected from sites where there was sufficient data to obtain accurate genotype likelihoods. 2185 redundant or non-variant sites were removed post-genotyping, yielding an estimate of the redundancy rate in the filtered call set.

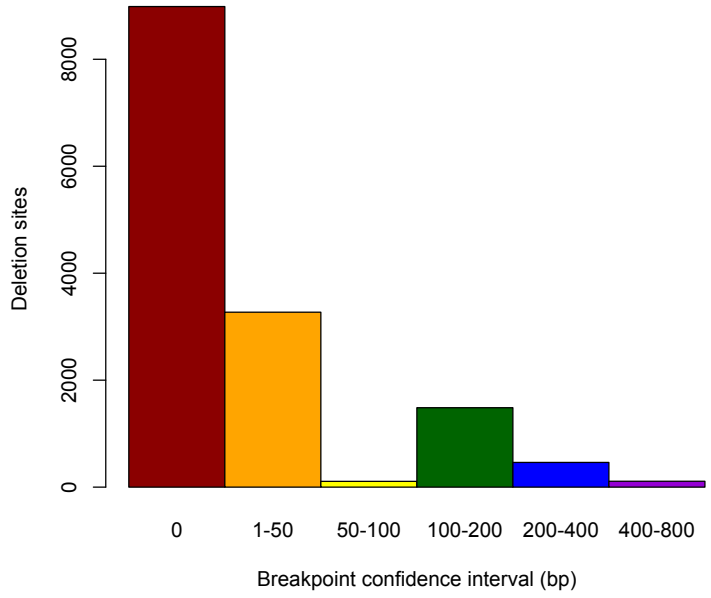
All three call sets and validation data are available as supplementary data files:
http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase1/analysis_results/experimental_validation/sv/

Ascertainment of deletion breakpoints

Breakpoint assembly from consensus of two methods

Tigra_SV assembly + CROSSMATCH alignment (*Ken Chen, U Texas*)

Tigra_SV assembly + AGE alignment (*Alexej Abyzov, Yale*)



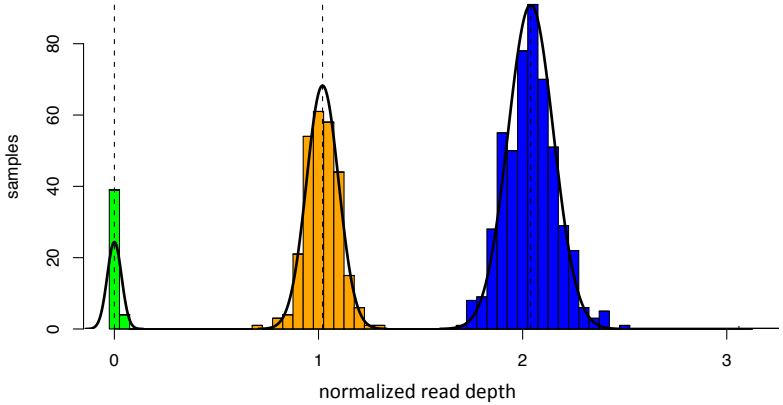
- 63% of the deletions have assembled breakpoints
- 85% have confidence intervals within 50bp

Deletion genotyping

Goal: Accurate genotype likelihoods across all samples

Genome STRiP was used for genotyping

An integrated likelihood framework incorporates evidence from read depth and read pairs; split reads were not used for genotyping in 1000G Phase 1



Sites were selected for genotyping when there was sufficient data to calculate accurate likelihoods. Duplicate sites were removed post- genotyping based on overlap and the calculated genotype likelihoods.

Genotyping accuracy

Sites	Evaluation Data	# Sites Evaluated	HOMREF (Conrad)	HET (Conrad)	HOMALT (Conrad)	OVERALL
14,422	Conrad et al. 80% RO 248 samples	1,092	99.92%	99.01%	99.47%	99.82%

Handsaker, R.E., Korn, J.M., Nemesh, J. & McCarroll, S.A. Discovery and genotyping of genome structural polymorphism by sequencing on a population scale. *Nat Genet* **43**, 269-76 (2011)

Structural Variation in 1000 Genomes

What are the goals?

Create a comprehensive catalog
of human structural polymorphism

Create a reference panel
for imputing structural polymorphisms

Creating integrated haplotypes

Reference panel for imputation

- SNPs, indels, large deletions were integrated and phased using beagle + MaCH on 1092 samples

  Hyun Min Kang, University of Michigan

  Brian Browning, University of Washington

- Integrated call set available in VCF format
  - Main product of 1000 Genomes Phase 1
- Notes for power users
  - Large deletions were treated as point events
  - Large deletions were only genotyped in 946 samples

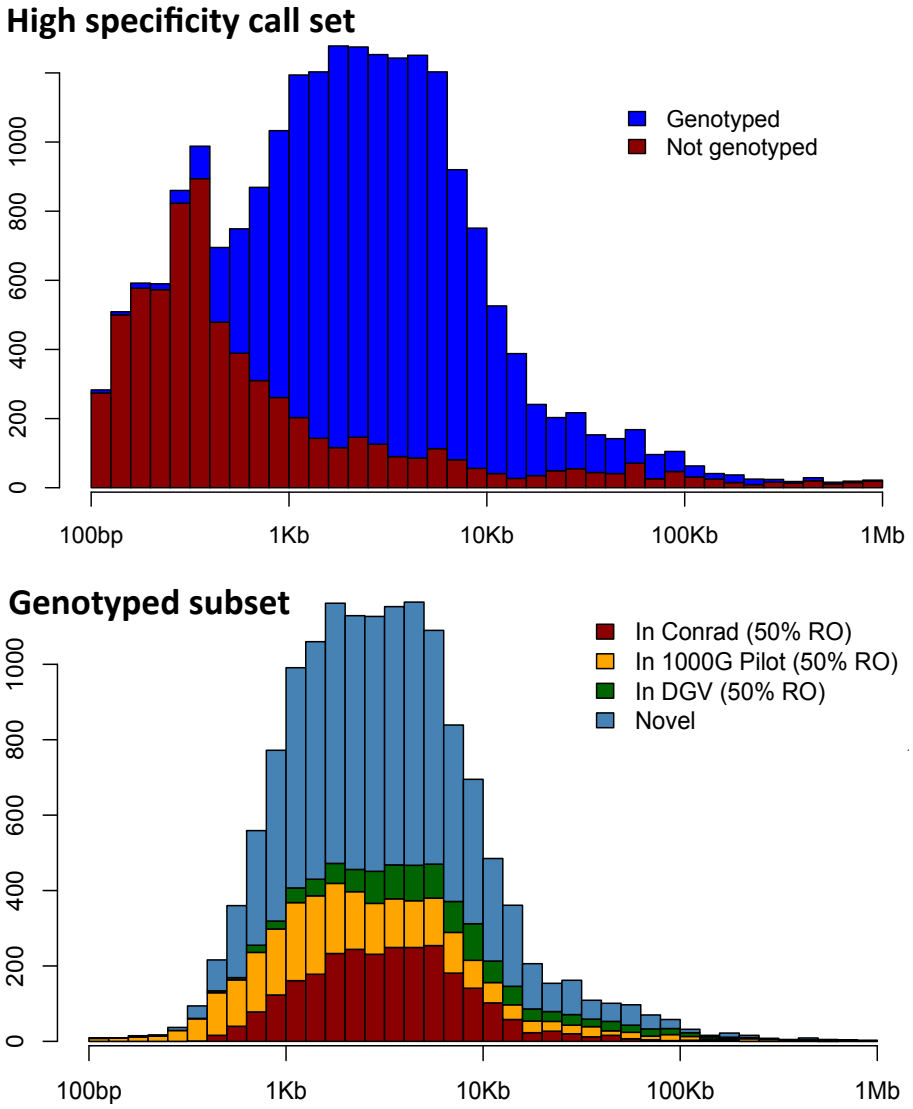
    (genotypes imputed in remaining 146 from nearby SNPs / indels)

Length distribution and novelty

The set of genotyped variants is enriched for longer polymorphisms (median length 2,974).

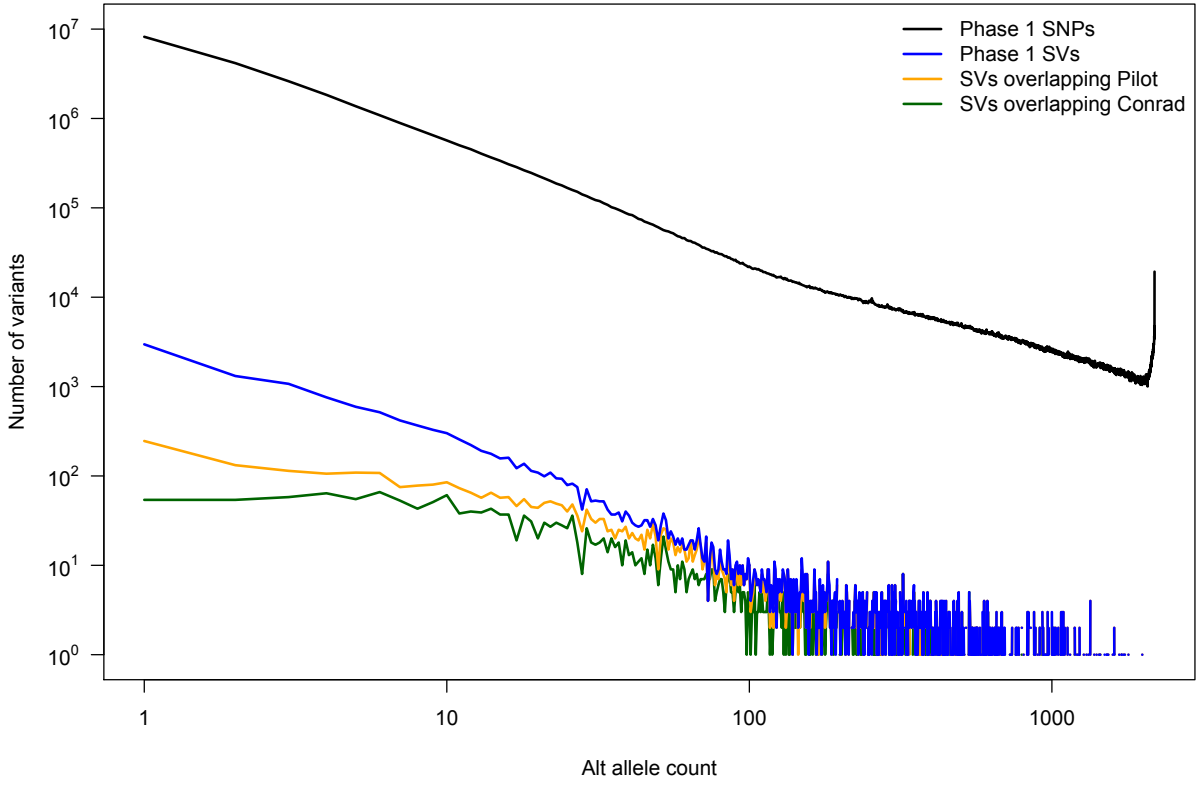
Selection of genotyped sites adds additional QC (e.g. filtering non-variant sites).

Among the genotyped variants, 57% are novel.



Deletion site frequency spectrum

Site frequency spectrum is roughly linear on log-scale plot, matching population genetic expectation. The phase 1 call set is ascertaining more rare variants than 1000G Pilot or older array-based studies (Conrad, 2010).



Variant Catalog - Deletions

Union of raw calls (113,694 sites)

ftp://ftp-trace.ncbi.nih.gov/1000genomes/ftp/phase1/analysis_results/input_call_sets/ALL.wgs.merged_5_del_call_sets_bps.20101123.sv_dels.low_coverage.sites.vcf.gz

High specificity subset (23,594 sites, estimated FDR < 5%)

- Same file, use all records with FILTER != "NONVAL"

Genotyped subset (site list only, 14,422 sites)

- Same file, use records with FILTER == "PASS"
- Other filter values:

NONAUTX	Site not on autosome or chrX
NONGT	Insufficient data to genotype site
DUPLICATE	Site determined to be duplicate post-genotyping
NONVARIANT	Site determined to be non-variant post-genotyping

Genotypes are in the integrated call set

ftp://ftp-trace.ncbi.nih.gov/1000genomes/ftp/phase1/analysis_results/integrated_call_sets/

Variant Call Format
Conventions used for large deletions in 1000G

VCF file format specification: http://vcftools.sourceforge.net/specs.html

```
CHROM POS      ID       REF      ALT   QUAL FILTER INFO

1     2371985  DEL_455  CGCTGG…  C       .    .     END=2374312;HOMLEN=5;HOMSEQ=GCTGG;CIPOS=-25,9;CIEND=-10,25
1     2918690  DEL_833  G        <DEL>   .    .     END=2919922;CIPOS=-18,19;CIEND=-17,33
```

Precise variants	*Example*
REF gives the entire reference allele (may be long)	CGCTGGCCT…
ALT gives the entire alternate allele (usually short)	C
HOMSEQ identical sequence at breakpoint (if any)	GCTGG
HOMLEN length of HOMSEQ	5
CIPOS confidence interval on POS (before bkpt assy)	-25,9
CIEND confidence interval on END (before bkpt assy)	-10,25

Imprecise variants	
REF is a single base (at POS)	G
ALT will be <DEL>	<DEL>
END gives best-estimate of end coordinate	2919922
CIPOS confidence interval on POS	-18,19
CIEND confidence interval on END	-17,33

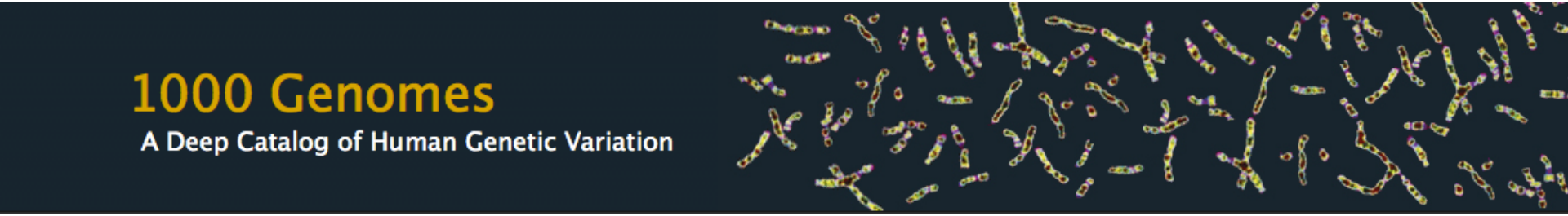
Summary

- 1000 Genomes Phase 1
  - High quality deletion call set
  - Low false discovery rate; high genotype accuracy
  - Integrated reference panel for imputation
- Goals for 1000 Genomes Phases 2 & 3
  - Build on and expand variant catalog
  - Larger reference panel for large deletions
  - Imputation resources for other variant types

Software Links

- 1000 Genomes Project data provides a rich data set for developing and assessing detected structural variants
- 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/software/genomestrip/genome-strip
  - Includes the SVVariantAnnotator, which can utilize microarray intensities to measure the efficacy of SV discovery algorithms

Acknowledgements



1000 Genomes Structural Variation Analysis Group

WashU – Asif Chinwalla

WT Sanger Institute – Klaudia Walter, Manuela Zanda, Sarah Lindsay, Thomas Keane

Yale – Alexej Abyzov, Mark Gerstein, Jasmine Mu, Ekta Khurana

EMBL – Adrian Stuetz, Tobias Rausch, Andreas Schlattl, Markus Fritz

Univ of Washington – Can Alkan, Peter Sudmant, Art Ko, Fereydoun Hormozdiari

Oxford – Zamin Iqbal, Gil McVean

LSU – Miriam Konkel, Jerilyn Walker, Mark Batzer

MSSM – Seungtai Yoon, Vlad Makarov, Jayon Lihm

AECOM – Kenny Ye

BC – Chip Stewart, Gabor Marth, Deniz Kural, Michael Stromberg, Alistair Ward, Jiantao Wu

Broad Institute – Josh Korn, Jim Nemesh, Bob Handsaker, Steve McCarroll

HMS – Mindy Shi, Marcin von Grotthuss

UMich – Ryan Mills

UCSD – Jonathan Sebat, Doug Greer

UT / MC Anderson – Ken Chen

Co-chairs: Evan Eichler, Jan Korbel, Charles Lee
Matt Hurles (emeritus)