Discovering and genotyping deletions using Genome STRiP

Bob Handsaker
Medical and Population Genetics, Broad Institute
Department of Genetics, Harvard Medical School

NextGen Sequencing Workshop
February 17, 2011

Genome STRucture in Populations

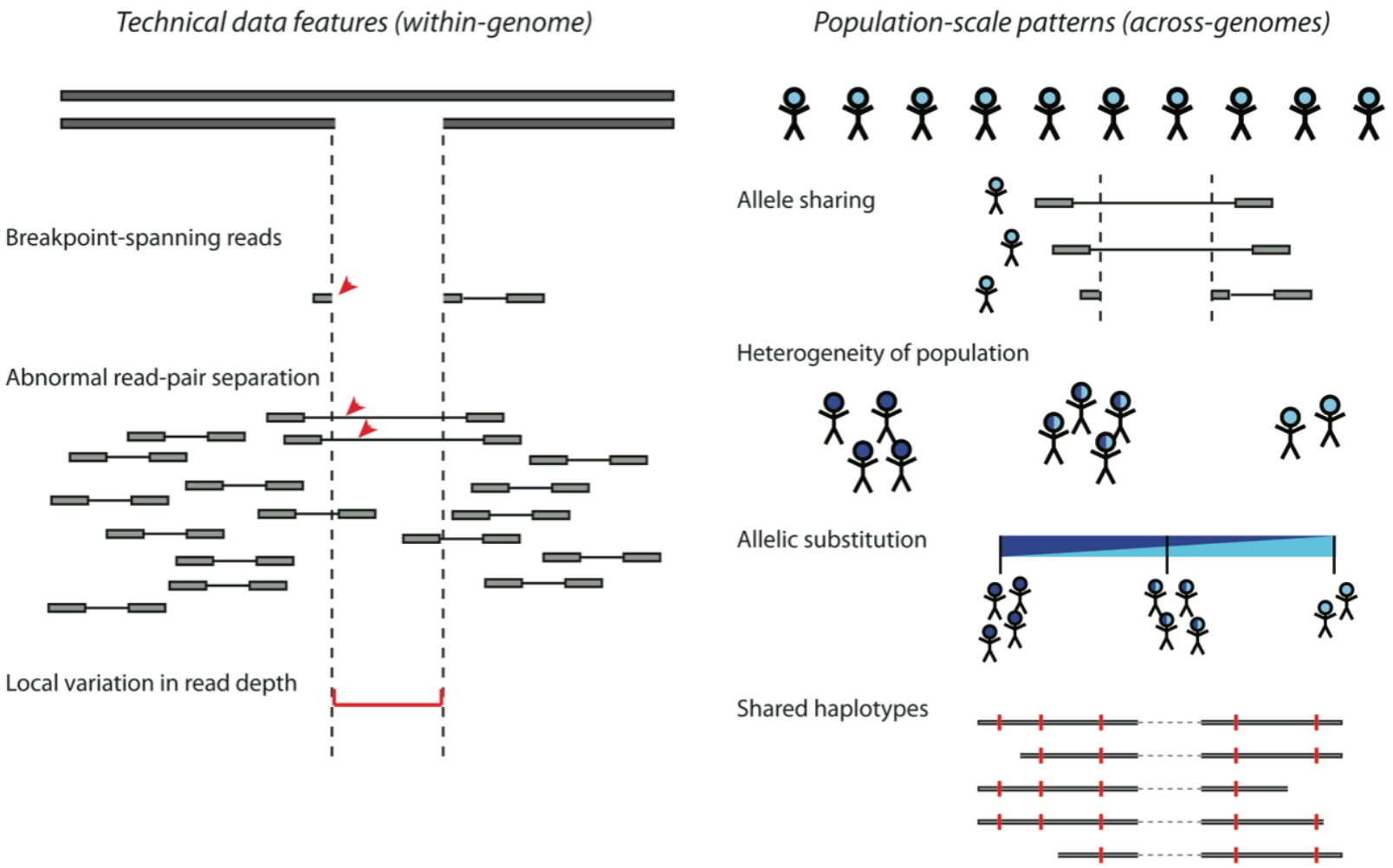
What is it?

Method used for discovering and genotyping deletions (100bp – 1Mb) in the 1000 Genomes Project pilot

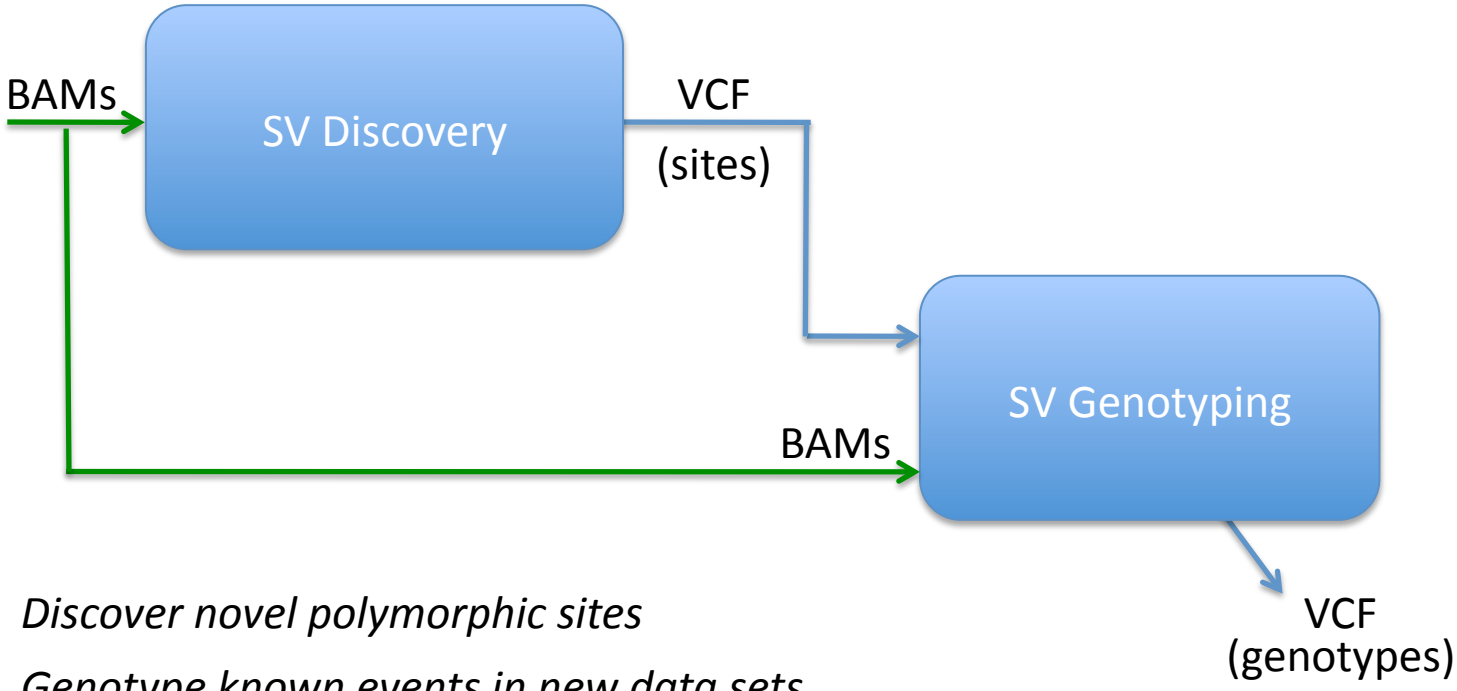
- 168 samples @ 2x – 8x coverage, Illumina paired and single end
- Called 7,015 deletions with estimated FDR of 3.7%
- Best overall sensitivity of all algorithms evaluated using low coverage sequencing data
- Genotyped 13,826 deletions with estimated overall accuracy exceeding 99%

Handsaker, et al., Nature Genetics, 2011

Discovery and genotyping are enhanced by combining technical and population-level features of a data set



Discovery and genotyping are two distinct modules in Genome STRiP

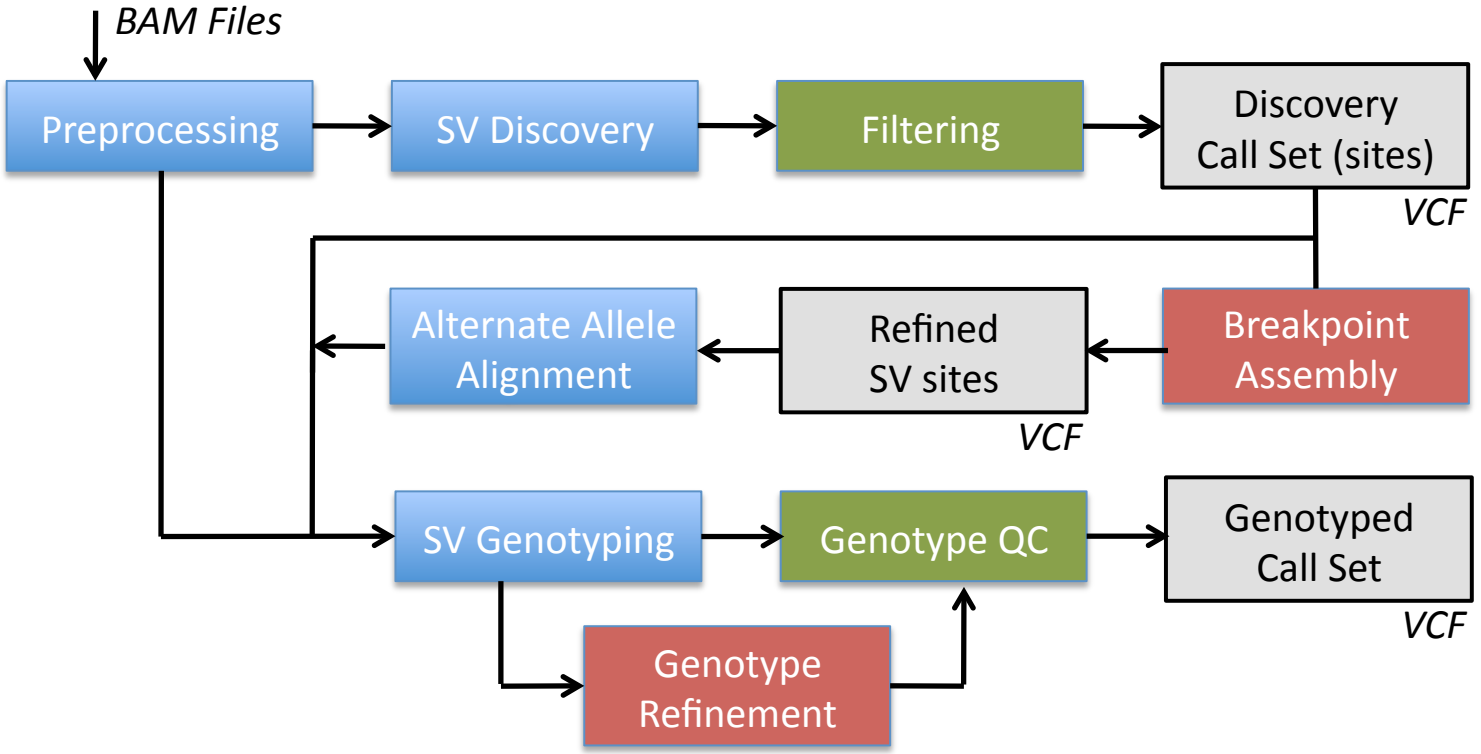


Discover novel polymorphic sites

Genotype known events in new data sets

Genotype call sets from multiple discovery methods

Detailed processing pipeline



Genome STRiP module

Third party tool (e.g. TIGRA or velvet for assembly, BEAGLE for genotype refinement)

Manual process currently (VQSR under evaluation for discovery filtering)

Queue script for preprocessing

```
java –Xmx4g org.broadinstitute.sting.queue.QCommandLine
    -cp SVToolkit.jar:GenomeAnalysisTK.jar:Queue.jar
    -S ${SV_DIR}/qscript/SVPreprocess.q
    -S ${SV_DIR}/qscript/SVQScript.q
    -md metadata
    -configFile ${SV_DIR}/conf/genstrip_parameters.txt
    -tempDir /high/performance/temp
    -gatk ${SV_DIR}/lib/gatk/GenomeAnalysisTK.jar
    -R /humgen/1kg/reference/human_g1k_v37.fasta
    -genomeMaskFile human_g1k_v37.mask.36.fasta
    -I bam1.bam -I bam2.bam
```

Output directory for metadata

Alignability mask

Inputs: BAM files, reference sequence, alignability mask

Outputs: aggregate statistics on data set (insert sizes, coverage depth, etc.)

Example: metadata/isd.hist.stats.dat contains insert size distribution statistics

SAMPLE	LIBRARY	READGROUP	NPAIRS	MEDIAN	RSD
HG00098	g1k-sc-HG00098-1	NA	23410435	456	42.06
HG00100	g1k-sc-HG00100-A	NA	36251941	379	37.03

Alignability masks

Align k-mers centered on each base position back to reference using bwa
Mask is 1 if k-mer aligns multiple places, 0 if k-mer aligns uniquely

Mask is a function of the reference sequence and K
If you have multiple read lengths, use the smallest
Pre-computed masks for common genomes can be downloaded
ftp://ftp.broadinstitute.org/pub/svtoolkit/svmasks

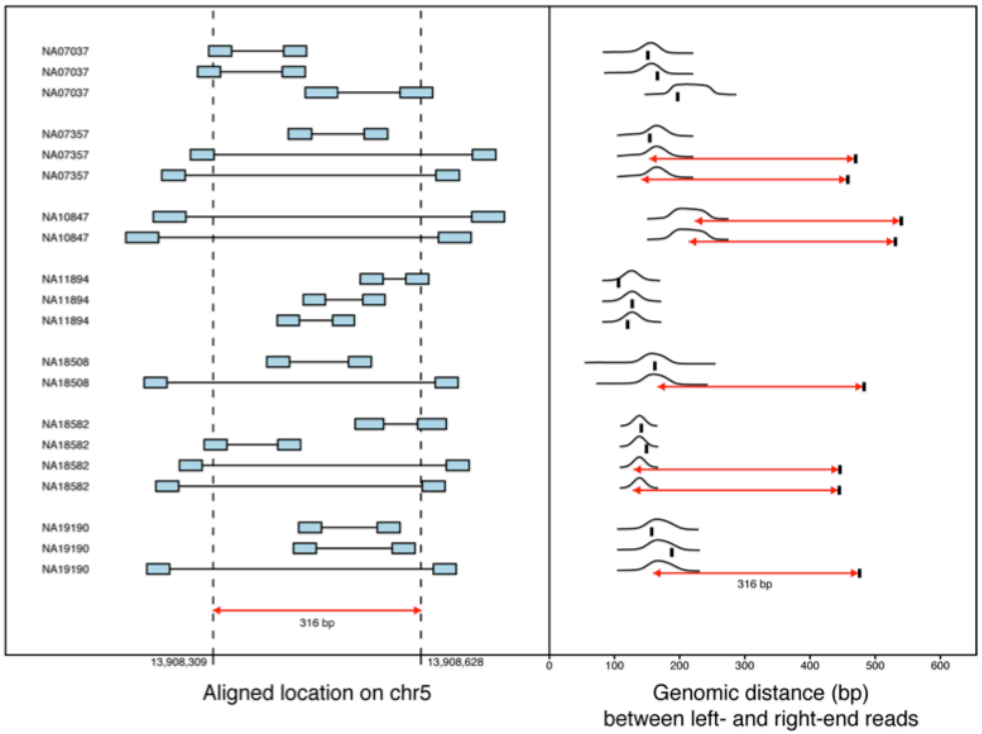
Utility exists to build your own masks for different genomes / read lengths

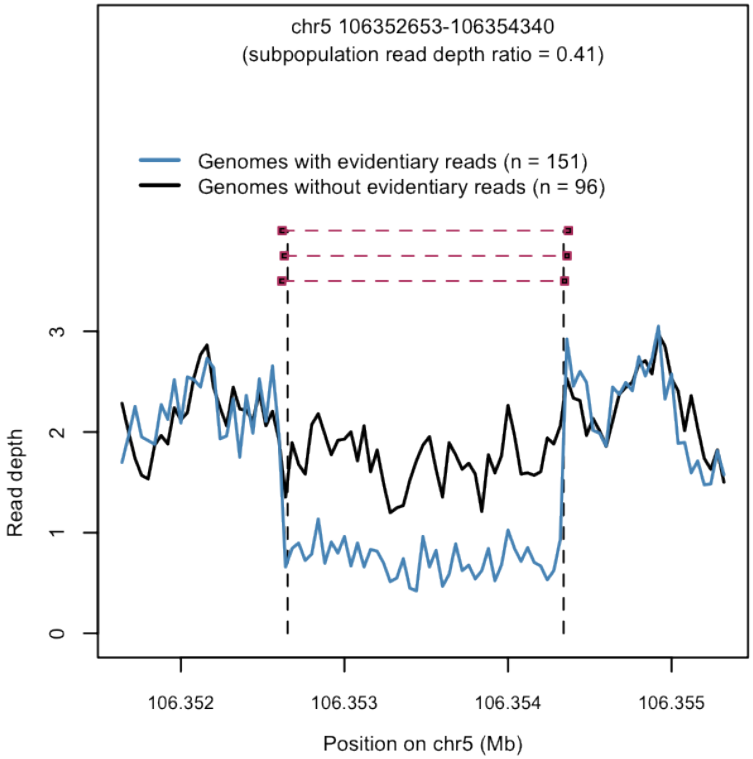
```
java –Xmx4g –cp SVToolkit.jar:GenomeAnalysisTK.jar
  org.broadinstitute.sv.apps.ComputeGenomeMask
  -R /humgen/1kg/reference/human_g1k_v37.fasta
  -O human_g1k_v37.mask.101.fasta
  -readLength 101
```

This can be parallelized – see documentation on the web site

SV Discovery

Deletion discovery integrates diverse features of the sequencing data, including aberrantly spaced read pairs, differential read depth, and distribution of evidence across multiple samples.





Queue script to run deletion discovery

Note: some detail elided

```
java -Xmx4g org.broadinstitute.sting.queue.QCommandLine
    ...
    -S ${SV_DIR}/qscript/SVDiscovery.q
    -R /humgen/1kg/reference/human_g1k_v37.fasta
    -genomeMaskFile human_g1k_v37.mask.36.fasta
    -genderMapFile sample_gender.map
    -runDirectory run1
    -minimumSize 100
    -maximumSize 1000000
    -I bam1.bam -I bam2.bam
    -O run1/deletions.discovery.vcf
    -jobProject 1KG
    -jobQueue week
    -jobLogDir run1/logs
    -windowSize 10000000
    -windowPadding 100000
```

Directory for intermediate files

Can parallelize based on event size

Output VCF file

Optional arguments for parallelization on a compute cluster

VCF with ***raw calls*** and metrics

List of many attributes

```
CHR  POS     ID     REF ALT   QUAL FILTER INFO
1    55461   DEL_3  G   <DEL> .    .      END=56280,SVLEN=-845,…
1    86007   DEL_4  T   <DEL> .    .      END=92142,SVLEN=-6113,…
```

Most raw calls are not true structural variations

Filtering raw calls

Most raw calls are false discoveries

Current practice: Use GATK VariantEval with user-defined filters

requires significant expertise (and potentially weeks)

The default SVDiscovery script uses the 1000 Genomes pilot filters which may not be optimal for other data sets:

```
java –Xmx4g –jar GenomeAnalysisTK.jar
  -T VariantFiltration
  -B:variant,VCF deletions.discovery.unfiltered.vcf
  -o deletions.discovery.vcf
  -R /humgen/1kg/reference/human_g1k_v37.fasta
  -filterName COVERAGE -filter GSDEPTHCALLTHRESHOLD >= 1.0
  -filterName COHERENCE -filter GSCOHPVALUE <= 0.01
  -filterName DEPTHPVAL -filter GSDEPTHPVALUE >= 0.01
  -filterName DEPTH -filter GSDEPTHRATIO > 0.8 ||
                           (GSDEPTHRATIO > 0.63 && GSMEMBPVALUE >= 0.01)
```

Future goal: Automate using GATK Variant Quality Score Recalibrator (VQSR)

- Train on highly confident known SVs to predict novel SVs
- Eliminates the need for manually setting filters for each data set
- Currently under evaluation for use with Genome STRiP

Evaluating discovery output

Strategies for evaluating deletion calls

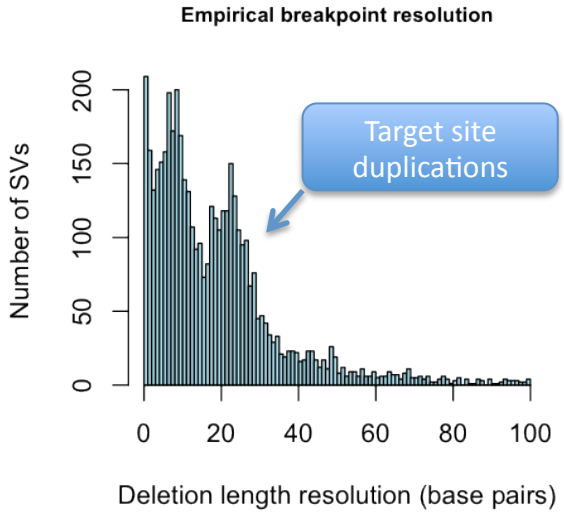
- Compare to previously ascertained data sets
  - 1000 Genomes pilot (22000 deletion events)
- Successful breakpoint assembly
- Lack of heterozygous SNPs
  - Individuals carrying deletions should be depleted for heterozygous SNP calls
- Utilize other data sets where possible
  - Array intensity data, array-based SNP data
- Genotyping QC measures
  - Call rate
  - Hardy-Weinberg equilibrium
- Pool data by genotype class and look with IGV

Breakpoint assembly

To determine precise breakpoints, use a third party tool (e.g. TIGRA, velvet)

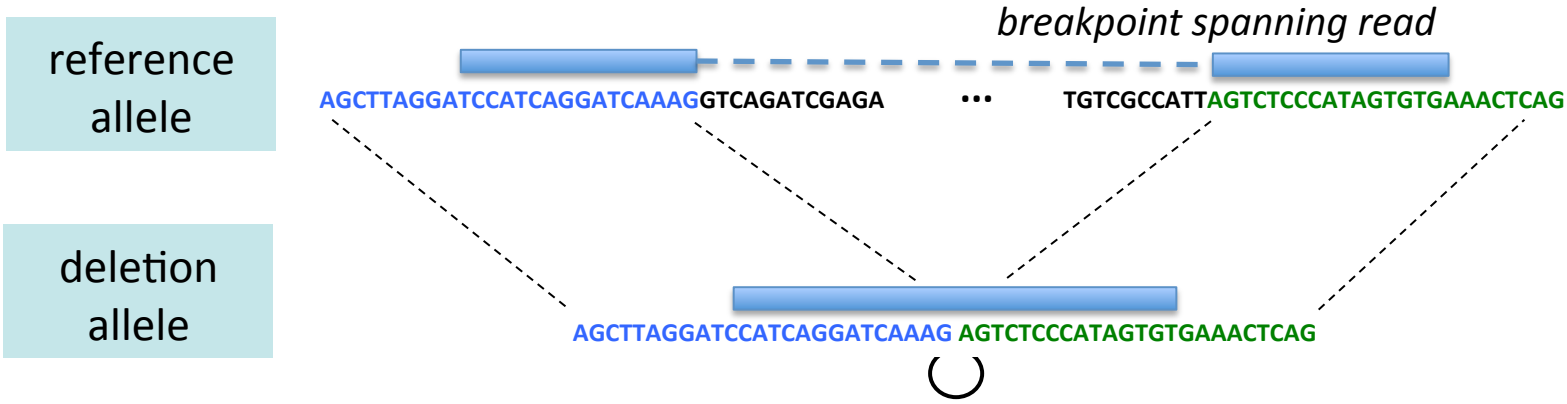
Genome STRiP generates calls with approximate coordinates (typically 10-20 bp resolution)

To utilize breakpoint-spanning reads in genotyping, you need exact breakpoint coordinates.



In the 1000 Genome pilot, Ken Chen (WashU) used the TIGRA assembler [L Chen] to assemble breakpoints for about half of the called deletions.
Web site: http://sourceforge.net/projects/tigrasv

Alternate allele alignment



When you have precise alleles, you can use breakpoint-spanning reads in genotyping. There are three sources:

Source	How handled
"in-place" reads aligned at the breakpoint	Realigned on-the-fly to alt allele during genotyping
unmapped mates where mate is aligned nearby	Realigned on-the-fly during genotyping
completely unmapped reads	Alternate allele aligner

Queue script for alt allele alignment

Runs BWA internally (as a library) to generate alt allele alignments

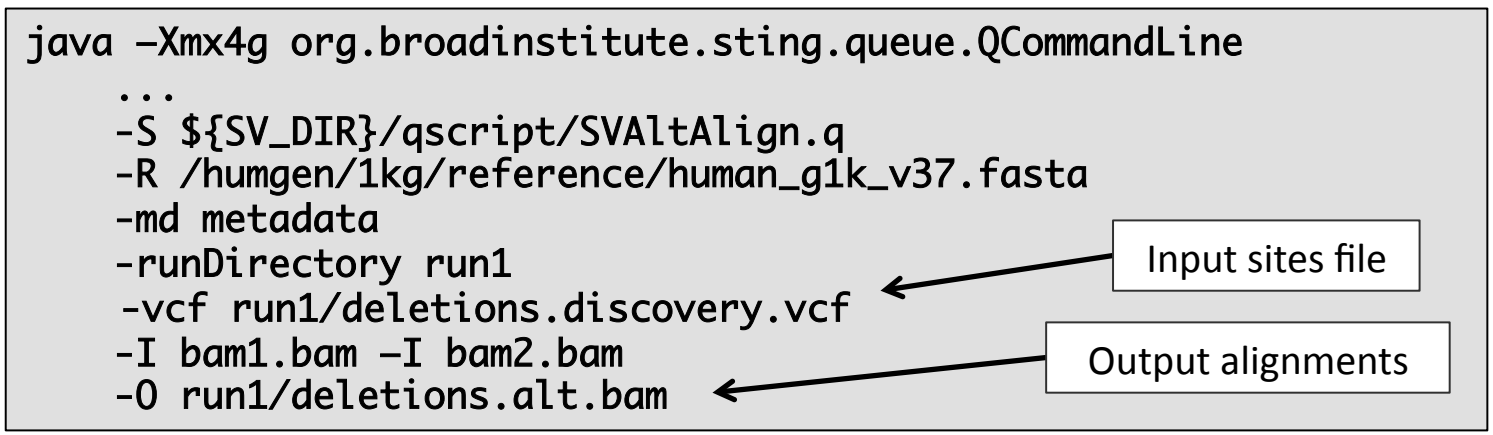
Inputs:

- VCF file containing SVs with exact alleles
- BAM files containing unmapped reads

Outputs:

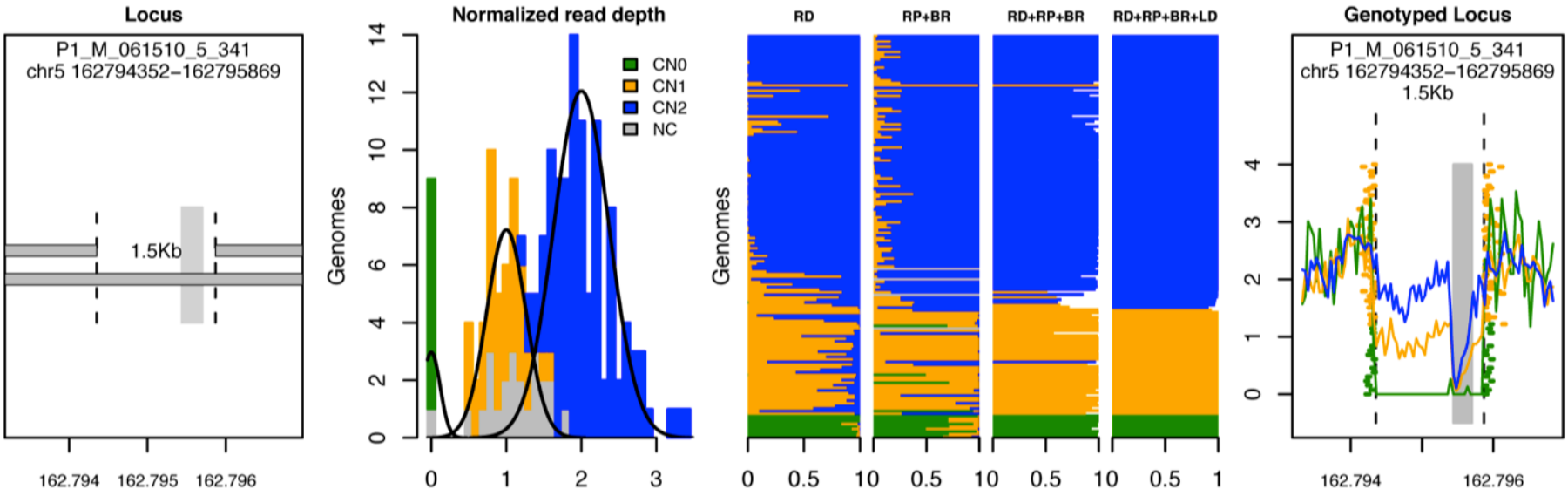
- BAM file containing alignments to alternate alleles

```
java –Xmx4g org.broadinstitute.sting.queue.QCommandLine
    ...
    -S ${SV_DIR}/qscript/SVAltAlign.q
    -R /humgen/1kg/reference/human_g1k_v37.fasta
    -md metadata
    -runDirectory run1
    -vcf run1/deletions.discovery.vcf
    -I bam1.bam –I bam2.bam
    -O run1/deletions.alt.bam
```



SV Genotyping

Genome STRiP integrated information from read depth, discordant read pairs and breakpoint spanning reads to genotype deletions.



Support for genotyping other types of variants (e.g. duplications) is under development.

Queue script for SV genotyping

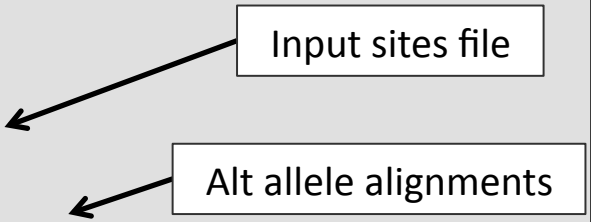
Inputs:

- VCF file containing polymorphic structural variation sites
- Aligned BAM files
- Optional pre-computed alternate allele alignments

Outputs:

- VCF file containing genotype likelihoods for every sample

```
java –Xmx4g org.broadinstitute.sting.queue.QCommandLine
    ...
    -S ${SV_DIR}/qscript/Genotyper.q
    -R /humgen/1kg/reference/human_g1k_v37.fasta
    -genomeMaskFile human_g1k_v37.mask.101.fasta
    -genderMapFile sample_gender.map
    -md metadata
    -runDirectory run1
    -vcf run1/deletions.discovery.vcf
    -I bam1.bam –I bam2.bam
    -altAlignments run1/deletions.alt.bam
    -O run1/deletions.genotypes.vcf
```



Genotype refinement and QC

1000 Genomes pilot used BEAGLE for genotype refinement

Exploits LD between deletions and Hapmap SNPs

There is currently not an automated module in Genome STRiP to perform genotype refinement using LD.

Genotype QC

Should perform typical genotype QC (call rate, Hardy-Weinberg equilibrium). Not all sites are genotypable by Genome STRiP.

For 1000 Genomes pilot, we used two criteria:

- At least 50% of the samples called with 95% confidence.
- Genotypes in HWE ($p > 0.01$) in each of the three populations (CEU, YRI, CHB+JPT)
- *Most sites were genotypable unless they were short, repetitive (< 200bp of unique sequence) and had no precise breakpoints.*

Usage Scenarios

De novo deletion discovery and genotyping
Genotyping known events in new samples

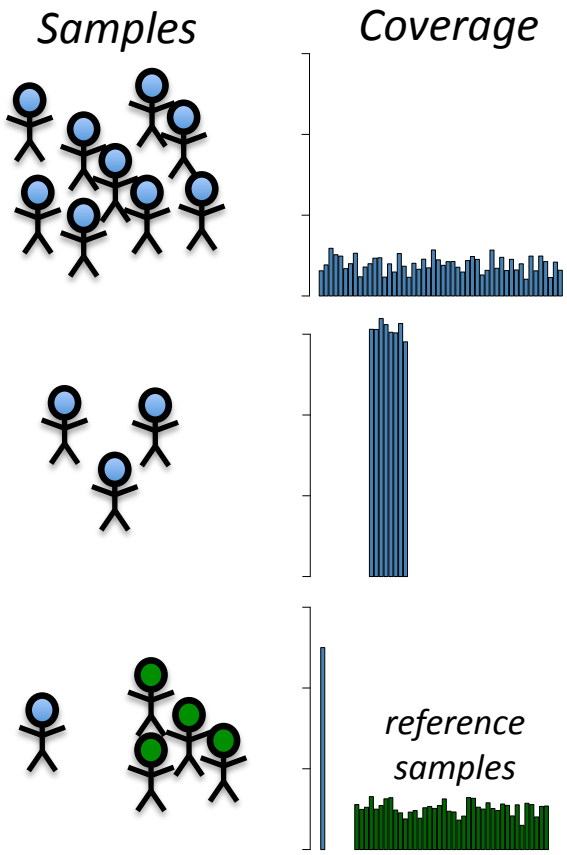
<u>Whole Genome Population Sequencing</u>

Need 20-30+ samples for good results
Low or high coverage, can be variable

<u>Future Goals</u>

Targeted resequencing

Deep coverage single individual
using 1000G reference samples
as background population



Resource requirements

Performance on some sample analyses
All steps are highly parallel, designed for compute farms

Algorithm Step	Data Set Size	Run time (CPU days)
Preprocessing	672x (168 x 4x) 2.3Tb	11
Discovery	672x (168 x 4x) 2.3Tb	5 *
Alt allele alignment	672x (168 x 4x) 2.3Tb	4
Genotyping	22,000 events	4
Preprocessing	4000x (1000 x 4x) 17Tb	86
Discovery	4000x (1000 x 4x) 17Tb	150
Preprocessing	716x (179 x 4x)	13
Alt allele alignment	716x (179 x 4x)	5
Genotyping	22,000 events	7

** Older version, not representative*

Availability

Available now for experienced GATK users

Web site

- Documentation and pre-compiled releases
- http://www.broadinstitute.org/gsa/wiki/index.php/Genome_STRiP
- Includes installation test that performs a 5-minute analysis
- *Source code release available soon*

Support mailing list

- http://sourceforge.net/projects/svtoolkit/support

Summary

- Genome STRiP performed well in the 1000 Genomes pilot on deletion discovery and genotyping
- Now available for general use by experienced GATK users
- Usage scenarios

  De novo deletion discovery and genotyping in sequencing-based GWAS

  Genotyping known deletions (e.g. from 1000 Genomes) in new samples

Acknowledgements

Broad / HMS

Josh Korn
Jim Nemesh
Nick Patterson
Jared Maguire
Steve McCarroll

GATK / Queue

Khalid Shakir
Aaron McKenna
Matt Hanna
Kiran Garimella
Eric Banks
Mark DePristo

1000 Genomes Structural Variation Group

Ryan Mills
Klaudia Walter
Chip Stewart
Ken Chen
Can Alkan
Alex Abyzov
Chris Yoon
Kai Ye
Yujun Zhang
Zhengdong Zhang
Don Conrad
Jeff Kidd
Zam Iqbal
Mindy Shi
Kenny Ye
Ekta Khurana
Jasmine Mu
Michael Stromberg

Matt Hurles
Jan Korbel
Evan Eichler
Jonathan Sebat
Charles Lee
Mark Gerstein