Introduction to Genome STRiP for discovery and genotyping of deletions

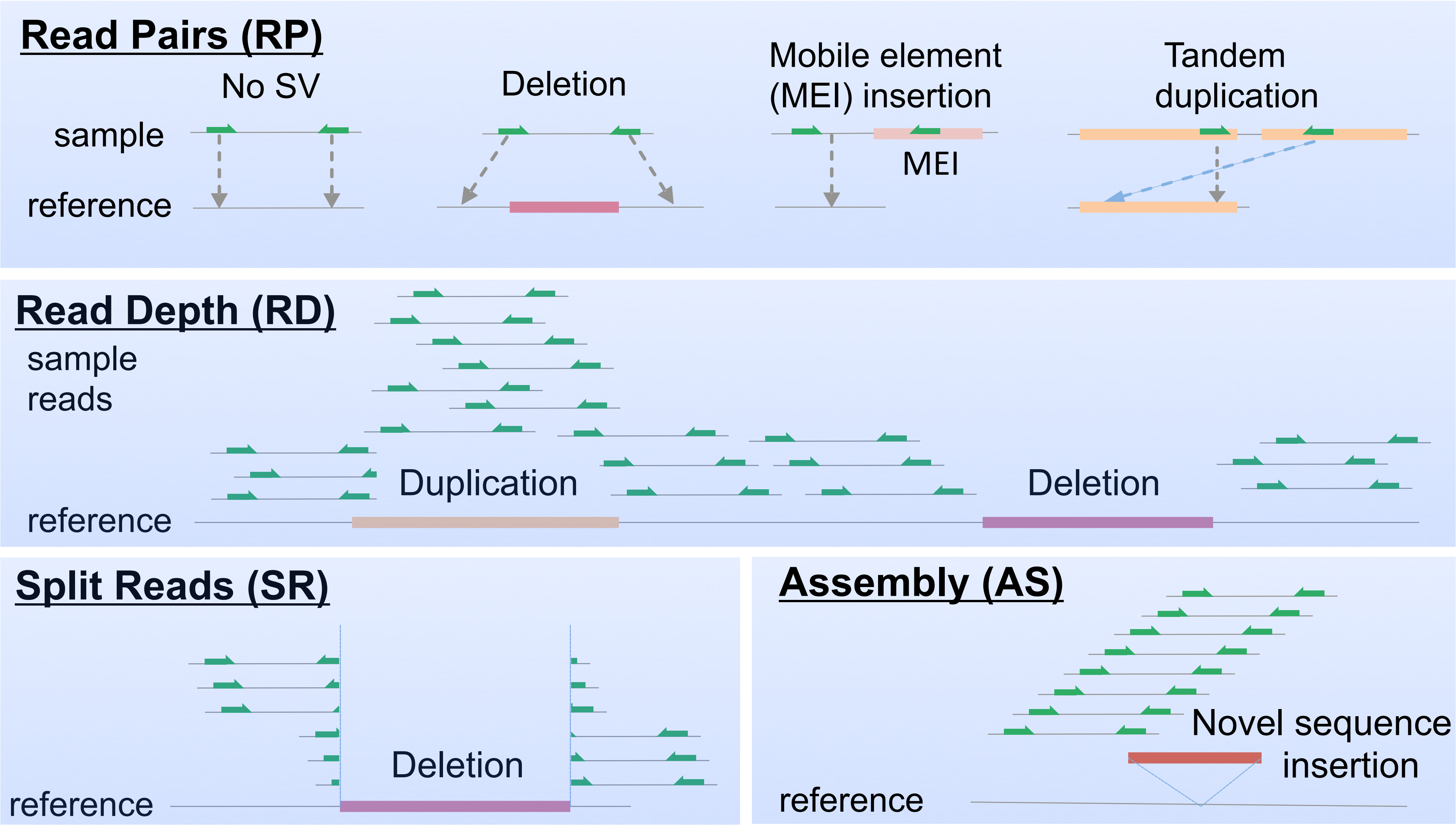
Bob Handsaker
Medical and Population Genetics, Broad Institute
Department of Genetics, Harvard Medical School

December 4, 2012

Outline

- Overview of structural variation calling
- Genome STRiP processing pipelines
- Techniques for quality control
- Software and support

Ascertaining large variants from small reads



Read Pairs (RP)

No SV

sample

reference

Deletion

Mobile element (MEI) insertion

MEI

Tandem duplication

Read Depth (RD)

sample reads

Duplication

Deletion

reference

Split Reads (SR)

Deletion

reference

Assembly (AS)

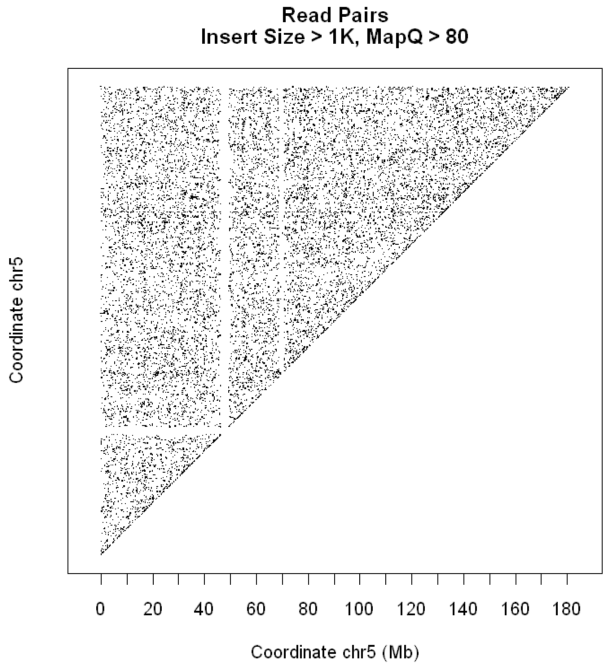
Novel sequence insertion

reference

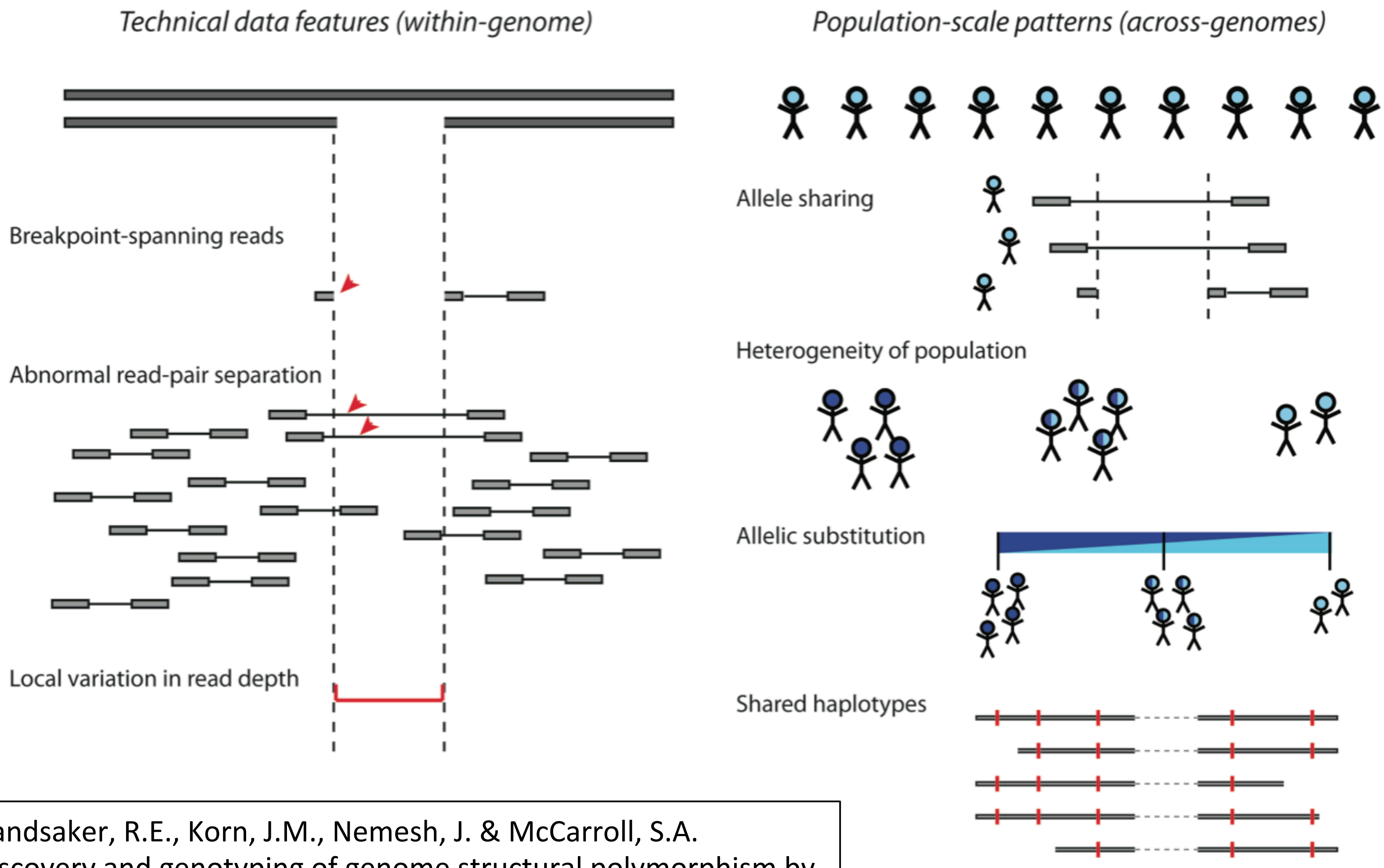
Graphics courtesy of Jan Korbel, Ryan Mills

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
- Deeply sequenced genomes
  - Increased depth can help, but methodology is more important



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


Technical data features (within-genome)
Population-scale patterns (across-genomes)
Breakpoint-spanning reads
Abnormal read-pair separation
Local variation in read depth
Allele sharing
Heterogeneity of population
Allelic substitution
Shared haplotypes

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)

Genome STRucture in Populations

What is it?

Methods for discovering and genotyping large deletions from sequencing data

Our Focus

Whole genome sequencing (shallow or deep)
Using populations to inform calls in individuals
Germline/somatic DNA (not tumor/normal)

Genome STRiP in 1000 Genomes Project

Discovery specificity

Consistently low false discovery rate (1.5% - 4.2%)

Lowest FDR in 1000G pilot and in phase 1

Discovery sensitivity

Best overall sensitivity on low coverage sequencing data (Mills, 2011)

Contributed over 80% of phase 1 deletion call set

Genotyping accuracy

Genotyping algorithm of choice for pilot and phase 1

Genotyping accuracy, 1000 Genomes Phase 1

Genotyped Sites	Evaluation Data	# Sites Evaluated	HOMREF (Conrad)	HET (Conrad)	HOMALT (Conrad)	OVERALL
14,422	Conrad 2010 80% RO 248 samples	1,092	99.92%	99.01%	99.47%	99.82%

Other large projects using Genome STRiP

Genome of the Netherlands (GoNL) *University of Groningen*
250 whole genomes in trios at 12x coverage (9,000x)

GoT2D (Type 2 Diabetes) *Oxford University*
2800 cases/controls at 4x coverage (11,000x)

UK10K Cohorts Project *Sanger Institute*
2453 individuals (so far, 4000 planned) at 6x coverage (14,000x)

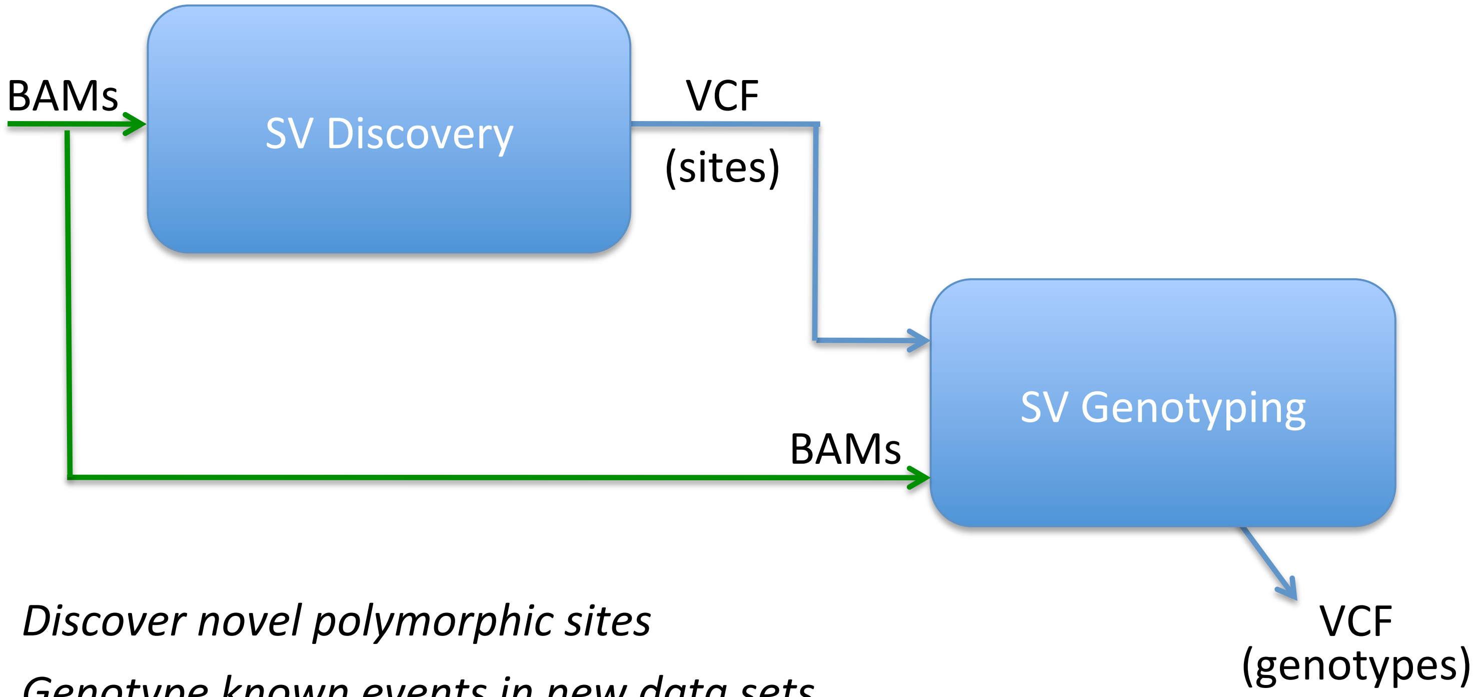
PROCESSING PIPELINE

Required inputs

Processing phases

- Preprocessing, Discovery, Genotyping

Discovery and genotyping are two distinct modules in Genome STRiP

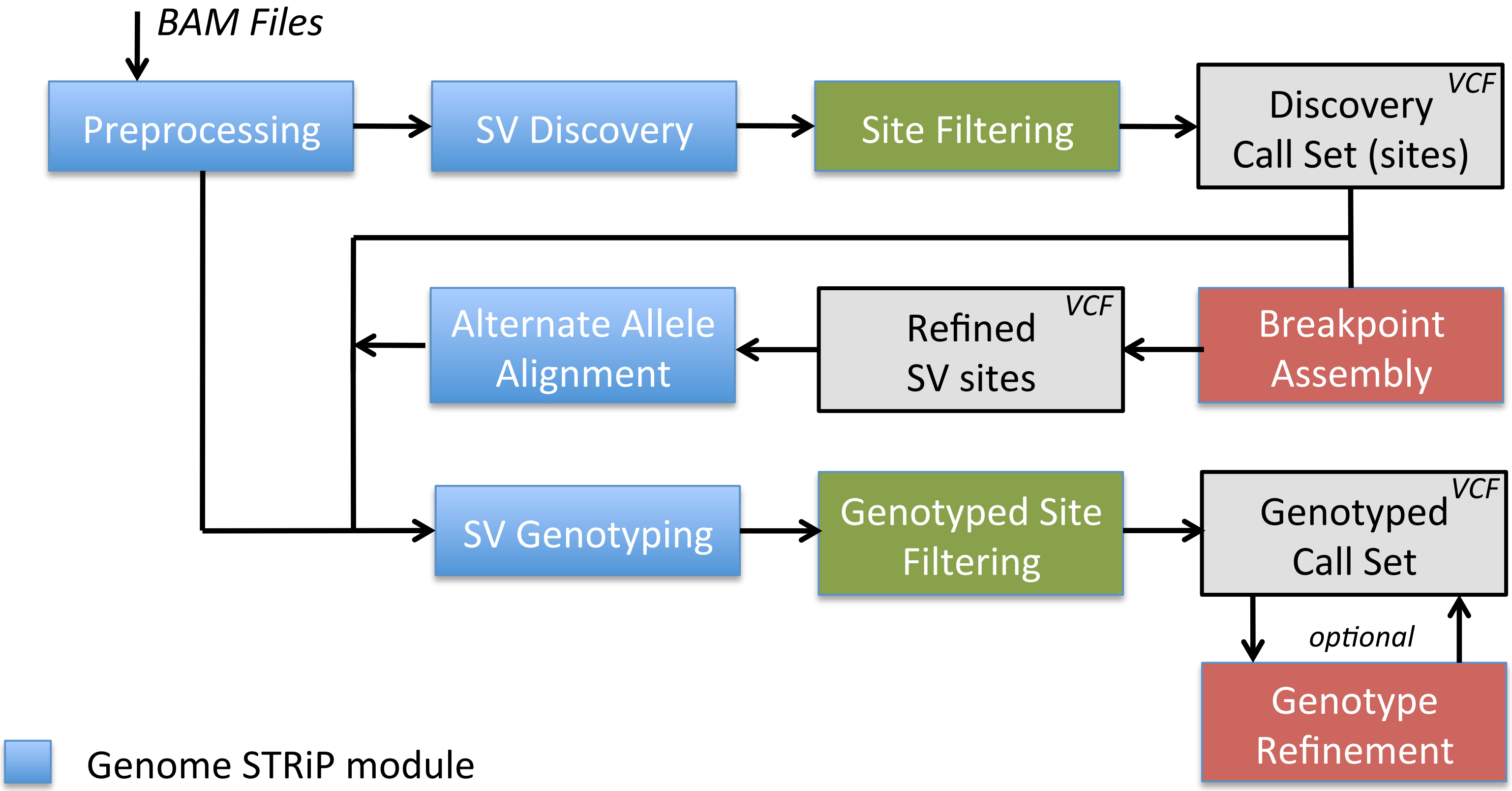


Discover novel polymorphic sites

Genotype known events in new data sets

Genotype call sets from multiple discovery methods

Detailed processing pipeline



🟦 Genome STRiP module

🟩 Requires user intervention, defaults or best-practices are available

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

⬜ VCF data file

What inputs are needed to run Genome STRiP?

- "Analysis-ready" BAM files
  - Whole genome sequencing
  - Aligned, sorted, indexed, duplicates marked or removed
- Reference sequence
  - Indexed fasta file, must match *exactly* the reference used for alignment
- Alignability mask
  - Indicates which reference positions are uniquely alignable
  - Must be based on the same reference you are using
  - Commonly used masks are available for download
- CN2 mask
  - Flags regions unlikely to be copy-number polymorphic, used for estimating GC-bias
  - CN2 masks for common reference sequences are available for download
- Ploidy map
  - Required to process sex chromosomes
  - Indicates expected ploidy of positions on the reference, stratified by gender
- Gender map
  - Gender of each sample, required to call on sex chromosomes

Analysis-ready BAM files

- Reads aligned to reference sequence
- Sorted by coordinate and indexed
- MarkDuplicates is essential
- Indel realignment does not matter (with/without is ok)
- Key headers and tags must be present and consistent
  - Read group (RG tag), e.g. Illumina lane
  - Library (LB tag)
  - Sample (SM tag)
  - Platform (PL tag)
- GATK ReduceReads compression is not supported

Alignability Mask

- What is it?
  - An alignability mask indicates all sites on the reference that are uniquely alignable by a single, error-free read of length *k*
  - Generated by aligning *k*-mers centered on each base position back to the reference using bwa, test if *k*-mer aligns uniquely
  - Function of reference sequence and *k*
  - If you have multiple read lengths in your data, use smallest as *k*
- Where do I get it?
  - Mask files available for download (hg19, 1000G b36/b37)

    ftp://ftp.broadinstitute.org/pub/svtoolkit/svmasks/
- Building your own
  - See documentation for ComputeGenomeMask
  - Can be parallelized for scalability
  - Mask format is currently an indexed fasta file, but subject to change

CN2 Mask

- What is it?
  - Indicates sites on the reference unlikely to be copy-number polymorphic in most individuals
  - We use this when measuring GC-bias in read depth, following an approach similar to that described in Sudmant *et al.*, Science 2010
  - Excludes chrX, chrY, chrM, all unplaced contigs
  - Excludes regions with 200bp of UCSC-annotated repeats, segmental duplications or copy number variants from DGV
- Where do I get it?
  - Versions available for 1000G b36, b37

    ftp://ftp.broadinstitute.org/pub/svtoolkit/cn2masks/
  - Format is indexed fasta file, with 0 or 1 for each position
  - Bed files are also provided for convenience (or viewing), but not used during processing
- Building your own
  - No tools provided to build your own
  - Liftover might be an option for human sequence

Ploidy Map

- What is it?
  - Simple text file of expected ploidy on reference by gender
  - Used in newer versions of Genome STRiP
  - Not strictly necessary when processing autosome only (code generally assumes ploidy 2 if missing), but should be supplied
  - Beware: incorrect results on sex chromosomes if ploidy file is missing, may lead to incorrect QC statistics
- Where do I get it?
  - Example available for 1000G b37

    ftp://ftp.broadinstitute.org/pub/svtoolkit/ploidymaps/
- Building your own
  - Simple text file, whitespace delimited
  - Columns: chrom, start, end, gender, ploidy
  - Lines are matched in order, asterisks are wildcards

Example (1000G b37):

```
X  2699521  154931043  F  2
X  2699521  154931043  M  1
Y        1   59373566  F  0
Y        1   59373566  M  1
*        *          *  *  2
```

Gender Map

- What is it?
  - Lists the gender of each sample in your dataset
  - Genome STRiP does not attempt to infer the gender of samples
- Where to I get it?
  - You have to generate it
- File format
  - Tab delimited text file, no header
  - Columns: sample ID, gender
  - Gender can be M/F, Male/Female or 1 (male) and 2 (female)
  - Sample ID in the file must match the sample ID in your BAM files

Example:

```
SAMPLE1   M
SAMPLE2   F
SAMPLE3   F
```

Configuration file / Genome sizes

- What is it?
  - Specifies default values for many algorithm parameters used in Genome STRiP
- Why do I care?
  - Usually you don't, with one exception…
  - Currently, the "genome size" parameters in this file should be changed based on your reference sequence

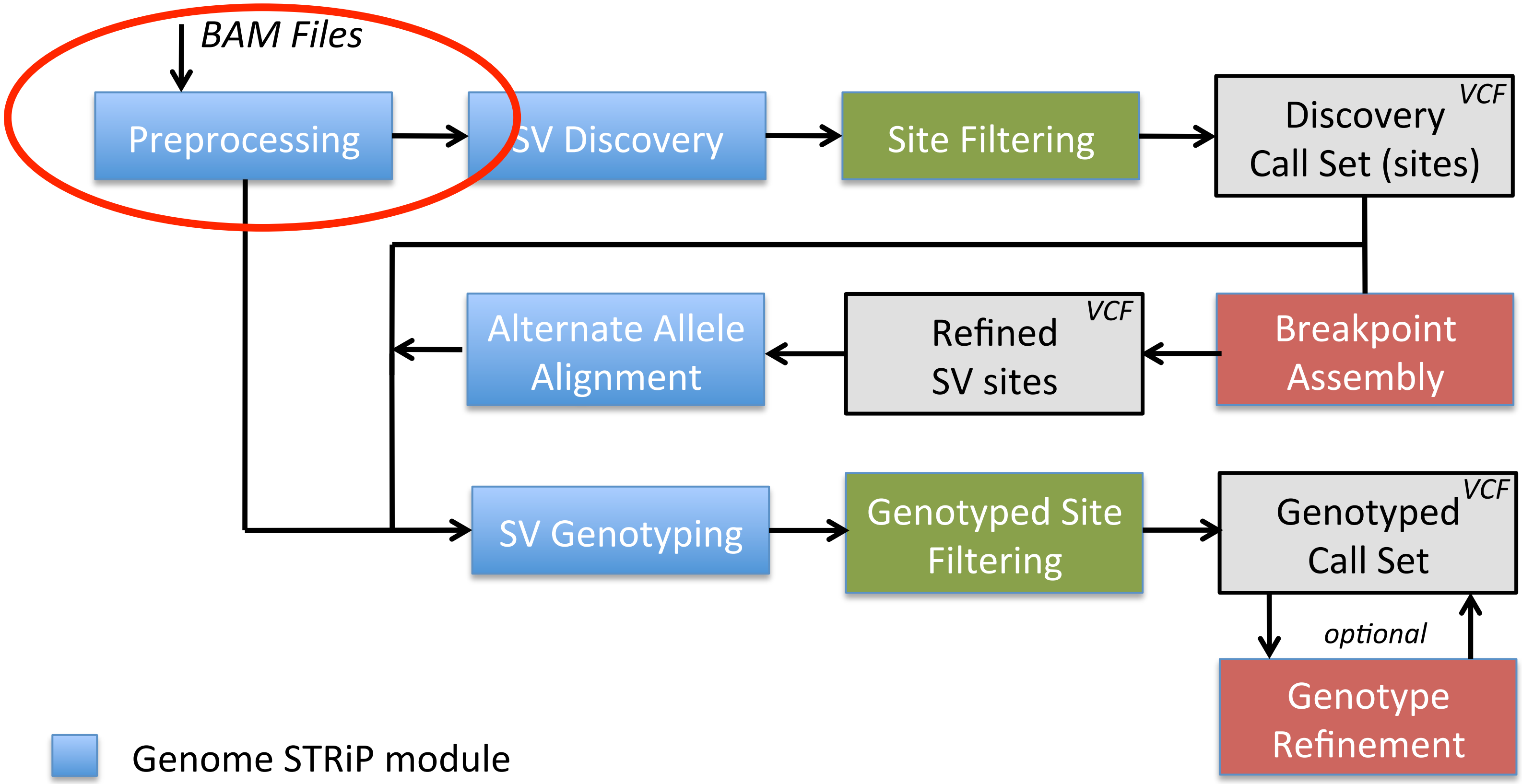
```
input.genomeSize:         2491855540
input.genomeSizeMale:     4845440507
input.genomeSizeFemale:   4960757950
```

input.genomeSize: Length of alignable reference sequence, without gaps
input.genomeSizeMale: Length of alignable reference x ploidy at each site for males
input.genomeSizeFemale: Length of alignable reference x ploidy at each site for females

- Where do I get it?
  - For pre-computed alignability masks, download the corresponding ".sizes" file and update these values in your configuration file
  - If you build your own alignability mask, see instructions on the FTP site

    ftp://ftp.broadinstitute.org/pub/svtoolkit/svmasks/README

Preprocessing



🟦 Genome STRiP module

🟩 Requires user intervention, defaults or best-practices are available

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

⬜ VCF data file

Preprocessing

- What is pre-computed?
  - Insert size distributions (per library)
  - Sequencing depth (per readgroup/library/sample)

    Expectation of fragments per base
  - Span coverage (per readgroup/library/sample)

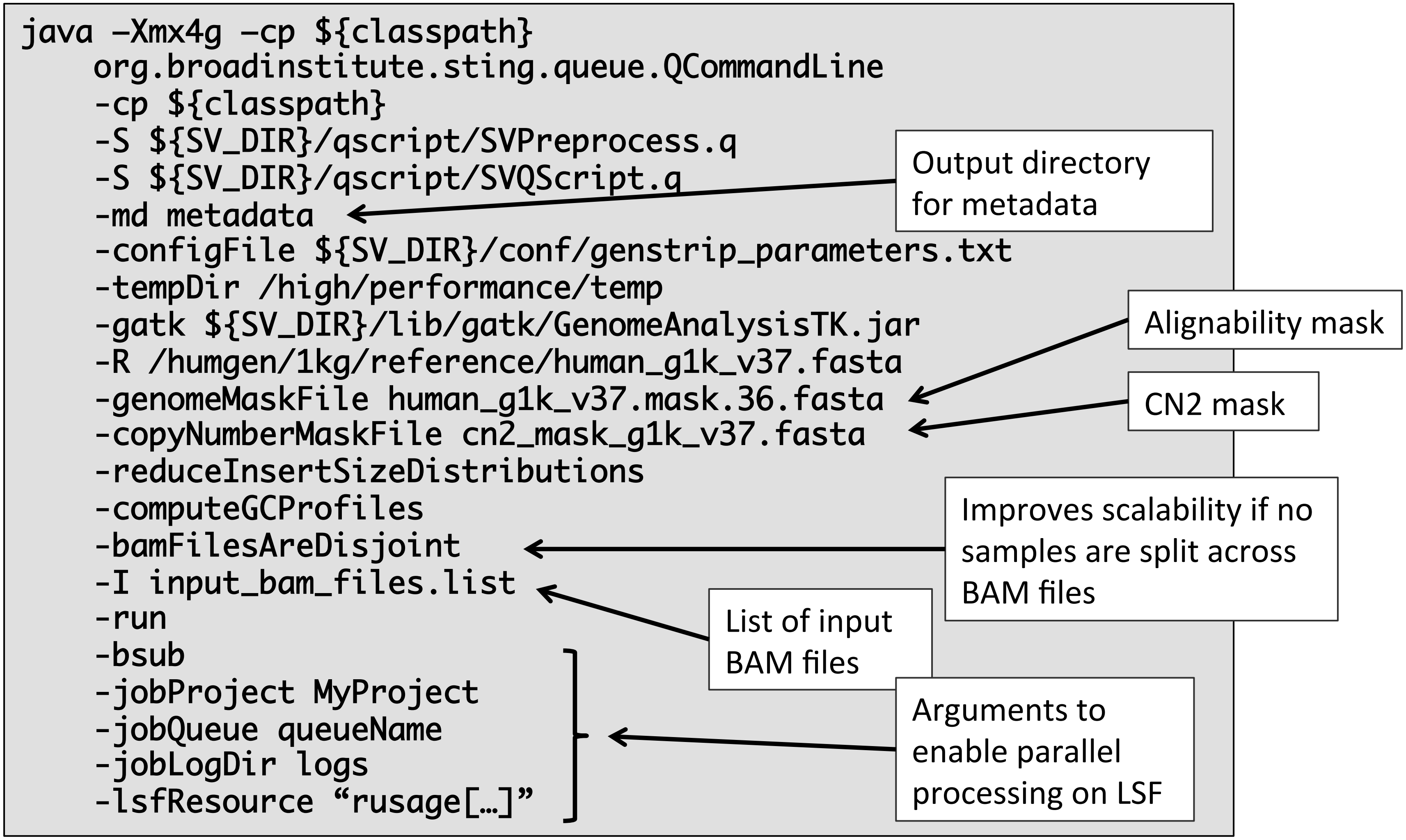
    Expectation of paired reads crossing a breakpoint
  - GC bias (per library)
- Inputs: BAM files
- Outputs: Multiple files in metadata directory
- Workflow: Parallel per-BAM, then merged

Metadata directory contents

- Insert size (fragment length) distributions
  - isd.spans.dat (text file of per-library statistics, useful for QC)
  - isd.hist.bin (full histograms, bulky but lossless)
  - isd.dist.bin (uses lossy accuracy-dependent compression)
  - Use *-reduceInsertSizeDistributions* to generate isd.dist.bin

    Important for scalability on large data sets
- Sequencing depth
  - Sub-directory metadata/depth and depth.dat summary file
- Span coverage
  - Measures total distance "spanned" between the two ends of paired-end reads
  - Sub-directory metadata/spans and spans.dat summary file
- GC-bias profiles
  - Enable with *-computeGCProfiles* command line argument
  - Sub-directory metadata/gcprofile and summary gcprofiles.zip file

Running Queue script for preprocessing



```
java –Xmx4g –cp ${classpath}
    org.broadinstitute.sting.queue.QCommandLine
    -cp ${classpath}
    -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
    -copyNumberMaskFile cn2_mask_g1k_v37.fasta
    -reduceInsertSizeDistributions
    -computeGCProfiles
    -bamFilesAreDisjoint
    -I input_bam_files.list
    -run
    -bsub
    -jobProject MyProject
    -jobQueue queueName
    -jobLogDir logs
    -lsfResource “rusage[…]”
```

Output directory for metadata

Alignability mask

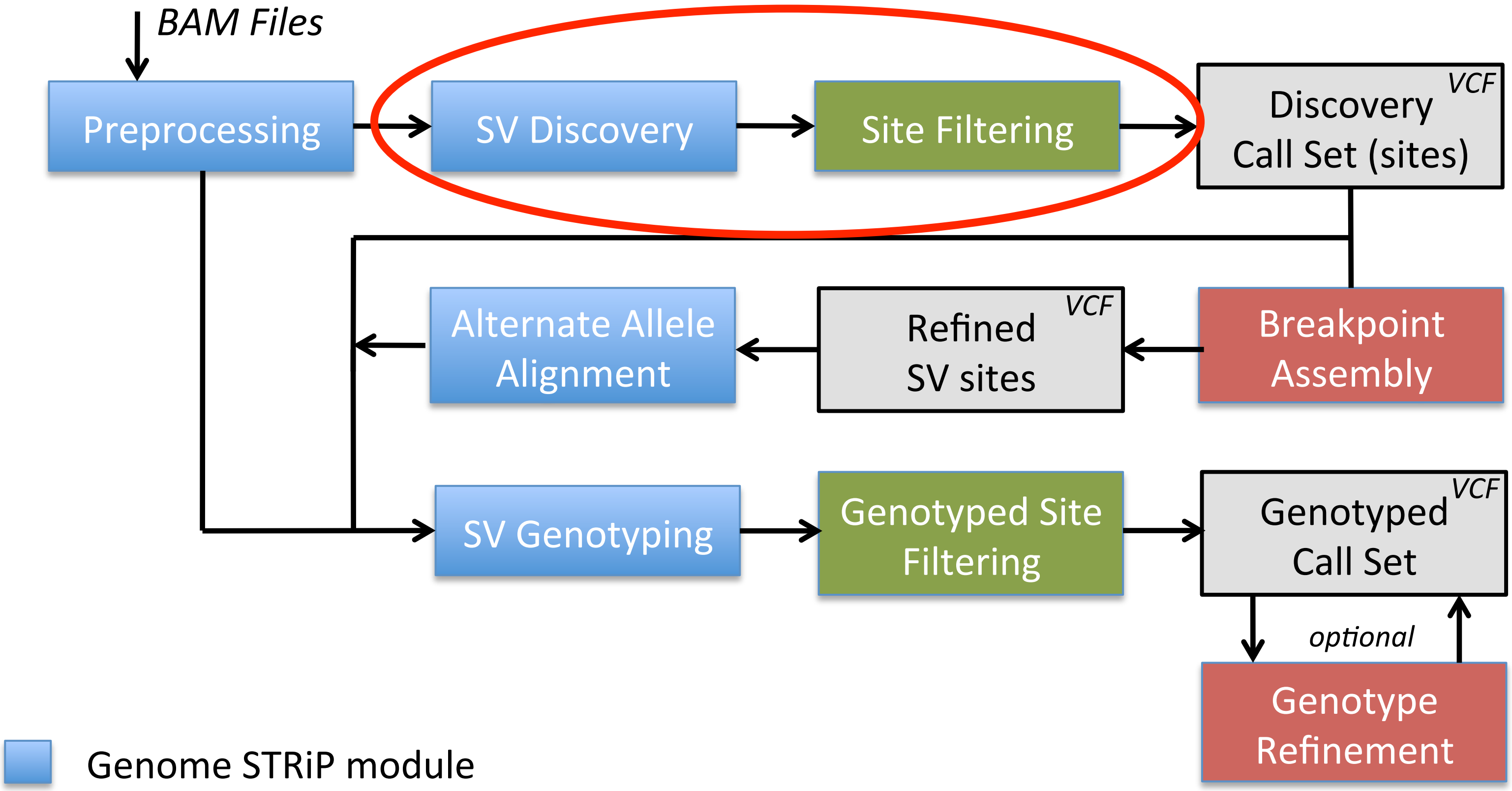
CN2 mask

Improves scalability if no samples are split across BAM files

List of input BAM files

Arguments to enable parallel processing on LSF

Discovery



BAM Files

Preprocessing

SV Discovery

Site Filtering

Discovery Call Set (sites) *VCF*

Alternate Allele Alignment

Refined SV sites *VCF*

Breakpoint Assembly

SV Genotyping

Genotyped Site Filtering

Genotyped Call Set *VCF*

optional

Genotype Refinement

🟦 Genome STRiP module

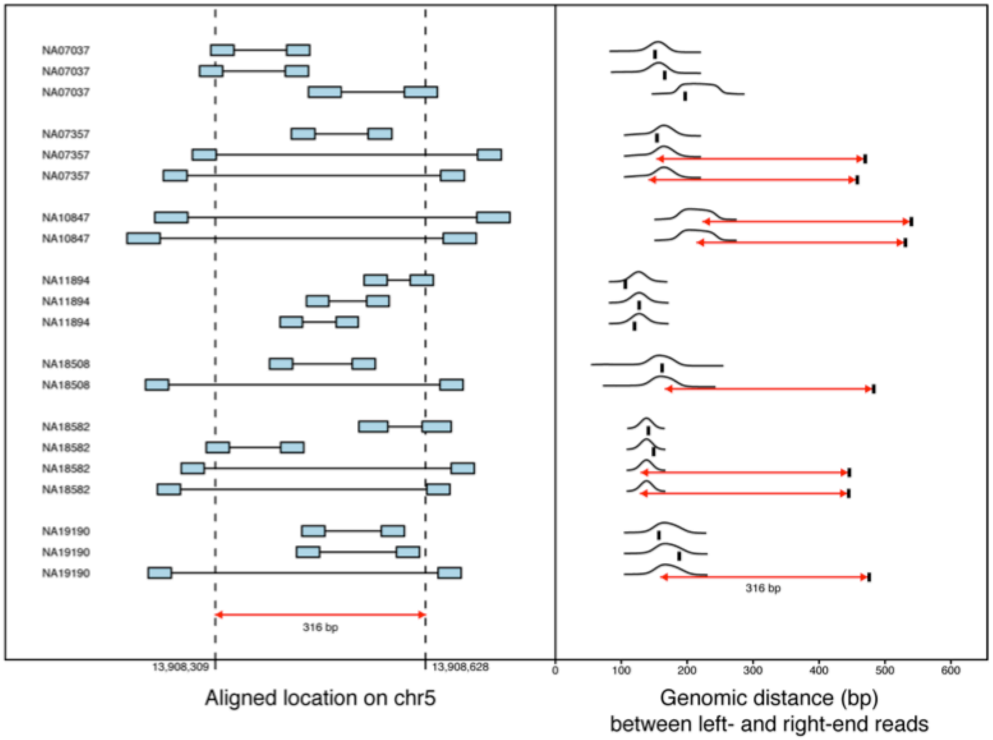
🟩 Requires user intervention, defaults or best-practices are available

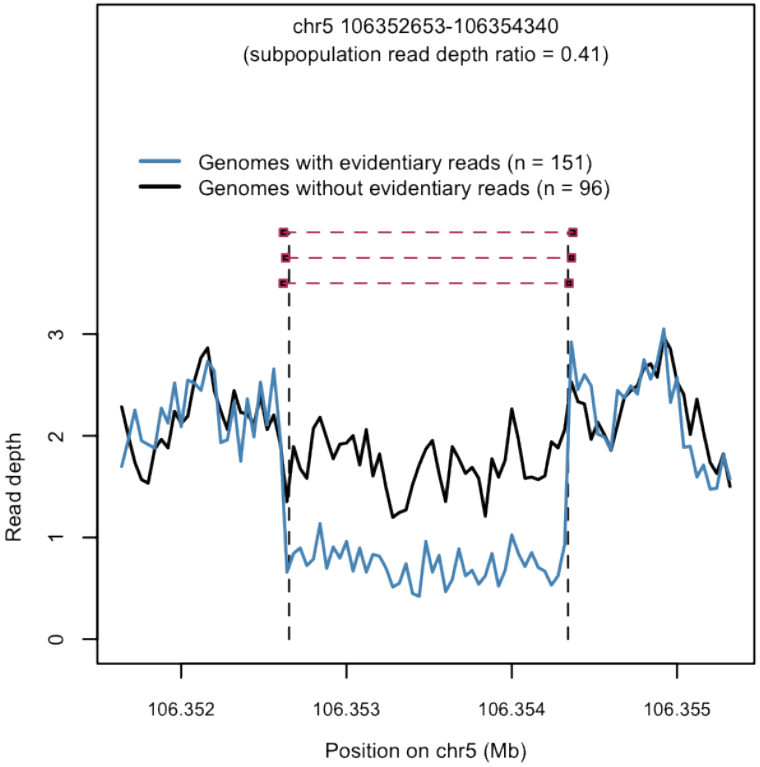
🟥 Third party tool (e.g. TIGRA or velvet for assembly, BEAGLE for genotype refinement)

⬜ VCF data file

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.

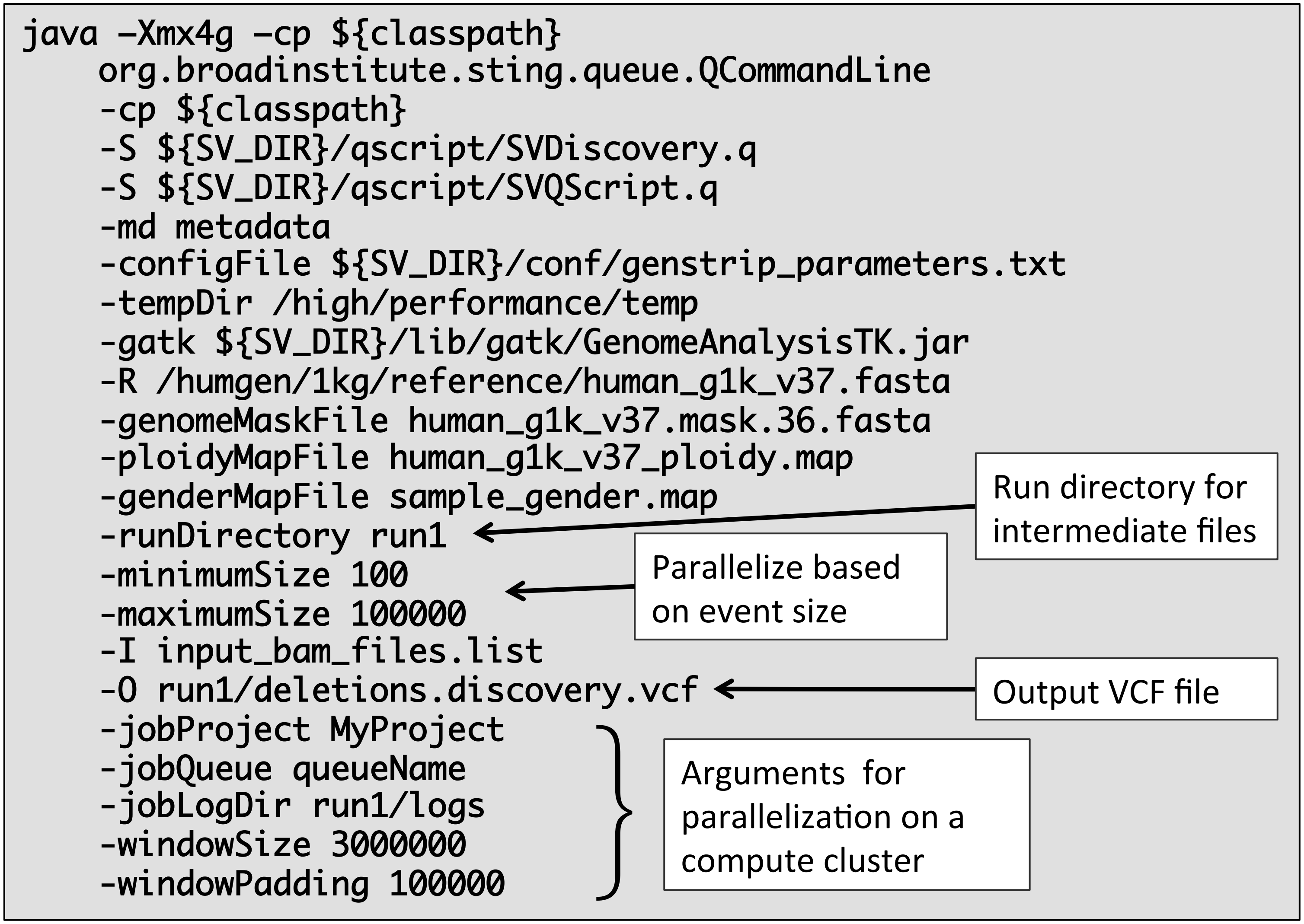




SV Discovery

- Inputs: BAM Files, metadata directory
- Outputs: Site VCF file
  - Contains records for all *evaluated* sites
  - The FILTER field tells whether a site is called as a true variant
  - Most evaluated sites are typically *not* called
  - The INFO field contains other quality metrics
- Auxilliary outputs: Multiple files in run directory
  - Used for QC and filtering
- Workflow: Parallel per-genome-locus and per-length-range, then merged
  - Tip: In the current implementation, discovery runs that include sites larger than 100Kb are significantly slower. If you have a large data set, it is recommended to do separate runs for events shorter than 100Kb and larger than 100Kb (your throughput will be more uniform)

Running Queue script for deletion discovery



```
java –Xmx4g –cp ${classpath}
    org.broadinstitute.sting.queue.QCommandLine
    -cp ${classpath}
    -S ${SV_DIR}/qscript/SVDiscovery.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
    -ploidyMapFile human_g1k_v37_ploidy.map
    -genderMapFile sample_gender.map
    -runDirectory run1
    -minimumSize 100
    -maximumSize 100000
    -I input_bam_files.list
    -O run1/deletions.discovery.vcf
    -jobProject MyProject
    -jobQueue queueName
    -jobLogDir run1/logs
    -windowSize 3000000
    -windowPadding 100000
```

Run directory for intermediate files

Parallelize based on event size

Output VCF file

Arguments for parallelization on a compute cluster

Discovery site filtering

Default filters

Filter name	Description
COVERAGE	Site has excessive read pileup
COHERENCE	Read pairs spacing is not consistent with a single segregating event
DEPTH	Read depth is not consistent with the read pair evidence across samples
DEPTHPVAL	Read depth differences are not significant

Not in default filter list, but recommended best-practices

ALPHASAT	Call is in regions of mostly alpha satellite repeat
PAIRSPERSAMPLE	Read pair evidence is thinly distributed across samples

Additional discovery site filters

For alpha satellite filtering, first run SVAnnotator to add annotations to the VCF about the repeat content of each evaluated site.

```
java –Xmx4g –cp ${classpath} org.broadinstitute.sv.main.SVAnnotator
  -R /humgen/1kg/reference/human_g1k_v37.fasta
  -A MobileElements
  -repeatTrackFile ucsc_repeats_g1k_v37.dat
  -vcf deletions.discovery.unfiltered.vcf
  -O deletions.discovery.annotated.vcf
```

Track file from UCSC browser, also available from our ftp site

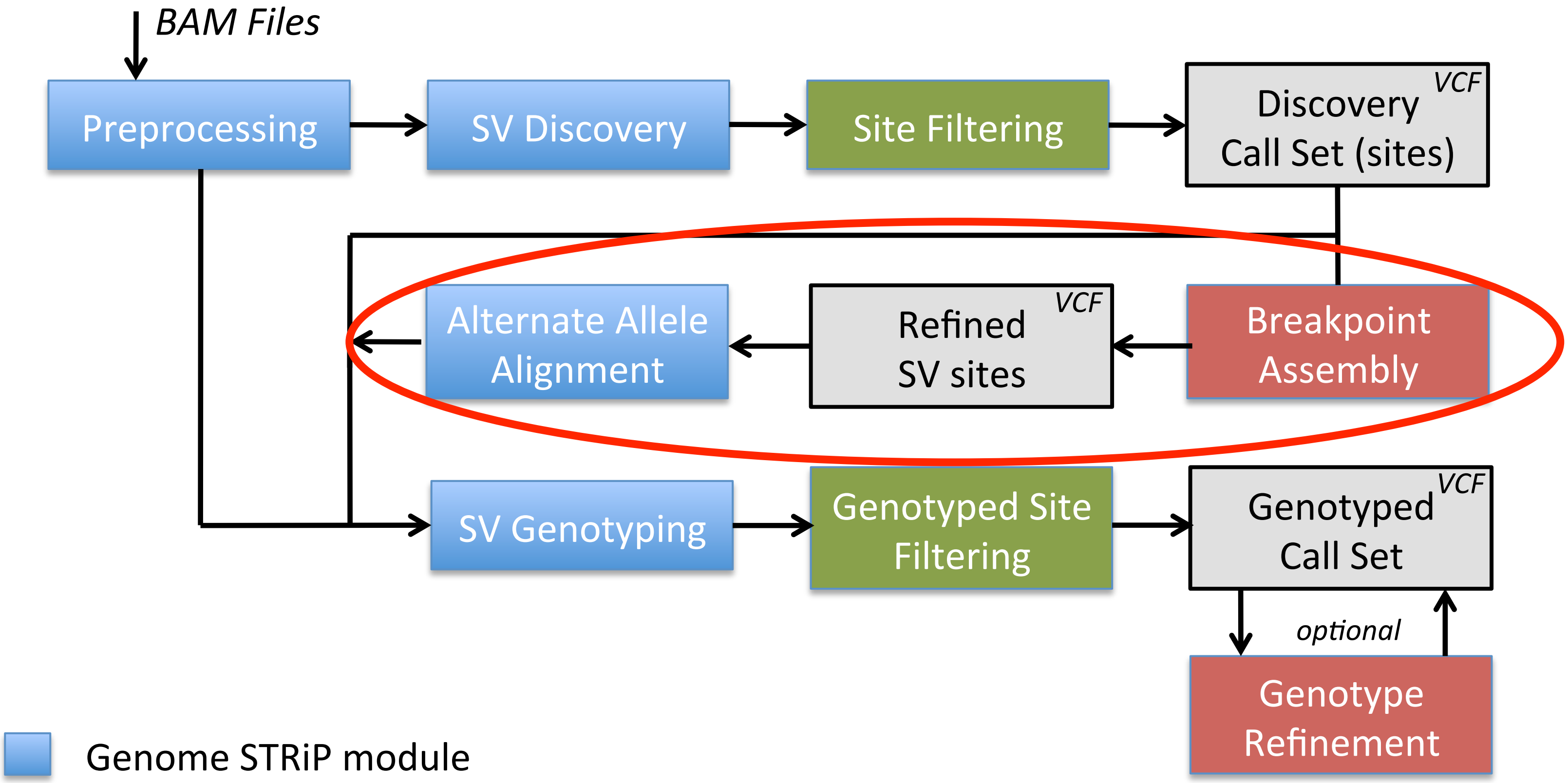
Both filters (separately or together) can be run via GATK VariantFiltration. This can also be combined with the default filters run from SVDiscovery.q.

```
java –Xmx4g –jar GenomeAnalysisTK.jar
  -T VariantFiltration
  -B:variant,VCF deletions.discovery.annotated.vcf
  -o deletions.discovery.vcf
  -R /humgen/1kg/reference/human_g1k_v37.fasta
  -filterName PAIRSPERSAMPLE -filter "GSNPAIRS < 1.1 * GSNSAMPLES"
  -filterName ALPHASAT –filter "GSALPHASATFRACTION > 0.90"
```

Best practice filters

If you have deep sequencing, you might consider a higher threshold than 1.1 on PAIRSPERSAMPLE, but this threshold has worked well for both 4x sequencing (1000 Genomes) and 12x sequencing (GoNL).

Breakpoint assembly



🟦 Genome STRiP module

🟩 Requires user intervention, defaults or best-practices are available

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

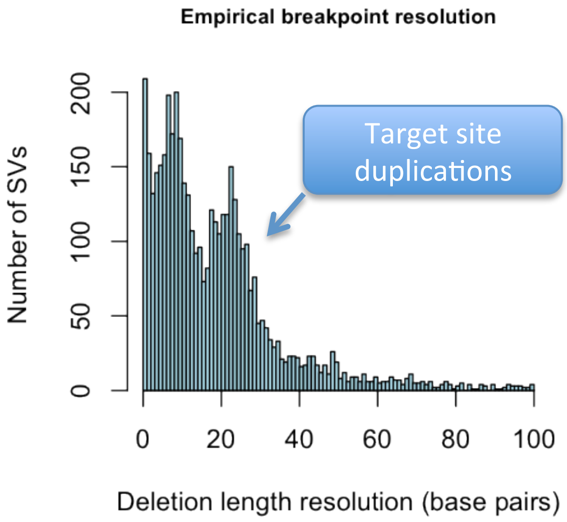
⬜ VCF data file

Breakpoint assembly

To determine precise breakpoints, use a third party tool (e.g. TIGRA-SV, velvet) or a catalog of known breakpoints (e.g. 1000 Genomes)

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

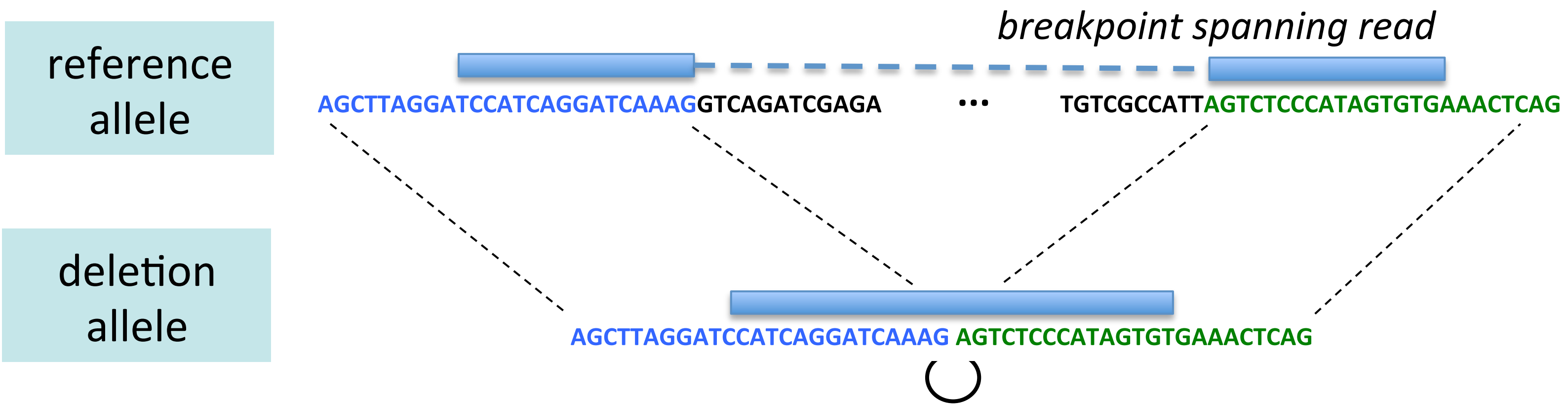
To utilize breakpoint-spanning reads in genotyping, you need exact breakpoint coordinates.



In the 1000 Genomes project, we used TIGRA-SV (WashU) and AGE (Yale) to assemble breakpoints for over half of the discovered deletion sites.

Alternate allele alignment

When precise alleles are available, we use breakpoint-spanning reads in genotyping.



reference allele

deletion allele

breakpoint spanning read

AGCTTAGGATCCATCAGGATCAAAGGTCAGATCGAGA ••• TGTCGCCATTAGTCTCCCATAGTGTGAAACTCAG

AGCTTAGGATCCATCAGGATCAAAG AGTCTCCCATAGTGTGAAACTCAG

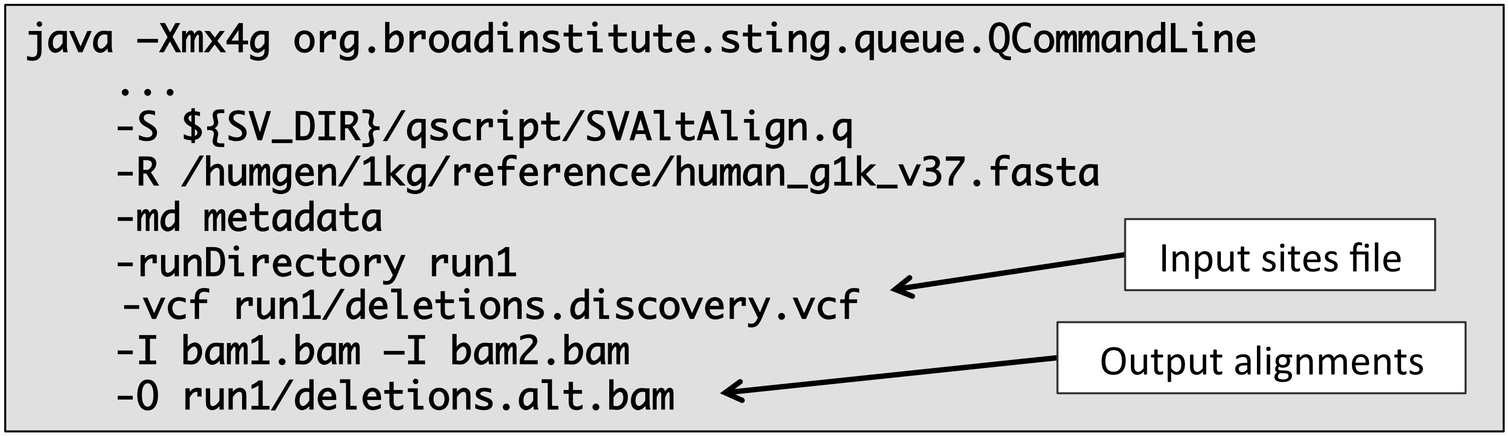
There are three sources for breakpoint-spanning reads:

Source	How handled
“in-place” reads aligned at the breakpoint	Automatically realigned on-the-fly to alt allele during genotyping
unmapped mates in same BAM file where mate is aligned nearby	Automatically realigned on-the-fly during genotyping
completely unmapped reads	Requires alternate allele aligner

Queue script for alt allele alignment

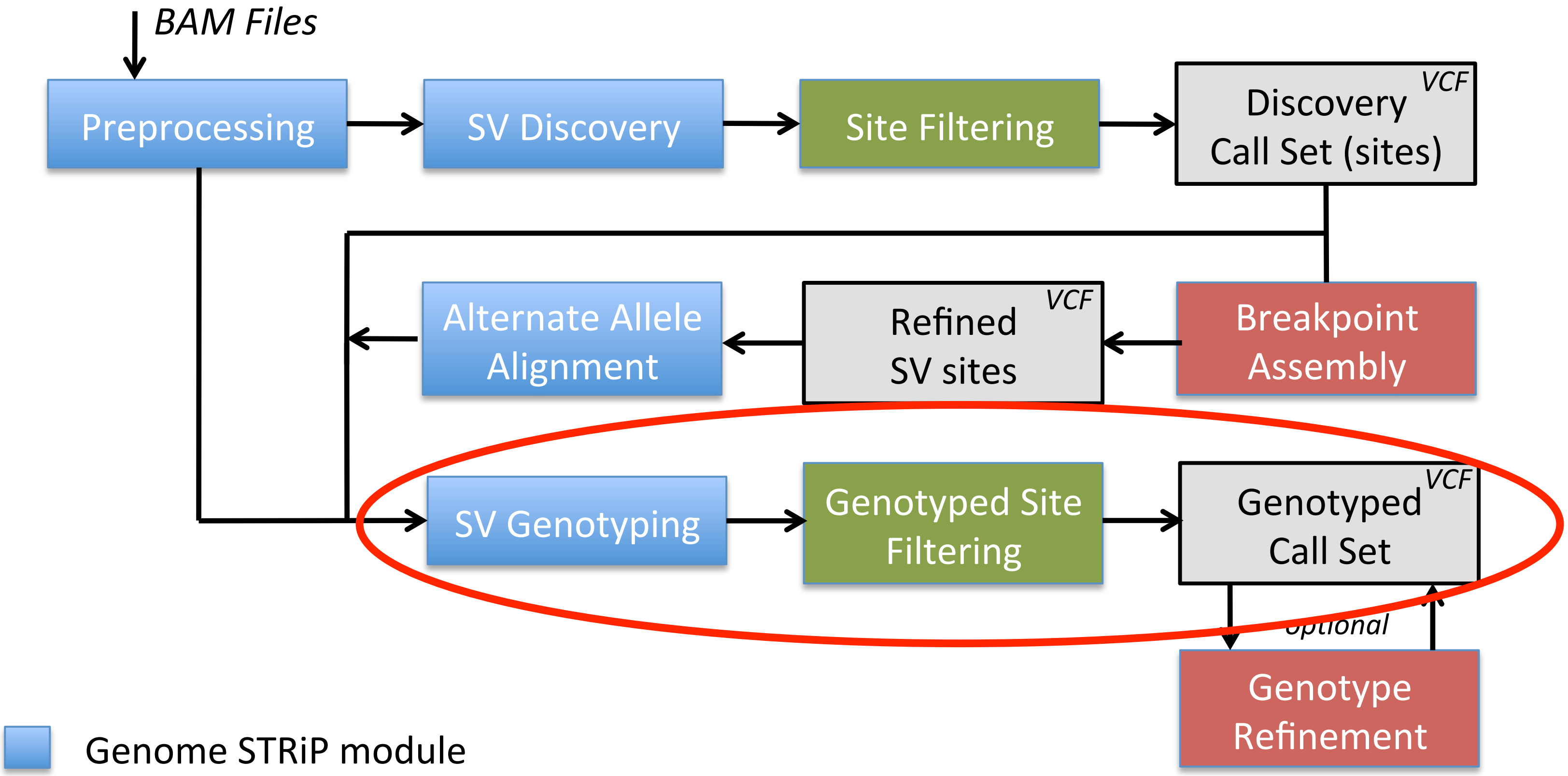
Inputs: VCF file containing SVs with exact alleles
BAM files containing unmapped reads
Outputs: BAM file containing alignments to alternate alleles

With longer reads (e.g. 100bp) there is only marginal benefit in running the alternate allele aligner step. Most reads will be in the main BAM files (often soft-clipped) and will be used automatically for breakpoint genotyping.



```
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
```

Genotyping



BAM Files

Preprocessing

SV Discovery

Site Filtering

Discovery Call Set (sites) *VCF*

Breakpoint Assembly

Refined SV sites *VCF*

Alternate Allele Alignment

SV Genotyping

Genotyped Site Filtering

Genotyped Call Set *VCF*

optional

Genotype Refinement

🟦 Genome STRiP module

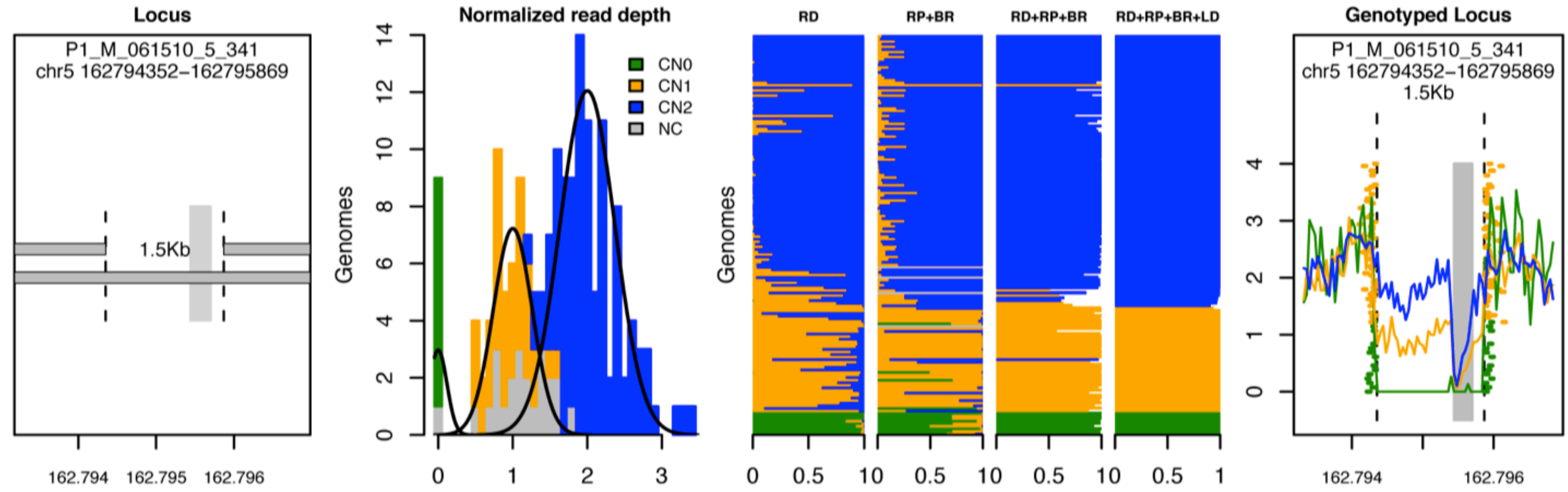
🟩 Requires user intervention, defaults or best-practices are available

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

⬜ VCF data file

SV Genotyping

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

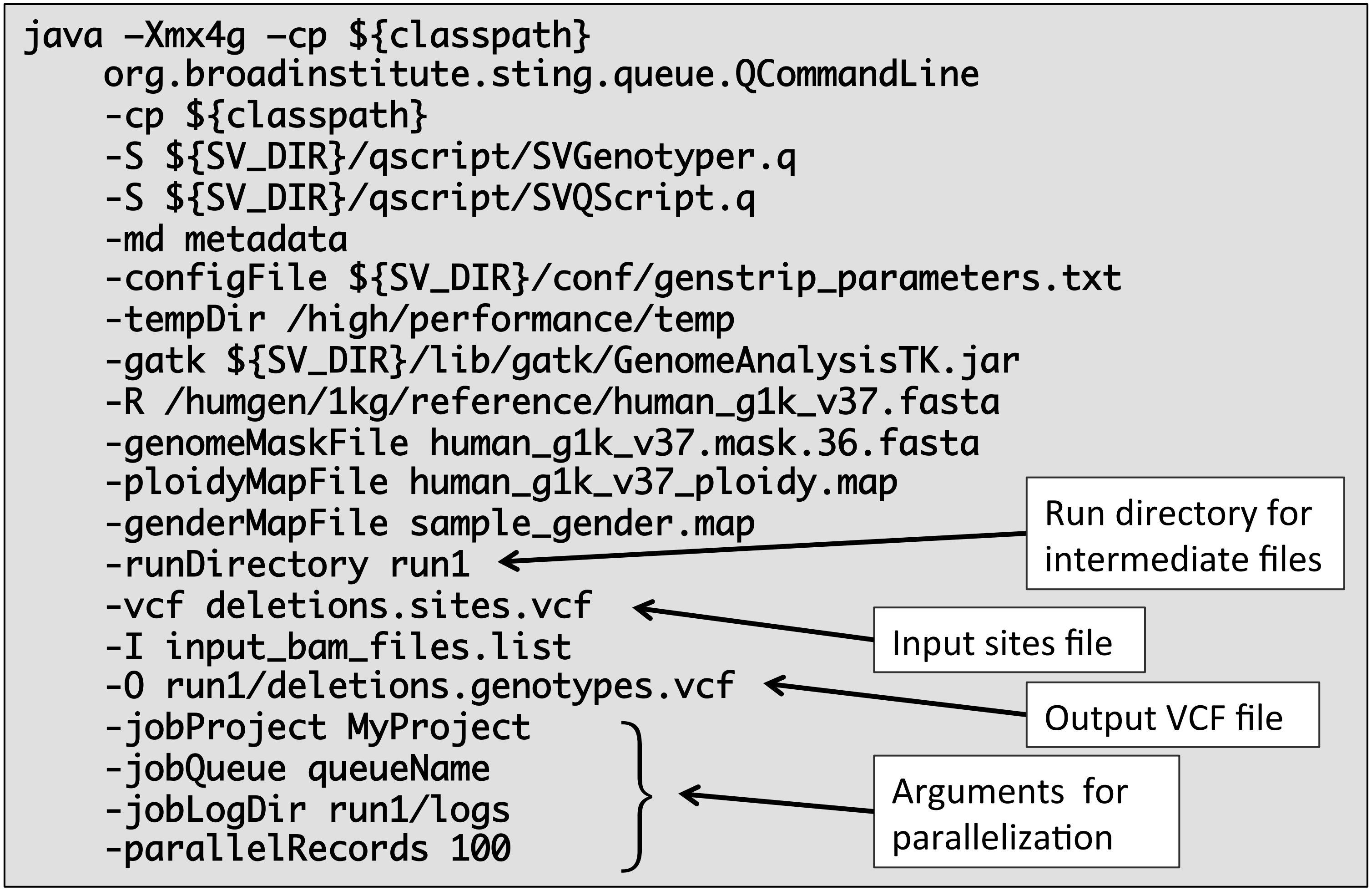


Support for genotyping other types of variants (e.g. duplications) is under development.

SV Genotyping

- Inputs: BAM Files, VCF site file, metadata directory
  - Typically you want to extract only the PASS sites for genotyping if you are using sites generated from Genome STRiP discovery
  - Reformat VCF records to include specific alleles (rather than <DEL>) to enable breakpoint-based genotyping
- Outputs: Genotype VCF file
  - Contains records for all input sites
  - The FILTER field tells whether a site passes genotyping site filters
  - Non-passing sites will still contain genotype calls
- Auxilliary outputs: Multiple files in run directory
  - Used for QC and filtering
- Workflow: Parallel per-input-site from VCF (in batches), then merged

Running Queue script for deletion genotyping



```
java -Xmx4g -cp ${classpath}
    org.broadinstitute.sting.queue.QCommandLine
    -cp ${classpath}
    -S ${SV_DIR}/qscript/SVGenotyper.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
    -ploidyMapFile human_g1k_v37_ploidy.map
    -genderMapFile sample_gender.map
    -runDirectory run1
    -vcf deletions.sites.vcf
    -I input_bam_files.list
    -O run1/deletions.genotypes.vcf
    -jobProject MyProject
    -jobQueue queueName
    -jobLogDir run1/logs
    -parallelRecords 100
```

Run directory for intermediate files

Input sites file

Output VCF file

Arguments for parallelization

Genotype VCF for SVs

Sample VCF file (INFO field not shown)

```
CHROM   POS ID      REF    ALT QUAL FILTER INFO  FORMAT          NA10000

1  2918690 DEL_833 G   <DEL>  .    PASS    …    GT:FT:GL:GL0:GQ  0/1:PASS:-9.3,-0.0,-135.2:-11.1,-0.0,-13.0:93
```

FORMAT tag	Examples	Description
GT	0/1	Genotype (0 = reference, 1 = alt) But if FT is not PASS, should consider as no-call
FT	PASS, LowQual	Filter field for genotypes LowQual for sites with GQ < 13 (95% confidence)
GL	-9.3,-0.0,-135.2	Genotype likelihoods (uses population frequency information extracted from read depth mixture model)
GL0	-11.1,-0.0,-13.0	Genotype likelihoods with no frequency prior Use these values for LD-based genotype refinement
GQ	93	Genotype quality (phred scaled)

Genotyped Site Filtering

Current best practices are to apply a set of site-level filters after genotyping in the following categories*

- Data sufficiency
  - Filter sites with insufficient data to genotype accurately
- Genotype accuracy
  - Filter sites with low apparent genotype accuracy
- Non-variant sites
  - Flag sites that appear non-variant post-genotyping
- Duplicate sites
  - Flag apparent duplicate calls based on genotype likelihoods
  - Filter, retaining the site with the highest genotype quality

* Running these post-genotyping site filters is not done by default in the latest released version of Genome STRiP, but is a planned option for future releases.

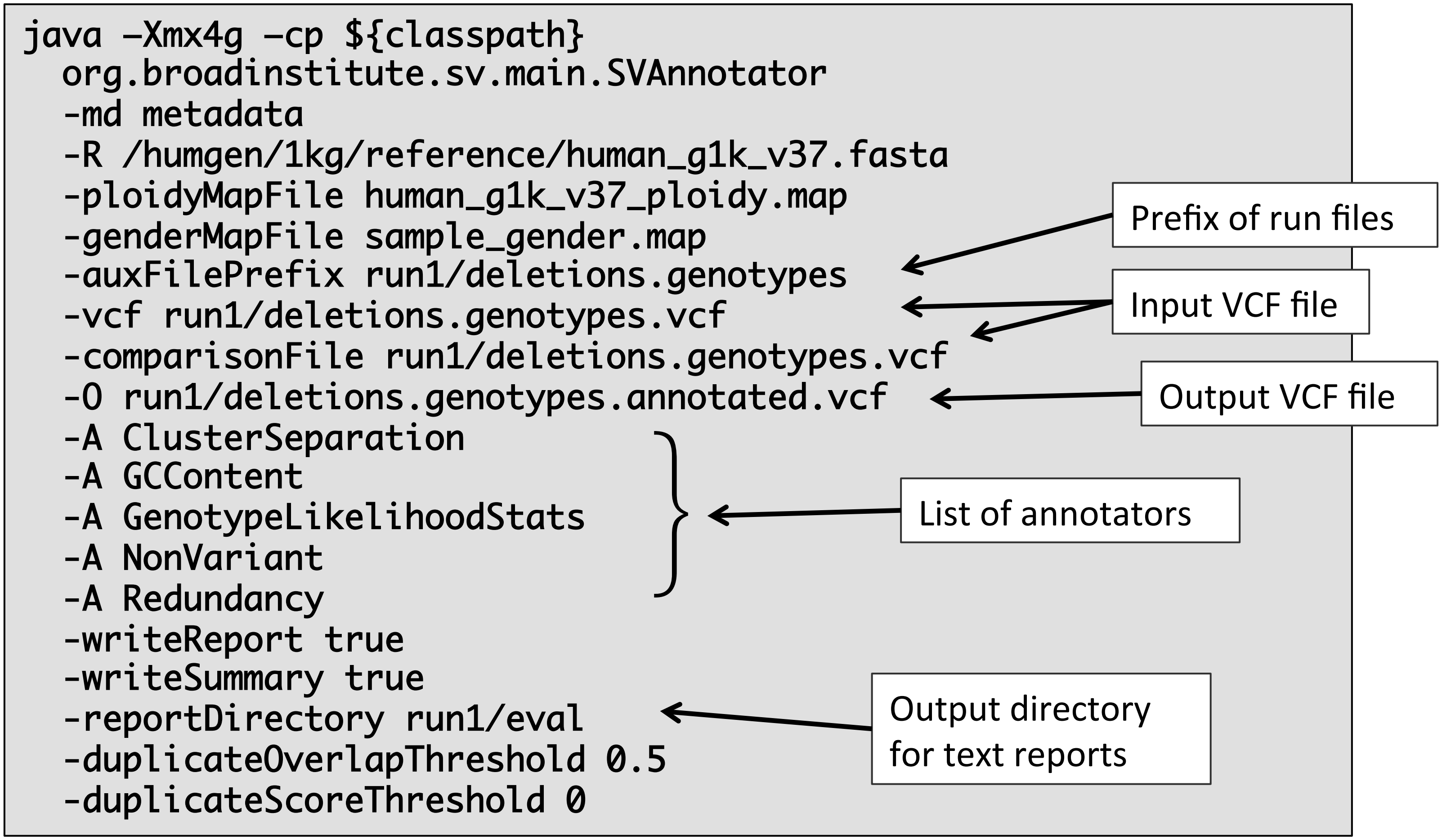
Genotyped site filtering

Current best-practice filters (from 1000G and other large projects)

Filter name	Description
ALIGNLENGTH	Site has insufficient alignable bases (default 200)
CLUSTERSEP	Read depth model shows insufficient cluster separation
GTDEPTH	Read depth at the site is too low or too high
INBREEDINGCOEFF	Filter site due to excess number of het calls
DUPLICATE	Site is apparent duplicate of another site
NONVARIANT	Site is likely non-variant based on genotype likelihoods

These filters are currently implemented by running annotators post-genotyping and then using GATK VariantFiltration. This strategy is applicable only when breakpoints are not being used for genotyping.

Annotations used for filtering



```
java –Xmx4g –cp ${classpath}
  org.broadinstitute.sv.main.SVAnnotator
  -md metadata
  -R /humgen/1kg/reference/human_g1k_v37.fasta
  -ploidyMapFile human_g1k_v37_ploidy.map
  -genderMapFile sample_gender.map
  -auxFilePrefix run1/deletions.genotypes
  -vcf run1/deletions.genotypes.vcf
  -comparisonFile run1/deletions.genotypes.vcf
  -O run1/deletions.genotypes.annotated.vcf
  -A ClusterSeparation
  -A GCContent
  -A GenotypeLikelihoodStats
  -A NonVariant
  -A Redundancy
  -writeReport true
  -writeSummary true
  -reportDirectory run1/eval
  -duplicateOverlapThreshold 0.5
  -duplicateScoreThreshold 0
```

Prefix of run files

Input VCF file

Output VCF file

List of annotators

Output directory for text reports

Genotyped site filters

These post-genotyping filters are applicable only when breakpoints are not being used for genotyping. If you are genotyping with breakpoints, the depth-based filters should be applied to the read depth signal only, which currently requires custom post-processing.

Example of using GATK VariantFiltration to run best-practices filters, based on experiences from 1000 Genomes, GoNL and other projects.

```
java –Xmx4g –jar GenomeAnalysisTK.jar
  -T VariantFiltration
  -B:variant,VCF deletions.discovery.annotated.vcf
  -o deletions.discovery.vcf
  -R /humgen/1kg/reference/human_g1k_v37.fasta
  -filterName ALIGNLENGTH –filter “GSELENGTH < 200”
  -filterName CLUSTERSEP –filter “GSCLUSTERSEP == NA || GSCLUSTERSEP <= 2.0”
  -filterName GTDEPTH –filter “GSM1 == NA || GSM1 <= 0.5 || GSM1 >= 2.0”
  -filterName INBREEDINGCOEFF –filter “GLINBREEDINGCOEFF != NA &&
                                       GLINBREEDINGCOEFF < -0.15”
  -filterName NONVARIANT –filter “GSNONVARSCORE != NA && GSNONVARSCORE >= 13.0”
  -filterName DUPLICATE –filter “GSDUPLICATESCORE != NA && GSDUPLICATESCORE >= 0”
```

Genotype refinement

1000 Genomes used BEAGLE & MaCH for genotype refinement

Exploits LD between deletions and SNPs / small indels
Generated complete phased haplotypes at all sites
Can be computationally demanding

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

QUALITY CONTROL

Brief overview of several QC methods

Illustrative examples from GoNL and 1000 Genomes Projects

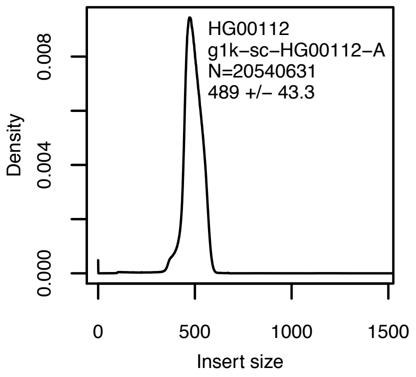
Sample QC: Insert size distributions

Review insert size distribution summary statistics: metadata/isd.stats.dat

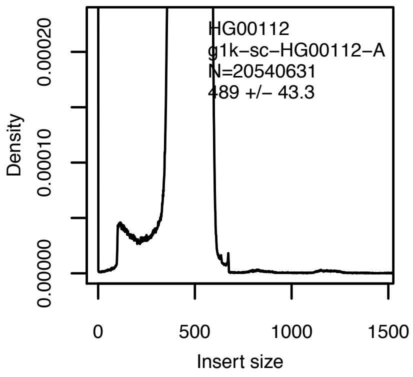
```
SAMPLE   LIBRARY           READGROUP  NPAIRS  NBINS  MEDIAN   RSD
HG00112  g1k-sc-HG00112-A     NA    20540631  49437    489   43.28
HG00112  g1k-sc-HG00112-D     NA    32894100  48657    417   34.03
```

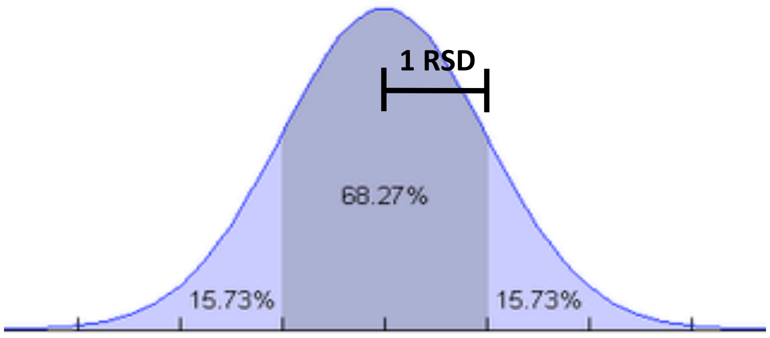
Distributions for individual libraries can be plotted using PlotInsertSizeDistributions. Two plots are generated for each library, one full scale and one expanded. Distributions with excess mass in the *right* tail are particularly problematic for discovery. The example below is relatively clean in the right tail.

Full scale



Expanded





RSD is a variance measure more robust to outliers.
RSD = 0.5 x width of middle 68.27% of the distribution.

Discovery QC
Filter summary from GoNL Project

The Genome of the Netherlands Project (GoNL) performed 12x whole-genome sequencing on 250 trios/quartets of Dutch ancestry

Discovery sites evaluated:	32,759
Passing discovery sites:	14,103 (43%)

FILTER	COUNTSINGLE	COUNTTOTAL
ALPHASAT	331	2060
COHERENCE	823	2461
COVERAGE	257	6846
DEPTH	3341	16829
DEPTHPVAL	2092	10834
PAIRSPERSAMPLE	251	7424
PASS	14103	14103

The rate of passing sites (43%) is higher than 4x sequencing projects like 1000 Genomes. The PAIRSPERSAMPLE filter rejected relatively few sites on its own, as would be expected for 12x sequencing.

Genotyping QC in GoNL
Filters, genotyped sites rate, genotype call rate

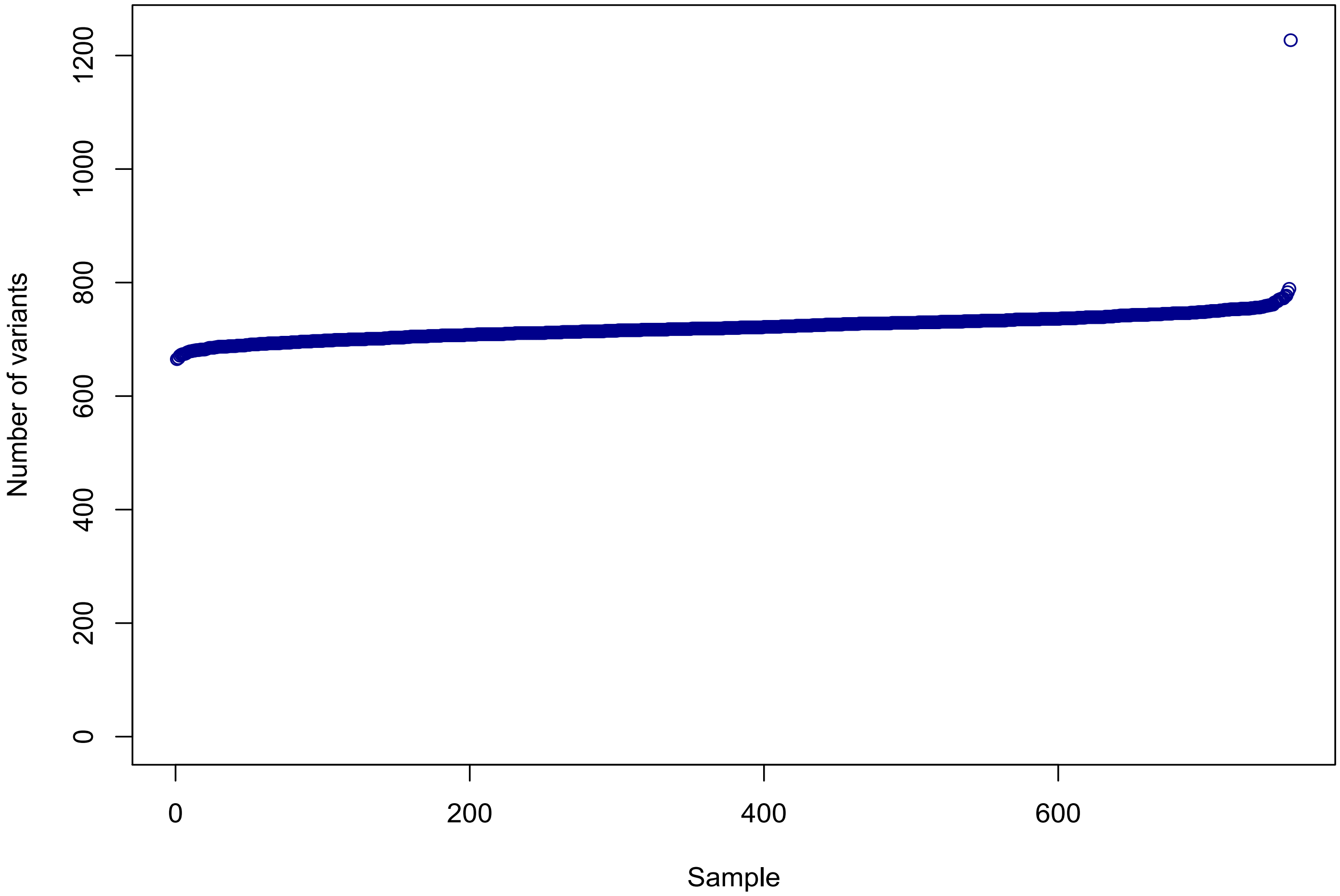
Passing discovery sites: 14,103
Genotyped sites: 9,352 (66%)

FILTER	COUNTSINGLE	COUNTTOTAL
ALIGNLENGTH	608	3351
CLUSTERSEP	252	3025
GTDEPTH	94	863
INBREEDINGCOEFF	477	2467
DUPLICATE	30	30
NONVARIANT	55	62
PASS	9352	9352

The genotyping rate (66%) is higher than 1000G Phase 1 (61%) – in both projects breakpoints were not used in genotyping, which means many shorter sites are called but not genotyped. At passing sites the genotype call rate (at 95% confidence) is 98.9% before genotype refinement, substantially higher than 1000G which used 4x sequencing.

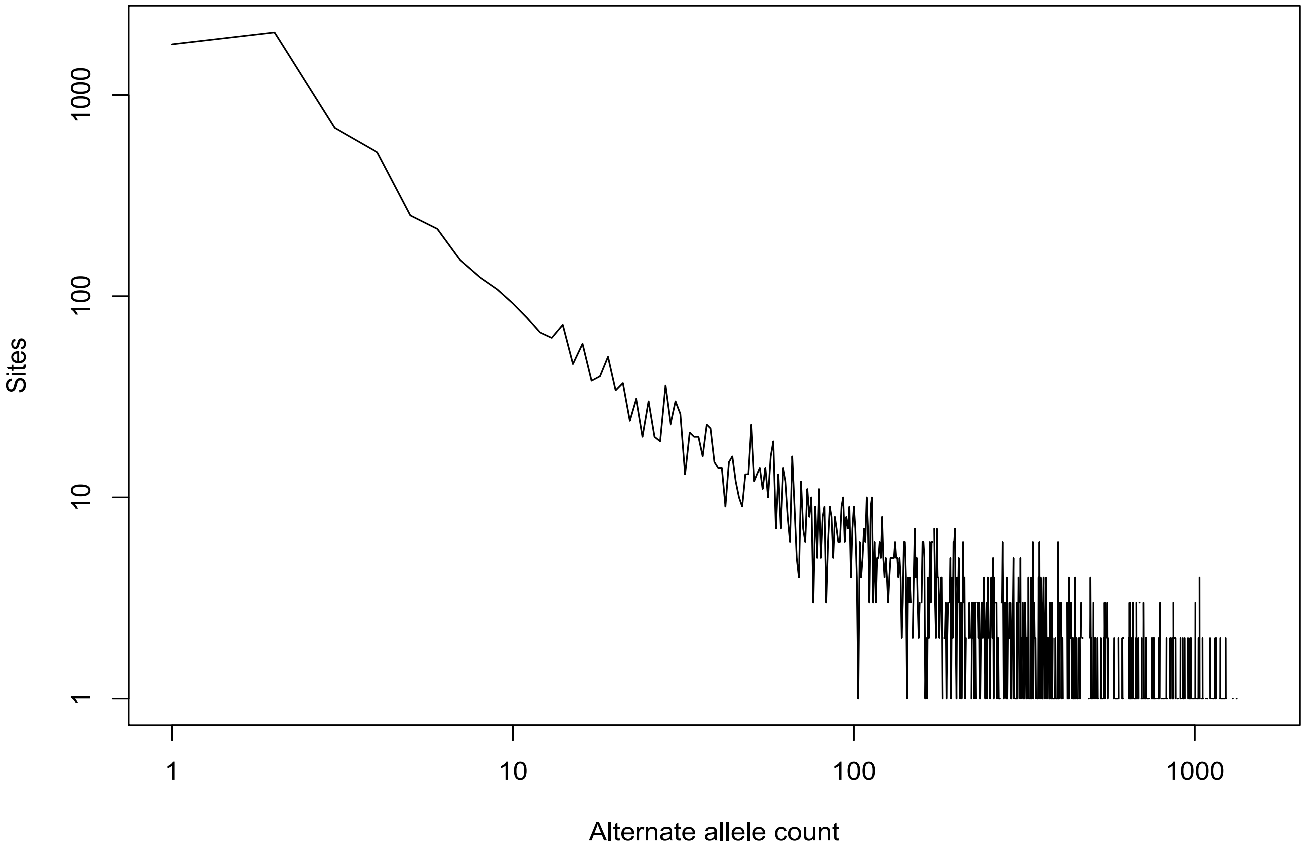
Genotype QC: Variants per sample

Data from GoNL: The distribution of variants per sample looks very uniform in this relatively homogeneous European cohort, with the exception of one clear outlier, which was removed for investigation.



Genotyping QC: Allele frequency spectrum

The overall spectrum looks roughly linear on this log/log plot. The apparently reduced power to call singletons is likely due to the trio design (the offspring should never have singletons).



Genotype QC

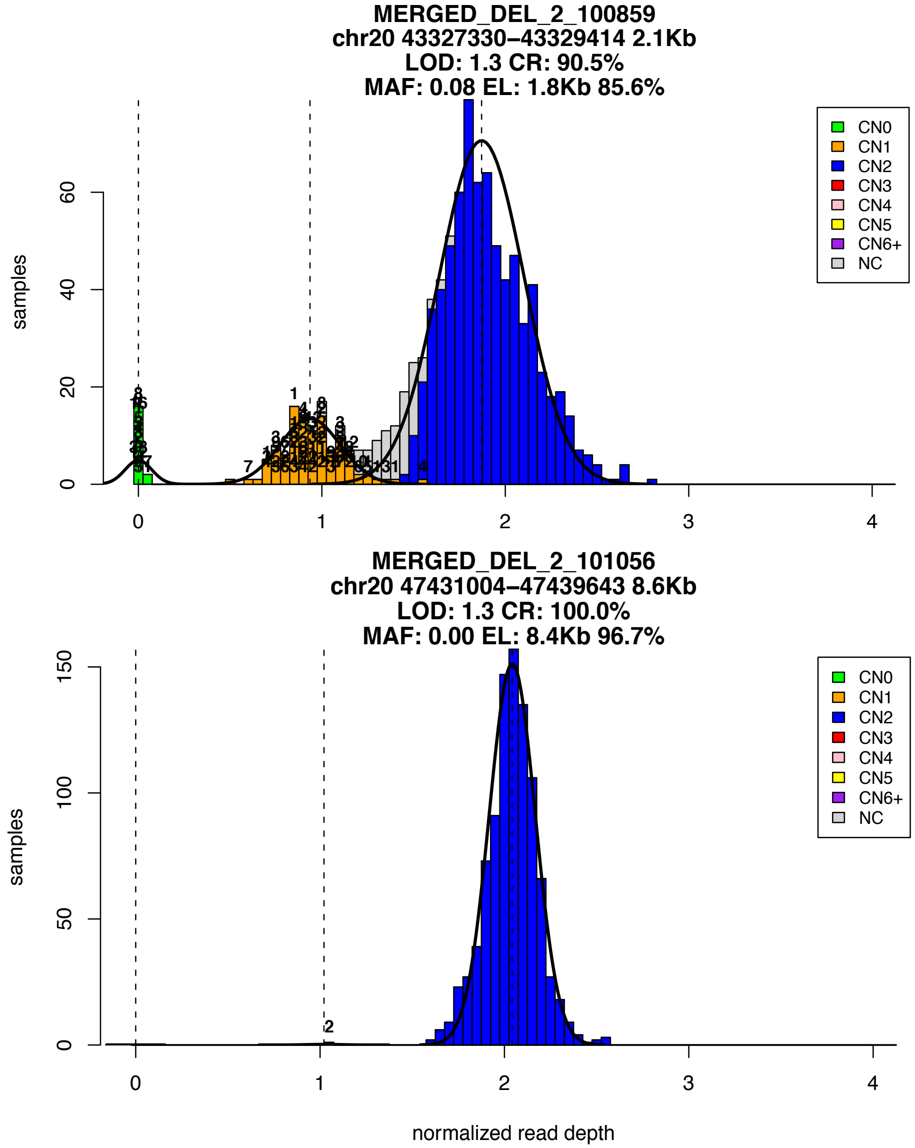
- Comparison to previously ascertained call sets
  - 1000 Genomes
  - ReciprocalOverlap annotator is a useful tool (compares two SV vcf files)
- Successful breakpoint assembly
- Lack of heterozygous SNPs in hemizygous regions
  - Individuals heterozygous for a deletion should not have heterzygous SNPs at the locus

Individual site QC: PlotGenotypingResults

PlotGenotypingResults is a utility program that generates PDF plots for individual sites, like the ones shown on the right (4x sequencing from 1000 Genomes). Inspection of individual sites is a valuable QC tool.

The top site is relatively short (2.1Kb) and high frequency. The small black numbers are supporting read pairs, seen in samples called het or homozygous deleted. Some samples (gray) are not confidently called in 4x sequencing at this length scale.

The bottom site is a larger (8.6Kb) singleton, with a call rate of 100%.



SOFTWARE AND SUPPORT

Software availability

Usage scenarios

Resource requirements

Software availability and support

Web site

- http://www.broadinstitute.org/software/genomestrip
- Documentation, FAQ
- Software downloads
  - You need to register in order to download (name, email, organization)
  - Production release: Corresponds to 1000 Genomes Pilot
  - Interim releases: More recent updates, supported, limited documentation
- Most of the functionality discussed here is in the interim releases.

Support mailing list

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

GATK Support Forum

http://gatkforums.broadinstitute.org

Installation test

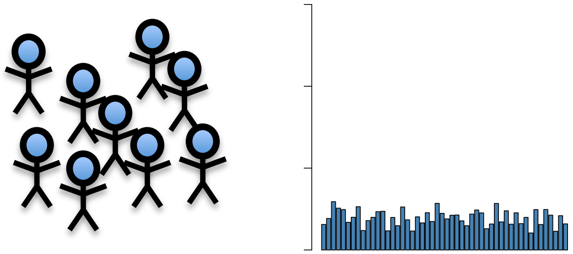
Used to validate correct installation

- Ten minute example run on toy data set
- People also use this as a recipe for production analyses
- Most common pitfalls
  - Installtest runs single threaded, not parallel
  - Does not use -reduceInsertSizeDistributions, add this for scalability
  - Does not use -computeGCProfiles, add this for much better accuracy
  - Uses -L 1 (restrict analysis to chromosome 1) to speed up the test, remove this for whole genome analysis (or if your reference has "chr1")
  - Modify configuration file with appropriate genome sizes based on the alignability mask you are using

Usage Scenarios

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

Samples *Coverage*



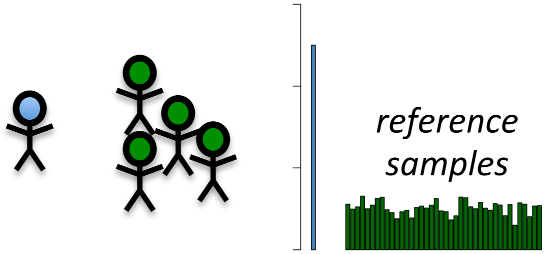
<u>Whole Genome Population Sequencing</u>

Need 20-30+ samples for good results

Low or high coverage, can be variable

<u>Future Goals</u>

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



Resource requirements

Performance on some sample analyses (1000 Genomes pilot / phase 1)
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 sites x 168 samples	4

Preprocessing	3800x (946 x 4x) 17Tb	86
Discovery	3800x (946 x 4x) 17Tb	150
Genotyping	113,000 sites x 946 samples	360

A number of scalability improvements are under development

Summary

- Genome STRiP has performed well in the 1000 Genomes Project on deletion discovery and genotyping
- Genome STRiP has been used successfully in other large projects
- Common usage scenarios
  - De novo deletion discovery and genotyping in sequencing-based GWAS
  - Genotyping known deletions (e.g. from 1000 Genomes) in new samples
- Improvements are ongoing
  - Usability and scalability
  - Best-practices and tools for calling and QC
  - Pipelines for new variant types and usage scenarios

Acknowledgements

Broad / HMS / McCarroll Lab

Josh Korn Jim Nemesh Nick Patterson Seva Kashin Steve McCarroll

GSA / GATK / Queue

Khalid Shakir Chris Hartl Aaron McKenna Matt Hanna
Ryan Poplin Maricio Carneiro Guillermo del Angel
Geraldine Van der Auwera Eric Banks Mark DePristo

Genome of the Netherlands

Laurent Francioli Kai Ye Paul deBakker

1000 Genomes Structural Variation Group

Ryan Mills
Klaudia Walter
Chip Stewart
Ken Chen
Can Alkan
Miriam Konkel
Jonathan Sebat
Mark Gerstein
Matt Hurles
Alex Abyzov
Chris Yoon
Kai Ye
Yujun Zhang
Zhengdong Zhang
Adrian Stuetz
Tobias Rausch
Mark Batzer
Jan Korbel
Don Conrad
Jeff Kidd
Zam Iqbal
Mindy Shi
Kenny Ye
Peter Sudmant
Thomas Keane
Gabor Marth
Evan Eichler
Ekta Khurana
Jasmine Mu
Michael Stromberg
Marcin Von Grotthuss
Jiantao Wu
Scott Devine
Fereydoun Hormozdiari
Mark Batzer
Charles Lee