Indel Calling Pipeline in the GATK

Guillermo del Angel, Ph.D.

Genome Sequencing and Analysis Group
Medical and Population Genetics Program

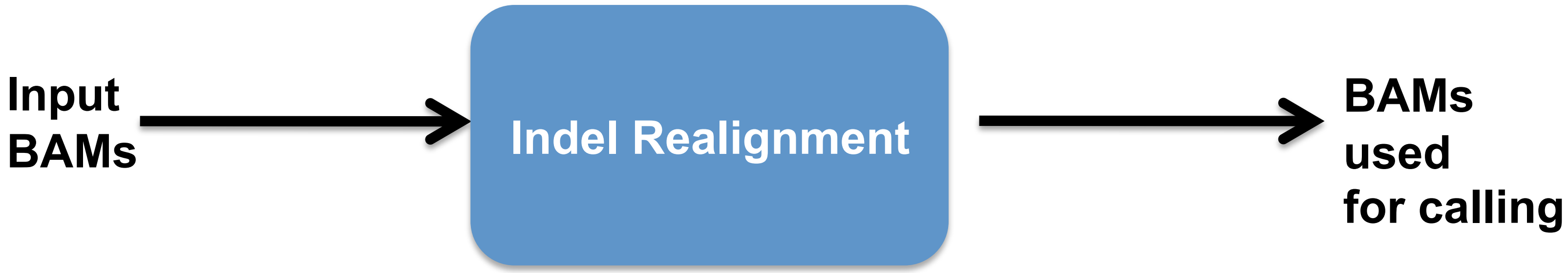
Feb 17, 2011

BROAD
INSTITUTE

What are the GATK’s indel processing abilities?

GATK Tool	Function
IndelRealigner	Runs multiple sequence alignment on reads and forms consensus indels suitable for variant genotyping.
UnifiedGenotyper	Determines consensus alternate alleles, optimal allele frequency distribution, determines whether sites should be called, assigns genotypes and annotations.
VariantFiltration	Filters calls based on given expressions.
VariantEval	Indel metrics and stratifications for analysis

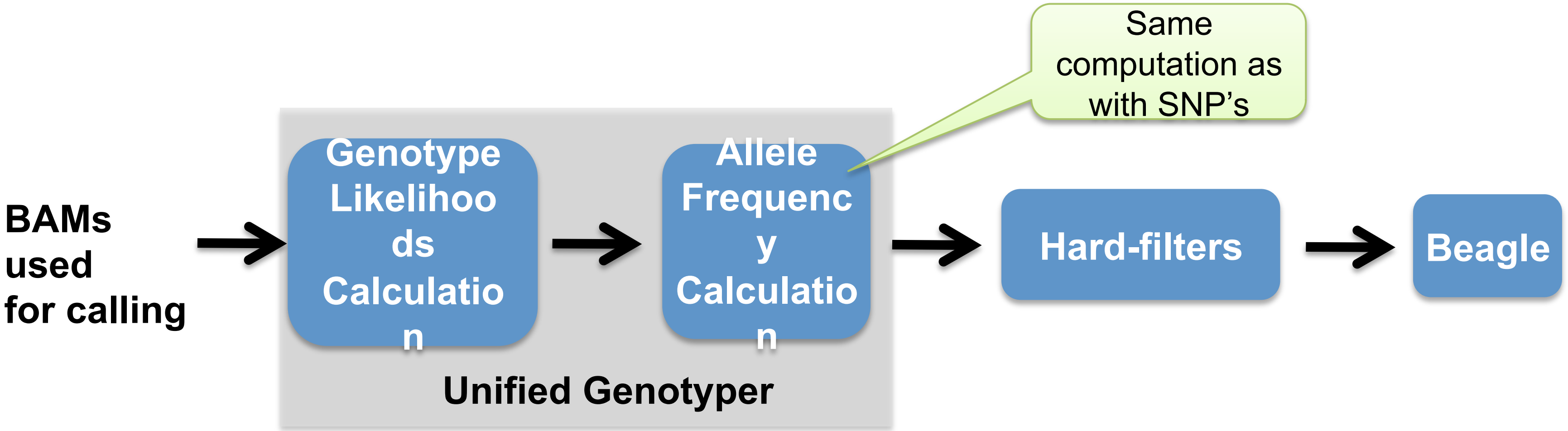
Step 1: BAM data processing



- Indel realignment is a critical step in preparing BAM’s for indel calling.
- We recommend full indel realigning (Smith Waterman) at all sites, **realignment using only known sites is not enough!**

Note: Exome BAM’s coming out of Picard have already been fully indel-realigned!

Step 2: Indel discovery



Same computation as with SNP’s

BAMs used for calling

Genotype Likelihoods Calculation

Allele Frequency Calculation

Unified Genotyper

Hard-filters

Beagle

- The genotype likelihoods calculation is inspired by Dindel (with kind permission from C Albers and R Durbin).
- Typical command line:

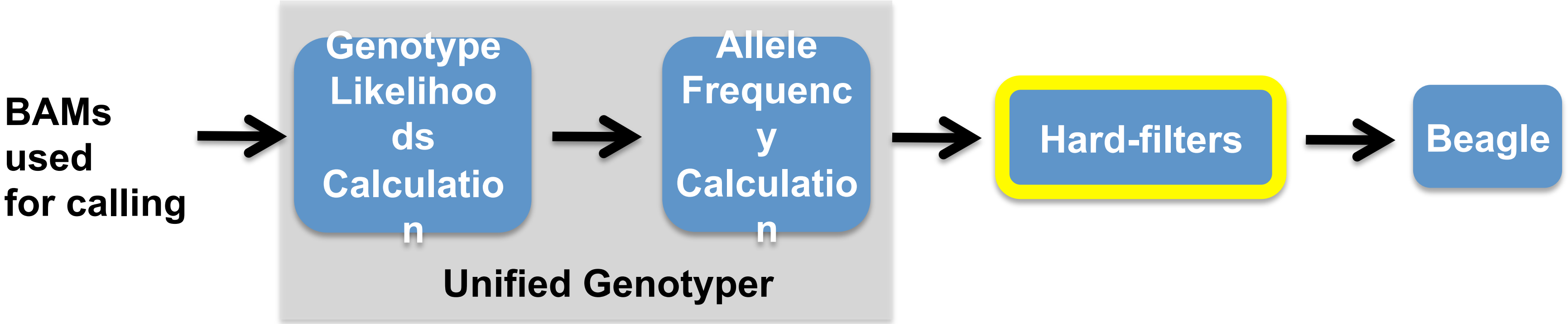
```
java -jar GenomeAnalysisTK.jar –R ref.fasta -T
  UnifiedGenotyper –L mytargets.list –I myreads.bam
  –o mycalls.vcf -B:dbsnp,VCF dbsnp.vcf -glm DINDEL
```

Only difference with SNP calling!

Some details and caveats…

- All standard parameters used in UG for SNP calling are also valid for indels!
  - E.G. `–stand_call_conf` for a calling threshold.
- Heuristic for controlling sensitivity:
  - We'll only consider indels for genotyping if they are present in N reads, controlled by `–minIndelCnt` parameter. Default value: 5, may want lower value for higher sensitivity in lowpass samples.
- Limitations:
  - Only bi-allelic sites considered. If more than 2 alt. alleles detected at a site, the one with most supporting reads taken.
- NOTE: Application of BAQ will severely degrade indel caller performance. Make sure argument `–baq` is either not included or set to OFF!

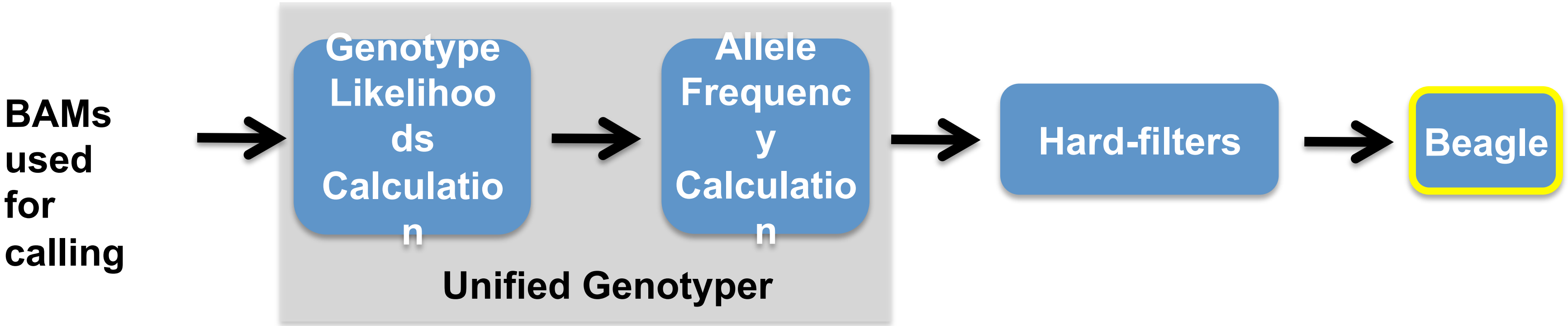
Step 3: variant filtration (indels)


BAMs used for calling
Genotype Likelihoods Calculation
Allele Frequency Calculation
Unified Genotyper
Hard-filters
Beagle

- Hard filters are needed for eliminating calls coming from read artifacts.
- This is an ongoing area of improvement, stay tuned on the GATK Wiki for best practice recommendations!
- Example command line with current best practice:

```
java –jar ./dist/GenomeAnalysisTK.jar –T VariantFiltration  ref.fasta –o
out.vcf -B:variant,VCF input.vcf \
--filterExpression "QUAL<30.0" --filterName "LowQual” \
--filterExpression "SB>=-1.0" --filterName "StrandBias” \
--filterExpression "QD<1.0" --filterName "QualByDepth” \
--filterExpression "(MQ0 >= 4 && ((MQ0 / (1.0 * DP)) > 0.1))" --filterName
"HARD_TO_VALIDATE” \
--filterExpression "HRun>=15" --filterName "HomopolymerRun"
```

Step 4 (Optional): Genotype refinement



BAMs used for calling

Genotype Likelihoods Calculation

Allele Frequency Calculation

Unified Genotyper

Hard-filters

Beagle

- Beagle can be used to refine genotypes of indel calls. Current recommended best practice is to merge Indel and SNP calls and running Beagle on combined set. More details our Wiki page.

Assessing indel callsets

- How do we know if the callset that we have is of high sensitivity and high specificity?
- How many variants should we typically get?
- How should indels be distributed in size, allele frequency and types of indels?

VariantEval’s support for Indels

```
java –jar GenomeAnalysisTK.jar -B:eval,V
CF mycalls.vcf  -T VariantEval –R reffile.fasta -EV
IndelMetricsByAC  -EV IndelStatistics -B:dbsnp,VCF
dbsnp.vcf -o output.txt
```

Key module! Produces indel size distributions as well as classification tables

This produces a GATK report file with aggregated statistics.

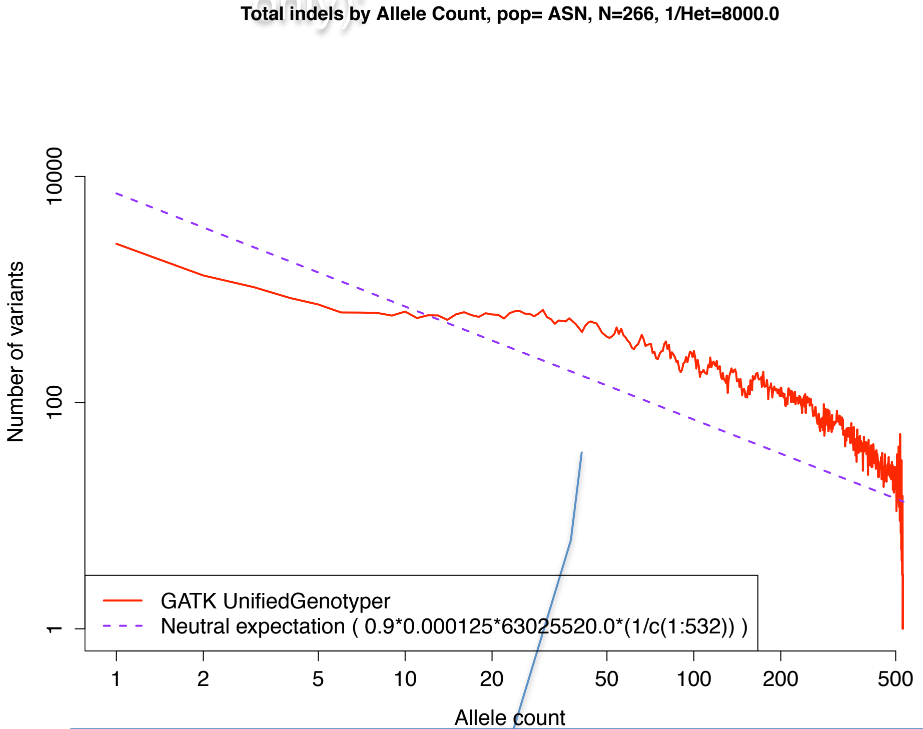
```
##:GATKReport.v0.1 CountVariants : Counts different classes of variants in the sample
CountVariants  CompRod  CpG      EvalRod  JexlExpression  Novelty  nProcessedLoci  nCalledLoci   nRefLoci  nVariantLoci  vari
antRate            variantRatePerBp  nSNPs  nInsertions  nDeletions  nComplex  nNoCalls  nHets  nHomRef  nHomVar  nSingleton
s  heterozygosity         heterozygosityPerBp  hetHomRatio         indelRate              indelRatePerBp  deletionInsertionR
atio
CountVariants  dbsnp    CpG      eval     none            all      63025520        1215          0         1215          0.00
001928             51872.00000000    0      611          604         0         0         724    0       491      0
   0.00001149             87051.00000000       1.47454175          0.00001928             51872.00000000  0.98854337

CountVariants  dbsnp    CpG      eval     none            known    63025520        1000          0         1000          0.00
001587             63025.00000000    0      491          509         0         0         567    0       433      0
   0.00000900             111156.00000000      1.30946882          0.00001587             63025.00000000  1.03665988

CountVariants  dbsnp    CpG      eval     none            novel    63025520        215           0         215           0.00
000341             293141.00000000   0      120          95          0         0         157    0       58       0
   0.00000249             401436.00000000      2.70689655          0.00000341             293141.00000000  0.79166667

CountVariants  dbsnp    all      eval     none            all      63025520        13580         0         13580         0.00
021547             4641.00000000     0      6649         6931        0         0         8852   0       4728     0
   0.00014045             7119.00000000        1.87225042          0.00021547             4641.00000000   1.04241239
```

How many indels should I get?

Published estimates In Whole Genome ~ 1 indel/8000 bp

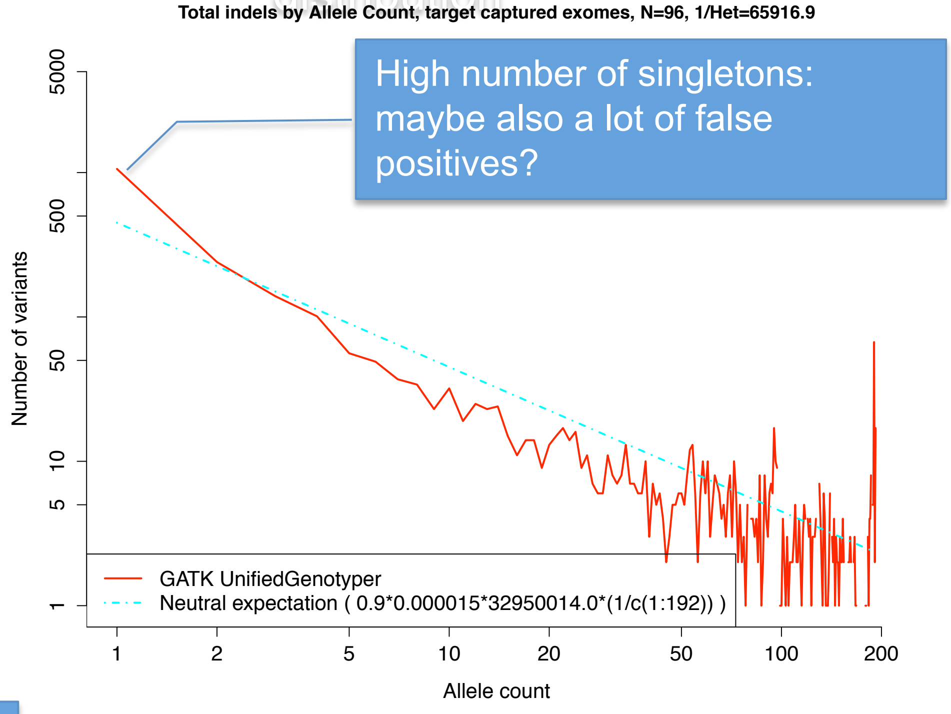
Empirical exome estimate: ~500 indels/exome (33 Mbp)

Lowpass example AC distribution (Chr20 only):



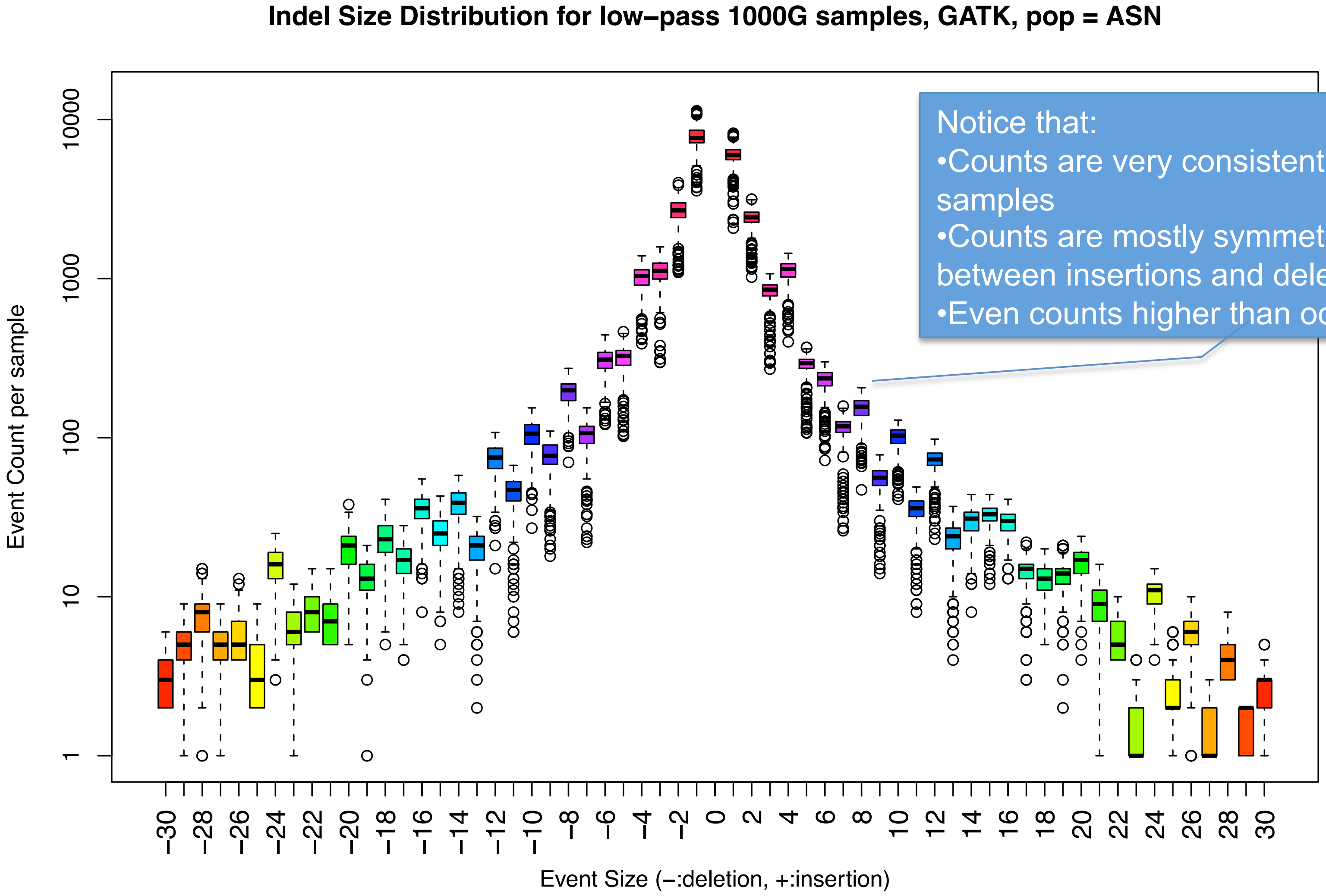
High heterozygosity in lowpass call set is leading us to focus on improving specificity of our calls.

Exome example AC distribution:



High number of singletons: maybe also a lot of false positives?

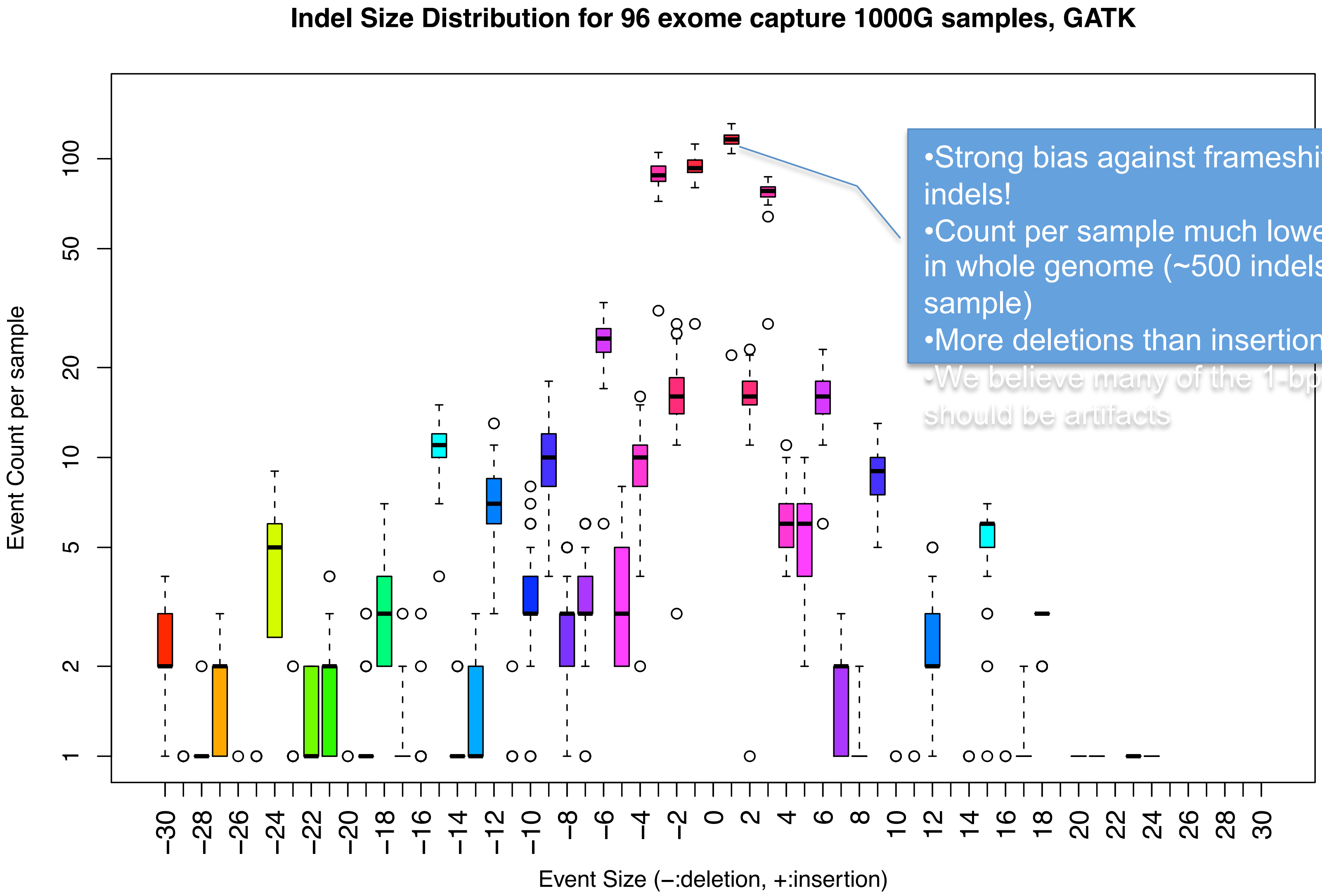
A typical plot of indel size distribution in whole genome sets



Notice that:

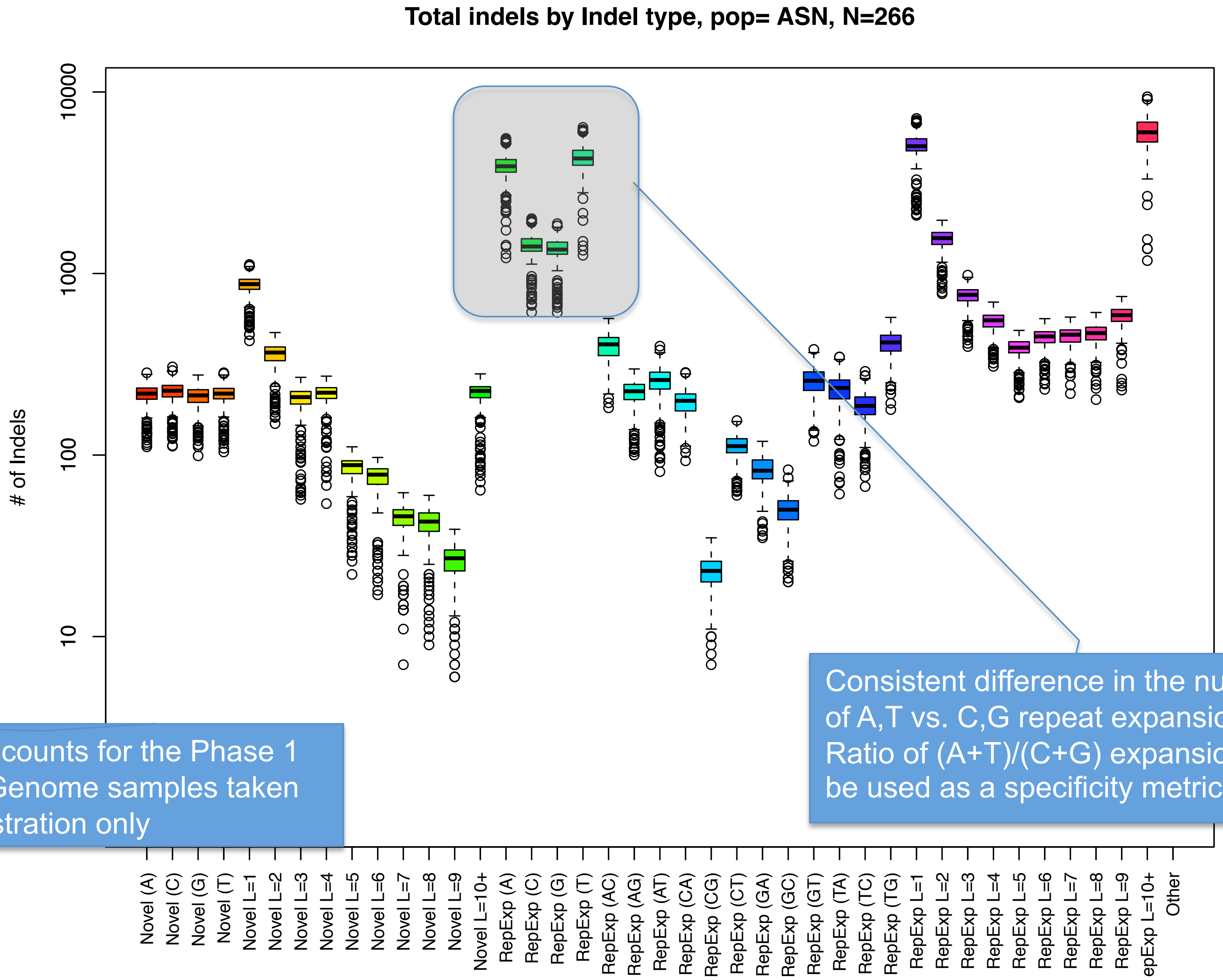
- Counts are very consistent among samples
- Counts are mostly symmetric between insertions and deletions
- Even counts higher than odd counts

Indel size distribution in exomes



- •Strong bias against frameshift indels!
- •Count per sample much lower than in whole genome (~500 indels per sample)
- •More deletions than insertions
- •We believe many of the 1-bp events should be artifacts

Different indel types come at different rates



These counts for the Phase 1 1000 Genome samples taken for illustration only

Consistent difference in the number of A,T vs. C,G repeat expansions: Ratio of (A+T)/(C+G) expansions can be used as a specificity metric!

Other callers

- Aside from the GATK, SAMTools and DINDEL can be alternatively used for indel calling.
- Example command line using SAMTools' mpileup caller:

```
samtools mpileup  -ugf ref.fasta reads.bam | ../samtools/
  bcftools/bcftools view -vc - > myout.vcf
```

- More info at:
  - http://samtools.sourceforge.net/mpileup.shtml
  - http://www.sanger.ac.uk/resources/software/dindel/