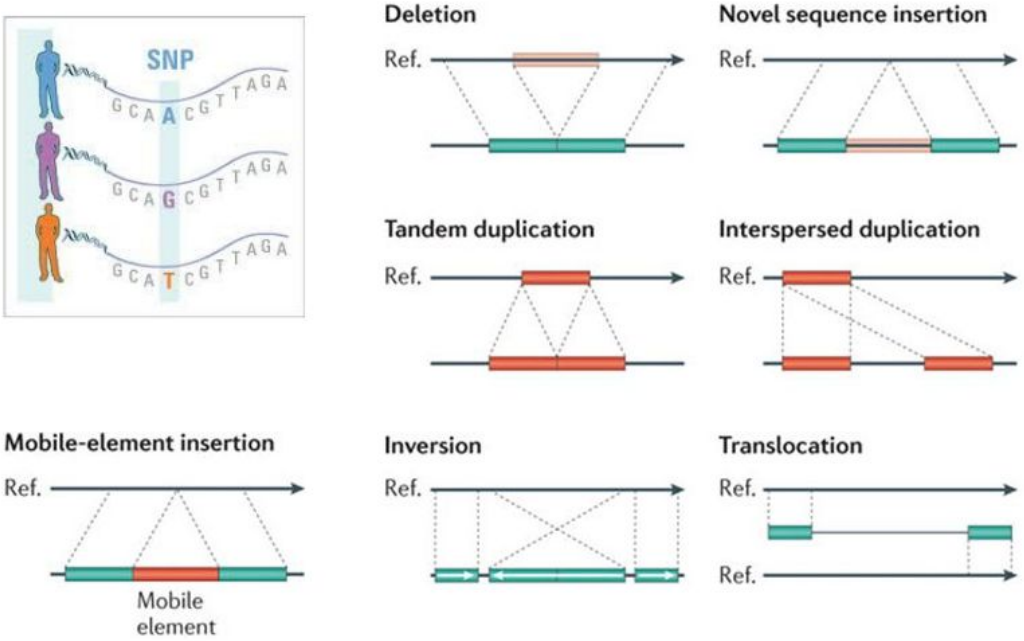
# Variant Callers

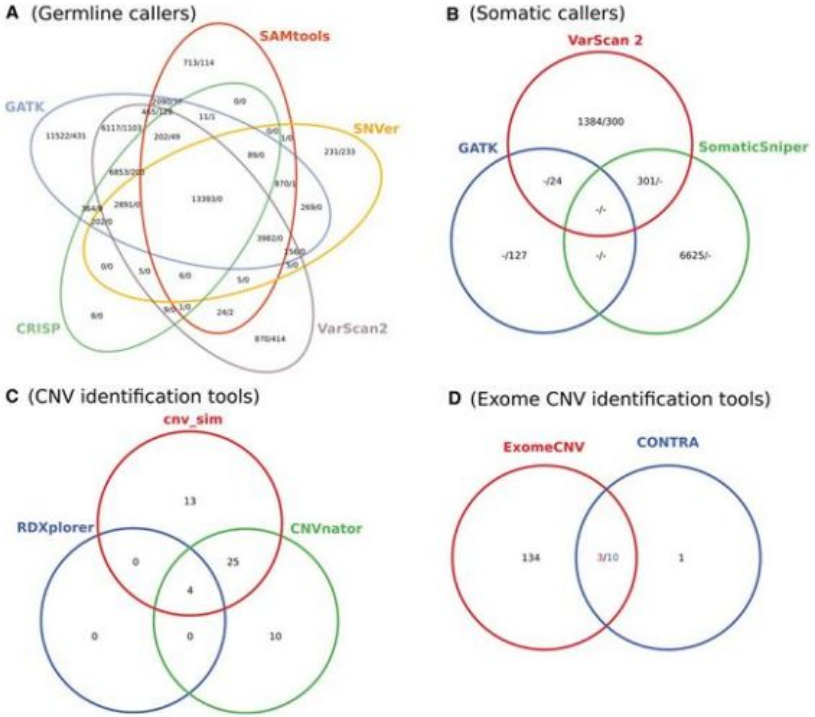
J Fass | 24 August 2017

# Variant Types




Adapted from Alkan et al, Nature Reviews Genetics 2011

# Caller Consistency




Pabinger (2014) Briefings Bioinformatics 15:256

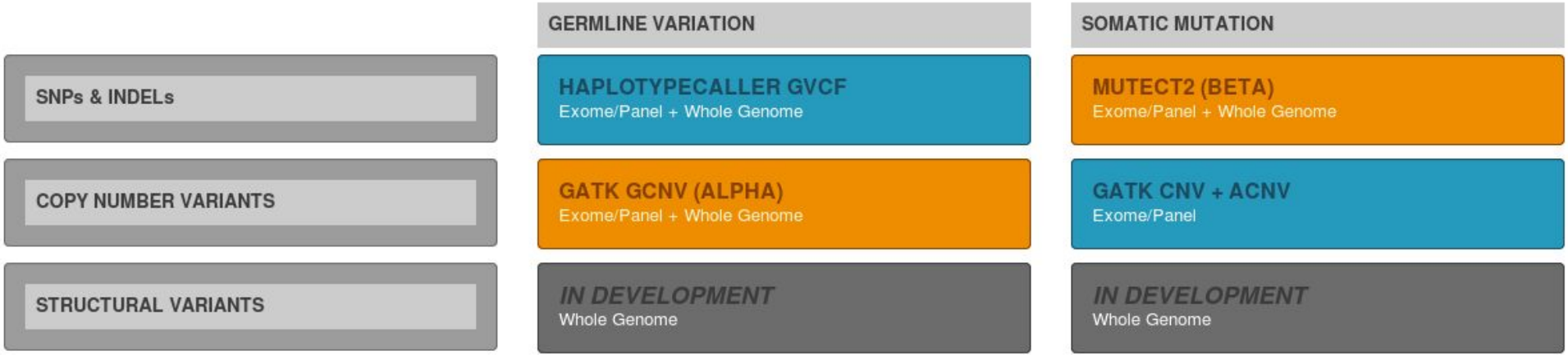
# Freebayes

Bayesian haplotype caller that can call SNPs, short CNVs / duplications, complex multi-nucleotide MNPs … smaller than the Illumina reads used in alignment. Can adapt to copy number variation (depth variation) to call haplotypes.

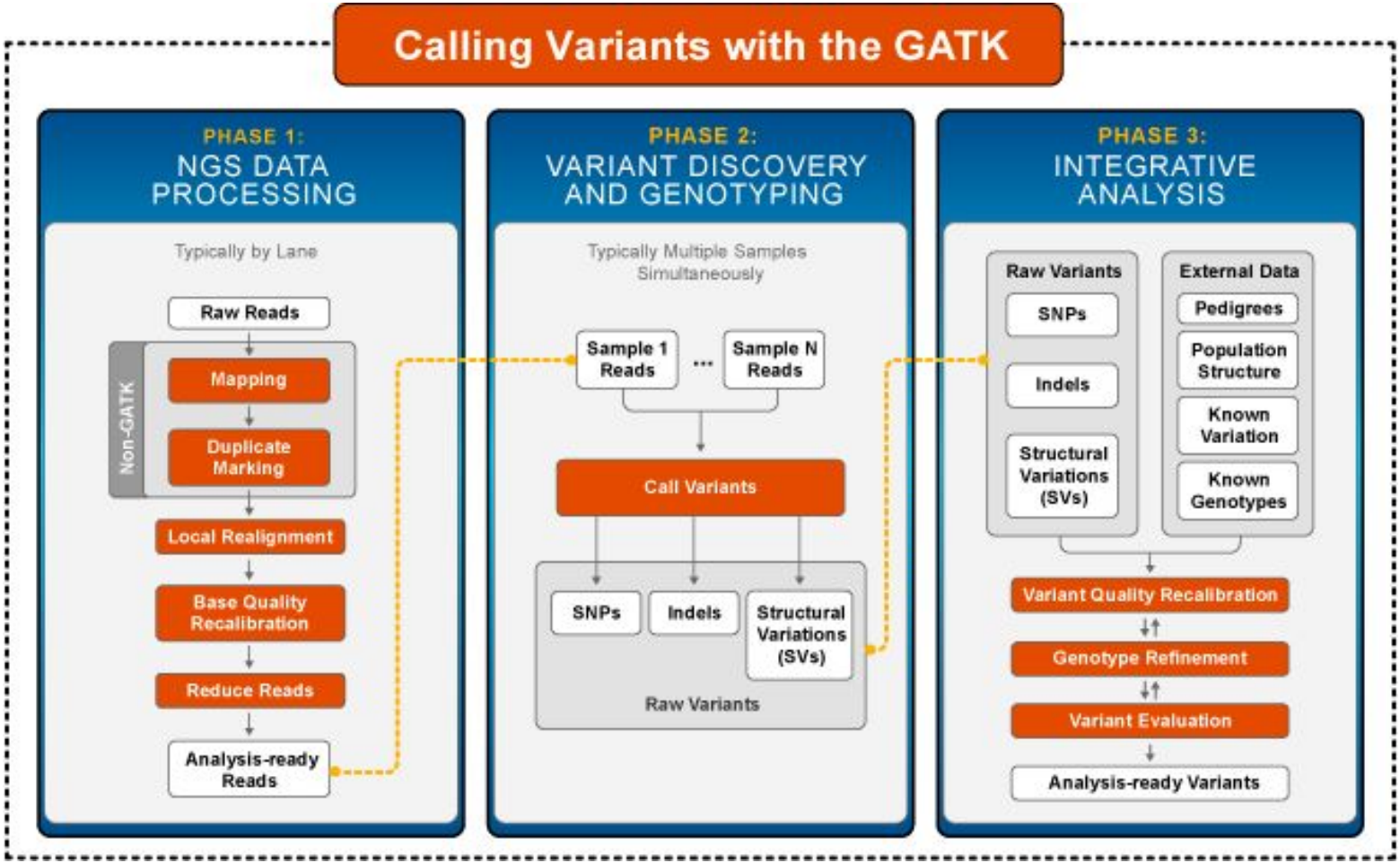
*Phases* haplotypes across multiple position (determines haplotype blocks) using Illumina reads (they must contain multiple variants).

# GATK

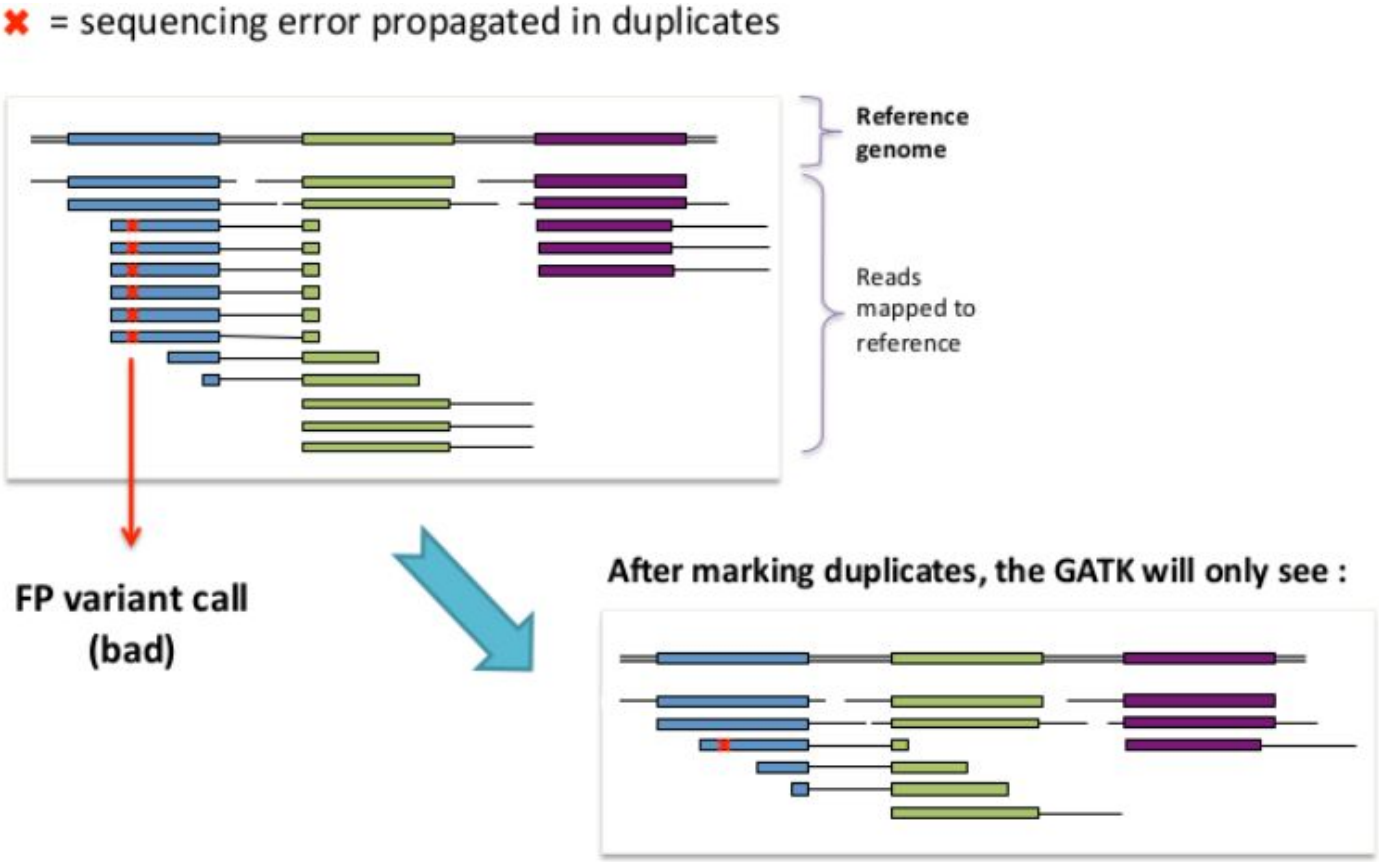
Broad Institute's variant caller, includes exhaustive pipeline of multiple types of recalibration, error correction, etc.

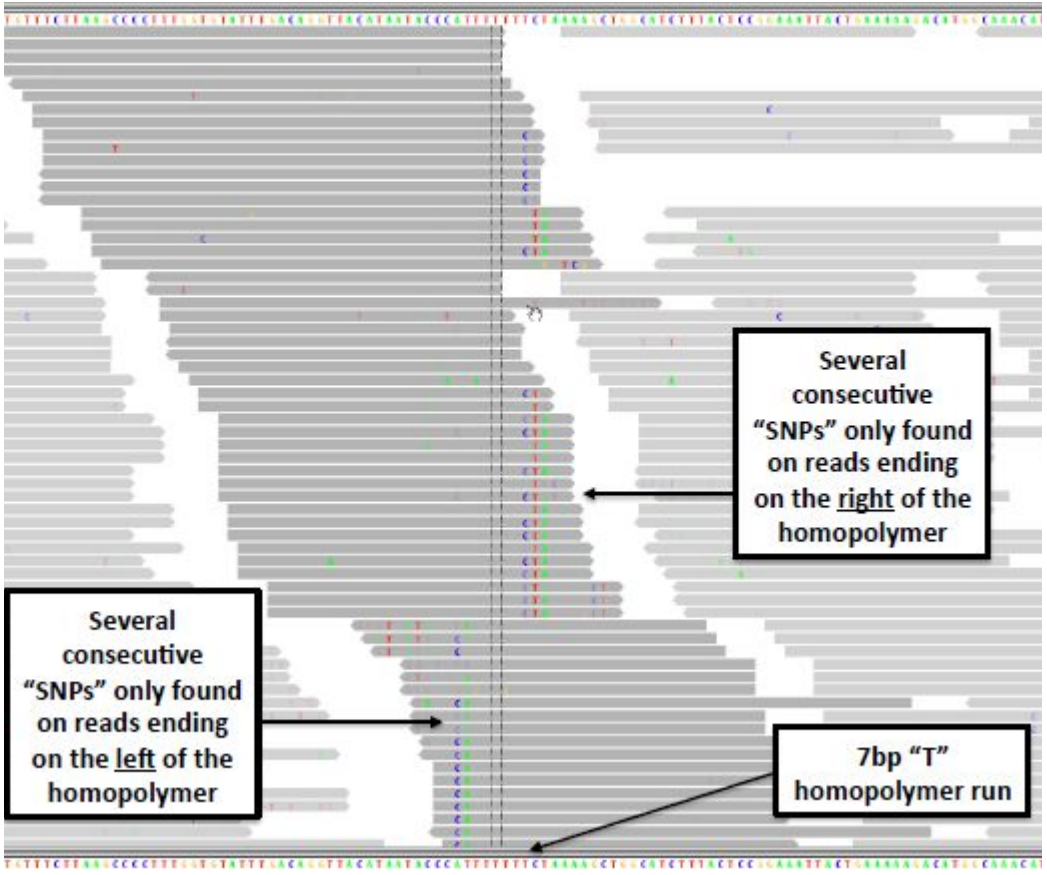


# GATK


Calling Variants with the GATK
PHASE 1:
NGS DATA PROCESSING
Typically by Lane
Raw Reads
Non-GATK
Mapping
Duplicate Marking
Local Realignment
Base Quality Recalibration
Reduce Reads
Analysis-ready Reads
PHASE 2:
VARIANT DISCOVERY AND GENOTYPING
Typically Multiple Samples Simultaneously
Sample 1 Reads
...
Sample N Reads
Call Variants
SNPs
Indels
Structural Variations (SVs)
Raw Variants
PHASE 3:
INTEGRATIVE ANALYSIS
Raw Variants
SNPs
Indels
Structural Variations (SVs)
External Data
Pedigrees
Population Structure
Known Variation
Known Genotypes
Variant Quality Recalibration
Genotype Refinement
Variant Evaluation
Analysis-ready Variants

# GATK *duplicates*






Indel Realignment

# GATK Indel Realignment

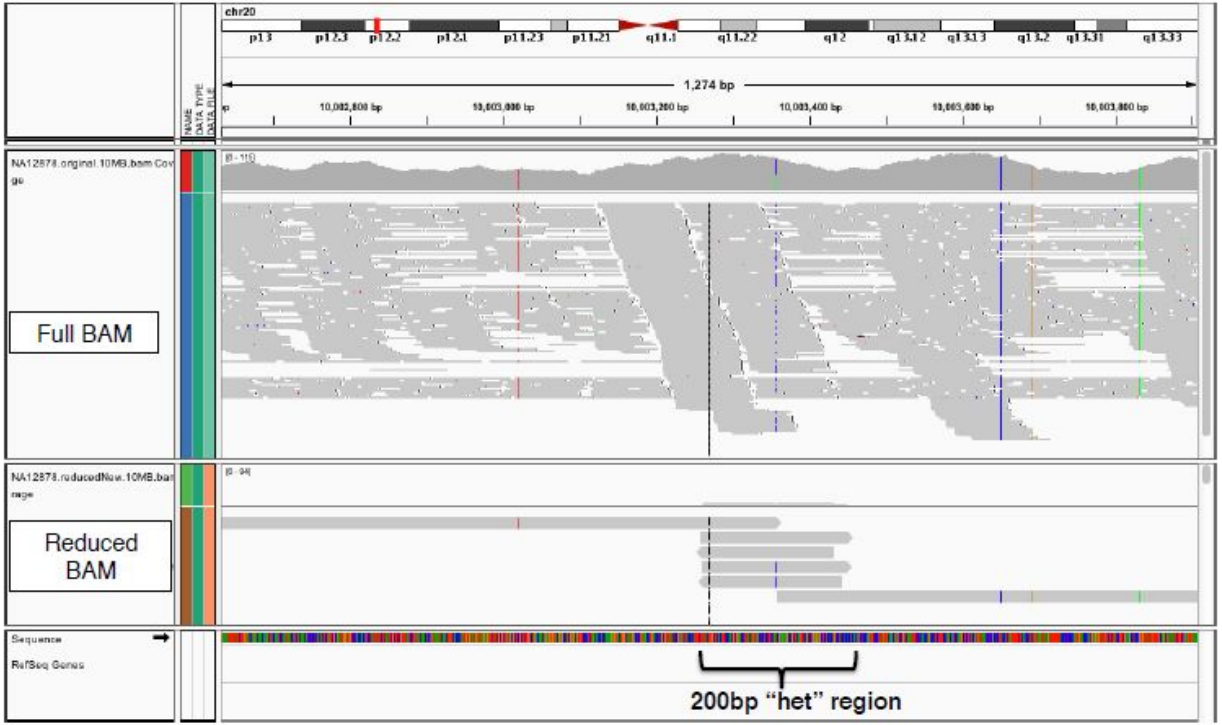
Uses information from:

- Known sites (dbSNP, 1000 Genomes)
- Indels present in original alignments (in CIGARs)
- Sites where evidence suggests a hidden indel

# GATK Base Quality Recalibration

- Critical for downstream analysis
- Scores assigned by sequencers are inaccurate and biased
- Recalibration information is obtained by analyzing covariation among several features of a base, including:
  - Reported quality score
  - Position within the read (machine cycle)
  - Preceding and current nucleotide (sequencing chemistry effect)
  - Known variants are used to discount most of the real genetic variation present in the sample
  - All other differences from the reference are assumed to be sequencing errors
  - Indel Realignments first reduces noise from misalignments

# GATK Read Compression



# GATK Haplotype Caller

- Calls SNPs, indels, and some structural variants simultaneously by performing a local de-novo assembly
- Distinguishes genetic variant and random machine noise
- Uses “active regions” for variant calling, based on significant evidence for variation
- Determines likelihoods of the haplotypes given the read data
- Assigns sample genotypes based on Bayesian likelihoods

# GATK Variant Quality Score Recalibration (VQSR)

- Initial variant calling has very large set that is full of false positives
- Hand-tuned filtering requires time and expertise
- Statistical model could be used to recalibrate variants
- Each variant has a set of statistics associated with them that are called variant annotations
- Real variants tend to cluster together via these statistics
- SNPs and indels must be recalibrated separately
- Training resources:
  - SNP (HapMap, Omni, 1000G, dbSNP)
  - INDEL (Mills)

# GATK gVCFs for incremental calling

- VCF files with information for every position in the genome regardless of variant calls
- Used by GATK to perform variant discovery in a way that enables joint analysis of multiple samples, but decoupled from the initial individual variant calling step. I.e. you don't have to call variants on all your samples together to perform a joint analysis
- Drastically reduces run time and allows for easy incorporation of additional samples into the pipeline

# GATK Hard Filtering

- Based on some criteria relevant to your research
- Useful for:
  - Small data sets in terms of both number of samples or size of targeted regions
  - No database available with high confidence known variants

For SNPs:

QD < 2.0
MQ < 40.0
FS > 60.0
MQRankSum < -12.5
ReadPosRankSum < -8.0

For indels:

QD < 2.0
ReadPosRankSum < -20.0
InbreedingCoeff < -0.8
FS > 200.0

# Delly, Lumpy, Whamg ...

Delly, https://github.com/dellytools/delly

Lumpy, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4197822/

https://github.com/zeeev/wham … multiple SV types (deletions, duplications, inversions, small insertions > 50bp) from Illumina read pairs

Note tool to merge multiple sets (different callers) of SV calls: mergeSVcallers

# Genome STRIP (STRucture In Populations)

http://software.broadinstitute.org/software/genomestrip/

Targets large Multiallelic Copy Number Variations (mCNVs); implicitly duplications, deletions.

# CNVnator

https://github.com/abyzovlab/CNVnator

Another CNV caller; works off of read depth signal to call regions of fold depth changes.

# Sniffles

**Sniffles** … (larger SVs) from PacBio or Oxford Nanopore reads

(makes use of “cleaner” alignments from NGM-LR)