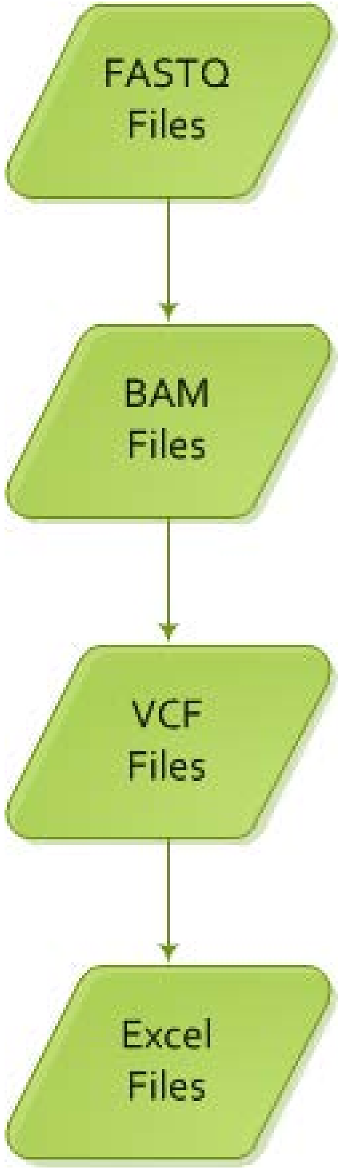
Variant Detection in Next Generation Sequencing Data

John Osborne
Sept 14, 2012

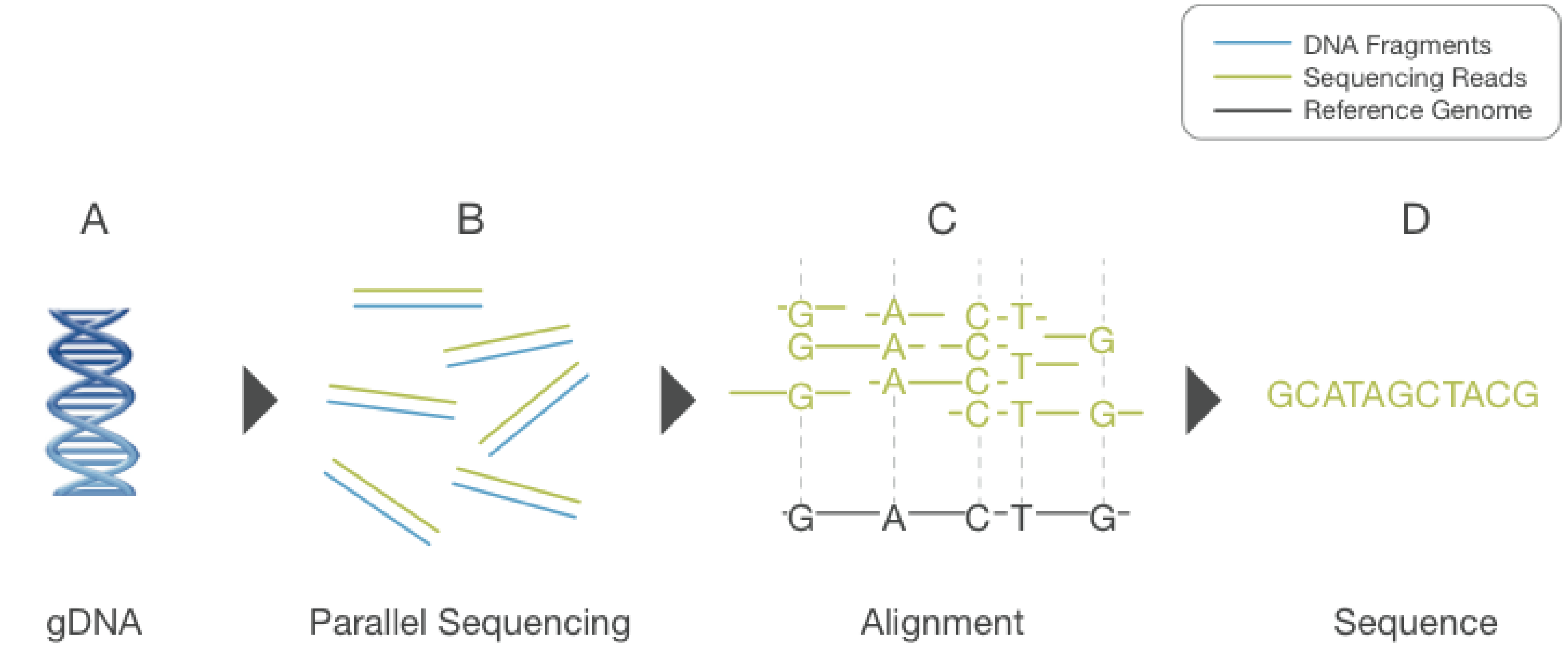
Overview

- My Bias
  - Talk slanted towards analyzing whole genomes using Illumina paired end reads with open source tools
- Background
- Alignment Software
- Detecting Variation
  - Nucleotide
  - Structural
- Analyzing and Interpreting Variation
- *Best practices change extremely rapidly*



Next Generation Sequencing

Figure 1: Conceptual Overview of Whole-Genome Resequencing

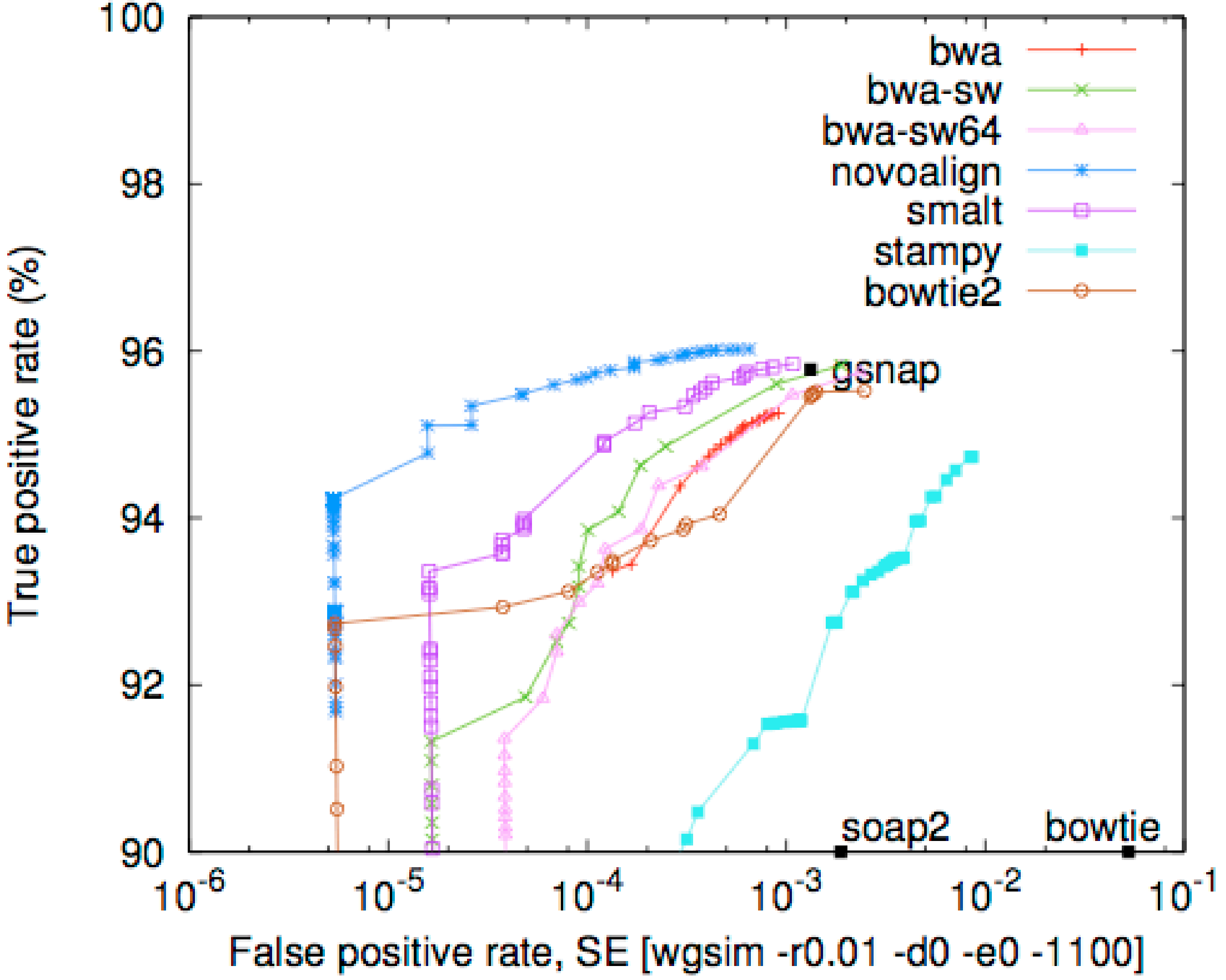


A. Extracted gDNA.
B. gDNA is fragmented into a library of small segments that are each sequenced in parallel.
C. Individual sequence reads are reassembled by aligning to a reference genome.
D. The whole-genome sequence is derived from the consensus of aligned reads.

Taken from Illumina website

Short Read Alignment

- *Making comparisons is very difficult!*
- Test Parameters
  - Read length size
  - Introduced errors
  - Paired versus single end reads
- Metrics
  - Discovery? Accuracy? Area under curve?
  - What is correct?
- Downstream analysis
- Comparisons are time consuming to do and are typically only done when somebody releases a new aligner

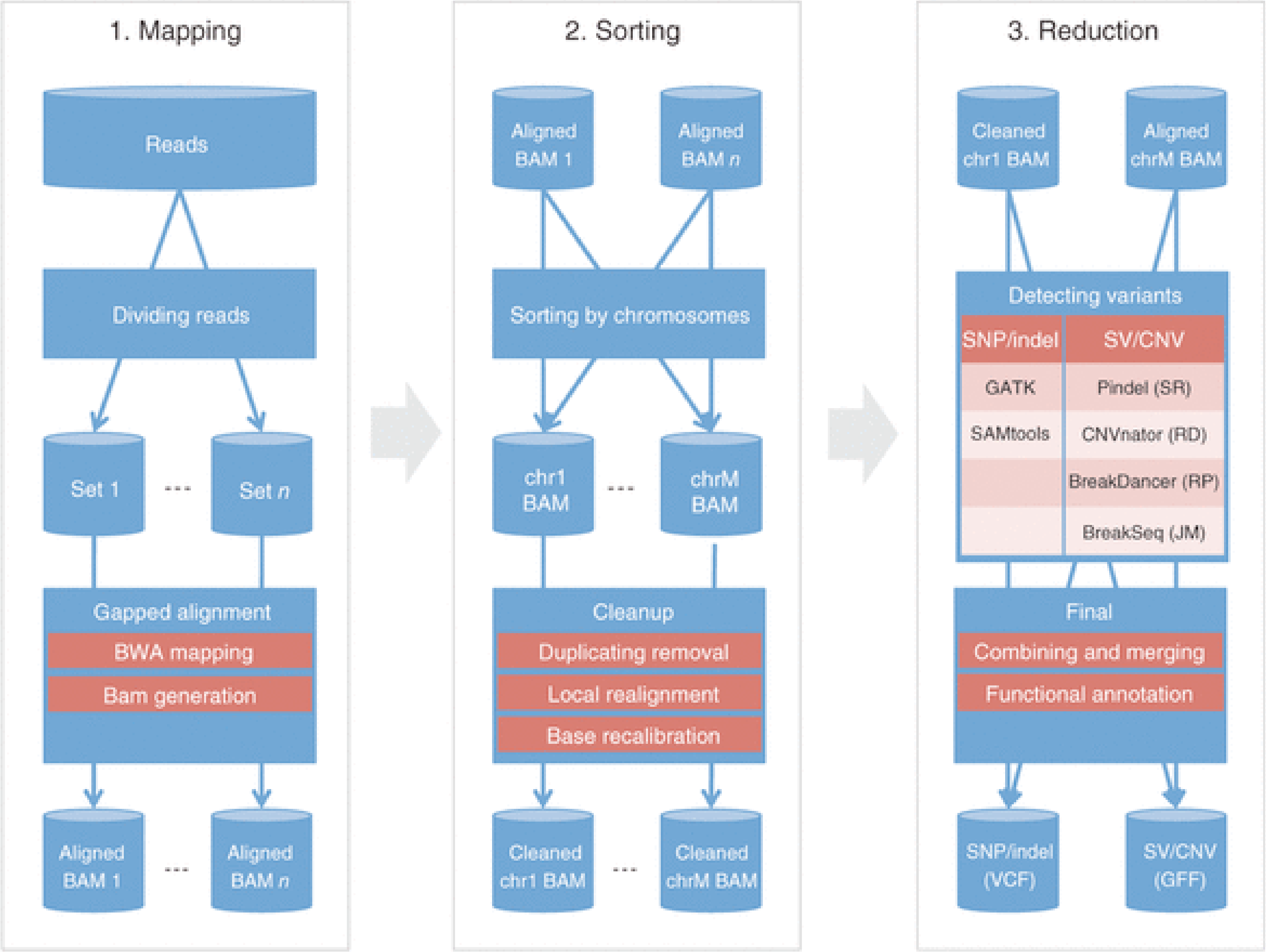


From homepage of Heng Li: http://lh3lh3.users.sourceforge.net/alnROC.shtml

Short Read Aligner Conclusions

- The differences between aligners are not that large anymore
  - BWA, Bowtie2 are all available on cheaha
- I currently recommend BWA, but I suspect it will be supplanted by something else
  - Bowtie2
  - Novoalign
  - SeqAlto or something newer
- For longer reads (>=200bp) I would recommend BWA-SW, Bowtie2 (long read version) or CUSHAW2 (new)
- Select your aligner based on your **downstream workflow**, for example use of BWA is recommended by GATK

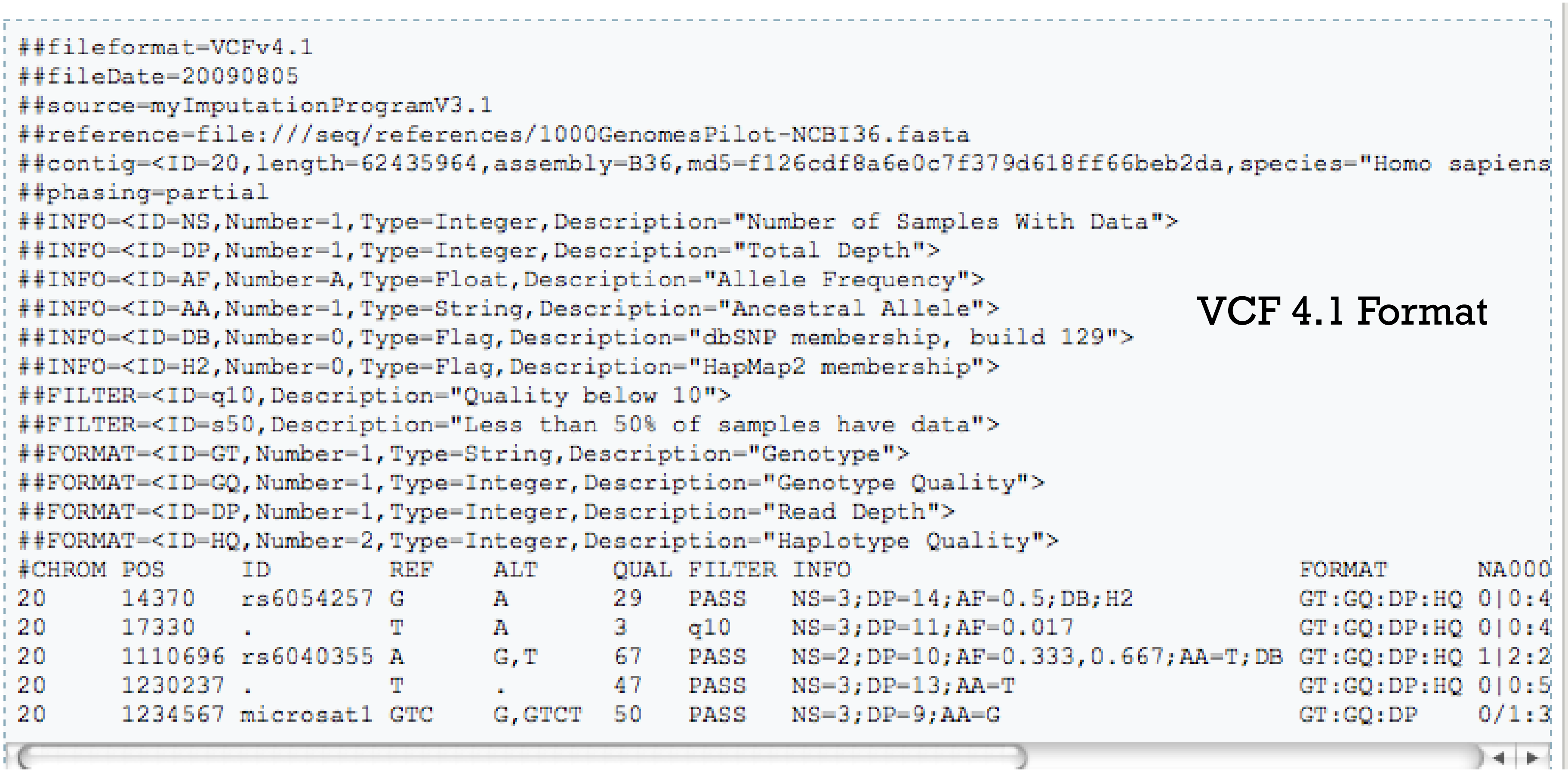
HugeSeq Workflow



From "Detecting and annotating genetic variations using the HugeSeq pipeline" Lam et al., 2012

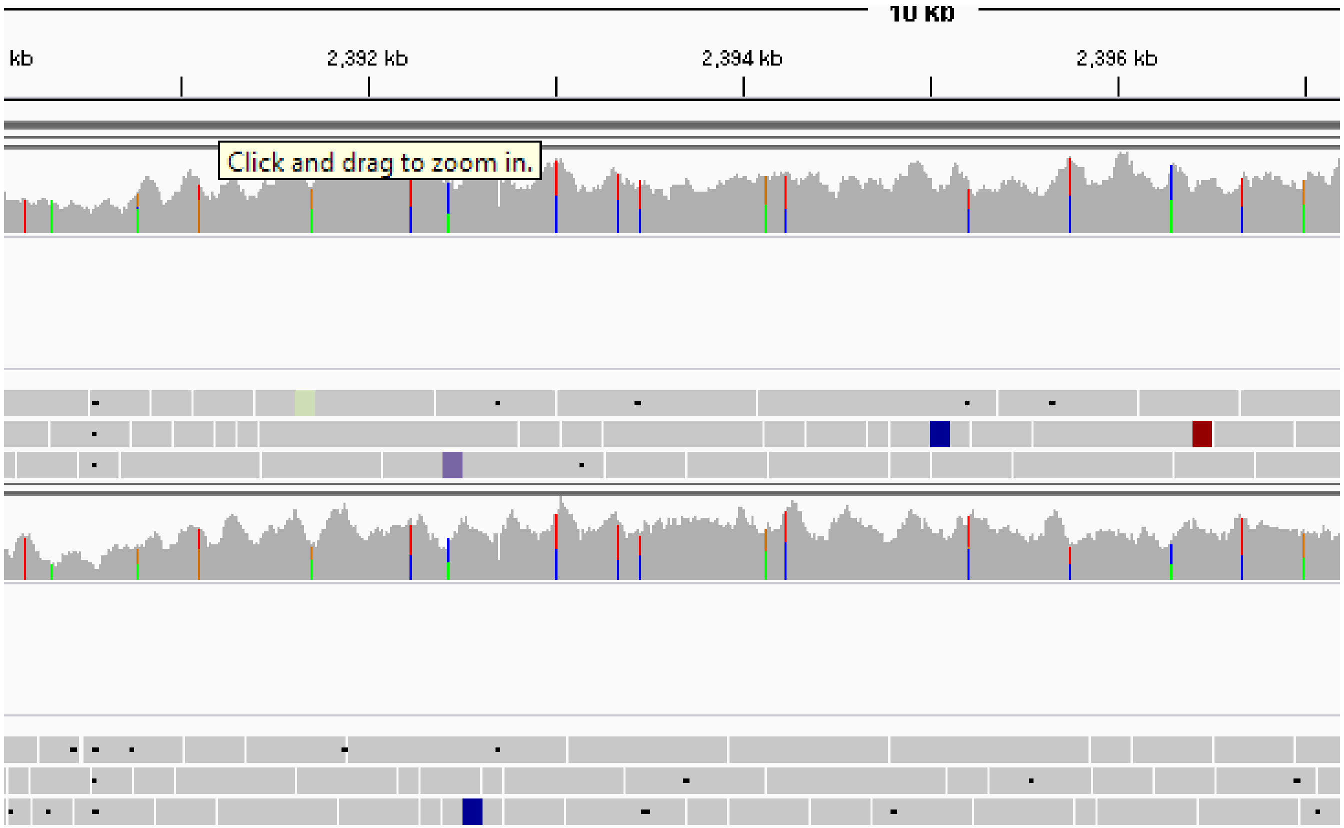
Variation Detection

- Nucleotide Polymorphisms
- "Structural Variants" / Rearrangements



```
##fileformat=VCFv4.1
##fileDate=20090805
##source=myImputationProgramV3.1
##reference=file:///seq/references/1000GenomesPilot-NCBI36.fasta
##contig=<ID=20,length=62435964,assembly=B36,md5=f126cdf8a6e0c7f379d618ff66beb2da,species="Homo sapiens
##phasing=partial
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of Samples With Data">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth">
##INFO=<ID=AF,Number=A,Type=Float,Description="Allele Frequency">
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=DB,Number=0,Type=Flag,Description="dbSNP membership, build 129">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##FILTER=<ID=q10,Description="Quality below 10">
##FILTER=<ID=s50,Description="Less than 50% of samples have data">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##FORMAT=<ID=HQ,Number=2,Type=Integer,Description="Haplotype Quality">
#CHROM POS     ID         REF    ALT     QUAL FILTER INFO                              FORMAT      NA000
20     14370   rs6054257  G      A       29   PASS   NS=3;DP=14;AF=0.5;DB;H2           GT:GQ:DP:HQ 0|0:4
20     17330   .          T      A       3    q10    NS=3;DP=11;AF=0.017               GT:GQ:DP:HQ 0|0:4
20     1110696 rs6040355  A      G,T     67   PASS   NS=2;DP=10;AF=0.333,0.667;AA=T;DB GT:GQ:DP:HQ 1|2:2
20     1230237 .          T      .       47   PASS   NS=3;DP=13;AA=T                   GT:GQ:DP:HQ 0|0:5
20     1234567 microsat1  GTC    G,GTCT  50   PASS   NS=3;DP=9;AA=G                    GT:GQ:DP    0/1:3
```

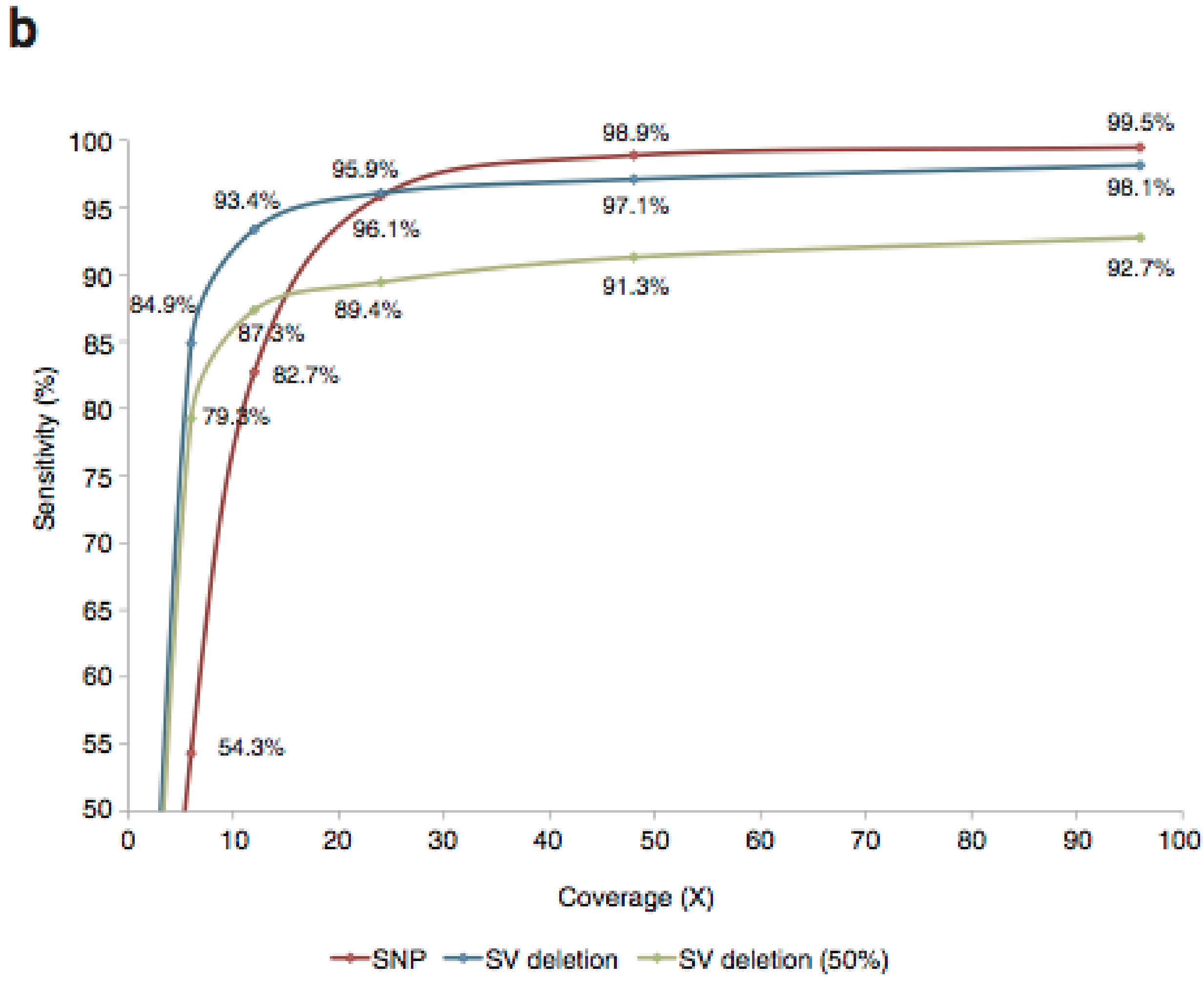
VCF 4.1 Format



SNP Detection

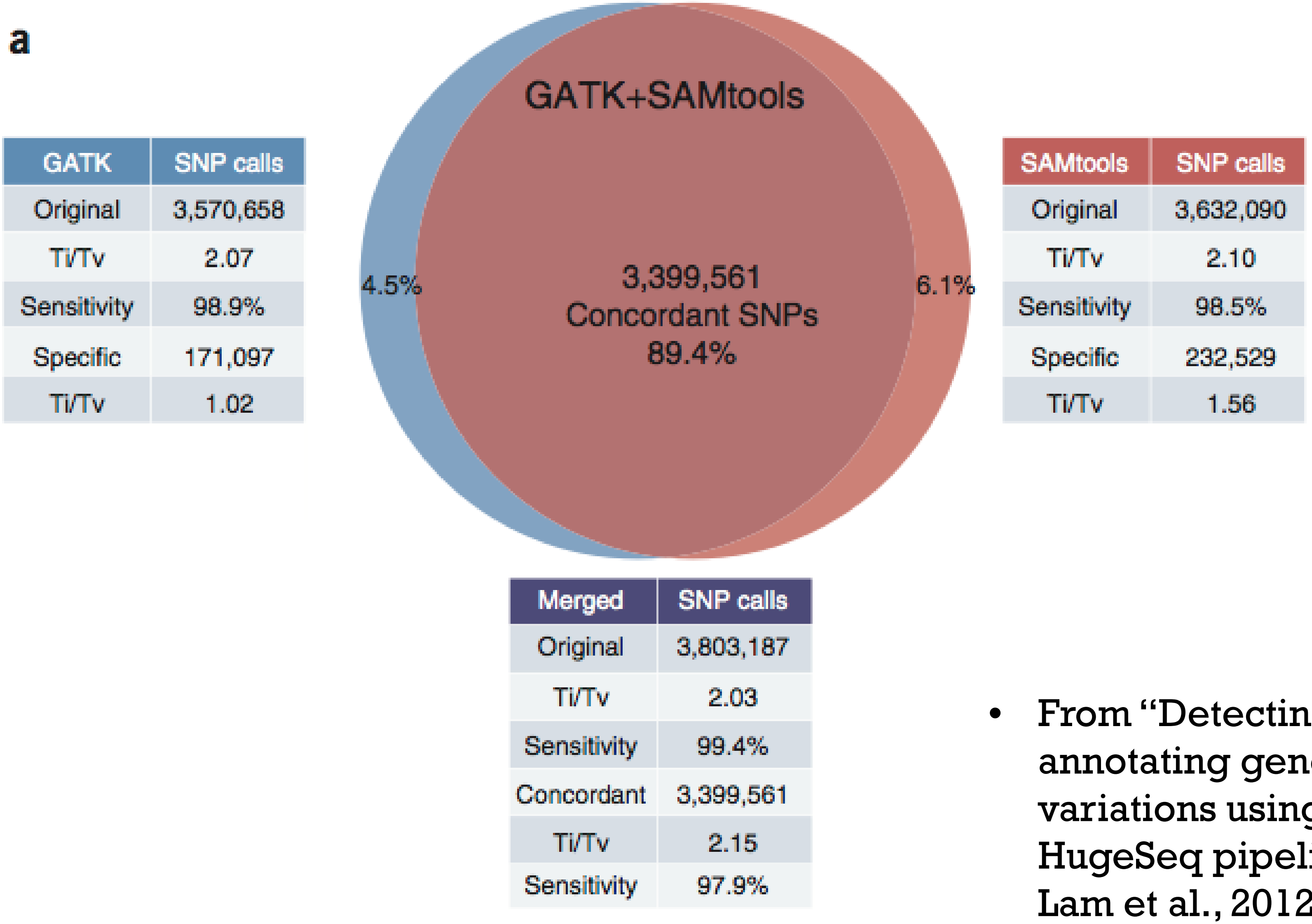
- Most advanced and reliable variant detection
  - New version of GATK can detect MNPs as well
- Coverage and Toolkit matter
- Problem isn't finding SNPs, it is finding the right SNPs

Coverage versus Sensitivity



- From "Detecting and annotating genetic variations using the HugeSeq pipeline" Lam et al., 2012

GATK and SAMTools Variant Calling



GATK	SNP calls
Original	3,570,658
Ti/Tv	2.07
Sensitivity	98.9%
Specific	171,097
Ti/Tv	1.02

SAMtools	SNP calls
Original	3,632,090
Ti/Tv	2.10
Sensitivity	98.5%
Specific	232,529
Ti/Tv	1.56

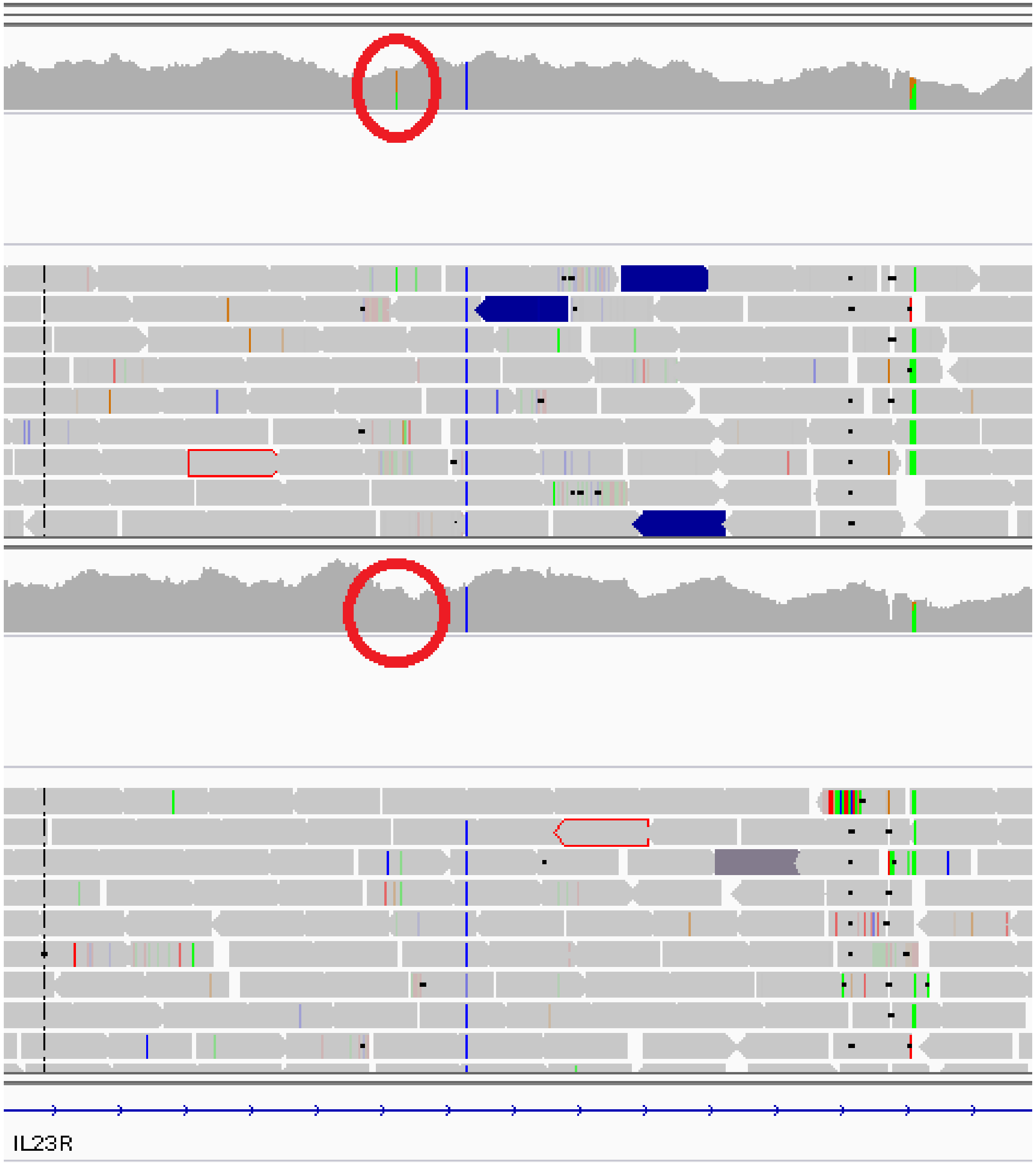
Merged	SNP calls
Original	3,803,187
Ti/Tv	2.03
Sensitivity	99.4%
Concordant	3,399,561
Ti/Tv	2.15
Sensitivity	97.9%

- From "Detecting and annotating genetic variations using the HugeSeq pipeline" Lam et al., 2012

False Positives

-Data from human monozygotic twins

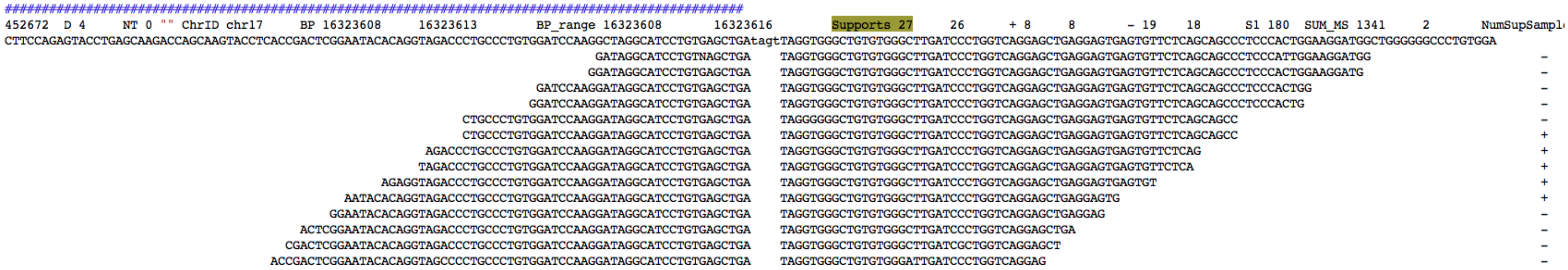
-Artifacts from borderline low coverage, top twin has 17 high quality reads (7 A) and the bottom has 23 high quality reads (2 A)



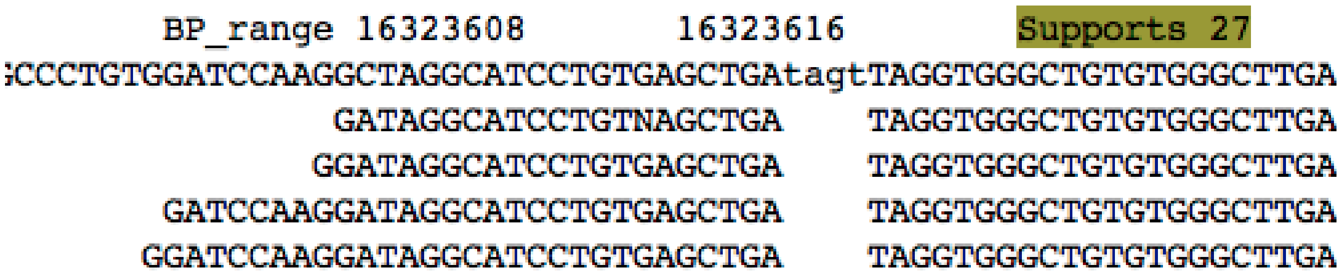
Structural Variation

- Methods
  - Small indels within single reads (GATK)
  - Discordant paired-end reads (Breakdancer, VariationHunter)
  - Depth of coverage (CNVnator, SegSeq)
  - Split reads (Pindel, ClipCrop)
- Very active area of research
  - Combined approaches becoming more common

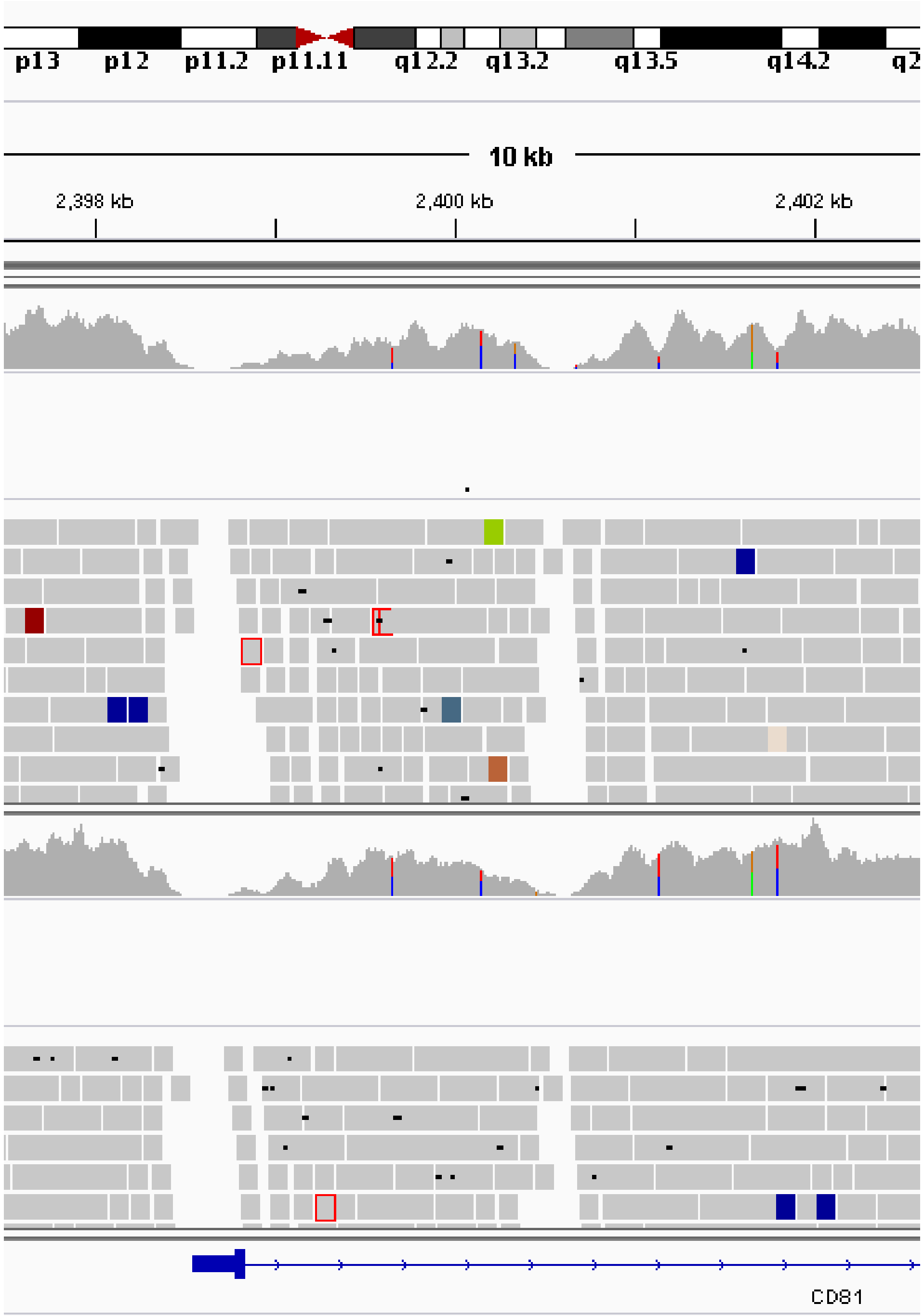
PINDEL Sample Output



```
#################################################################################################
452672  D 4     NT 0 "" ChrID chr17     BP 16323608     16323613        BP_range 16323608       16323616        Supports 27     26      + 8     8       - 19    18      S1 180  SUM_MS 1341     2       NumSupSampl
CTTCCAGAGTACCTGAGCAAGACCAGCAAGTACCTCACCGACTCGGAATACACAGGTAGACCCTGCCCTGTGGATCCAAGGCTAGGCATCCTGTGAGCTGAtagtTAGGTGGGCTGTGTGGGCTTGATCCCTGGTCAGGAGCTGAGGAGTGAGTGTTCTCAGCAGCCCTCCCACTGGAAGGATGGCTGGGGGCCCTGTGGA
                                                                                GATAGGCATCCTGTNAGCTGA    TAGGTGGGCTGTGTGGGCTTGATCCCTGGTCAGGAGCTGAGGAGTGAGTGTTCTCAGCAGCCCTCCCATTGGAAGGATGG                      -
                                                                               GGATAGGCATCCTGTGAGCTGA    TAGGTGGGCTGTGTGGGCTTGATCCCTGGTCAGGAGCTGAGGAGTGAGTGTTCTCAGCAGCCCTCCCACTGGAAGGATG                       -
                                                                        GATCCAAGGATAGGCATCCTGTGAGCTGA    TAGGTGGGCTGTGTGGGCTTGATCCCTGGTCAGGAGCTGAGGAGTGAGTGTTCTCAGCAGCCCTCCCACTGG                              -
                                                                       GGATCCAAGGATAGGCATCCTGTGAGCTGA    TAGGTGGGCTGTGTGGGCTTGATCCCTGGTCAGGAGCTGAGGAGTGAGTGTTCTCAGCAGCCCTCCCACTG                               -
                                                              CTGCCCTGTGGATCCAAGGATAGGCATCCTGTGAGCTGA    TAGGGGGGCTGTGTGGGCTTGATCCCTGGTCAGGAGCTGAGGAGTGAGTGTTCTCAGCAGCC                                        -
                                                              CTGCCCTGTGGATCCAAGGATAGGCATCCTGTGAGCTGA    TAGGTGGGCTGTGTGGGCTTGATCCCTGGTCAGGAGCTGAGGAGTGAGTGTTCTCAGCAGCC                                        +
                                                         AGACCCTGCCCTGTGGATCCAAGGATAGGCATCCTGTGAGCTGA    TAGGTGGGCTGTGTGGGCTTGATCCCTGGTCAGGAGCTGAGGAGTGAGTGTTCTCAG                                             +
                                                        TAGACCCTGCCCTGTGGATCCAAGGATAGGCATCCTGTGAGCTGA    TAGGTGGGCTGTGTGGGCTTGATCCCTGGTCAGGAGCTGAGGAGTGAGTGTTCTCA                                              +
                                                   AGAGGTAGACCCTGCCCTGTGGATCCAAGGATAGGCATCCTGTGAGCTGA    TAGGTGGGCTGTGTGGGCTTGATCCCTGGTCAGGAGCTGAGGAGTGAGTGT                                                   +
                                              AATACACAGGTAGACCCTGCCCTGTGGATCCAAGGATAGGCATCCTGTGAGCTGA    TAGGTGGGCTGTGTGGGCTTGATCCCTGGTCAGGAGCTGAGGAGTG                                                        +
                                            GGAATACACAGGTAGACCCTGCCCTGTGGATCCAAGGATAGGCATCCTGTGAGCTGA    TAGGTGGGCTGTGTGGGCTTGATCCCTGGTCAGGAGCTGAGGAG                                                          -
                                        ACTCGGAATACACAGGTAGACCCTGCCCTGTGGATCCAAGGATAGGCATCCTGTGAGCTGA    TAGGTGGGCTGTGTGGGCTTGATCCCTGGTCAGGAGCTGA                                                              -
                                      CGACTCGGAATACACAGGTAGACCCTGCCCTGTGGATCCAAGGATAGGCATCCTGTGAGCTGA    TAGGTGGGCTGTGTGGGCTTGATCGCTGGTCAGGAGCT                                                                -
                                    ACCGACTCGGAATACACAGGTAGCCCCTGCCCTGTGGATCCAAGGATAGGCATCCTGTGAGCTGA    TAGGTGGGCTGTGTGGGATTGATCCCTGGTCAGGAG                                                                  -
```



```
        BP_range 16323608       16323616        Supports 27
GCCCTGTGGATCCAAGGCTAGGCATCCTGTGAGCTGAtagtTAGGTGGGCTGTGTGGGCTTGA
                GATAGGCATCCTGTNAGCTGA    TAGGTGGGCTGTGTGGGCTTGA
               GGATAGGCATCCTGTGAGCTGA    TAGGTGGGCTGTGTGGGCTTGA
        GATCCAAGGATAGGCATCCTGTGAGCTGA    TAGGTGGGCTGTGTGGGCTTGA
       GGATCCAAGGATAGGCATCCTGTGAGCTGA    TAGGTGGGCTGTGTGGGCTTGA
```



Neither deletion was detected by Pindel..

Pitfalls of Structural Variant Detection with NGS

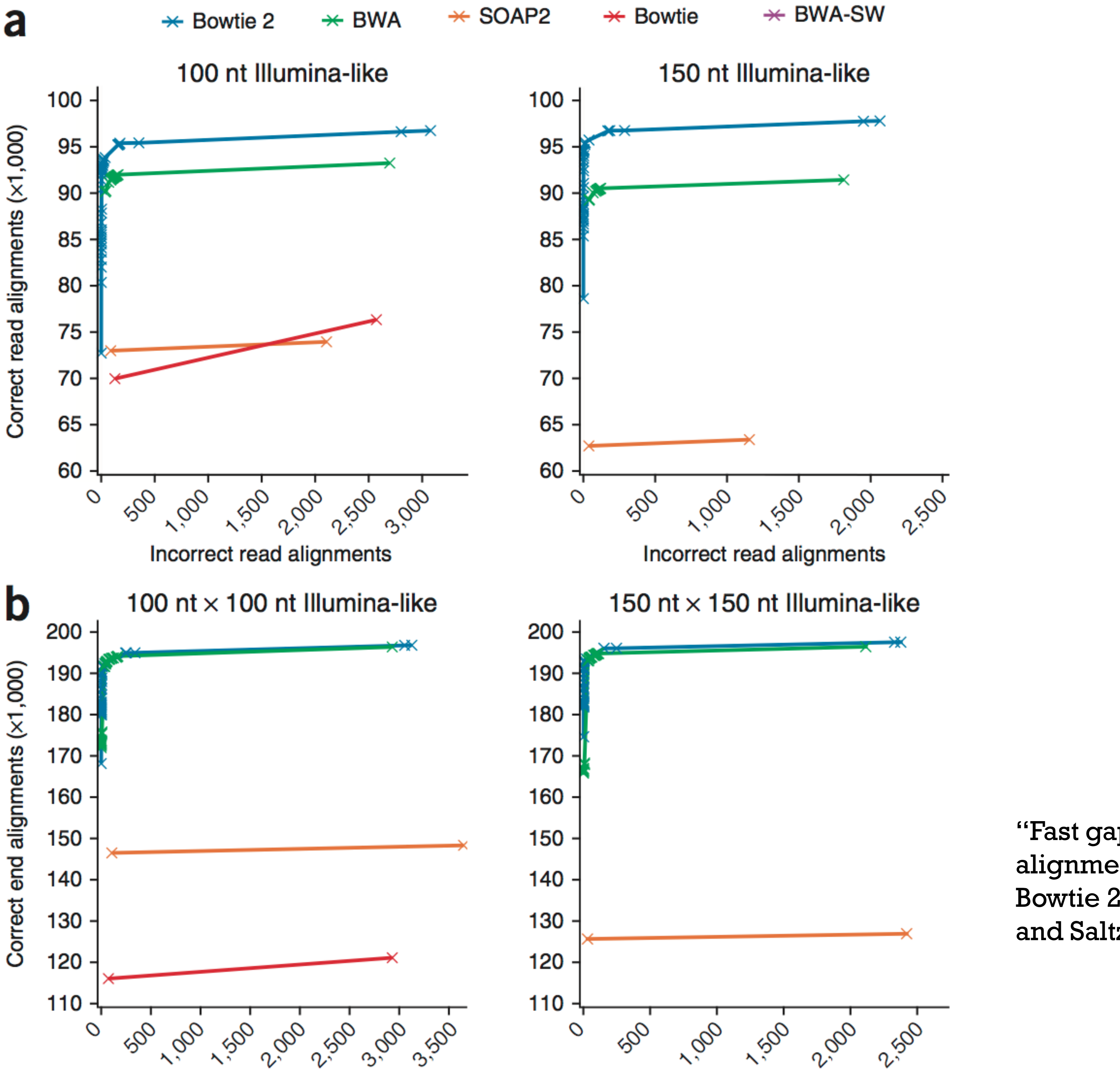
- Tips
  - Get as much coverage as possible
    - Not possible to find breakpoints with 5 fold coverage
  - Use multiple approaches
  - Remove duplicates
  - If it is important and you have time… look
    - In twin study, only 2 out of 12 SVs found by Pindel
- Personal Bias
  - GATK (small indels), Breakdancer (rearrangements), Pindel (split reads) and CNVator (repeat size estimation)

+

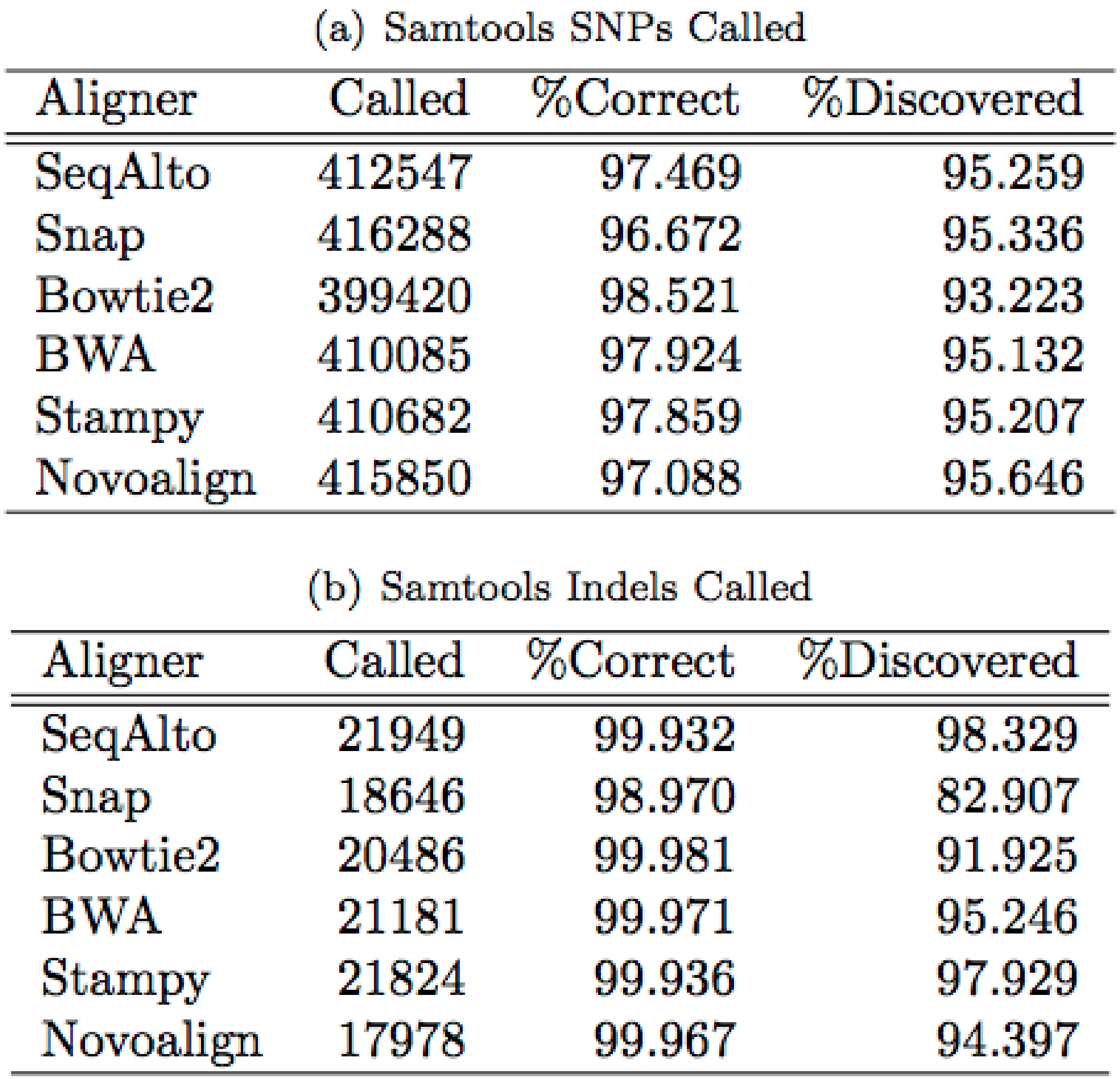
Interpreting Variation

- Getting some variants is easy, analyzing them is hard
- Commonly used tools in CCTS
  - IGV, BedTools, VCFTools, SNPEff
- Pipelines are becoming more popular
  - Annovar (Sift, Polyphen2)
- Online Resources

Questions?



"Fast gapped-read alignment with Bowtie 2", Langmead and Saltzburg (2012)



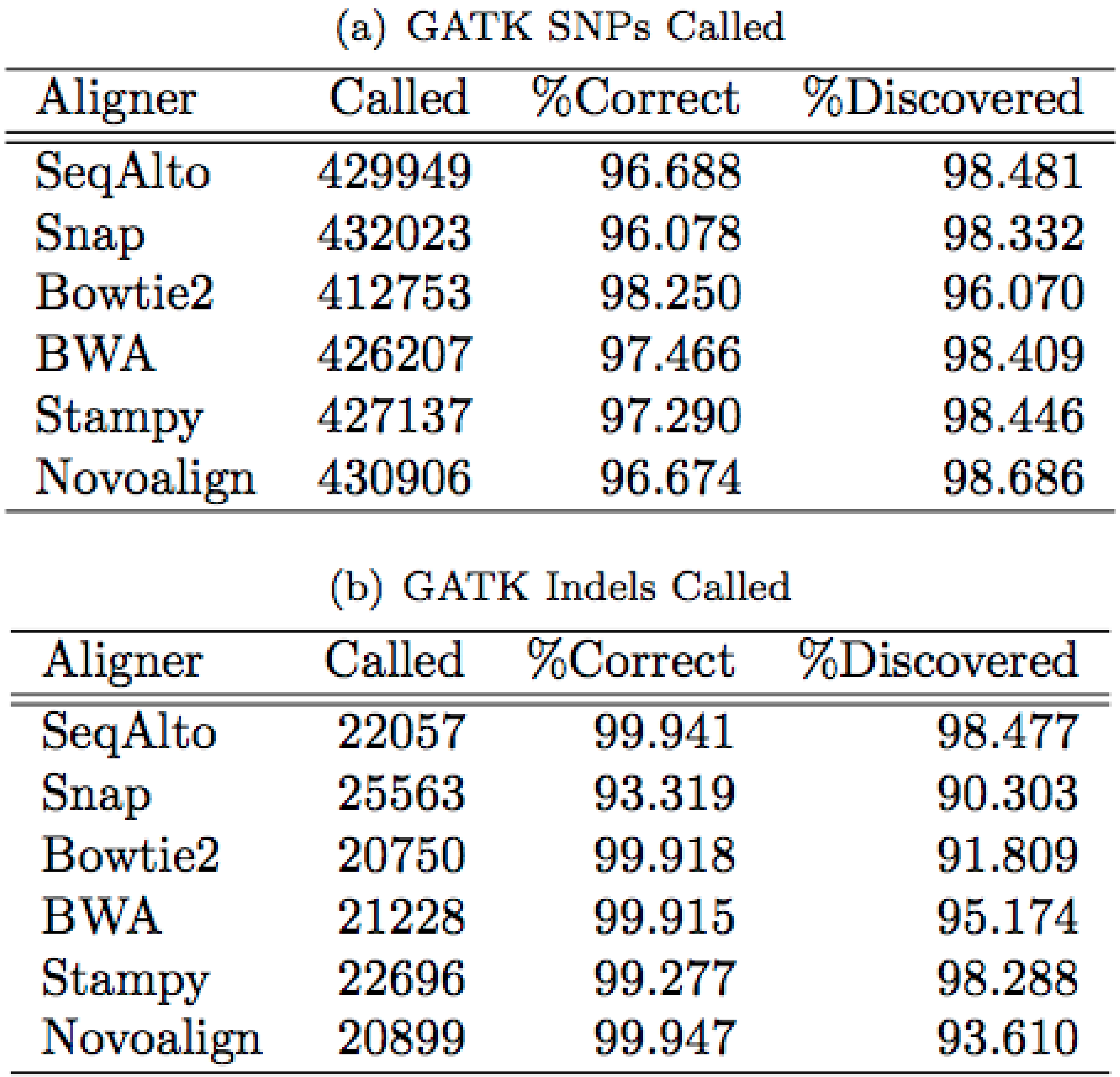
(a) Samtools SNPs Called

Aligner	Called	%Correct	%Discovered
SeqAlto	412547	97.469	95.259
Snap	416288	96.672	95.336
Bowtie2	399420	98.521	93.223
BWA	410085	97.924	95.132
Stampy	410682	97.859	95.207
Novoalign	415850	97.088	95.646

(b) Samtools Indels Called

Aligner	Called	%Correct	%Discovered
SeqAlto	21949	99.932	98.329
Snap	18646	98.970	82.907
Bowtie2	20486	99.981	91.925
BWA	21181	99.971	95.246
Stampy	21824	99.936	97.929
Novoalign	17978	99.967	94.397

Fast and Accurate Read Alignment for Resequencing, Mu et al, (2012)



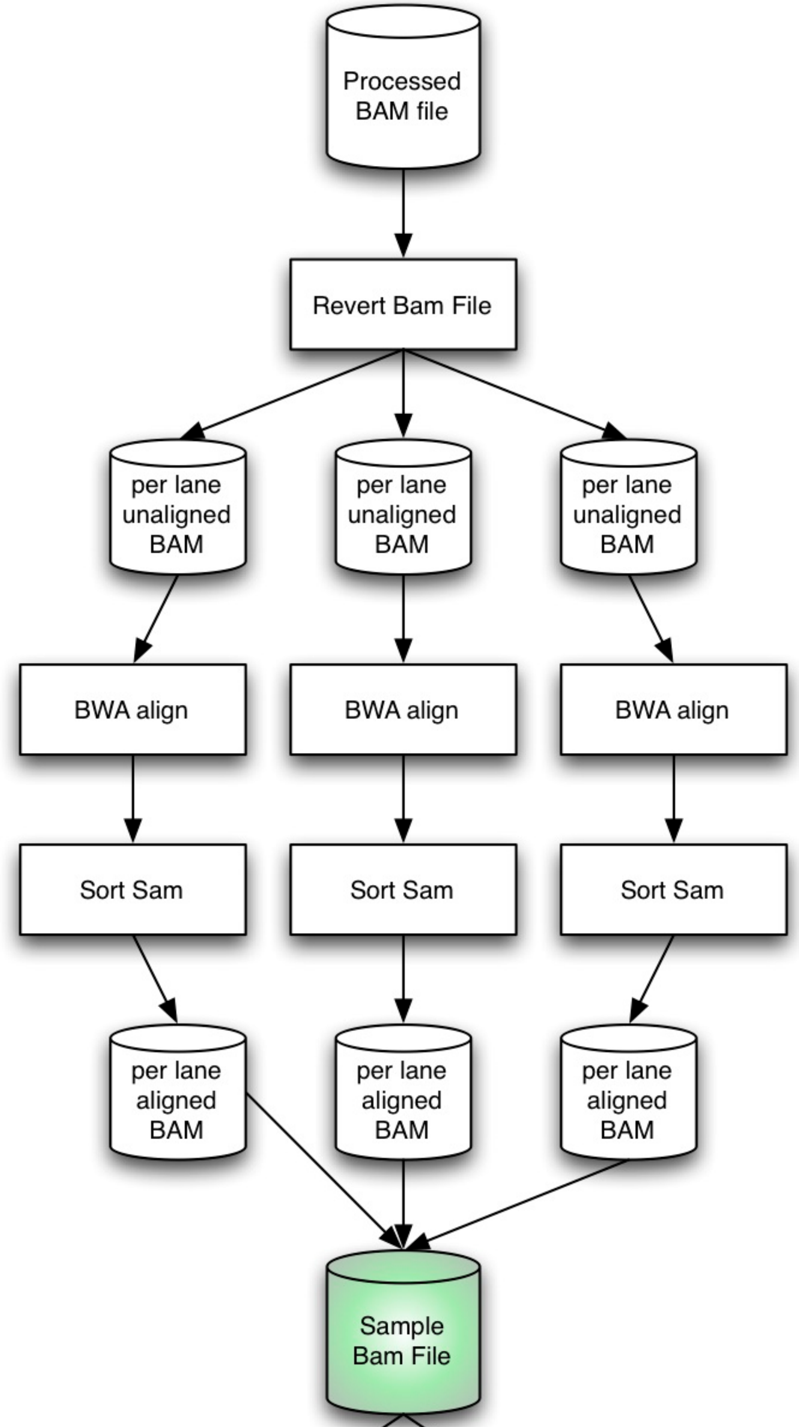
(a) GATK SNPs Called

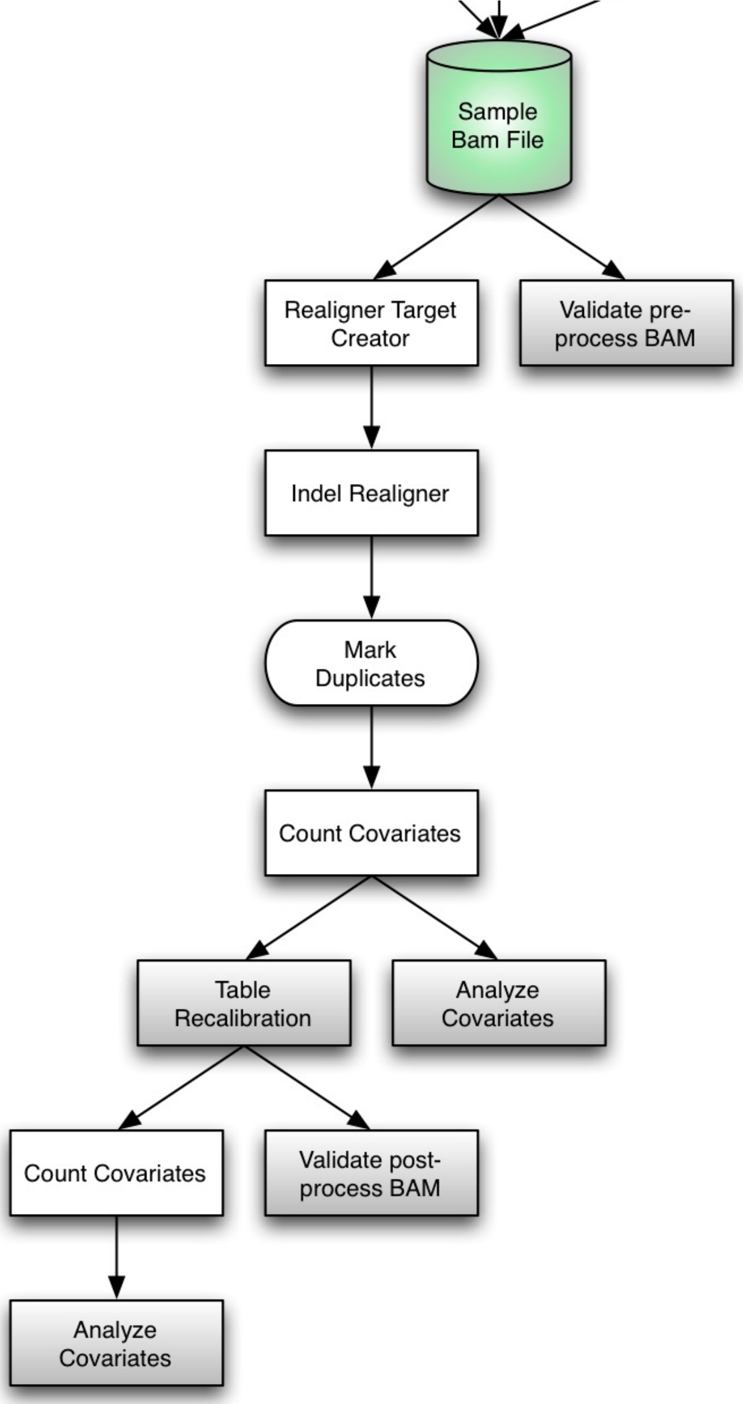
Aligner	Called	%Correct	%Discovered
SeqAlto	429949	96.688	98.481
Snap	432023	96.078	98.332
Bowtie2	412753	98.250	96.070
BWA	426207	97.466	98.409
Stampy	427137	97.290	98.446
Novoalign	430906	96.674	98.686

(b) GATK Indels Called

Aligner	Called	%Correct	%Discovered
SeqAlto	22057	99.941	98.477
Snap	25563	93.319	90.303
Bowtie2	20750	99.918	91.809
BWA	21228	99.915	95.174
Stampy	22696	99.277	98.288
Novoalign	20899	99.947	93.610

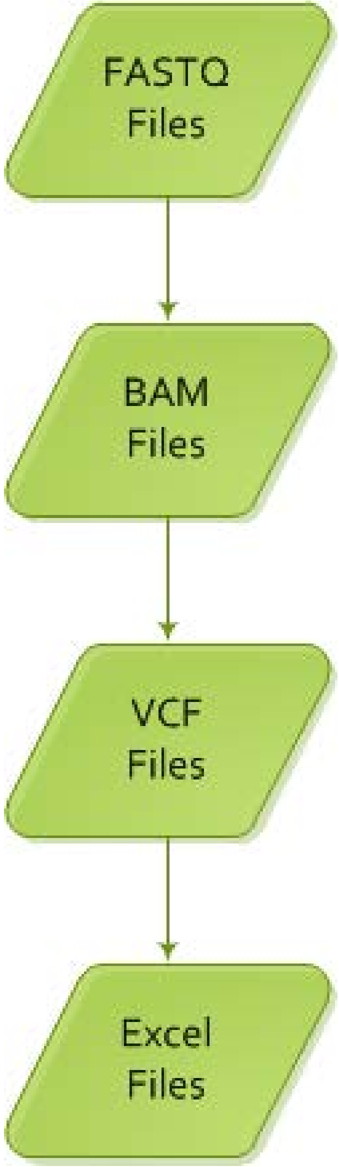
Fast and Accurate Read Alignment for Resequencing, Mu et al, (2012)


Processed BAM file
Revert Bam File
per lane unaligned BAM
per lane unaligned BAM
per lane unaligned BAM
BWA align
BWA align
BWA align
Sort Sam
Sort Sam
Sort Sam
per lane aligned BAM
per lane aligned BAM
per lane aligned BAM
Sample Bam File


Sample Bam File
Realigner Target Creator
Validate pre-process BAM
Indel Realigner
Mark Duplicates
Count Covariates
Table Recalibration
Analyze Covariates
Count Covariates
Validate post-process BAM
Analyze Covariates

Key Points

- Best practices change extremely rapidly
  - We don't know what the single best workflow is today
- Core variant toolset used by UAB CCTS
  - BWA for reference based alignment
  - Picard (duplicate removal)
  - GATK for SNP calling, realignment and recalibration
  - Breakdancer, Pindel for Structural Variant Detection
  - BedTools, VCFtools, IGV for interpretation



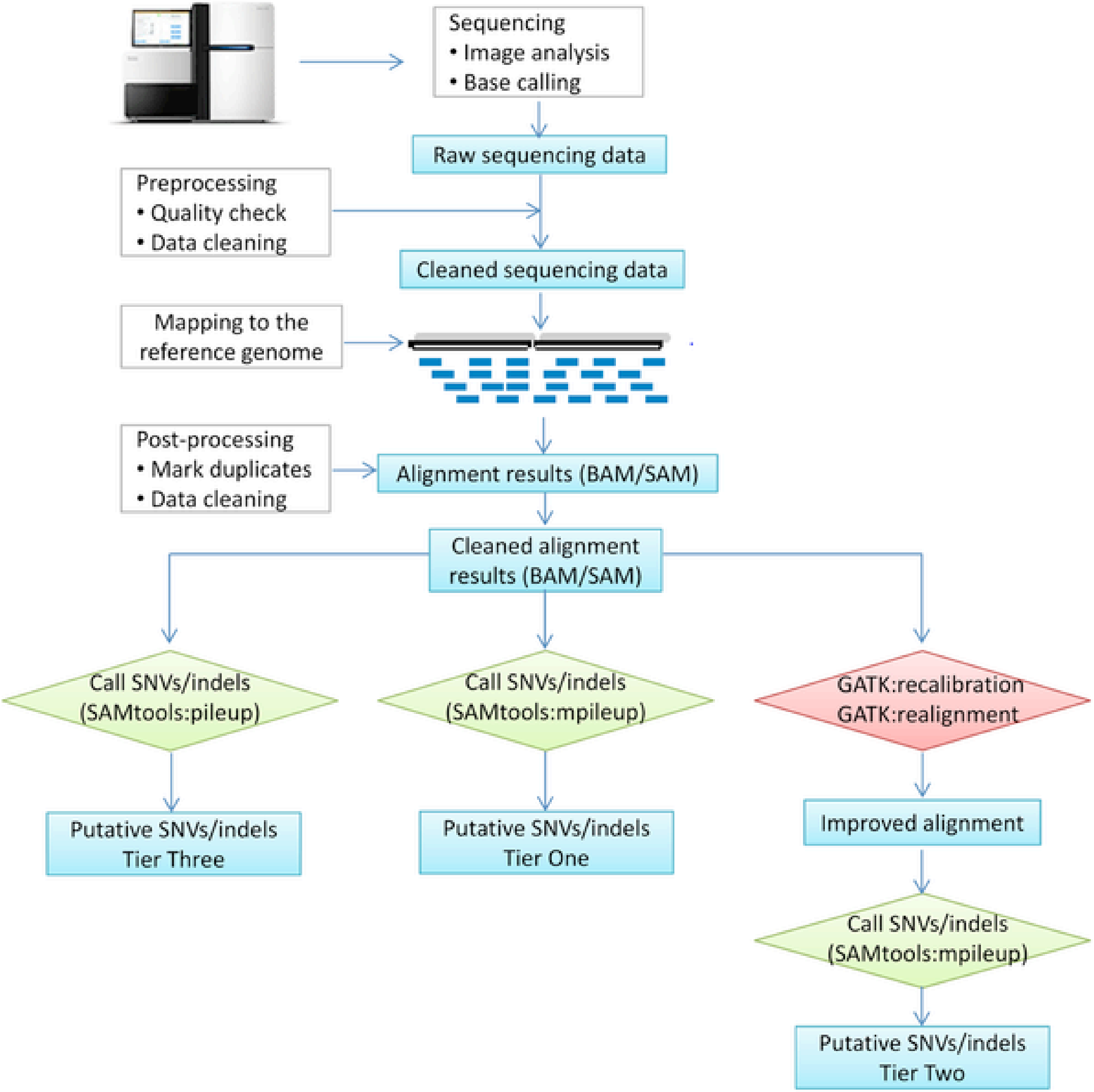
Actual GATK Data

```
chr1   802093   G   A   521.67   GT:AD:DP:GQ:PL   1/1:1,23:24:48.11:555,48,0

chr1   802191   G   A   54.33    GT:AD:DP:GQ:PL   0/1:31,12:43:84.36:84,0,458

chr1   802320   G   A   349.65   GT:AD:DP:GQ:PL   0/1:9,15:27:10.30:379,0,10
```

- 3 genotypes (0/0, 0/1, 1/1)
- GQ:PL
  - Genotype Quality
- AD:DP
  - Average Depth : Depth Quality

Workflow Overview

Workflow from "**Consensus Rules in Variant Detection from Next-Generation Sequencing Data**", Jia et al. (2012)



Sequencing
- Image analysis
- Base calling

Raw sequencing data

Preprocessing
- Quality check
- Data cleaning

Cleaned sequencing data

Mapping to the reference genome

Post-processing
- Mark duplicates
- Data cleaning

Alignment results (BAM/SAM)

Cleaned alignment results (BAM/SAM)

Call SNVs/indels (SAMtools:pileup)

Putative SNVs/indels Tier Three

Call SNVs/indels (SAMtools:mpileup)

Putative SNVs/indels Tier One

GATK:recalibration GATK:realignment

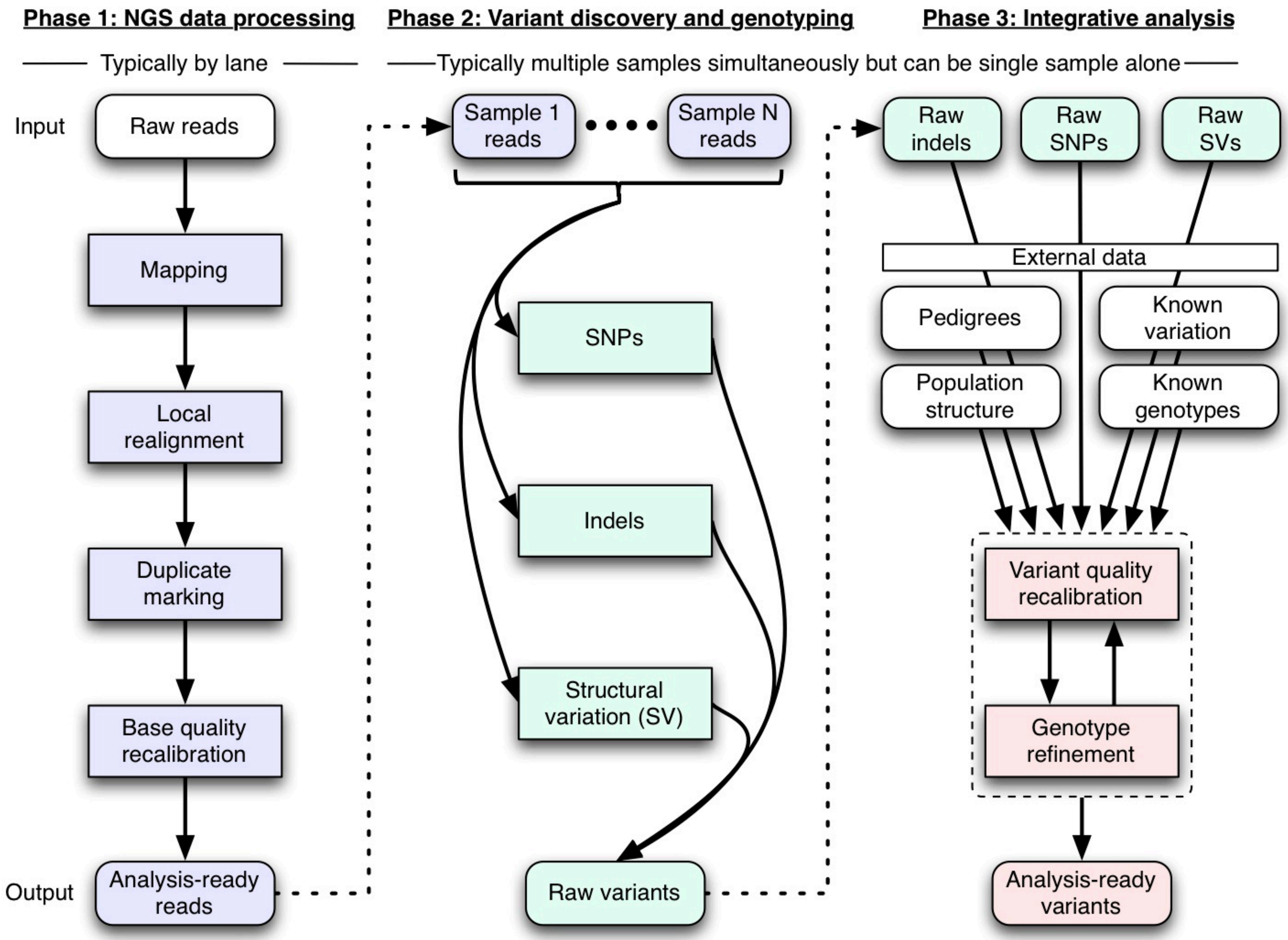
Improved alignment

Call SNVs/indels (SAMtools:mpileup)

Putative SNVs/indels Tier Two

Variation Detection

- Nucleotide Polymorphisms
  - SNPs
  - MNPs
- "Structural Variants" / Rearrangements
  - Insertions/Deletions (small and large)
  - Inversions
  - Tandem Duplications
  - Translocations


Phase 1: NGS data processing
Typically by lane
Input
Raw reads
Mapping
Local realignment
Duplicate marking
Base quality recalibration
Output
Analysis-ready reads
Phase 2: Variant discovery and genotyping
Typically multiple samples simultaneously but can be single sample alone
Sample 1 reads
Sample N reads
SNPs
Indels
Structural variation (SV)
Raw variants
Phase 3: Integrative analysis
Raw indels
Raw SNPs
Raw SVs
External data
Pedigrees
Known variation
Population structure
Known genotypes
Variant quality recalibration
Genotype refinement
Analysis-ready variants