Whole exome sequencing

Data analysis

Why exome sequencing

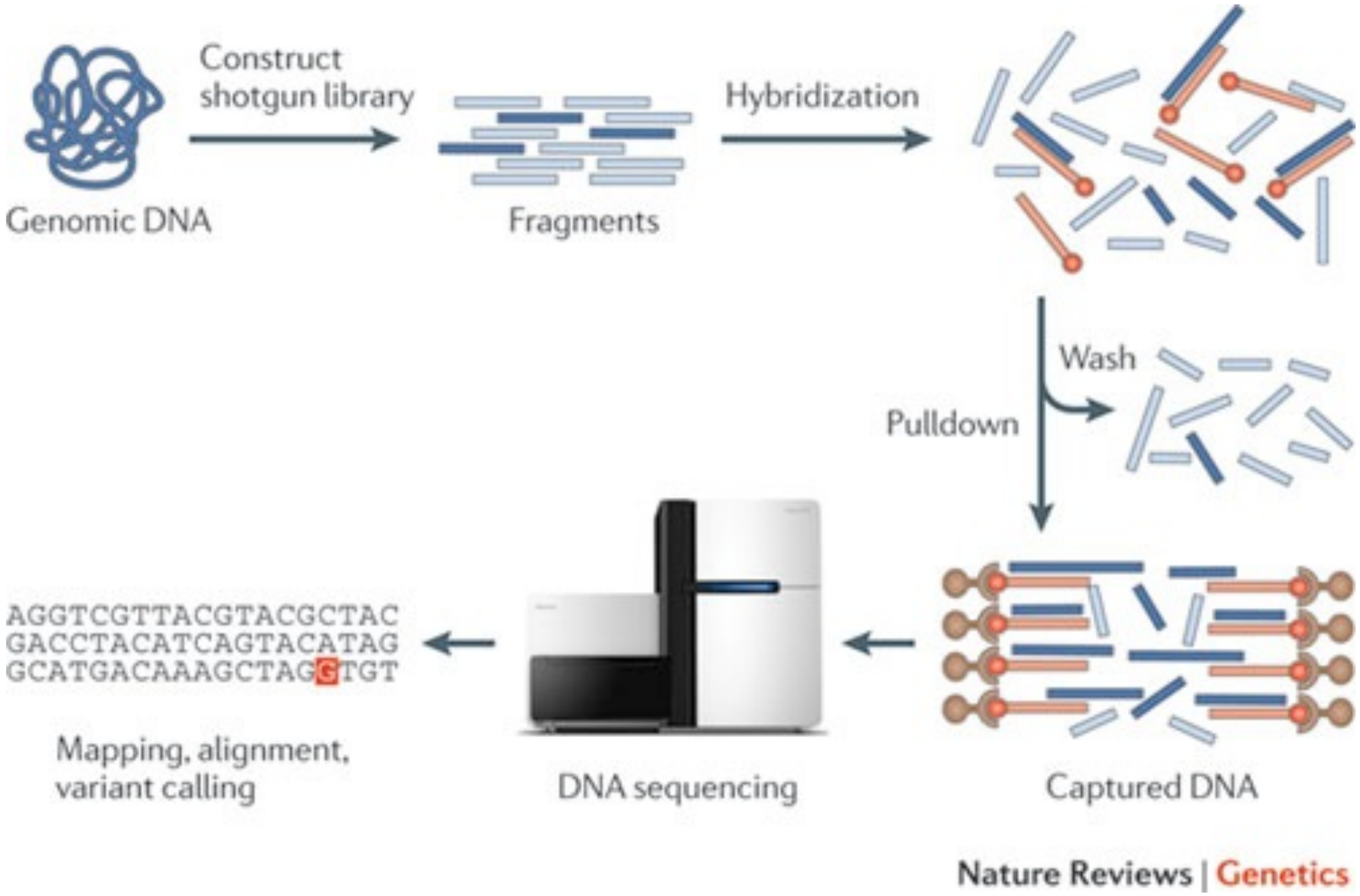
Cheaper than whole-genome sequencing

Exome is the **protein-coding** region of the genome – **1-2%**

Mutations in exome likely **cause protein expressions alterations** → causes for diseases

Provides molecular evidence to assist in clinical diagnosis

Exome sequencing - overview



Exome capture kits

NimbleGen

SeqCap EZ® Exome v3

Illumina TruSeq Exome

Exome capture kits

Kits **cover different regions**

Coverage variability

- Quantity and quality of the input DNA
- Library insert length and its distribution
- Repeat elements, tandem repeats and pseudogenes
- Extreme GC content
- Sequencing

How much coverage?

EdgeBio suggestions

Research studies - **30X** and **50X**
Clinical studies - at least **100X**

++ mean of coverage -> ++ % of the target region covered

Reads **on target**

- Agilent: 77-83%
- Illumina TruSeq: 69-70%
- NimbleGen: 84-86%

Applications of WES

Single gene disorders (Kabuki Syndrome, Kohlschütter-Tönz Syndrome)

Ng et al Nat Genet 2010

Schossig et al. Am J Hum Genet 2012

Genetic heterogenic disorders (autism, schizophrenia)

Girard et al. Nat Genet 2011

Yu et al. Neuron 2013

Cancer research (somatic mutations)

Varela et al. *Nature* 2011

…

Mendelian disorders

Rare, genetic disorders

40-82 per 1000 live births

Majority of patients **without** a diagnosis

>3000 disorders with **unknown** genetic causes

Stitziel et al, Genome Biol 2011

Sequencing study

Family (parents) often needs to be sequenced as well to find genetic cause

Recent publication

The NEW ENGLAND JOURNAL of MEDICINE

ORIGINAL ARTICLE

Clinical Whole-Exome Sequencing for the Diagnosis of Mendelian Disorders

Yaping Yang, Ph.D., Donna M. Muzny, M.Sc., Jeffrey G. Reid, Ph.D., Matthew N. Bainbridge, Ph.D., Alecia Willis, Ph.D., Patricia A. Ward, M.S., Alicia Braxton, M.S., Joke Beuten, Ph.D., Fan Xia, Ph.D., Zhiyv Niu, Ph.D., Matthew Hardison, Ph.D., Richard Person, Ph.D., Mir Reza Bekheirnia, M.D., Magalie S. Leduc, Ph.D., Amelia Kirby, M.D., Peter Pham, M.Sc., Jennifer Scull, Ph.D., Min Wang, Ph.D., Yan Ding, M.D., Sharon E. Plon, M.D., Ph.D., James R. Lupski, M.D., Ph.D., Arthur L. Beaudet, M.D., Richard A. Gibbs, Ph.D., and Christine M. Eng, M.D.

ABSTRACT

Setup

 250 patients

Blood samples

NibleGen kit
Illumina HiSeq 2000 or Genome Analyzer IIx (24 cases)

Coverage: avg **130X** with >**95%** of target bases at least **20X** coverage

Goal

Identify *known* mutations (not elucidating new ones)

Results

Identified 86 mutated alleles – causative in 62 patients

25% molecular diagnostic rate (33 autosomal dominant, 16 autosomal recessive, 9 X-linked) → many mendelian diseases have yet to be discovered

4 patients - **two non-overlapping molecular diagnoses** (will increase as more causing mutations are known)

Mutation types
small frameshift, in-frame, nonsense, splice missense mutations

Yang et al. *N Engl J Med (*2013)

Applications in cancer

Comparison of **normal** vs **tumor** tissue

Identification of **driver mutations**

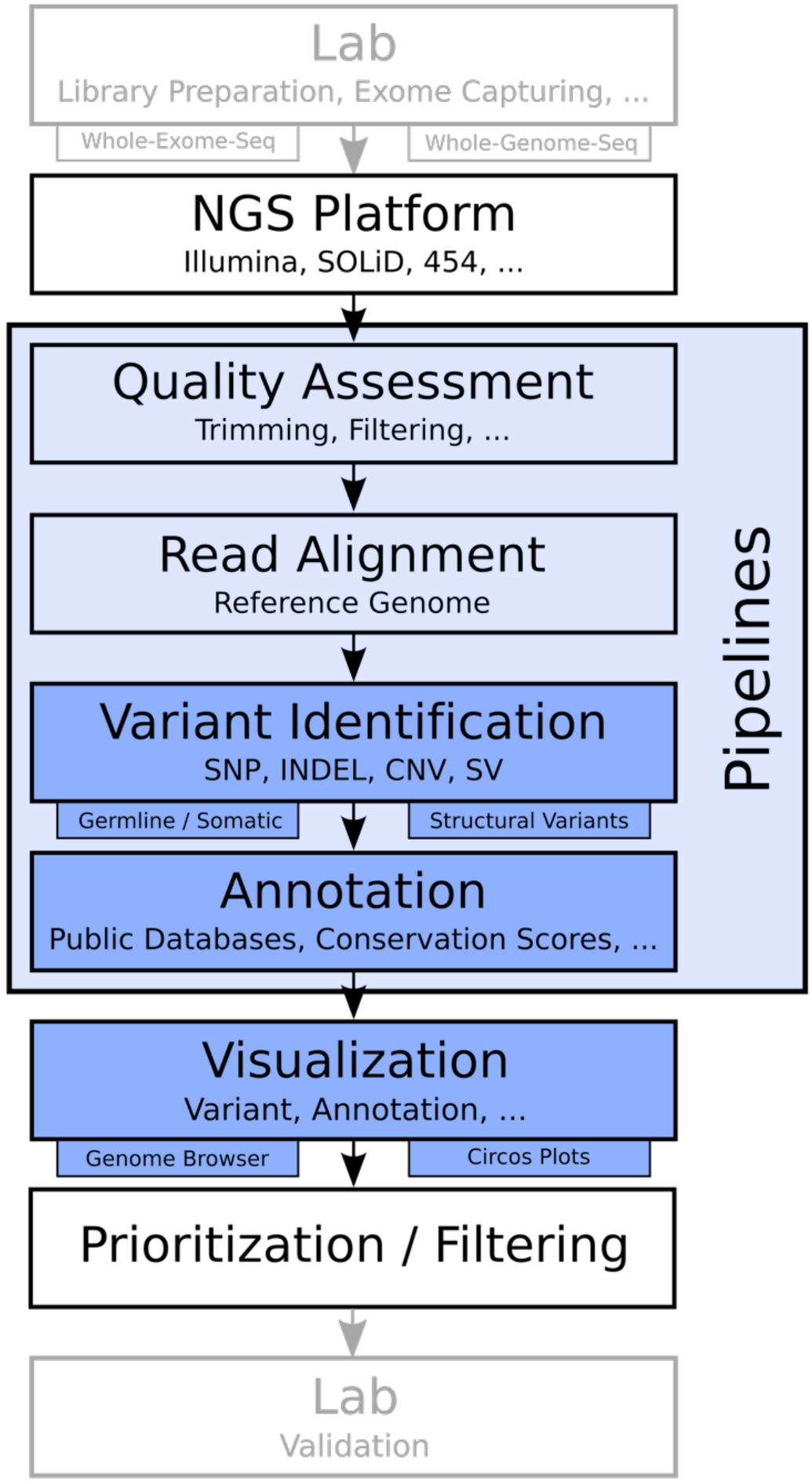
Find copy number variations (**CNVs**) and large structural variations (**SVs**)

Identify **germline mutations** that **increase risk** of getting cancer

...

Hill et al. *The genetics of melanoma: recent advances.* (2013)
Tenedini et al. *Targeted cancer exome sequencing reveals recurrent mutations in myeloproliferative neoplasms.* (2013)

Analysis


Lab
Library Preparation, Exome Capturing, ...
Whole-Exome-Seq
Whole-Genome-Seq
NGS Platform
Illumina, SOLiD, 454, ...
Quality Assessment
Trimming, Filtering, ...
Read Alignment
Reference Genome
Variant Identification
SNP, INDEL, CNV, SV
Germline / Somatic
Structural Variants
Annotation
Public Databases, Conservation Scores, ...
Pipelines
Visualization
Variant, Annotation, ...
Genome Browser
Circos Plots
Prioritization / Filtering
Lab
Validation

Quality control

Quality control

Problems

Base calling errors
Poor quality reads
Homopolymer issue
Reads and adapter contamination

Problem handling

Visualization of base quality scores and base distributions
Trimming & filtering of reads
- score and properties
- primer contaminations, N content, and GC characteristics

FastQC

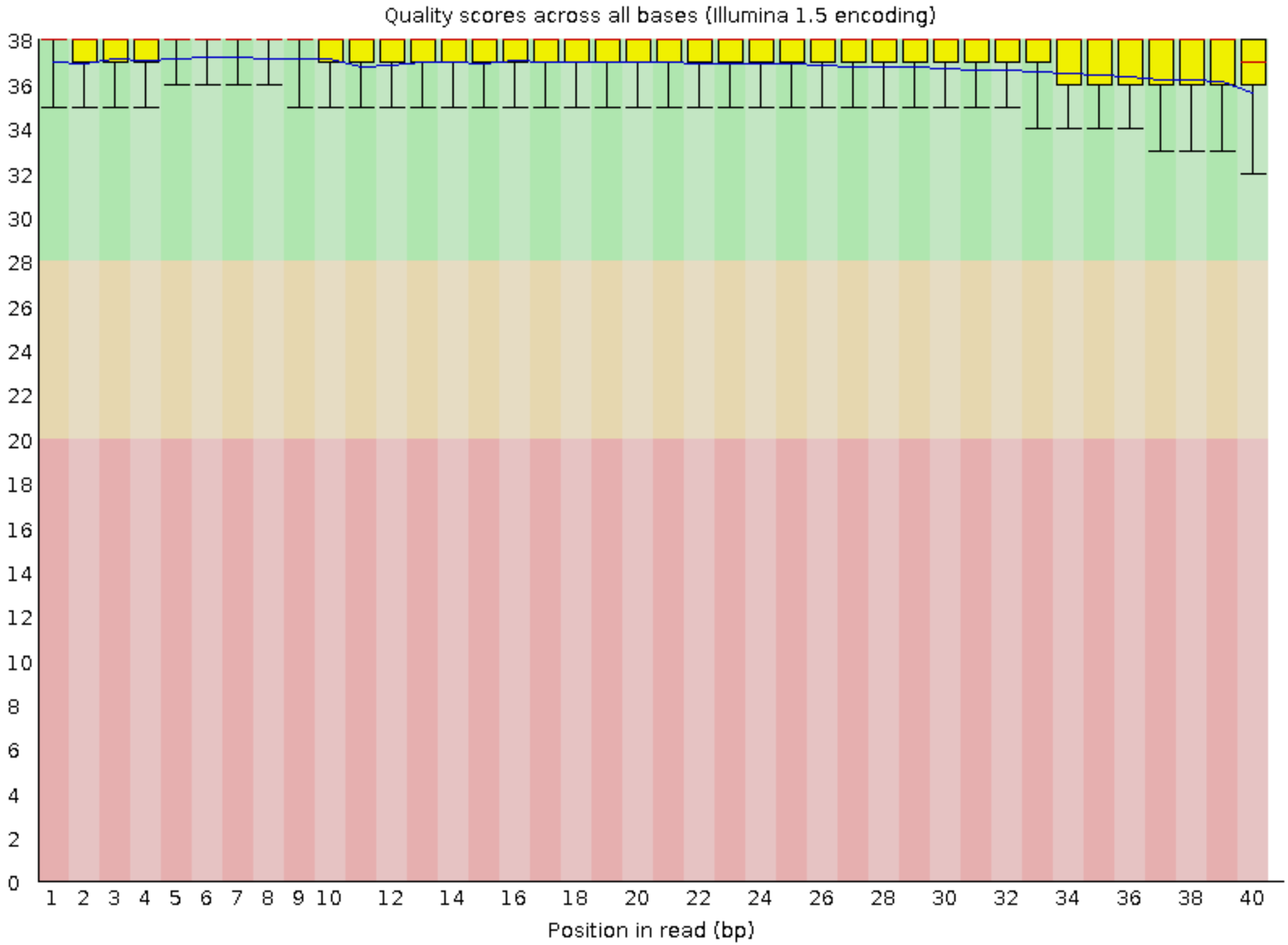
Java standalone tool

Fastq **quality control** (not aligned BAM,SAM)

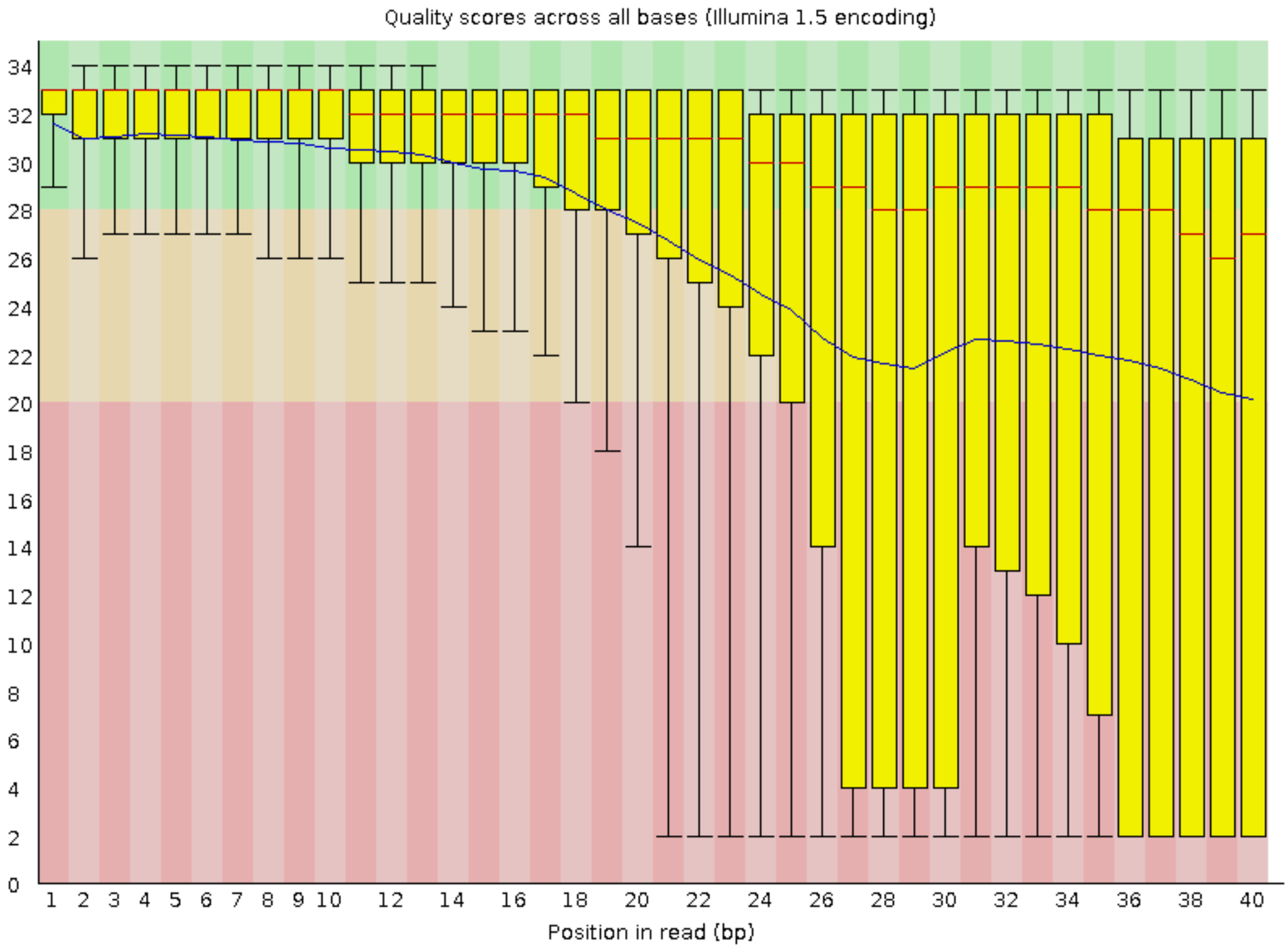
Outputs a report in HTML format
-> view in browser

http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

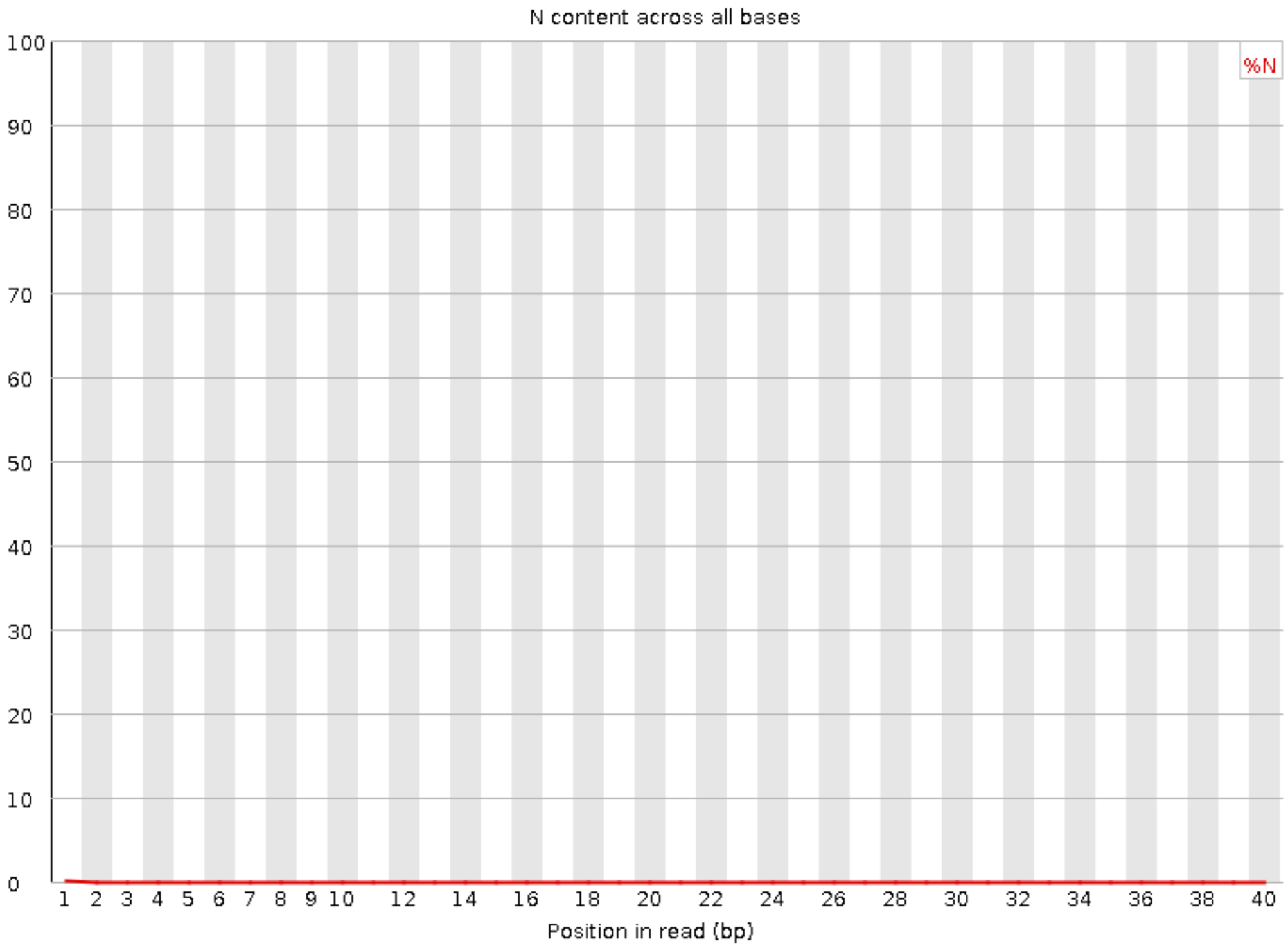
Good per base sequence quality



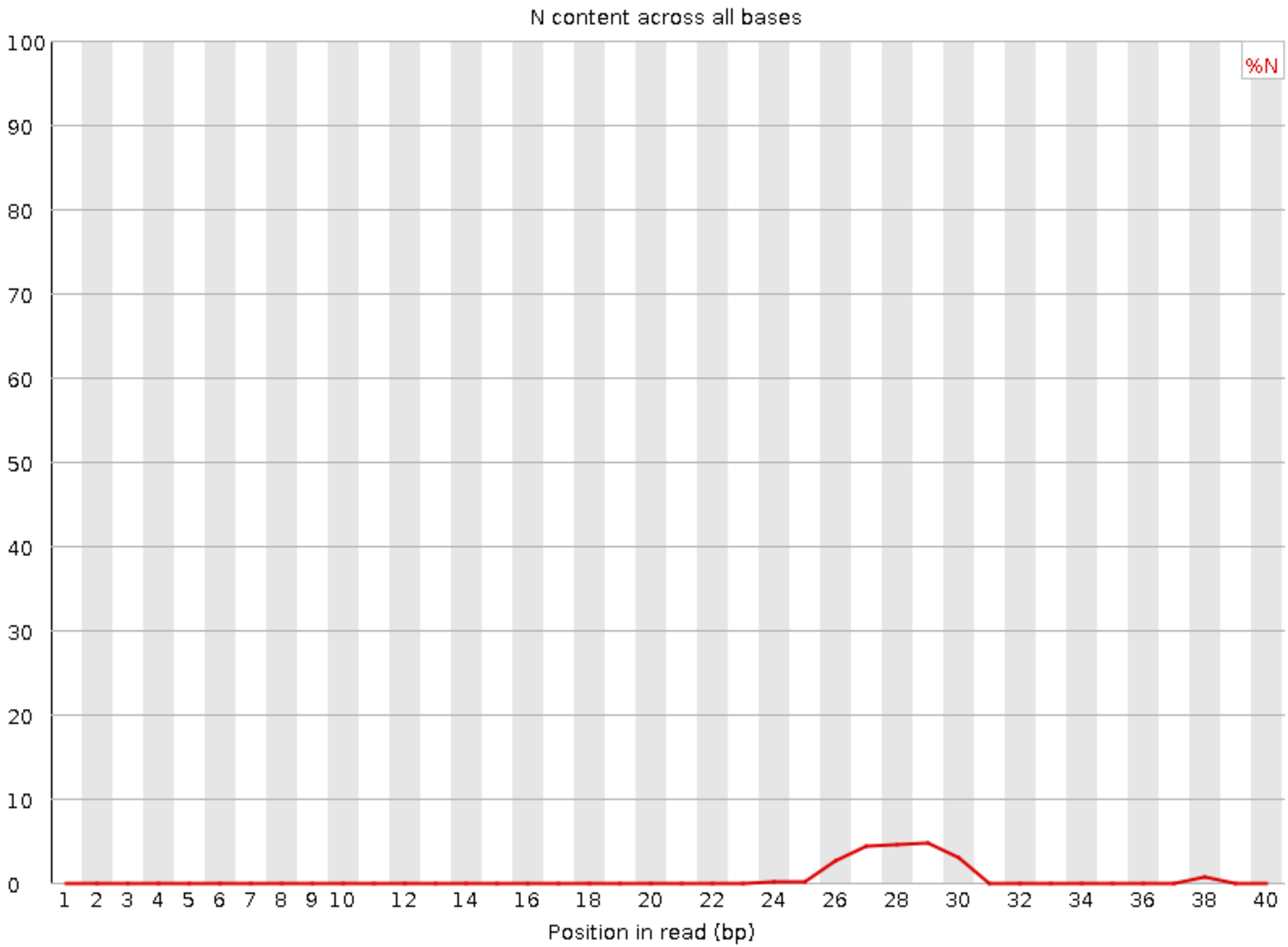
Poor per base sequence quality



Good per base N content



Poor per base N content



What to look for ?

- ☑ Base calling quality
- ☑ GC content
- ☑ N content
- ☑ Adapter contamination
- ☑ Trimming/filtering: FASTX-Toolkit, PRINSEQ

Alignment

Alignment

Current human reference genome
- hg19 (UCSC, chr prefix)
- GRCh37 (Genome reference consortium)

hg20

An issue has been encountered in the processing of the GRCh38 assembly. This issue is expected to delay the release of the assembly by several weeks. We will provide an updated release date estimate as it becomes available. If you have questions or concerns about this let us know.

Alignment

Current human reference genome
hg19 (UCSC, chr prefix)
GRCh37 (Genome reference consortium)

Tools
BWA, Bowtie, BFAST, ...

Colorspace support dropped
BWA > 1.6.0; Bowtie > 2.0

Index of reference genome has to be created prior to aligning

Decoy sequences

Sequences missed in hg19 assembly -- ~36Mb

Contains **sequences** of
- **Epstein-Barr virus** – often used in lymphoblast cell lines to “immortalize” cells
- **HuRef** (Craig Venters sequence)
- **de novo assembly** of NA12878 (from 1000 genomes project)
- **human DNA** sequences studied in bacteria

Decoy sequences

Sequences missed in hg19 assembly -- ~36Mb

Contains **sequences** of
- **Epstein-Barr virus** – often used in lymphoblast cell lines to "immortalize" cells
- **HuRef** (Craig Venters sequence)
- **de novo assembly** of NA12878 (from 1000 genomes project)
- **human DNA** sequences studied in bacteria

+ Avoid false forced alignment
+ Speed improvements

http://www.cureffi.org/2013/02/01/the-decoy-genome/

Alignment evaluation

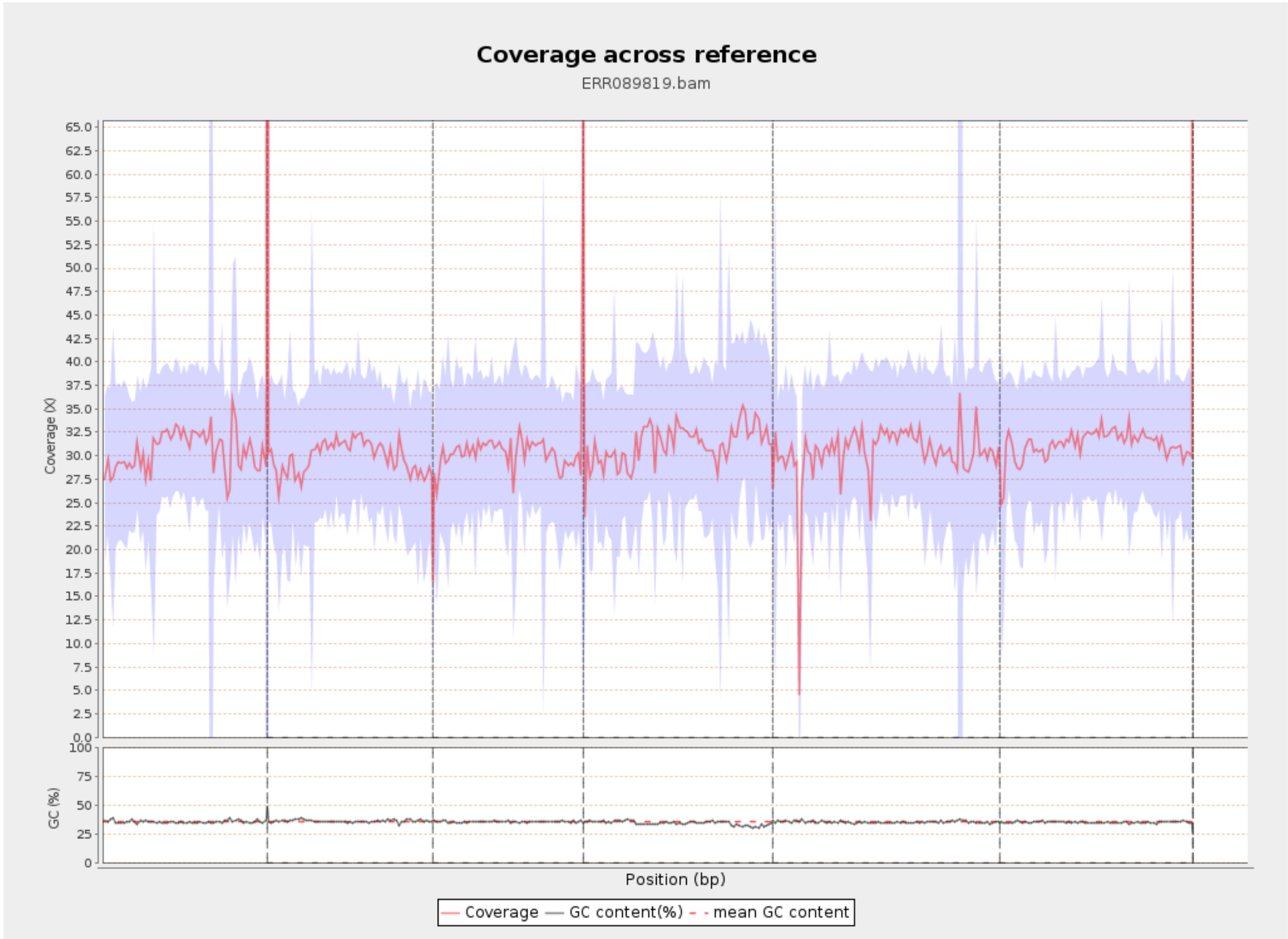
e.g. **Qualimap**

Implemented in JAVA

Features

- **mapping coverage** and nucleotide distribution
- main **properties** of the **alignment** data
- reads **mapped** inside/outside of the **regions** defined in an annotation reference
- sequencing **depth** statistics

Qualimap


Coverage across reference
ERR089819.bam
Coverage (X)
65.0
62.5
60.0
57.5
55.0
52.5
50.0
47.5
45.0
42.5
40.0
37.5
35.0
32.5
30.0
27.5
25.0
22.5
20.0
17.5
15.0
12.5
10.0
7.5
5.0
2.5
0.0
GC (%)
100
75
50
25
0
Position (bp)
Coverage
GC content(%)
mean GC content

What to look for ?

- ☑ Correct reference genome
- ☑ Alignment parameters - stringent vs. flexible
- ☑ Color space support?
- ☑ Sequence duplications
- ☑ Reads on target?
- ☑ Further information

- Benchmarking short sequence mapping tools (PMID: 23758764)
- Comparative analysis of algorithms for next-generation sequencing read alignment (PMID: 21856737)

Variant calling

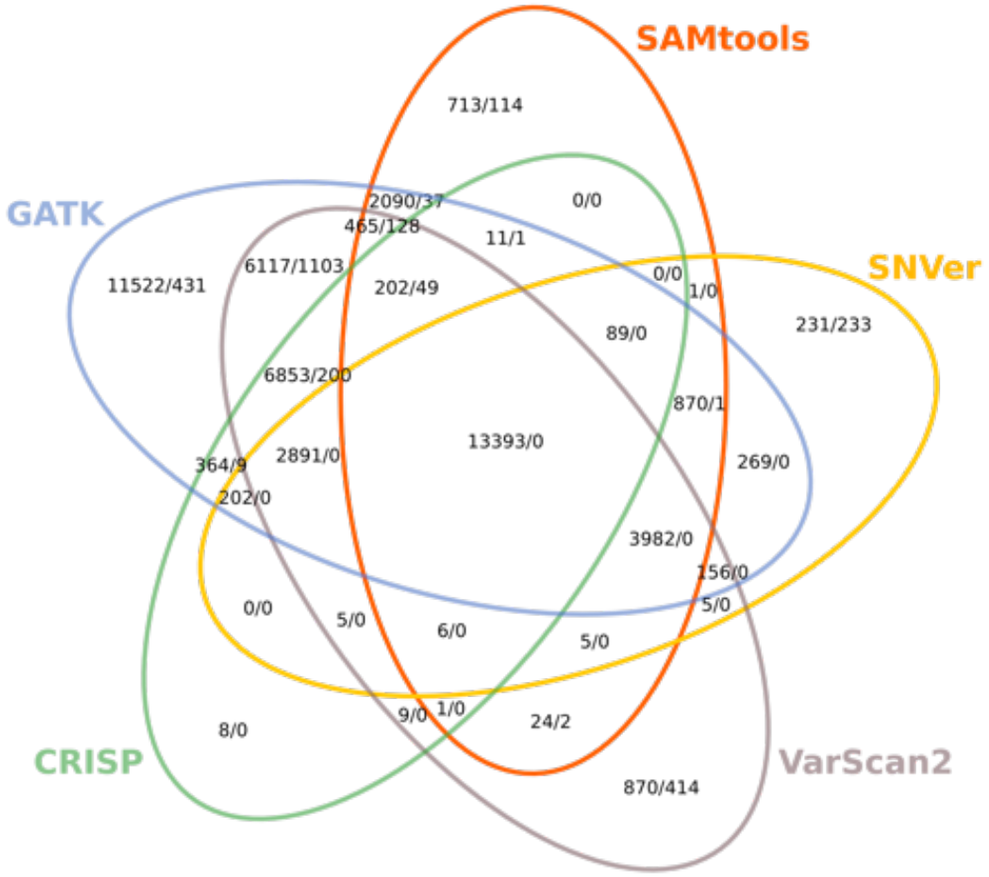
Variant calling

Identification of SNPs, INDELs, CNVs, SVs

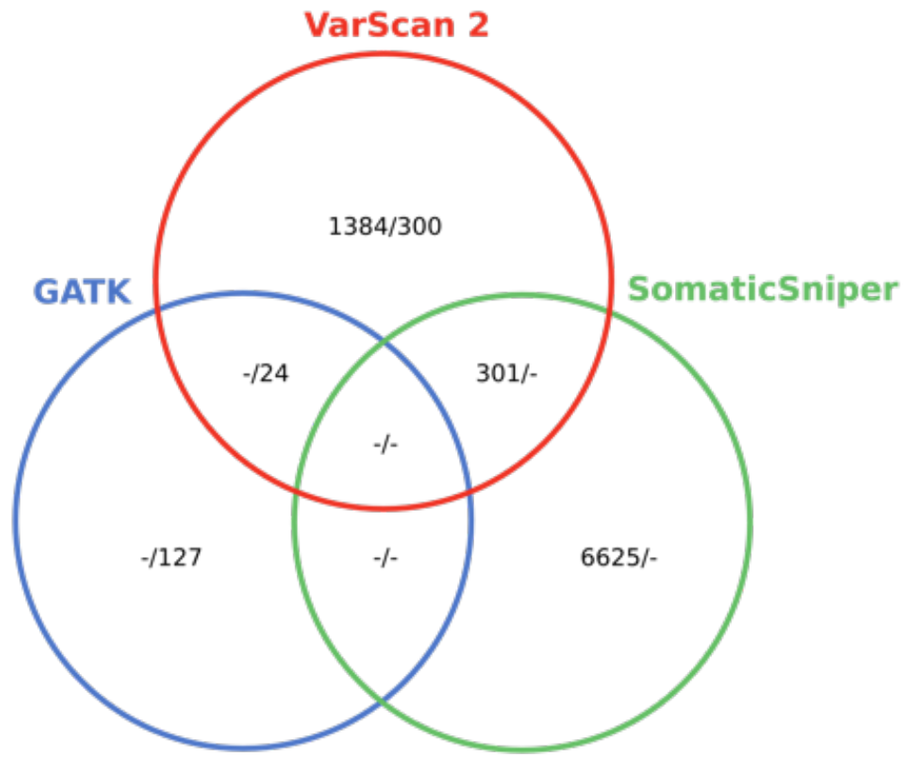
Germline mutations and **somatic** mutations

Tools
Atlas SNP & Atlas INDEL
Crisp
FreeBayes
GATK
SAMtools
SNVer
SomaticSniper
VarScan
...

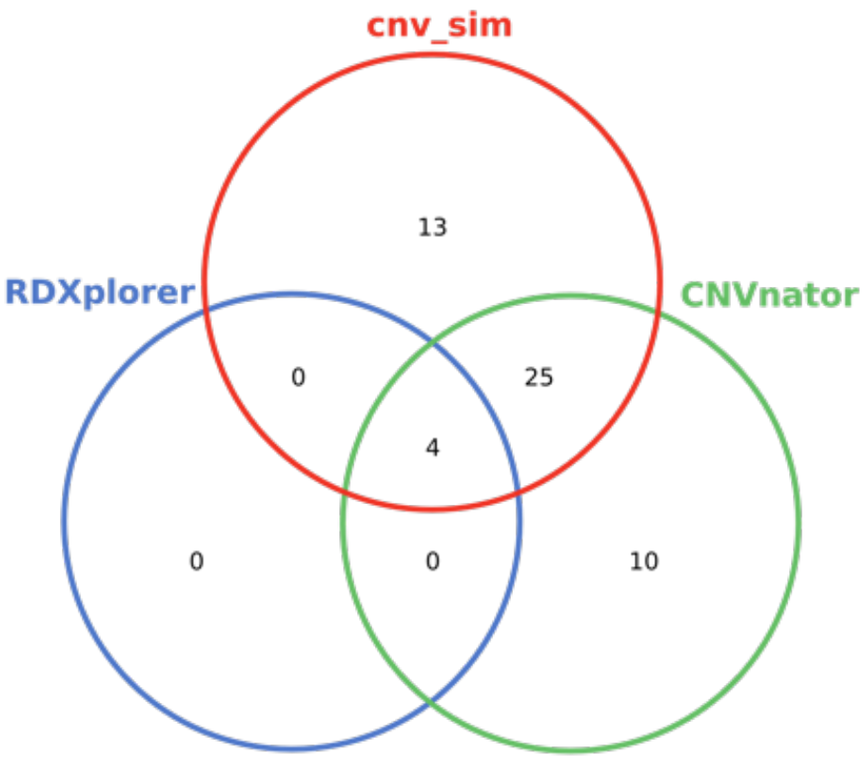
A (Germline callers)



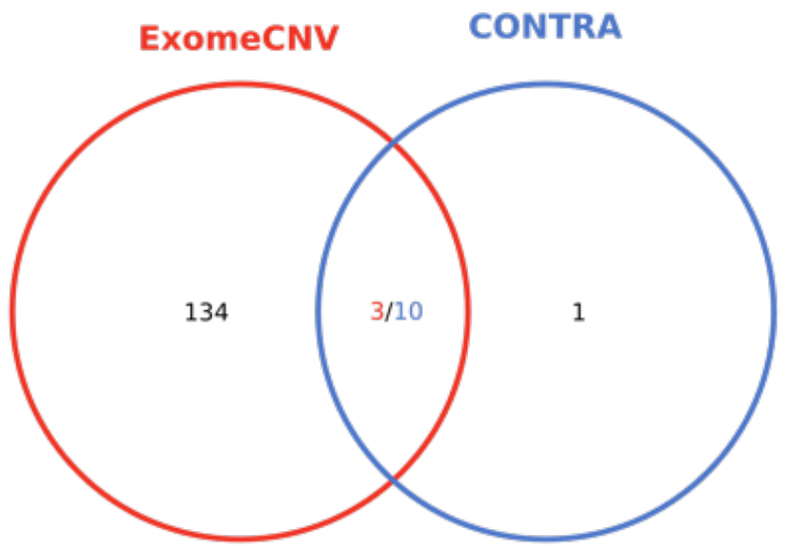
B (Somatic callers)



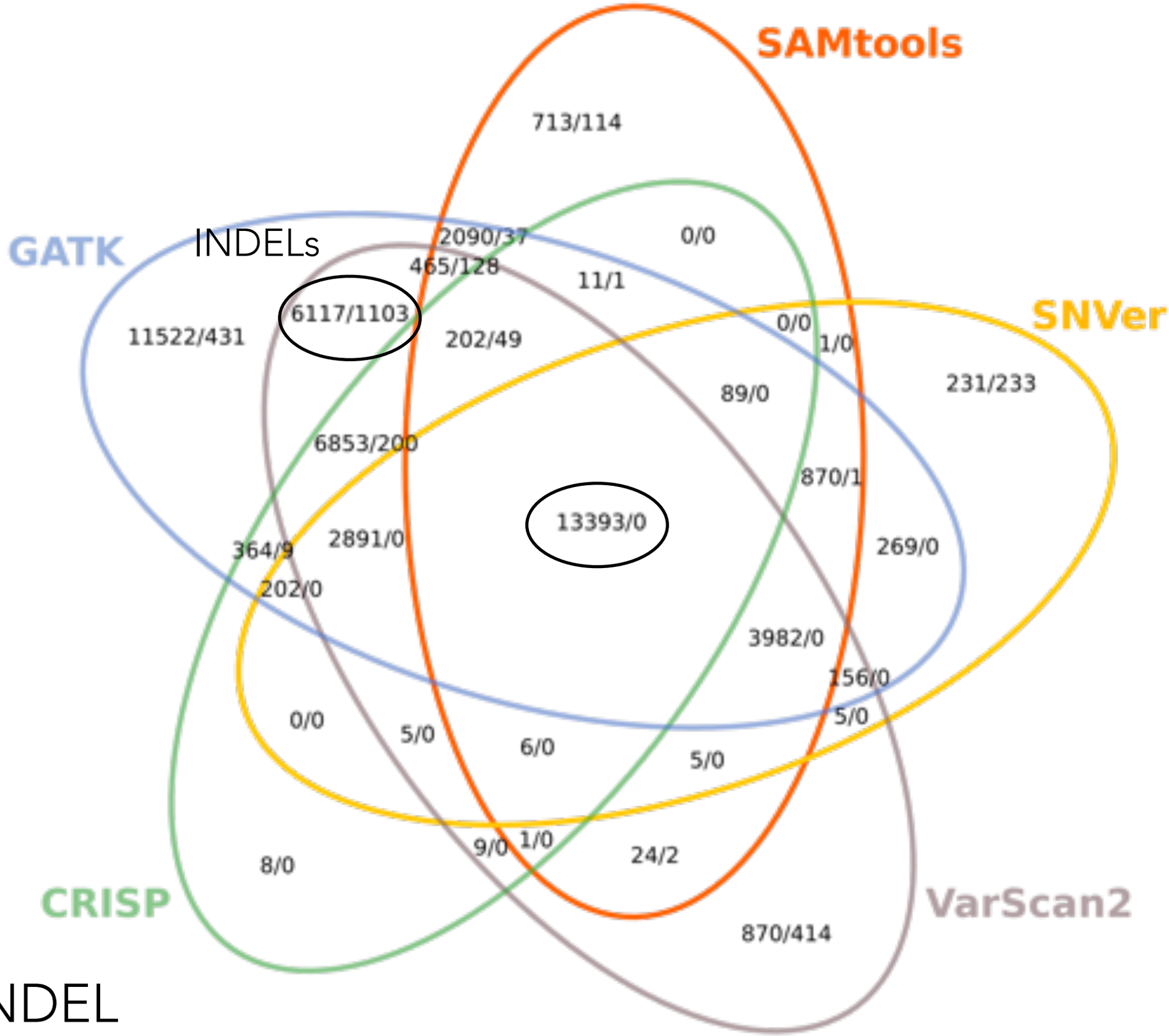
C (CNV identification tools)



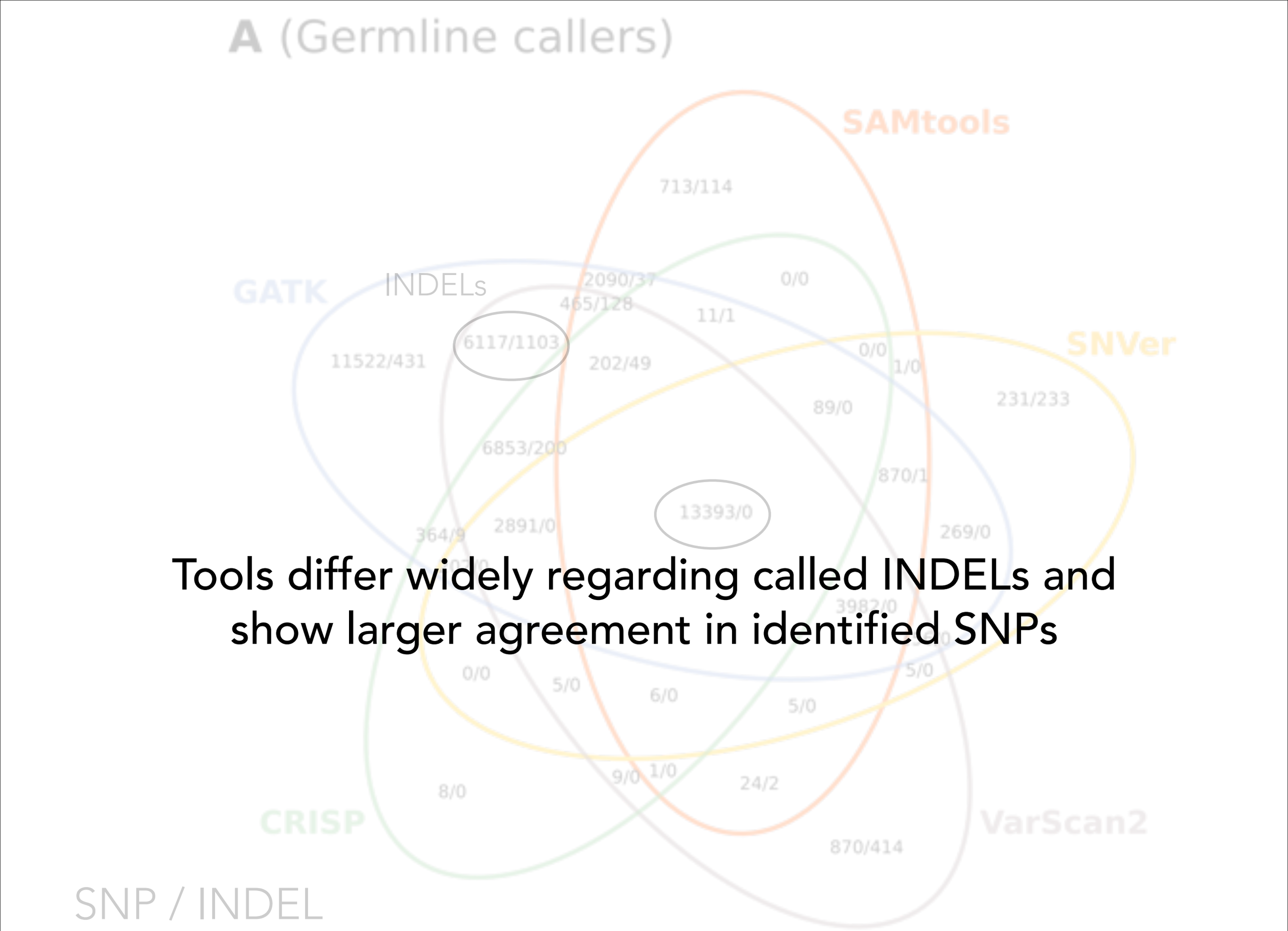
D (Exome CNV identification tools)



A (Germline callers)

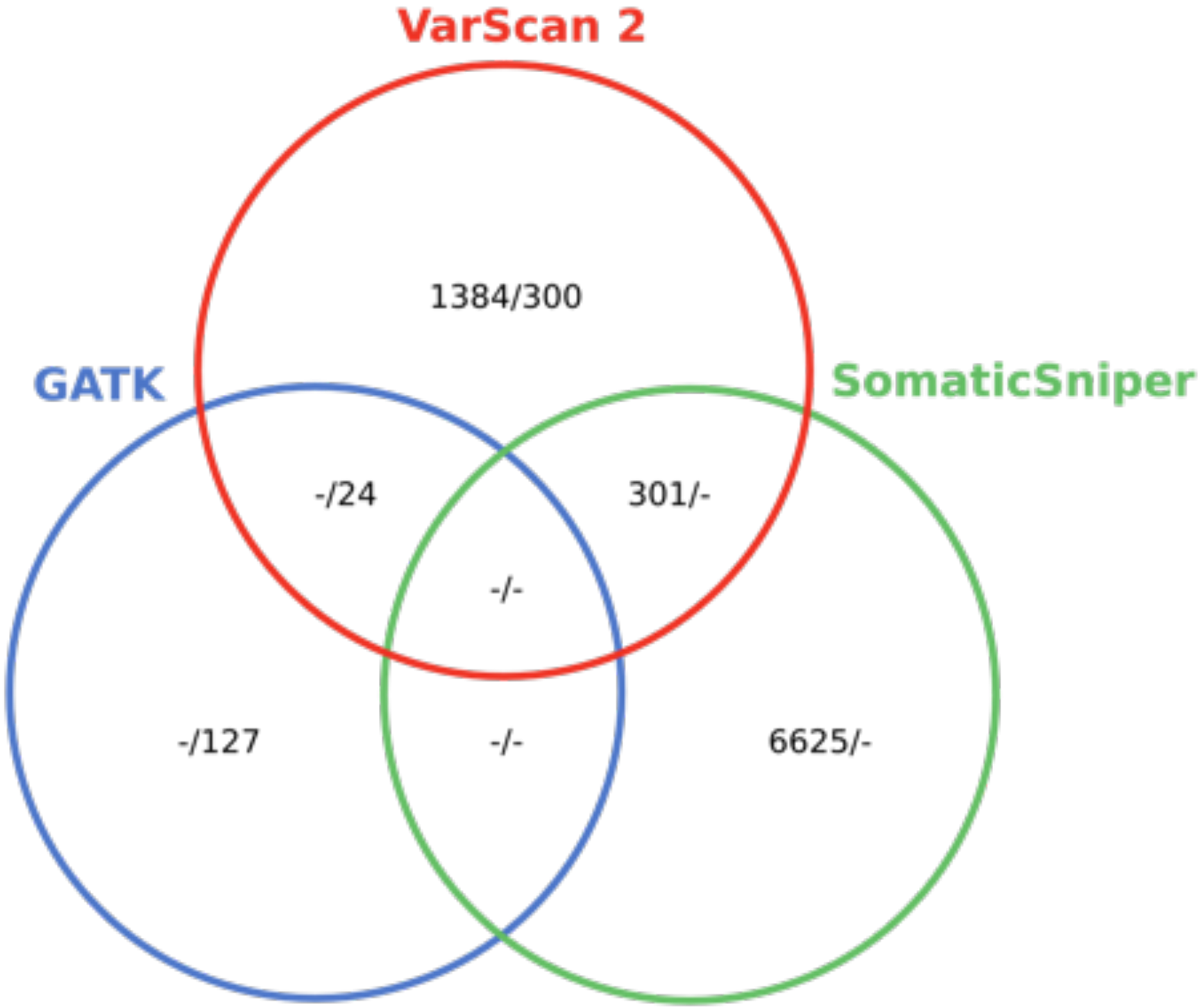


SNP / INDEL



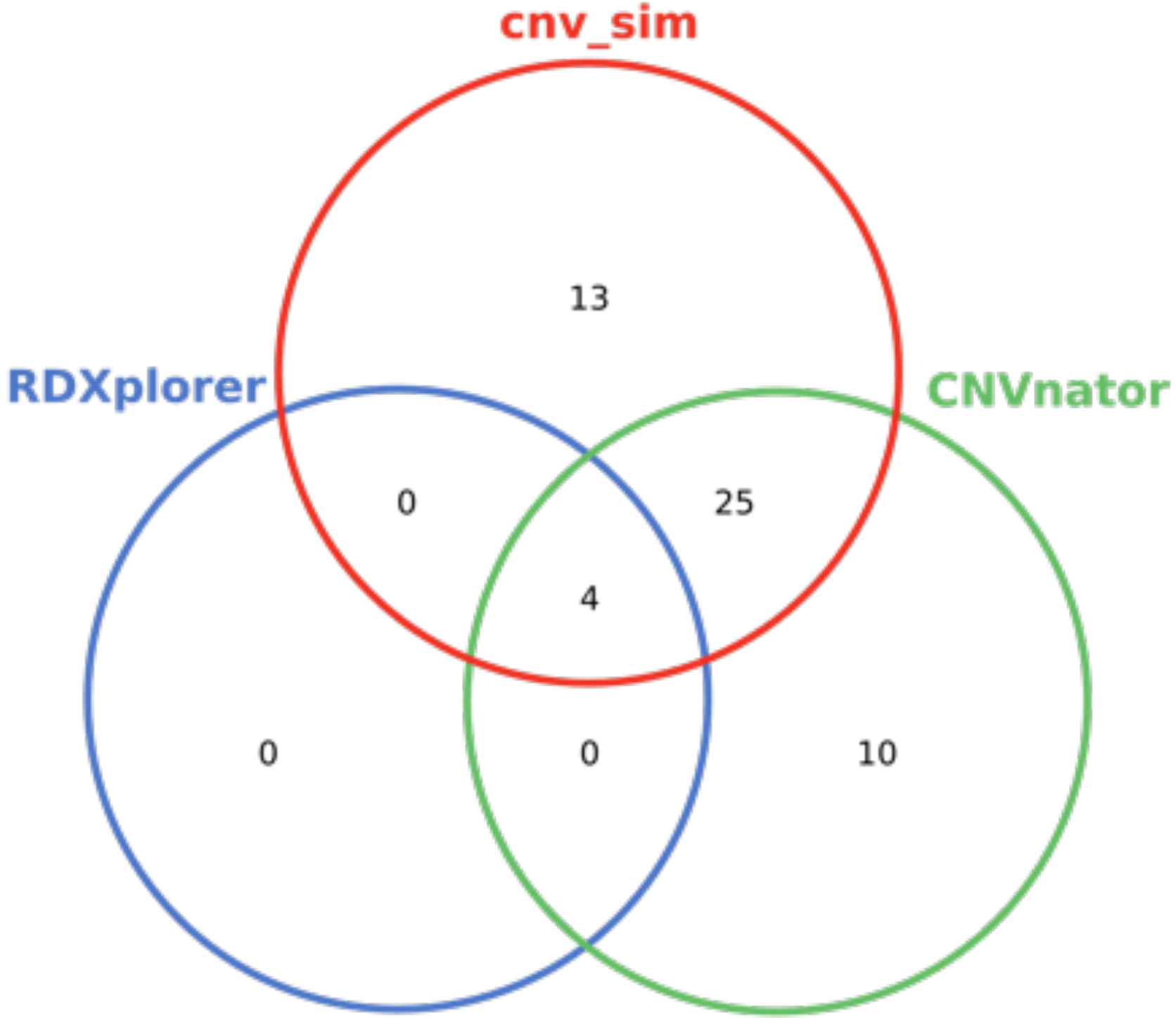
Tools differ widely regarding called INDELs and show larger agreement in identified SNPs

B (Somatic callers)



SNP / INDEL

C (CNV identification tools)



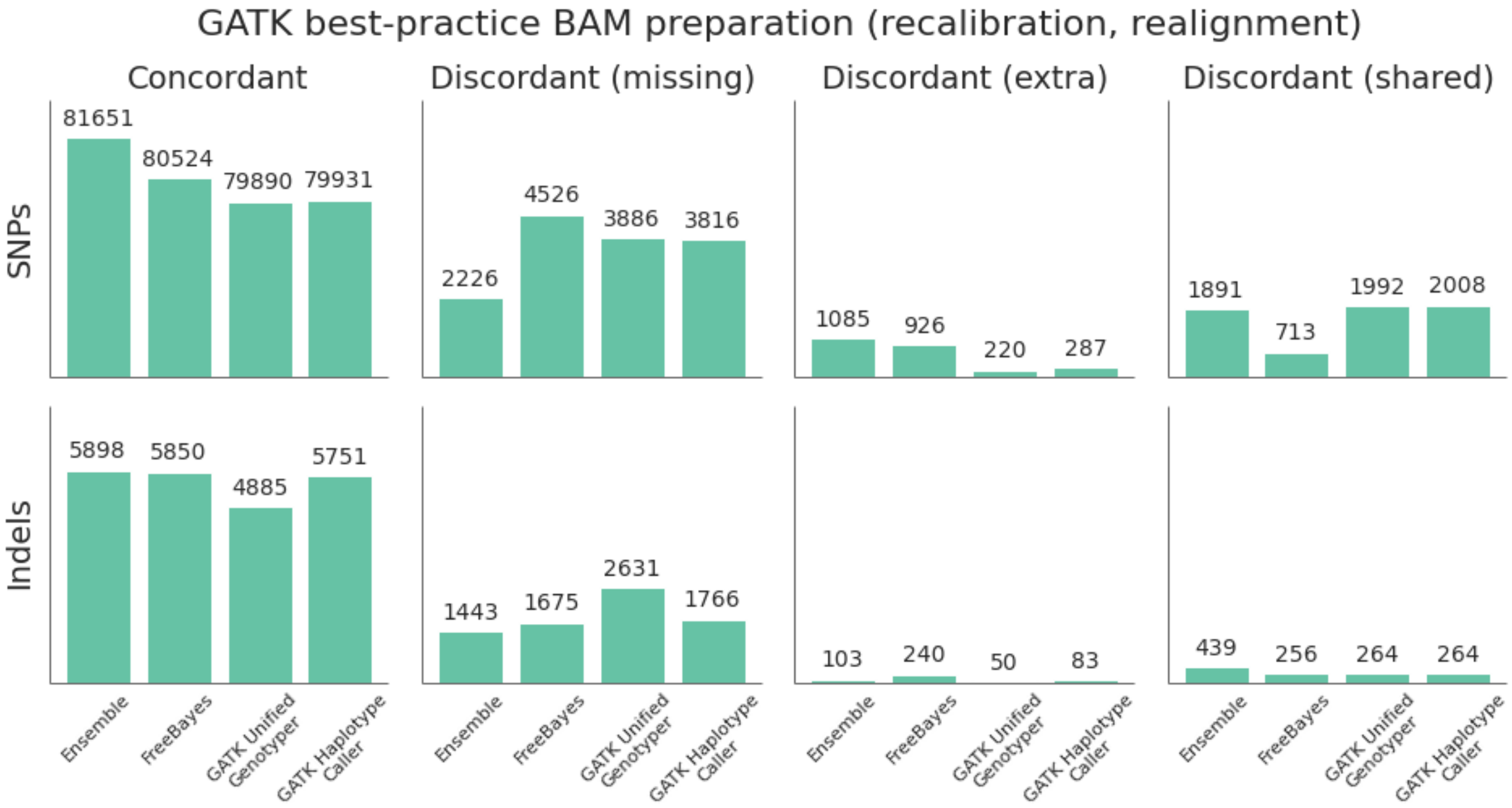
Variant caller evaluation

Highly confident set of reference (http://arxiv.org/abs/1307.4661)

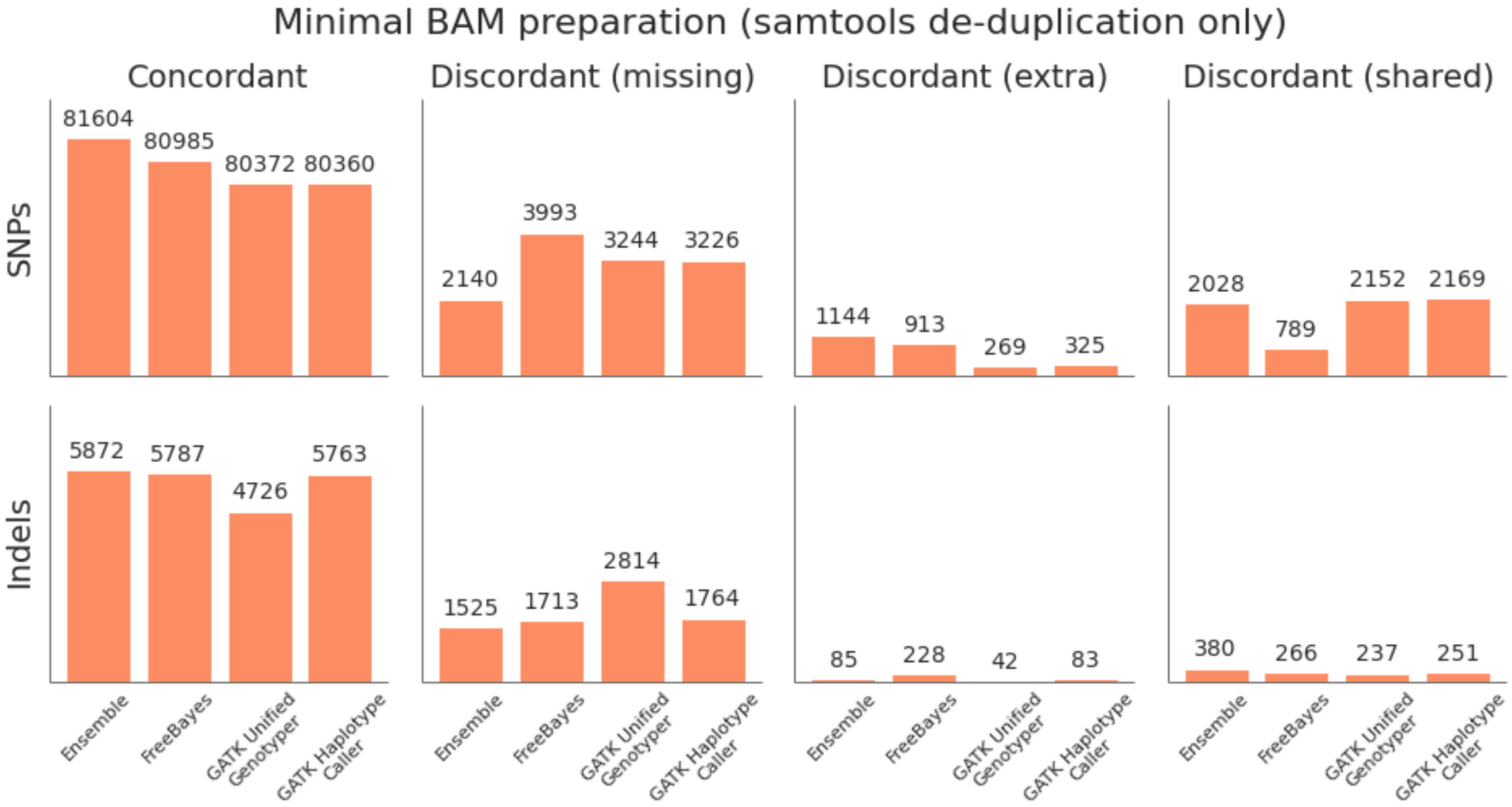
- GATK - UnifiedGenotyper
- GATK - Haplotype Caller
- FreeBayes
- Ensemble calling method (combining previous 3)

Post-processing methods of alignment (BAM files)
- *MarkDuplicates & GATK base QS recalibration & realignment around INDELs*
- Only de-duplication (samtools)

Results of variant caller evaluation



Results of variant caller evaluation



Summary of variant caller evaluation

FreeBayes outperforms the GATK callers on both SNP and INDEL calling; particularly resolving hom/het calls (discordant shared variants)

GATK **Haplotype Caller** is better than UnifiedGenotyper

Ensemble calling approach performs best - but takes the longest

What to look for ?

☑ What type of problem (Mendelian disorder, cancer, ...)

☑ Good coverage where variant is called

☑ Homopolymer problem ...

☑ Quality score recalibration and realignment around INDELs can be skipped in new versions - still good performance and huge speed improvements

Variant annotation

Variant annotation

Predict **functional impact** of variants
- query predefined scores
- list/combine different predictions

Basis for filtering variants
- Database links (dbSNP, refseq, GO)
- sequence/region-based functional annotation

Tools
ANNOVAR (command line / web)
SeattleSeq (web)
SnpEff (command line)
SVA (GUI)
VEP (web)

What to look for ?

☑ Annotating SNP & small INDELs works good

☑ Annotations for CNVs and SVs are limited

☑ Check versions of databases for filtering
e.g.: dbSNP 130 is considered contaminated with false positives

☑ Legal issues when using web applications

☑ Updates of underlying databases

Misc

Visualization

Visualization
IGV / Savant / UCSC genome browser

Locally installed pipelines & workflow tools
Galaxy (http://galaxy.i-med.ac.at)
Simplex - Java based exome seq pipeline (PMID: 22870267)

Tools
BEDtools, samtools, vcftools, cutadapt
UCSC tools (bedGraphToBigWig, wigToBigWig, ...)
Picard (insert size metrics, mark duplicates, ...)
...

Linux

most abundant sequence, its frequency, and percentage of total in file.fq:

```
cat myfile.fq | awk '((NR-2)%4==0){read=$1;total++;count[read]++}END{for(read in count){if(!max||count
```

Convert .bam back to .fastq:

```
samtools view file.bam | awk 'BEGIN {FS="\t"} {print "@" $1 "\n" $10 "\n+\n" $11}' > file.fq
```

Keep only top bit scores in blast hits (best bit score only):

```
awk '{ if(!x[$1]++) {print $0; bitscore=($14-1)} else { if($14>bitscore) print $0} }' blastout.txt
```

Keep only top bit scores in blast hits (5 less than the top):

```
awk '{ if(!x[$1]++) {print $0; bitscore=($14-6)} else { if($14>bitscore) print $0} }' blastout.txt
```

Trim leading whitespace in file.txt:

```
sed 's/^[ \t]*//' file.txt
```

Trim trailing whitespace in file.txt:

```
sed 's/[ \t]*$//' file.txt
```

https://github.com/stephenturner/oneliners

Whole exome sequencing

Data analysis