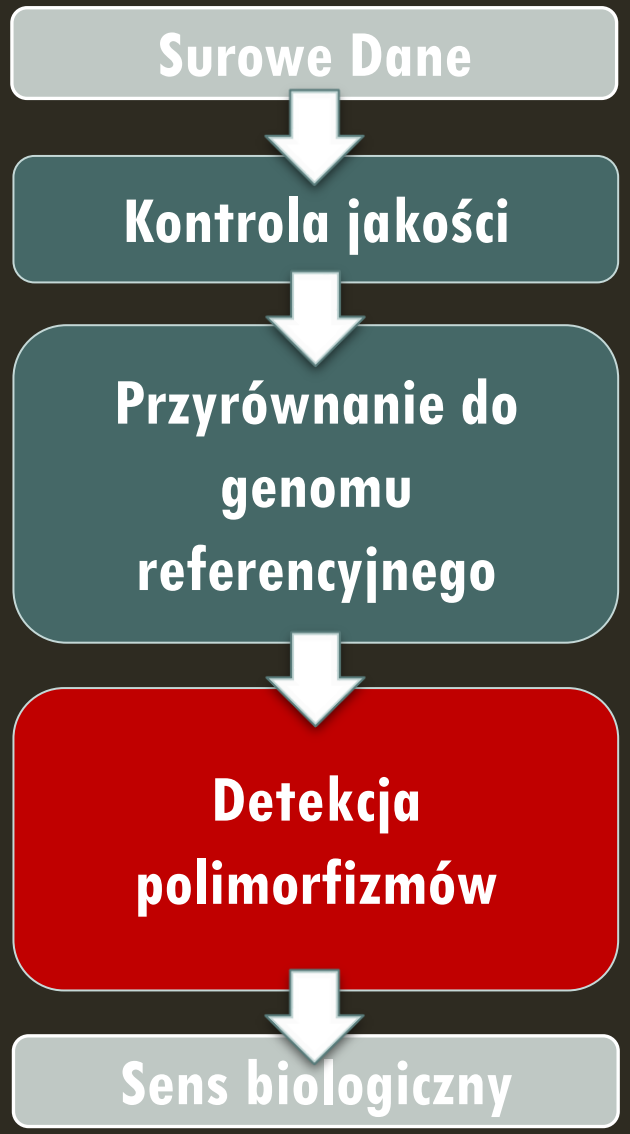
DETEKCJA POLIMORFIZMÓW

Analiza danych NGS
Wykład 6

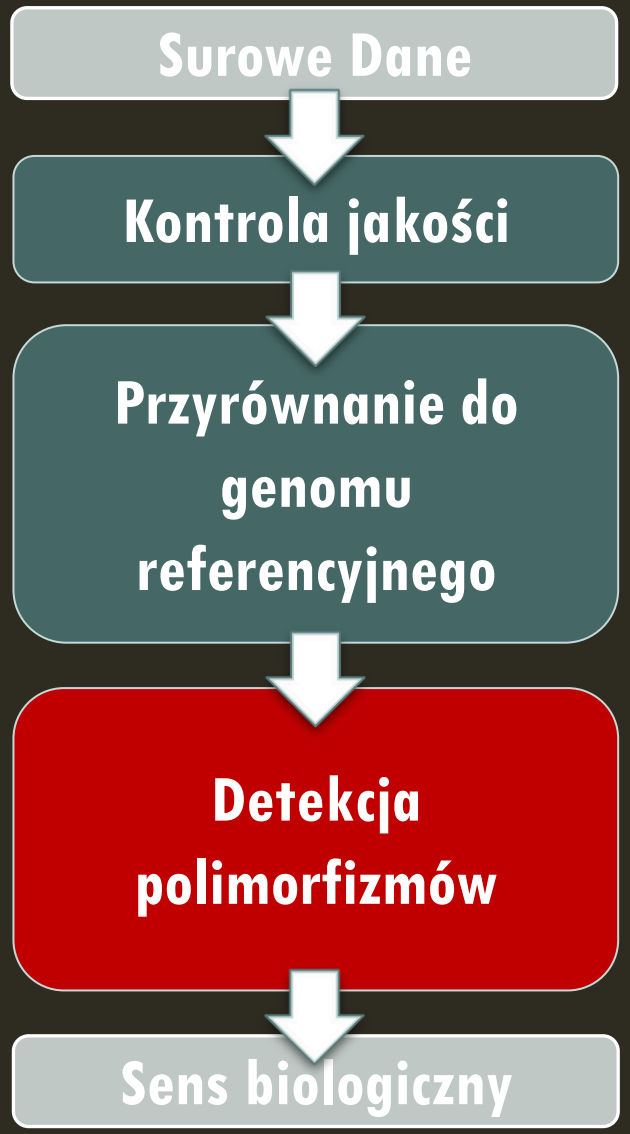
PIPELINE

Pipeline = łańcuch przetwarzania danych

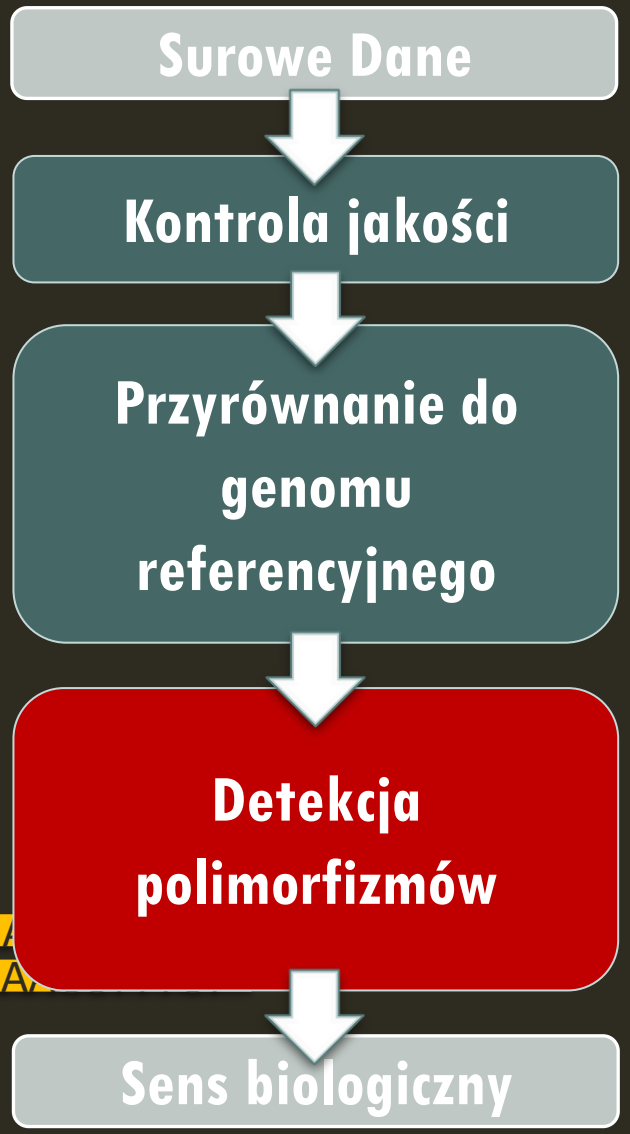


DETEKCJA POLIMORFIZMÓW

- Single Nucleotide Polymorphisms — SNP
- Insertions/Deletions — INDEL
- Copy Number Variations — CNV
- Loss of Heterozygosity — LOH
- Inversions — INV
- Translocations — TRANS

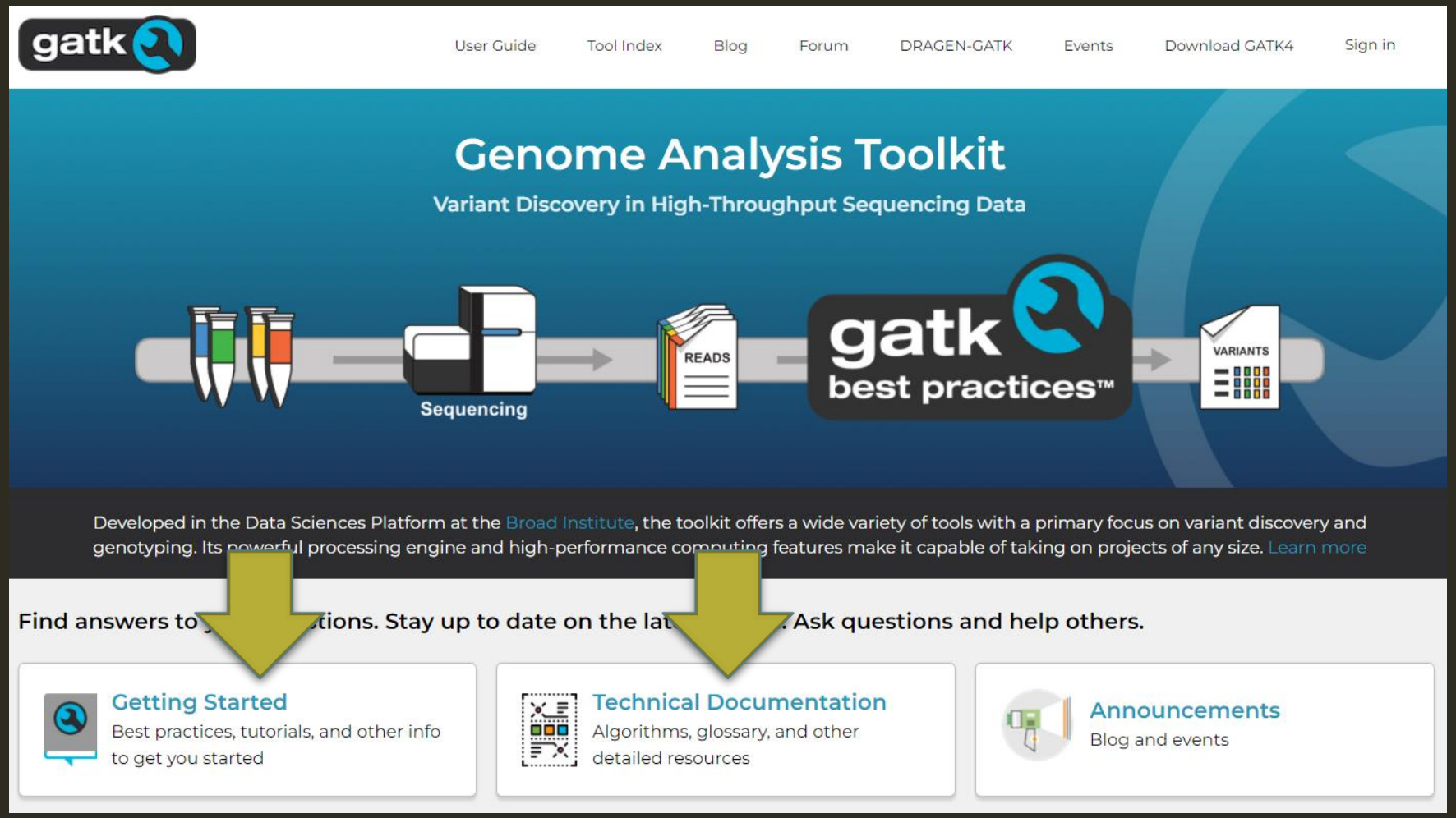


DETEKCJA POLIMORFIZMÓW



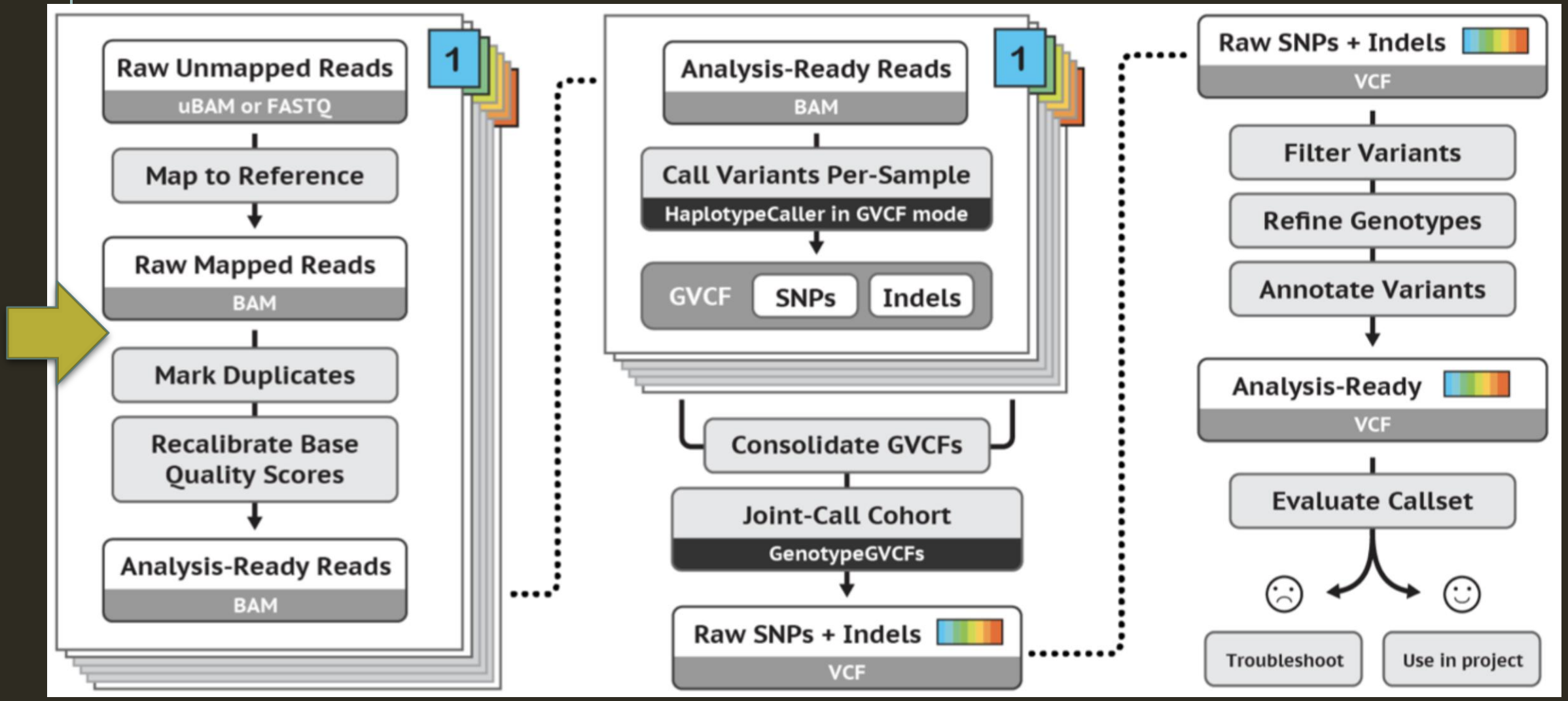
https://gatk.broadinstitute.org/hc/en-us

DETEKCJA POLIMORFIZMÓW – PAKIET GATK



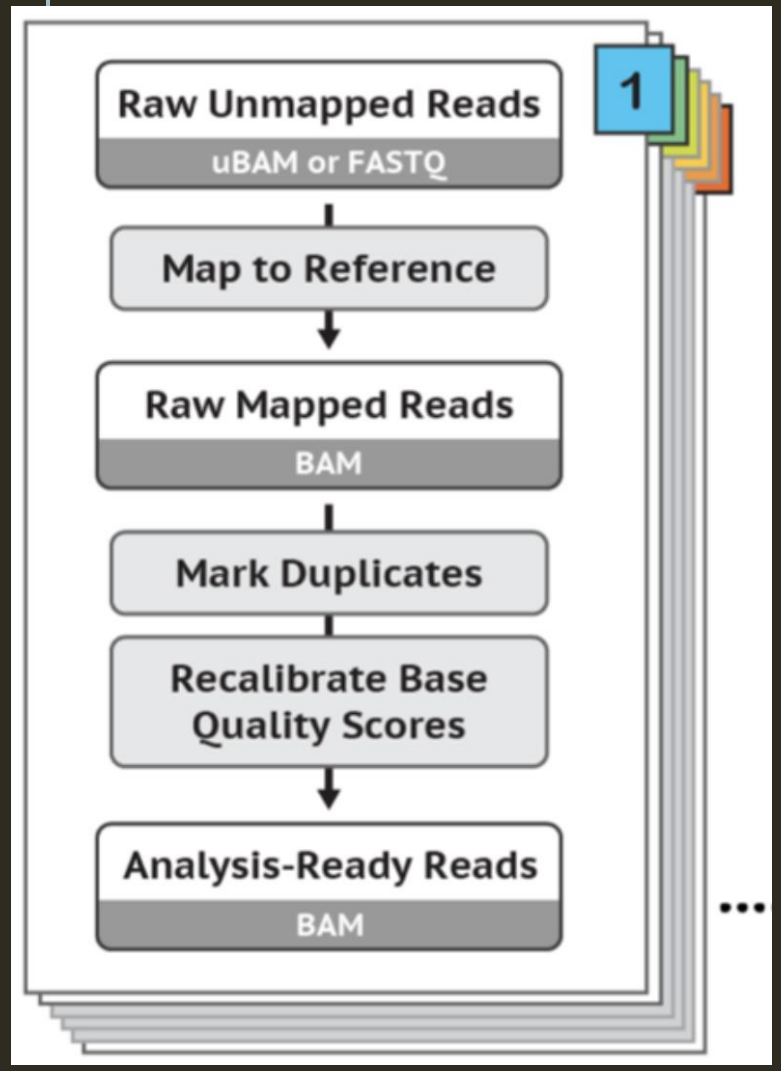
https://gatk.broadinstitute.org/hc/en-us

DETEKCJA POLIMORFIZMÓW – PAKIET GATK



https://gatk.broadinstitute.org/hc/en-us

DETEKCJA POLIMORFIZMÓW – PAKIET GATK

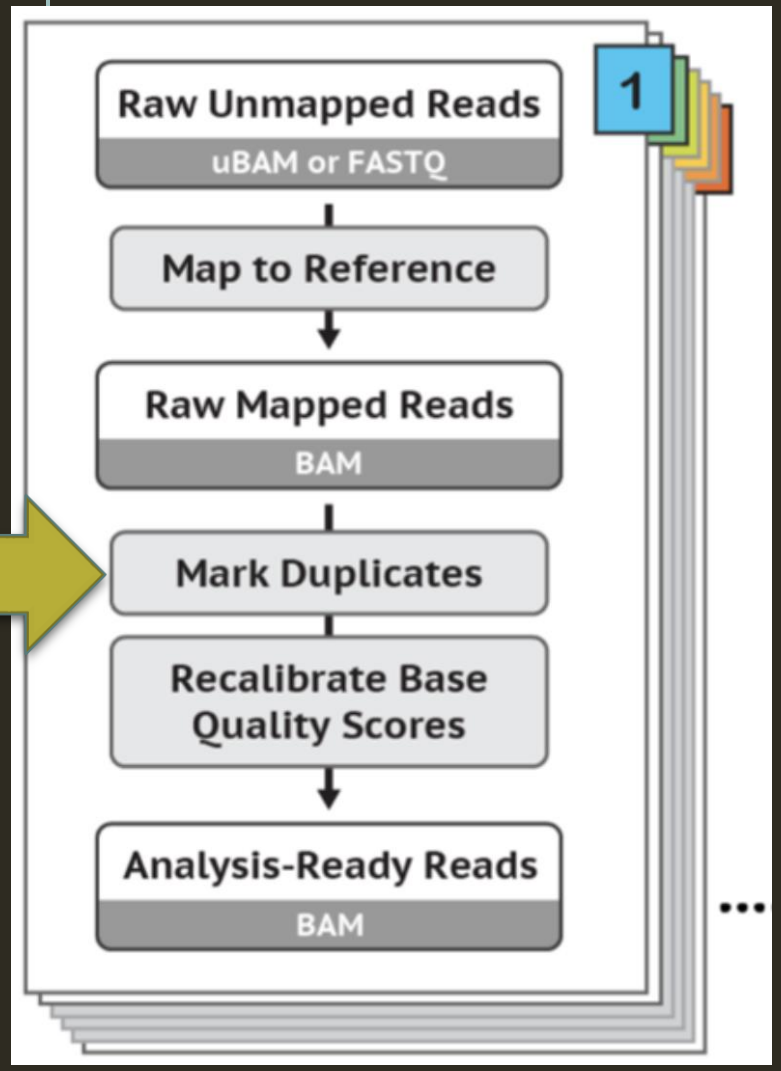


Mark Duplicates

„This tool locates and tags duplicate reads in a BAM or SAM file, where duplicate reads are defined as originating from a single fragment of DNA. Duplicates can arise during sample preparation e.g. library construction using PCR."

https://gatk.broadinstitute.org/hc/en-us

DETEKCJA POLIMORFIZMÓW – PAKIET GATK



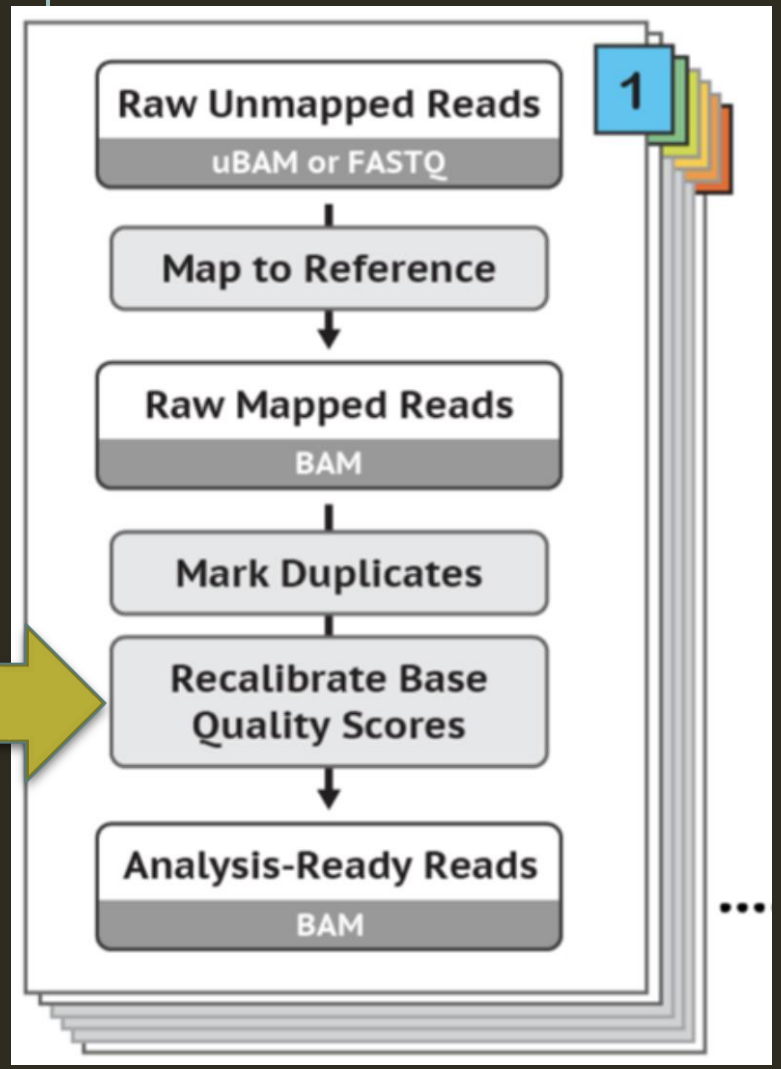
Mark Duplicates

„This tool locates and tags duplicate reads in a BAM or SAM file, where duplicate reads are defined as originating from a single fragment of DNA. Duplicates can arise during sample preparation e.g. library construction using PCR."

```
java -jar picard.jar MarkDuplicates \
      I=input.bam \
      O=marked_duplicates.bam \
      M=marked_dup_metrics.txt
```

https://gatk.broadinstitute.org/hc/en-us

DETEKCJA POLIMORFIZMÓW – PAKIET GATK



Base Quality Score Recalibration (BQSR)

„Base quality scores are per-base estimates of error emitted by the sequencing machines (…). For example, let's say the machine reads an A nucleotide, and assigns a quality score of Q20 in Phred-scale, that means it's 99% sure it identified the base correctly. This may seem high, but it does mean that we can expect it to be wrong in one case out of 100; so if we have several billion base calls (we get ~90 billion in a 30x genome), at that rate the machine would make the wrong call in 900 million bases -- which is a lot of bad bases. The quality score each base call gets is determined through some dark magic jealously guarded by the manufacturer of the sequencing machines."

DETEKCJA POLIMORFIZMÓW – PAKIET GATK

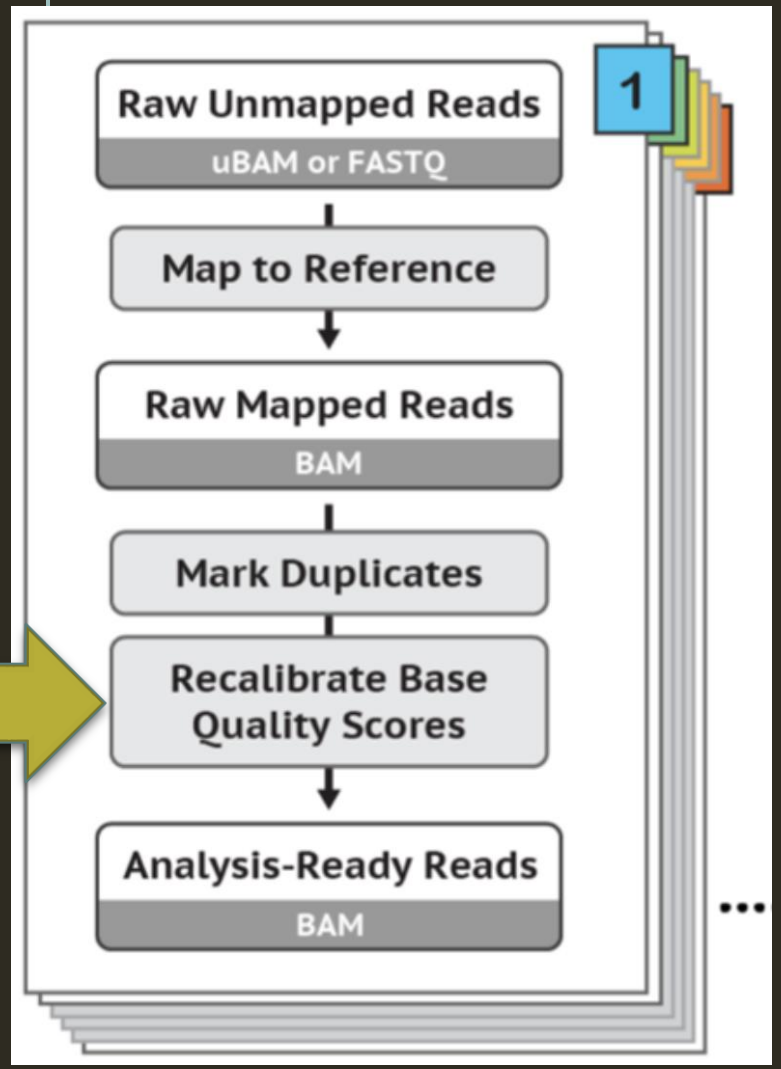
Base Quality Score Recalibration (BQSR)

„(…) the scores are subject to various sources of systematic (nonrandom) technical error, leading to over- or under-estimated base quality scores in the data. Some of these errors are due to the physics or the chemistry of how the sequencing reaction works, (…) o manufacturing flaws in the equipment. (…). We apply machine learning to model these errors empirically and adjust the quality scores accordingly. For example (…) whenever we called two A nucleotides in a row, the next base we called had a 1% higher rate of error. So any base call that comes after AA in a read should have its quality score reduced by 1%. We do that over several different covariates (mainly sequence context and position in read, or cycle) in a way that is additive. So the same base may have its quality score increased for one reason and decreased for another."

Szczegóły → gatk.broadinstitute.org/hc/en-us/articles/360035890531

https://gatk.broadinstitute.org/hc/en-us

DETEKCJA POLIMORFIZMÓW – PAKIET GATK



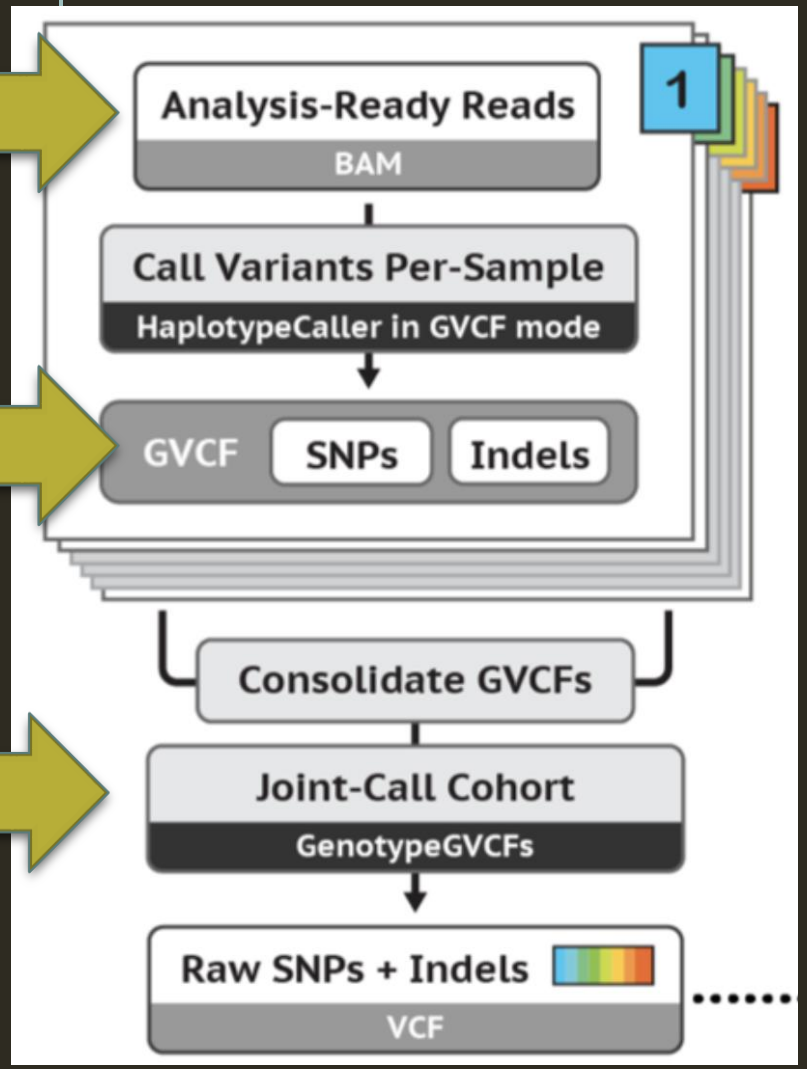
Base Quality Score Recalibration (BQSR)

```
gatk BaseRecalibrator \
  -I my_reads.bam \
  -R reference.fasta \
  --known-sites sites_of_variation.vcf \
  --known-sites another/optional/setOfSitesToMask.vcf \
  -O recal_data.table
```

```
gatk ApplyBQSR \
   -R reference.fasta \
   -I input.bam \
   --bqsr-recal-file recalibration.table \
   -O output.bam
```

https://gatk.broadinstitute.org/hc/en-us

DETEKCJA POLIMORFIZMÓW – PAKIET GATK



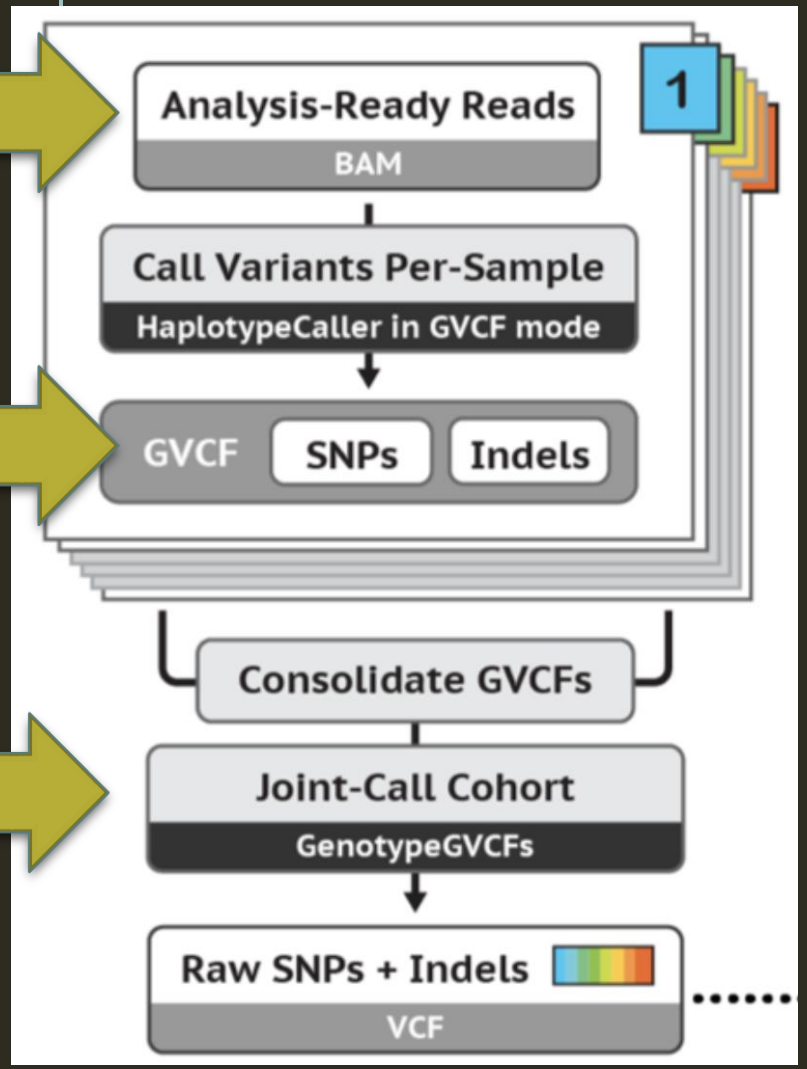
SNP and InDel calling

HaplotypeCaller „is capable of calling SNPs and indels simultaneously via local de-novo assembly of haplotypes in an active region. In other words, whenever the program encounters a region showing signs of variation, it discards the existing mapping information and completely reassembles the reads in that region. This allows the HaplotypeCaller to be more accurate when calling regions that are traditionally difficult to call, for example when they contain different types of variants close to each other."

CombineGVCFs „combine per-sample gVCF files produced by HaplotypeCaller into a multi-sample gVCF file."

https://gatk.broadinstitute.org/hc/en-us

DETEKCJA POLIMORFIZMÓW – PAKIET GATK



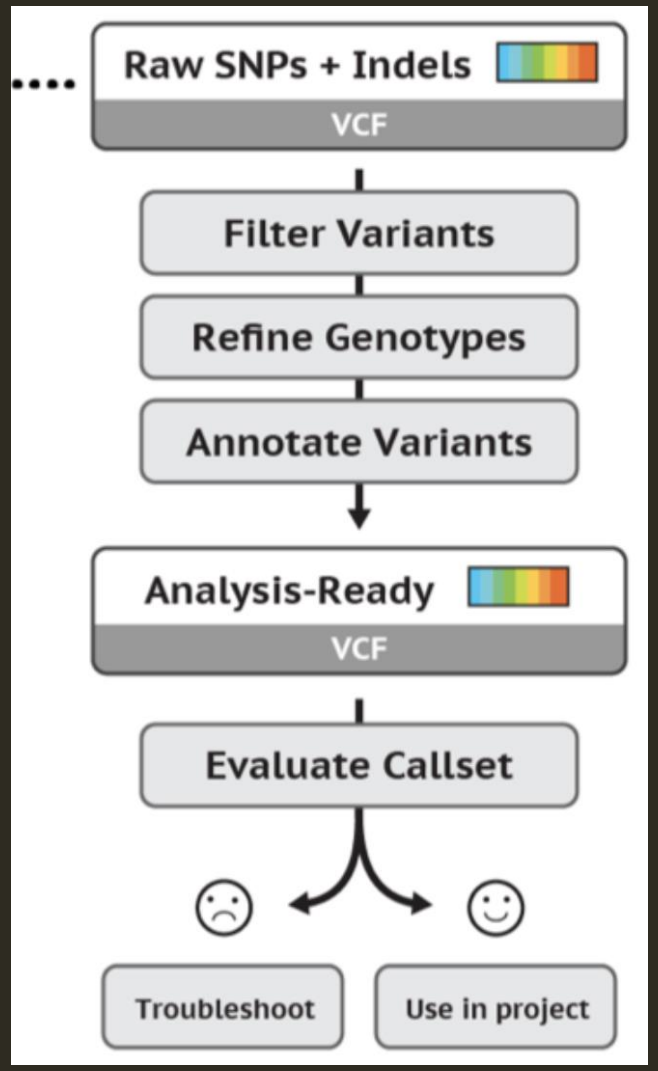
SNP and InDel calling

```
gatk --java-options "-Xmx4g" HaplotypeCaller  \
  -R Homo_sapiens_assembly38.fasta \
  -I input.bam \
  -O output.g.vcf.gz \
  -ERC GVCF
```

```
gatk CombineGVCFs \
  -R reference.fasta \
  --variant sample1.g.vcf.gz \
  --variant sample2.g.vcf.gz \
  -O cohort.g.vcf.gz
```

https://gatk.broadinstitute.org/hc/en-us

DETEKCJA POLIMORFIZMÓW – PAKIET GATK



Filtrowanie zbioru polimorfizmów
Adnotacja
Sens biologiczny

http://www.htslib.org/

DETEKCJA POLIMORFIZMÓW – SAMTOOLS

Samtools

Samtools is a suite of programs for interacting with high-throughput sequencing data. It consists of three separate repositories:

Samtools	Reading/writing/editing/indexing/viewing SAM/BAM/CRAM format
BCFtools	Reading/writing BCF2/VCF/gVCF files and calling/filtering/summarising SNP and short indel sequence variants
HTSlib	A C library for reading/writing high-throughput sequencing data

Samtools and BCFtools both use HTSlib internally, but these e packages contain their own copies of hts hey can be built independently.

Download

Source code releases can be downloaded from GitHub or Sourceforge:

Source release details

Workflows

We have described some standard workflows using Samtools:

- FASTQ to BAM / CRAM
- WGS/WES Mapping to Variant Calls
- Filtering of VCF Files
- Using CRAM within Samtools

Documentation

- Manuals
- HowTos
- Specifications
- Duplicate Marking
- Zlib Benchmarks
- CRAM Benchmarks
- Publications

Support

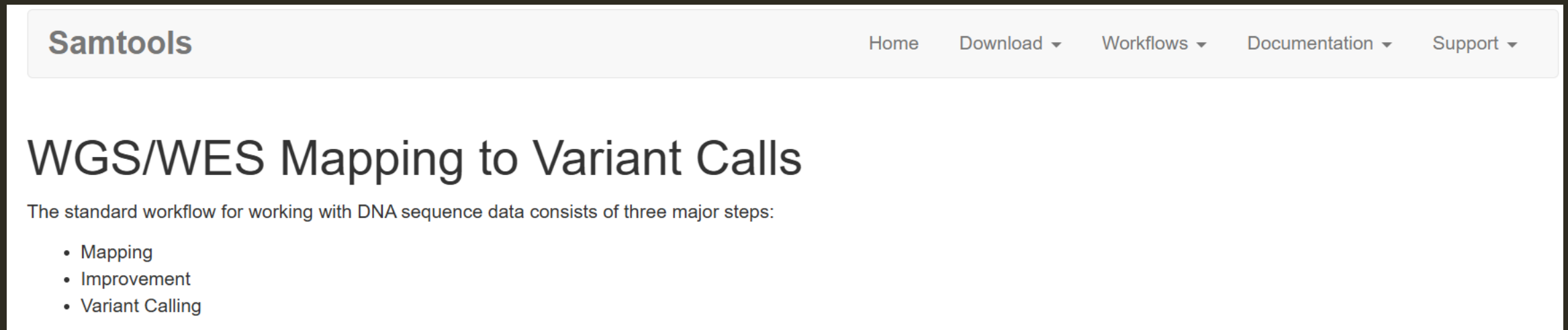
- Mailing Lists
- HTSlib issues
- BCFtools issues
- Samtools issues

DETEKCJA POLIMORFIZMÓW – SAMTOOLS

WGS/WES Mapping to Variant Calls

The standard workflow for working with DNA sequence data consists of three major steps:

- Mapping BWA
- Improvement GATK
- Variant Calling bcftools
- http://www.htslib.org/workflow/wgs-call.html



DETEKCJA POLIMORFIZMÓW – SAMTOOLS

WGS/WES Mapping to Variant Calls

The standard workflow for working with DNA sequence data consists of three major steps:

- Mapping BWA
- Improvement GATK
- Variant Calling bcftools

Variant Calling

To convert your BAM file into genomic positions we first use mpileup to produce a BCF file that contains all of the locations in the genome. We use this information to call genotypes and reduce our list of sites to those found to be variant by passing this file into bcftools call.

You can do this using a pipe as shown here:

```
bcftools mpileup -Ou -f <ref.fa> <sample1.bam> <sample2.bam> <sample3.bam> | bcftools call -vmO z -o <study.vcf.gz>
```

DETEKCJA POLIMORFIZMÓW – SAMTOOLS

WGS/WES Mapping to Variant Calls

The standard workflow for working with DNA seque

- Mapping BWA
- Improvement GATK
- Variant Calling bcftools

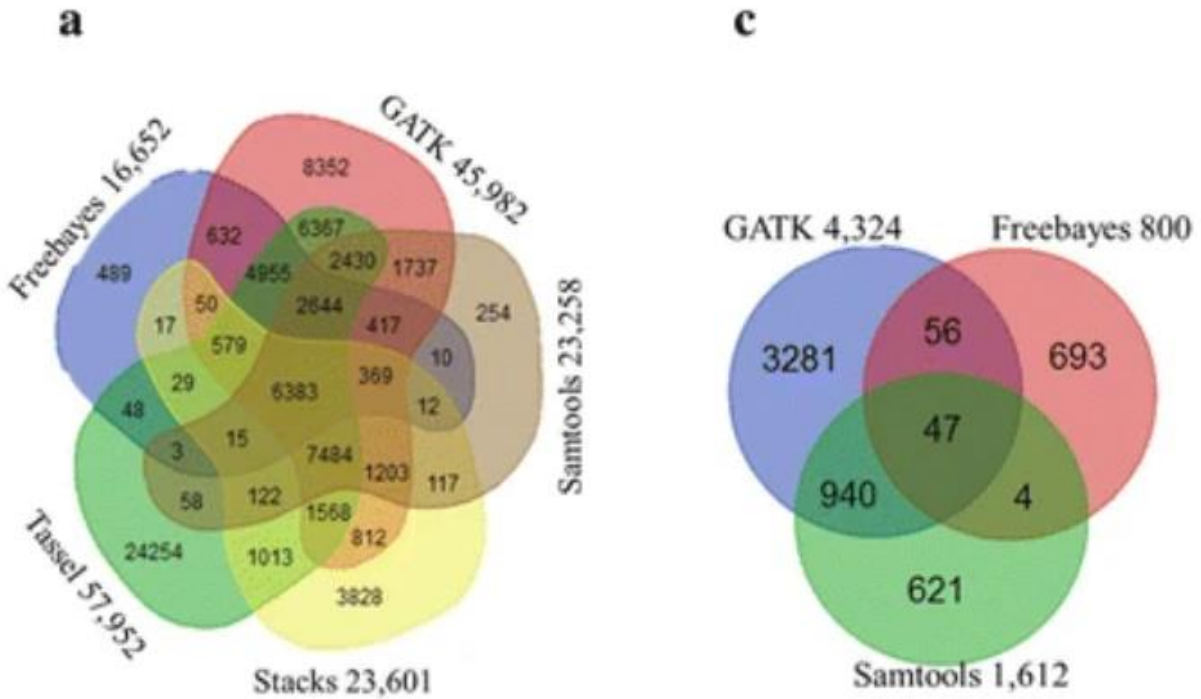
BCFtools

- Installation
  - Latest development version
- Calling
  - CNV calling
  - Consequence calling
  - Consensus calling
  - ROH calling
  - Variant calling and filtering
- Tips and Tricks
  - Converting formats
  - Extracting information
  - Filtering expressions
  - Plugins

SNP + INDELS

Mining sequence variations in representative polyploid sugarcane germplasm accessions

Xiping Yang [1], Jian Song [1], Qian You [1], Dev R Paudel [1], Jisen Zhang [2], Jianping Wang [3 4 5]



The number of variants called in the 14 sugarcane germplasm accessions. **a.** Venn diagram showing overlapping SNPs among five genome reference-based callers, Tassel, Stacks, Samtools, GATK, and Freebayes. (…) **c.** Venn diagram showing overlapping short insertions or deletions (InDels) among three reference-based callers, Samtools, GATK, and Freebayes

https://samtools.github.io/hts-specs/VCFv4.2.pdf

VCF

header section

polimorphisms (body)

```
##fileformat=VCFv4.2
##fileDate=20090805
##source=myImputationProgramV3.1
##reference=file:///seq/references/1000GenomesPilot-NCBI36.fasta
##contig=<ID=20,length=62435964,assembly=B36,md5=f126cdf8a6e0c7f379d618ff66beb2da,species="Homo sapiens",taxonomy=x>
##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      NA00001        NA00002        NA00003
20     14370   rs6054257 G      A       29   PASS   NS=3;DP=14;AF=0.5;DB;H2           GT:GQ:DP:HQ 0|0:48:1:51,51 1|0:48:8:51,51 1/1:43:5:.,.
20     17330   .         T      A       3    q10    NS=3;DP=11;AF=0.017               GT:GQ:DP:HQ 0|0:49:3:58,50 0|1:3:5:65,3   0/0:41:3
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|2:21:6:23,27 2|1:2:0:18,2   2/2:35:4
20     1230237 .         T      .       47   PASS   NS=3;DP=13;AA=T                   GT:GQ:DP:HQ 0|0:54:7:56,60 0|0:48:4:51,51 0/0:61:2
20     1234567 microsat1 GTC    G,GTCT  50   PASS   NS=3;DP=9;AA=G                    GT:GQ:DP    0/1:35:4       0/2:17:2       1/1:40:3
```

This example shows (in order): a good simple SNP, a possible SNP that has been filtered out because its quality is below 10, a site at which two alternate alleles are called, with one of them (T) being ancestral (possibly a reference sequencing error), a site that is called monomorphic reference (i.e. with no alternate alleles), and a microsatellite with two alternative alleles, one a deletion of 2 bases (TC), and the other an insertion of one base (T). Genotype data are given for three samples, two of which are phased and the third unphased, with per sample genotype quality, depth and haplotype qualities (the latter only for the phased samples) given as well as the genotypes. The microsatellite calls are unphased.

VCF

#CHROM POS ID REF ALT QUAL FILTER INFO FORMAT SAMPLE

Chr1 238 . C T 48 . DP=8;VDB=4.789490e-02;RPB=-1.551181e+00;AF1=0.5;AC1=1;DP4=4,1,3,0;MQ=44;FQ=51; PV4=1,1,0.16,0.41 GT:PL:GQ 0/1:78,0,135:81

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	NA00001	NA00002	NA00003
20	14370	rs6054257	G	A	29	PASS	NS=3;DP=14;AF=0.5;DB;H2	GT:GQ:DP:HQ	0|0:48:1:51,51	1|0:48:8:51,51	1/1:43:5:.,.
20	17330	.	T	A	3	q10	NS=3;DP=11;AF=0.017	GT:GQ:DP:HQ	0|0:49:3:58,50	0|1:3:5:65,3	0/0:41:3
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|2:21:6:23,27	2|1:2:0:18,2	2/2:35:4
20	1230237	.	T	.	47	PASS	NS=3;DP=13;AA=T	GT:GQ:DP:HQ	0|0:54:7:56,60	0|0:48:4:51,51	0/0:61:2
20	1234567	microsat1	GTC	G,GTCT	50	PASS	NS=3;DP=9;AA=G	GT:GQ:DP	0/1:35:4	0/2:17:2	1/1:40:3

This example shows (in order): a good simple SNP, a possible SNP that has been filtered out because its quality is below 10, a site at which two alternate alleles are called, with one of them (T) being ancestral (possibly a reference sequencing error), a site that is called monomorphic reference (i.e. with no alternate alleles), and a microsatellite with two alternative alleles, one a deletion of 2 bases (TC), and the other an insertion of one base (T). Genotype data are given for three samples, two of which are phased and the third unphased, with per sample genotype quality, depth and haplotype qualities (the latter only for the phased samples) given as well as the genotypes. The microsatellite calls are unphased.

#CHROM POS ID REF ALT QUAL FILTER INFO FORMAT SAMPLE

Chr1 238 . C T 48 . DP=8;VDB=4.789490e-02;RPB=-1.551181e+00;AF1=0.5;AC1=1;DP4=4,1,3,0;MQ=44;FQ=51; PV4=1,1,0.16,0.41 GT:PL:GQ 0/1:78,0,135:81

VCF

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	NA00001	NA00002	NA00003
20	14370	rs6054257	G	A	29	PASS	NS=3;DP=14;AF=0.5;DB;H2	GT:GQ:DP:HQ	0|0:48:1:51,51	1|0:48:8:51,51	1/1:43:5:.,.
20	17330	.	T	A	3	q10	NS=3;DP=11;AF=0.017	GT:GQ:DP:HQ	0|0:49:3:58,50	0|1:3:5:65,3	0/0:41:3
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|2:21:6:23,27	2|1:2:0:18,2	2/2:35:4
20	1230237	.	T	.	47	PASS	NS=3;DP=13;AA=T	GT:GQ:DP:HQ	0|0:54:7:56,60	0|0:48:4:51,51	0/0:61:2
20	1234567	microsat1	GTC	G,GTCT	50	PASS	NS=3;DP=9;AA=G	GT:GQ:DP	0/1:35:4	0/2:17:2	1/1:40:3

This example shows (in order): a good simple SNP, a possible SNP that has been filtered out because its quality is below 10, a site at which two alternate alleles are called, with one of them (T) being ancestral (possibly a reference sequencing error), a site that is called monomorphic reference (i.e. with no alternate alleles), and a microsatellite with two alternative alleles, one a deletion of 2 bases (TC), and the other an insertion of one base (T). Genotype data are given for three samples, two of which are phased and the third unphased, with per sample genotype quality, depth and haplotype qualities (the latter only for the phased samples) given as well as the genotypes. The microsatellite calls are unphased.

VCF

#CHROM POS ID REF ALT QUAL FILTER INFO FORMAT SAMPLE

Chr1 238 . C T 48 . DP=8;VDB=4.789490e-02;RPB=-1.551181e+00;AF1=0.5;AC1=1;DP4=4,1,3,0;MQ=44;FQ=51; PV4=1,1,0.16,0.41 GT:PL:GQ 0/1:78,0,135:81

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	NA00001	NA00002	NA00003
20	14370	rs6054257	G	A	29	PASS	NS=3;DP=14;AF=0.5;DB;H2	GT:GQ:DP:HQ	0|0:48:1:51,51	1|0:48:8:51,51	1/1:43:5:.,.
20	17330	.	T	A	3	q10	NS=3;DP=11;AF=0.017	GT:GQ:DP:HQ	0|0:49:3:58,50	0|1:3:5:65,3	0/0:41:3
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|2:21:6:23,27	2|1:2:0:18,2	2/2:35:4
20	1230237	.	T	.	47	PASS	NS=3;DP=13;AA=T	GT:GQ:DP:HQ	0|0:54:7:56,60	0|0:48:4:51,51	0/0:61:2
20	1234567	microsat1	GTC	G,GTCT	50	PASS	NS=3;DP=9;AA=G	GT:GQ:DP	0/1:35:4	0/2:17:2	1/1:40:3

This example shows (in order): a good simple SNP, a possible SNP that has been filtered out because its quality is below 10, a site at which two alternate alleles are called, with one of them (T) being ancestral (possibly a reference sequencing error), a site that is called monomorphic reference (i.e. with no alternate alleles), and a microsatellite with two alternative alleles, one a deletion of 2 bases (TC), and the other an insertion of one base (T). Genotype data are given for three samples, two of which are phased and the third unphased, with per sample genotype quality, depth and haplotype qualities (the latter only for the phased samples) given as well as the genotypes. The microsatellite calls are unphased.

VCF

#CHROM POS ID REF ALT QUAL FILTER INFO FORMAT SAMPLE

Chr1 238 . C T 48 . DP=8;VDB=4.789490e-02;RPB=-1.551181e+00;AF1=0.5;AC1=1;DP4=4,1,3,0;MQ=44;FQ=51; PV4=1,1,0.16,0.41 GT:PL:GQ 0/1:78,0,135:81

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	NA00001	NA00002	NA00003
20	14370	rs6054257	G	A	29	PASS	NS=3;DP=14;AF=0.5;DB;H2	GT:GQ:DP:HQ	0|0:48:1:51,51	1|0:48:8:51,51	1/1:43:5:.,.
20	17330	.	T	A	3	q10	NS=3;DP=11;AF=0.017	GT:GQ:DP:HQ	0|0:49:3:58,50	0|1:3:5:65,3	0/0:41:3
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|2:21:6:23,27	2|1:2:0:18,2	2/2:35:4
20	1230237	.	T	.	47	PASS	NS=3;DP=13;AA=T	GT:GQ:DP:HQ	0|0:54:7:56,60	0|0:48:4:51,51	0/0:61:2
20	1234567	microsat1	GTC	G,GTCT	50	PASS	NS=3;DP=9;AA=G	GT:GQ:DP	0/1:35:4	0/2:17:2	1/1:40:3

This example shows (in order): a good simple SNP, a possible SNP that has been filtered out because its quality is below 10, a site at which two alternate alleles are called, with one of them (T) being ancestral (possibly a reference sequencing error), a site that is called monomorphic reference (i.e. with no alternate alleles), and a microsatellite with two alternative alleles, one a deletion of 2 bases (TC), and the other an insertion of one base (T). Genotype data are given for three samples, two of which are phased and the third unphased, with per sample genotype quality, depth and haplotype qualities (the latter only for the phased samples) given as well as the genotypes. The microsatellite calls are unphased.

SINGLE- AND MULTI-SAMPLE

VCF

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	NA00001	NA00002	NA00003
20	14370	rs6054257	G	A	29	PASS	NS=3;DP=14;AF=0.5;DB;H2	GT:GQ:DP:HQ	0|0:48:1:51,51	1|0:48:8:51,51	1/1:43:5:.,.
20	17330	.	T	A	3	q10	NS=3;DP=11;AF=0.017	GT:GQ:DP:HQ	0|0:49:3:58,50	0|1:3:5:65,3	0/0:41:3
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|2:21:6:23,27	2|1:2:0:18,2	2/2:35:4
20	1230237	.	T	.	47	PASS	NS=3;DP=13;AA=T	GT:GQ:DP:HQ	0|0:54:7:56,60	0|0:48:4:51,51	0/0:61:2
20	1234567	microsat1	GTC	G,GTCT	50	PASS	NS=3;DP=9;AA=G	GT:GQ:DP	0/1:35:4	0/2:17:2	1/1:40:3

This example shows (in order): a good simple SNP, a possible SNP that has been filtered out because its quality is below 10, a site at which two alternate alleles are called, with one of them (T) being ancestral (possibly a reference sequencing error), a site that is called monomorphic reference (i.e. with no alternate alleles), and a microsatellite with two alternative alleles, one a deletion of 2 bases (TC), and the other an insertion of one base (T). Genotype data are given for three samples, two of which are phased and the third unphased, with per sample genotype quality, depth and haplotype qualities (the latter only for the phased samples) given as well as the genotypes. The microsatellite calls are unphased.

VARIANTS FILTERING
EXAMPLES

- Removal of variants with 2 or more alternative alleles
- Overall quality (QUAL)
- Setting minimum and maximum read depths for filtering (DP)
  - Set minimum number as 10 across all animals
  - Set maximum as: median read depth + 3 * standard deviation read depth
- Remove variants with the same basepair position
  - If two variants have the same bp position both are removed
  - Resolves issue with SNP and INDEL calls at same position

(…)

VARIANTS FILTERING

Toward better understanding of artifacts in variant calling from high-coverage samples

Heng Li [1]

Several universal filters applicable to most callers:

- Low-complexity (LC) filter: filtering variants overlapping with low-complexity regions (LCRs)
- Maximum depth (MD) filter: filtering sites covered by excessive number of reads. It should be noted that different callers may define the depth differently. For example, Platypus apparently only counts reads with unambiguous realignment. The read depth reported in the Platypus VCF is noticeably smaller in comparison with other callers.
- Quality filter (QU): filtering sites with the reported variant quality below a threshold.
- others

VARIANTS FILTERING

Next-generation data filtering in the genomics era

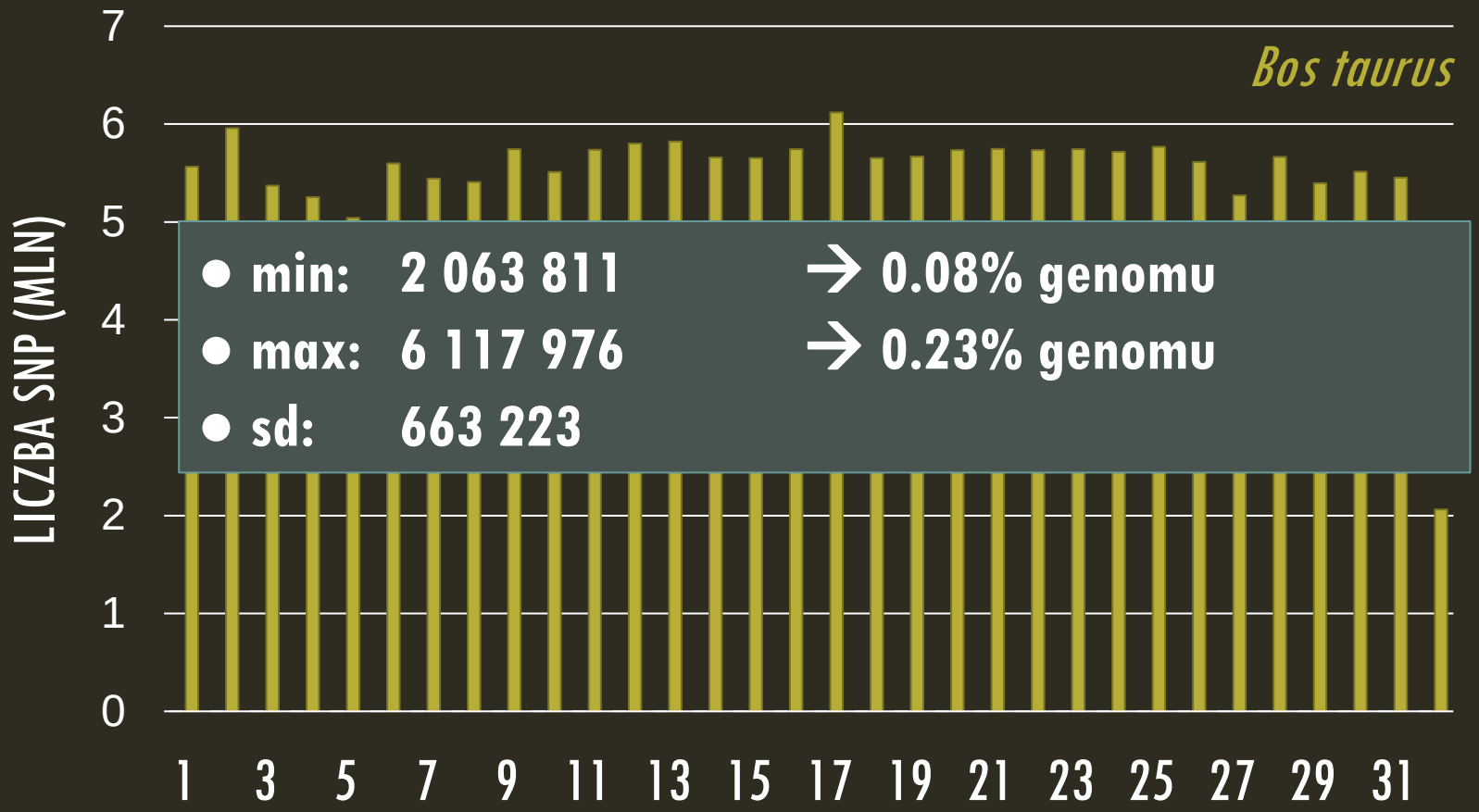
William Hemstrom, Jared A. Grummer, Gordon Luikart & Mark R. Christie

Nature Reviews Genetics **25**, 750–767 (2024) | Cite this article

Abstract

Genomic data are ubiquitous across disciplines, from agriculture to biodiversity, ecology, evolution and human health. However, these datasets often contain noise or errors and are missing information that can affect the accuracy and reliability of subsequent computational analyses and conclusions. A key step in genomic data analysis is filtering – removing sequencing bases, reads, genetic variants and/or individuals from a dataset – to improve data quality for downstream analyses. Researchers are confronted with a multitude of choices when filtering genomic data; they must choose which filters to apply and select appropriate thresholds. To help usher in the next generation of genomic data filtering, we review and suggest best practices to improve the implementation, reproducibility and reporting standards for filter types and thresholds commonly applied to genomic datasets. We focus mainly on filters for minor allele frequency, missing data per individual or per locus, linkage disequilibrium and Hardy–Weinberg deviations. Using simulated and empirical datasets, we illustrate the large effects of different filtering thresholds on common population genetics statistics, such as Tajima's D value, population differentiation (F_{ST}), nucleotide diversity (π) and effective population size (N_e).

DETEKCJA POLIMORFIZMÓW → SNP



Surowe Dane

Kontrola jakości

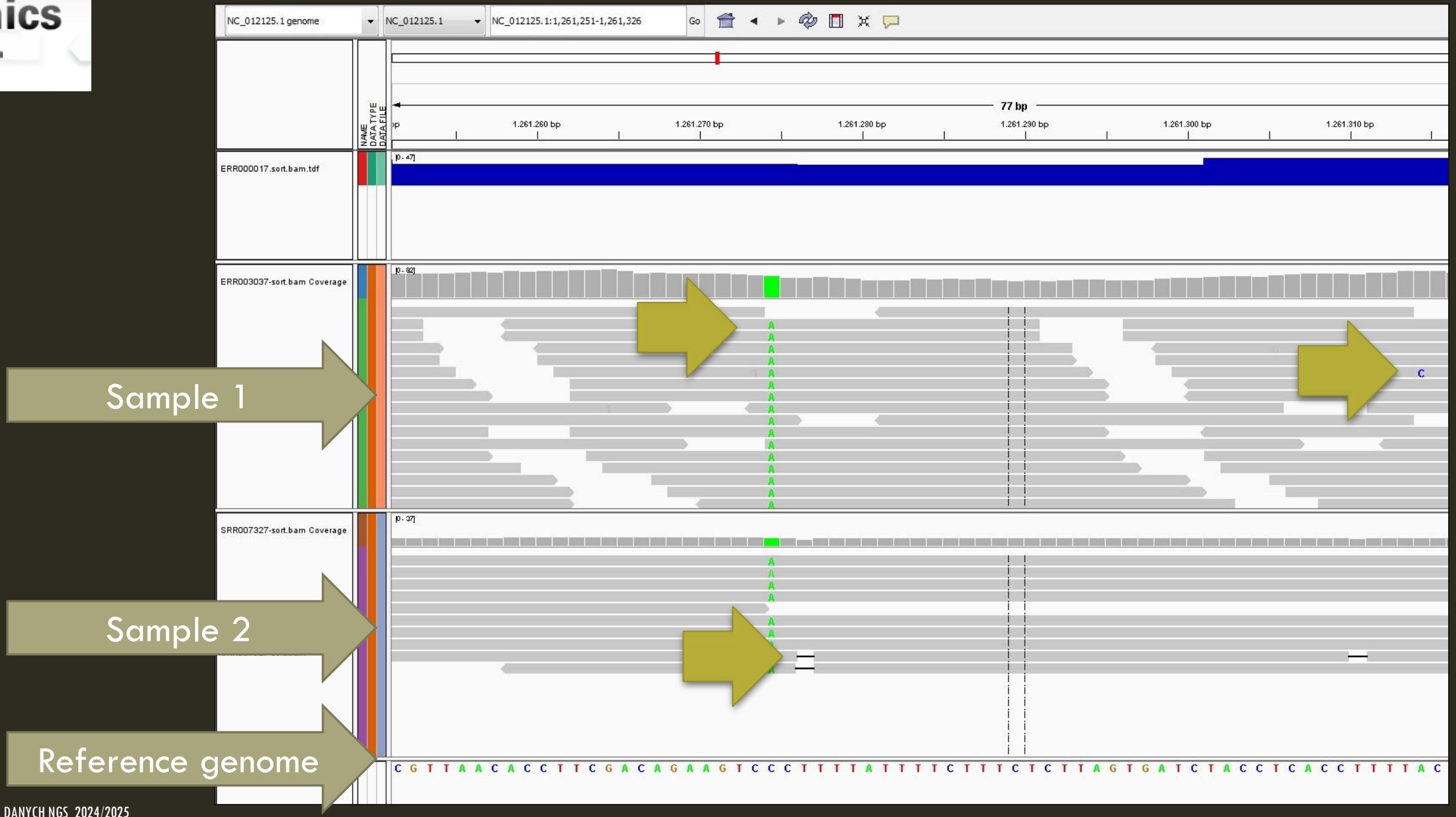
Przyrównanie do genomu referencyjnego

Detekcja polimorfizmów

Sens biologiczny

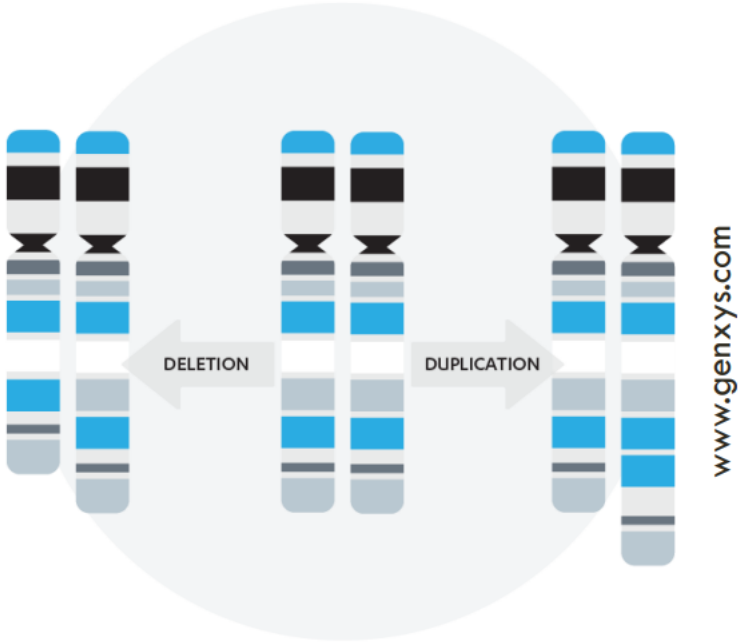
Integrative Genomics Viewer

http://software.broadinstitute.org/software/igv/



DETEKCJA CNV

DETEKCJA POLIMORFIZMÓW → CNV

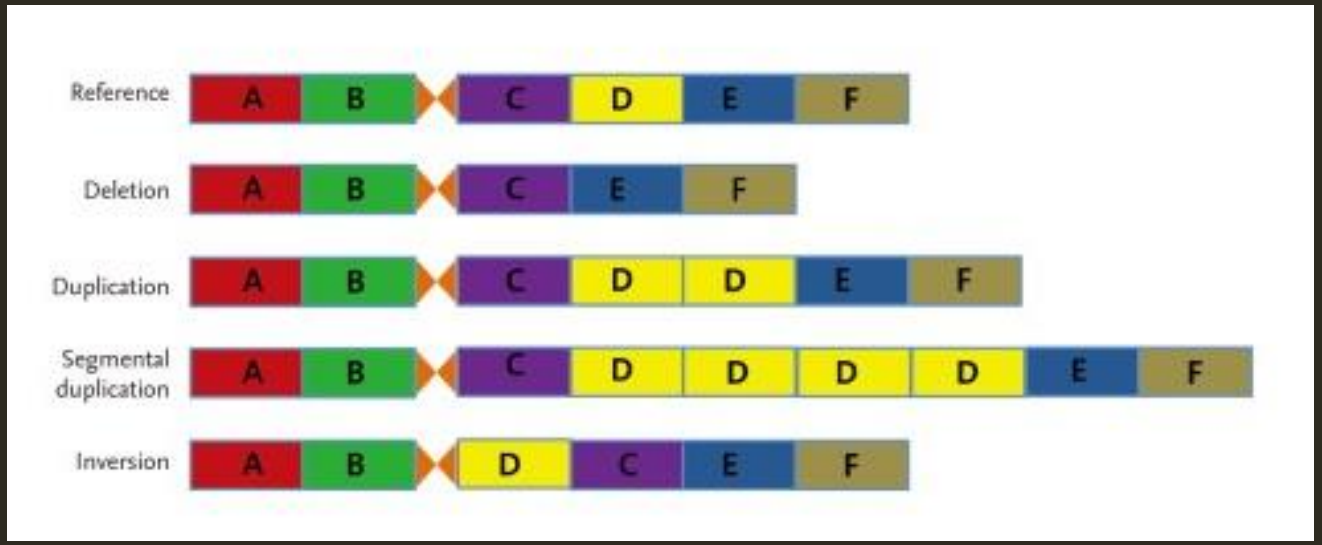


Rozmiar: bp - Mbp

Yim et al. 2015

COPY NUMER VARIATION (CNV)

Zmienność liczby kopii obejmuje duplikacje i delecje dłuższe niż 1 000 pz (rożne definicje).



„Different types of CNVs and an example of genome-wide detection of CNVs. The plot illustrates deletion, duplication, and multiple segmental duplication of the "D" locus compared with the reference genome. Inversion of "C" and "D" loci is also illustrated."

Mills et al. 2011

COPY NUMBER VARIATION (CNV)

CNVs → **gains** and **losses** (insertions and deletions) of genomic sequence greater than **50 bp** between two individuals of a species (Mills et al. 2011)

SNPs more frequent than CNVs
CNVs higher percentage of genomic sequence, greater effects, including the changing of gene structure and expression

Interrogation of the genome for both common and rare variations, including SNPs and CNVs, was proposed as an effective way to elucidate the causes of complex disease and traits (Manolio et al. 2009)

CNV → FENOTYP

CNVs:

- zmieniają strukturę i ekspresję genów
- są związane z wieloma chorobami
- mogą wpływać pozytywnie na proces adaptacji do zmieniającego się środowiska

Locus	CNV	Target gene	Disorder
2q31-q32	dup	*HOXD*	Mesomelic dysplasia Kantaputra type
2q35	dup	*IHH*	Syndactyly type 1, craniosynostosis Philadelphia type
7q36	dup	*SHH*	Triphalangeal thumb–polysyndactyly syndrome
	dup		Preaxial polydactyly type 2
	dup		Werner mesomelic syndrome
	dup		Haas polysyndactyly
	dup		Laurin–Sandrow syndrome
17q24.2-q24.3	dup		Congenital generalized hypertrichosis terminalis with or without gingival hyperplasia
17q24.3	dup	*SOX9*	Cooks syndrome
	del		Pierre Robin syndrome
	dup		Female-to-male sex reversal
20p12	dup	*BMP2*	Brachydactyly type A2
Xp	del	*SHOX*	Idiopathic short stature, Léri–Weill dyschondrosteosis
Xq22	del	*PLP1*	Pelizaeus–Merzbacher disease
	dup		Spastic paraplegia type 2
Xp21.2	del	*DAX1*	Male-to-female sex reversal

Klopocki and Mundlos 2011. Copy-Number Variations, Noncoding Sequences, and Human Phenotypes. Annual Review of Genomics and Human Genetics.12:1, 53-72.

CNV → FENOTYP

Table 2 Functional impacts of cattle CNV

References	Phenotype	Mechanism
Ohba et al. (2000) and Hirano et al. (2000)	Renal tubular dysplasia	Deletions of *CLDN-16* gene: Type 1—Deletion of 37 kb (exons 1 to 4); Type 2—56 kb (exons 1 to 4 and 821 bp of exon 5)
Drogemuller et al. (2001)	Anhidrotic ectodermal dysplasia	A deletion including exon 3 of the *EDA* gene
Sugimoto et al. (2003)	Myopathy of diaphragmatic muscles	A deletion of the *HSPA1B* gene
Dreger and Schmutz (2010)	Variant Red	Partially or completely duplicated *DEFB103* (β Defensin 3)
Meyers et al. (2010)	Osteopetrosis	A 2.8-kb deletion including exons 2 and 3 of the *SLC4A2* gene
Flisikowski et al. (2010)	Abortions and stillbirths	A 110-kb deletion of exons 3 and 4 of the *MIMT1* gene
Hou et al. (2011a, b)	Parasite resistance	A case–control study of CNV events associated with parasite resistance
Durkin et al. (2012)	Color sidedness	Serial translocation by means of circular intermediates encompassing *KIT*

METODY DETEKCJI CNV

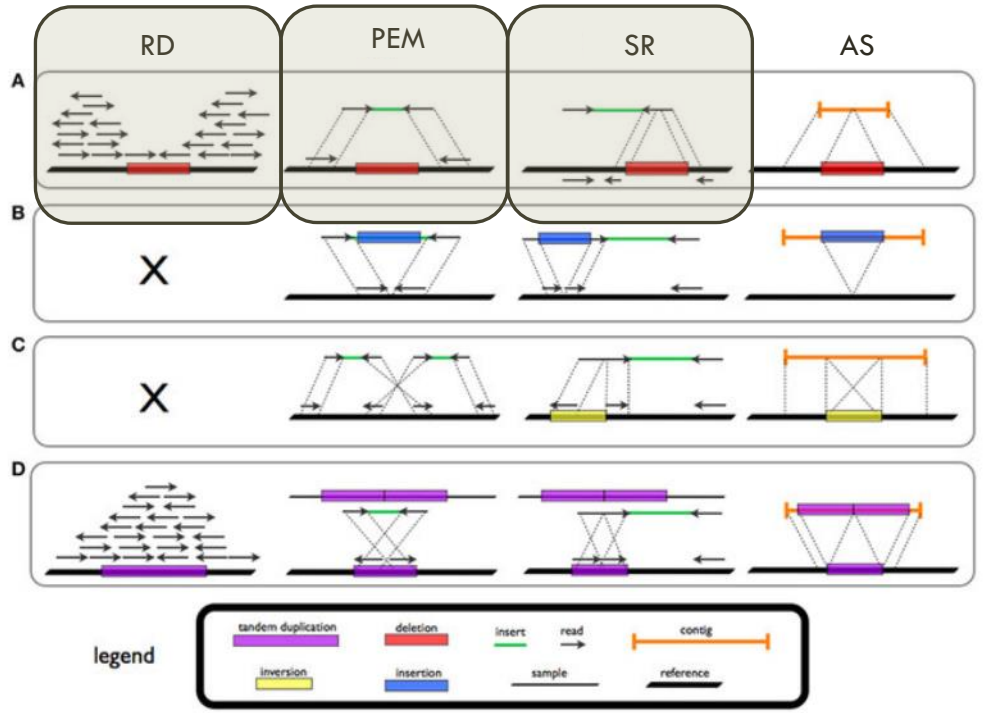
PEM/RP
Paired-end mapping

SR
Split-read

DOC/RD/RC
Depth of coverage/Read Depth/ Read Count

AS
Assembly- based

CB
Combinations of Methods



Figure

Caption

Figure 1: Signatures and patterns of SVs for deletion (A), novel sequence insertion (B), inversion (C), and tandem duplication (D) in read count (RC), read-pair (RP), split-read (SR), and de novo assembly (AS) methods.

METODY DETEKCJI CNV

PEM/RP

→ limited by the insert size

→ detects short CNVs

SR

→ applicable to the unique regions in the reference genome

DOC/RD/RC

→ detect larger CNVs in complex genomic region classes, which are difficult to detect using PEM and SR

AS

→ overwhelming demand on computational resources

→ perform poorly for repeated regions

CB

→ none of tools is able to detect the full spectrum of all types of CNVs with high sensitivity and specificity

RESEARCH ARTICLE Open Access

An evaluation of copy number variation detection tools for cancer using whole exome sequencing data

Fatima Zare[1], Michelle Dow[2], Nicholas Monteleone[1], Abdelrahm

WES

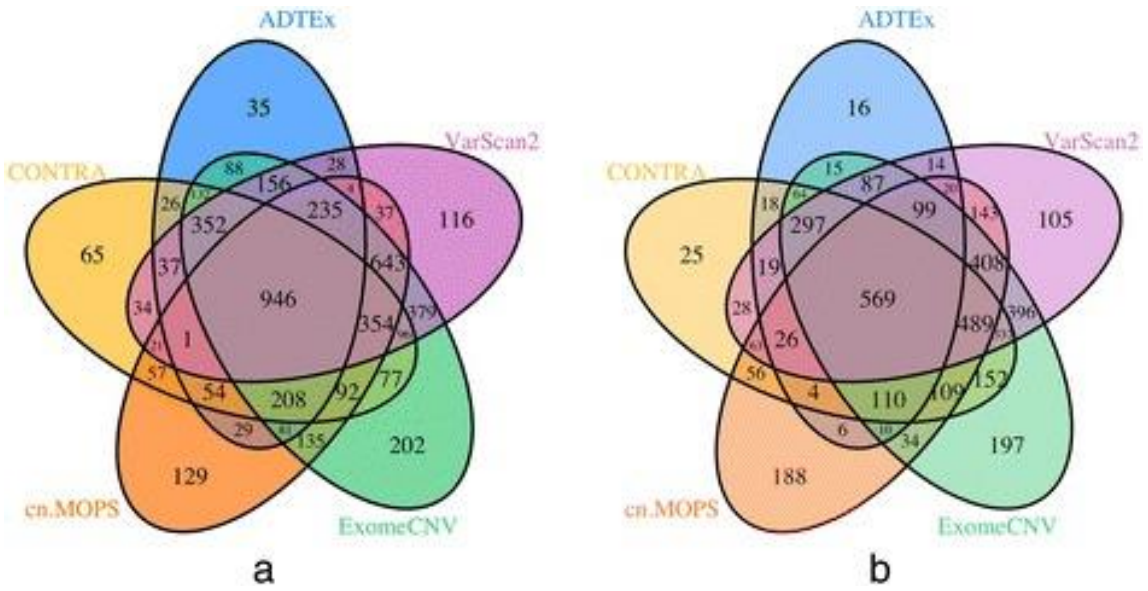
Abstract

Background: Recently copy number variation (CNV) has gained considerable interest as a type of genomic/genetic variation that plays an important role in disease susceptibility. Advances in sequencing technology have created an opportunity for detecting CNVs more accurately. Recently whole exome sequencing (WES) has become primary strategy for sequencing patient samples and study their genomics aberrations. However, compared to whole genome sequencing, WES introduces more biases and noise that make CNV detection very challenging. Additionally, tumors' complexity makes the detection of cancer specific CNVs even more difficult. Although many CNV detection tools have been developed since introducing NGS data, there are few tools for somatic CNV detection for WES data in cancer.

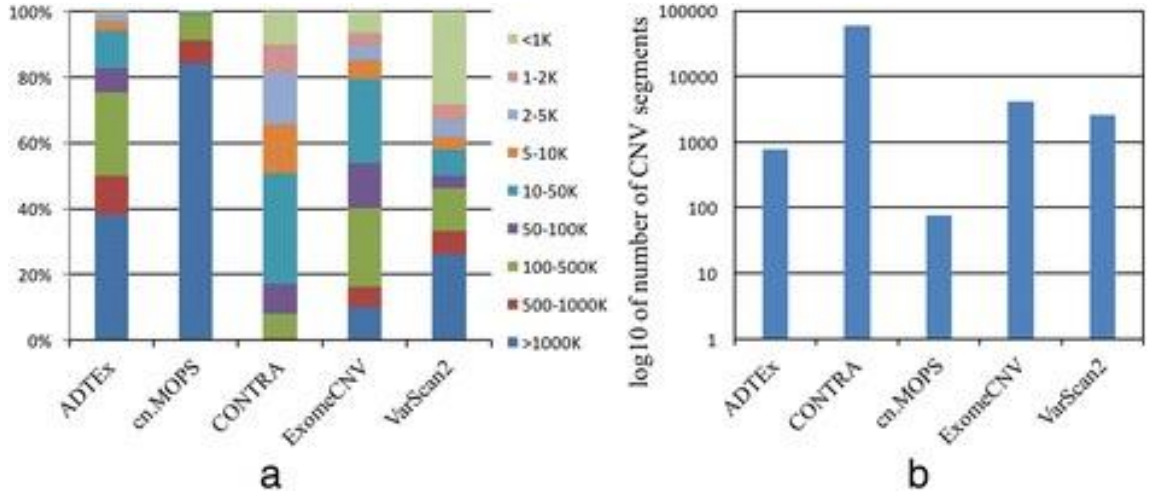
Results: In this study, we evaluated the performance of the most recent and commonly used CNV detection tools for WES data in cancer to address their limitations and provide guidelines for developing new ones. We focused on the tools that have been designed or have the ability to detect cancer somatic aberrations. We compared the performance of the tools in terms of sensitivity and false discovery rate (FDR) using real data and simulated data. Comparative analysis of the results of the tools showed that there is a low consensus among the tools in calling CNVs. Using real data, tools show moderate sensitivity (~50% - ~80%), fair specificity (~70% - ~94%) and poor FDRs (~27% - ~60%). Also, using simulated data we observed that increasing the coverage more than 10× in exonic regions does not improve the detection power of the tools significantly.

Conclusions: The limited performance of the current CNV detection tools for WES data in cancer indicates the need for developing more efficient and precise CNV detection methods. Due to the complexity of tumors and high level of noise and biases in WES data, employing advanced novel segmentation, normalization and de-noising techniques that are designed specifically for cancer data is necessary. Also, CNV detection development suffers from the lack of a gold standard for performance evaluation. Finally, developing tools with user-friendly user interfaces and visualization features can enhance CNV studies for a broader range of users.

Keywords: Copy number variation, Whole-exome sequencing, Somatic aberrations, Cancer

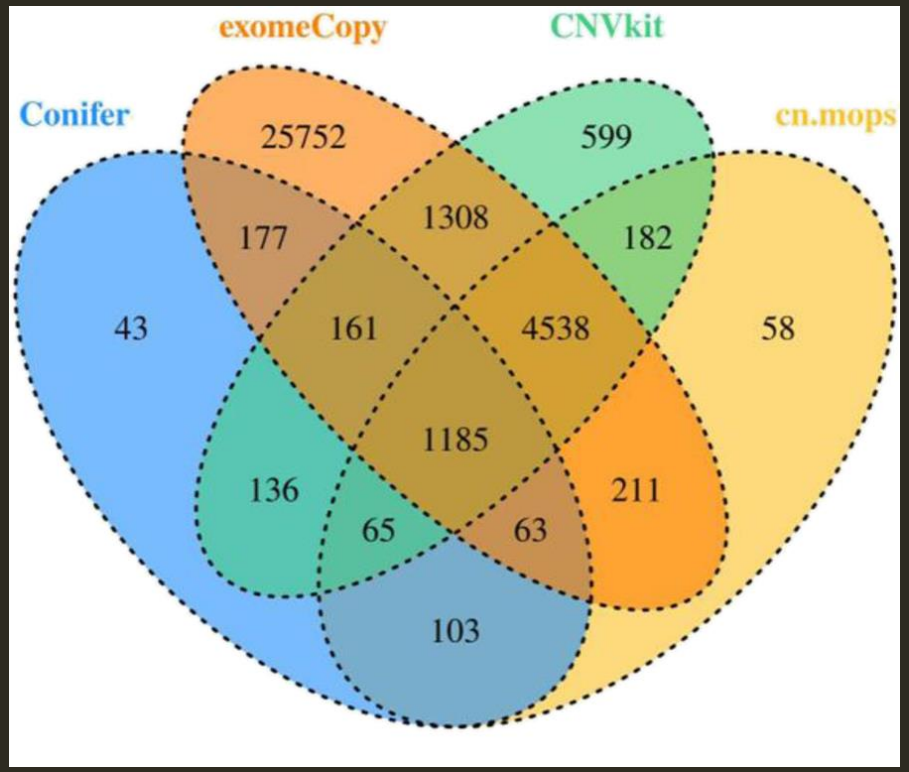


Venn diagrams of the average of the number of truly detected CNV genes from the 5 tools, (**a**) amplified genes, (**b**) deleted genes.



Characteristics of the detected CNV regions by the 5 tools. **(a)** Size distributions of CNV segments, **(b)** Number of detected CNV segments

ZGODNOŚĆ PROGRAMÓW DO DETEKCJI CNV



The overlapping consistency results. Fig shows the simulated data.

Zhao, L., Liu, H., Yuan, X. et al. Comparative study of whole exome sequencing-based copy number variation detection tools. BMC Bioinformatics 21, 97 (2020). https://doi.org/10.1186/s12859-020-3421-1

cancers MDPI

Cancers (Basel). 2021 Dec 14;13(24):6283. doi: 10.3390/cancers13246283

A Comparison of Tools for Copy-Number Variation Detec Whole Exome and Whole Genome Sequencing Data

Migle Gabrielaite [1,†], Mathias Husted Torp [1,†], Malthe Sebro Rasmussen [1], Sergio Andre Vieira [1], Christina Bligaard Pedersen [1,2], Savvas Kinalis [1], Majbritt Busk Madsen [1], Miya Demircan [1], Arman Simonyan [1], Christina Westmose Yde [1], Lars Rønn Olsen [1,2], Rasmu Maria Rossing [1,3], Finn Cilius Nielsen [1], Ole Winther [1,4,5], Frederik Otzen Bagger [1,6,7,*]

Editor: Dimitrios H Roukos

Author information · Article notes · Copyright and License information

PMCID: PMC8699073 PMID: 34944901

5. Conclusions

In summary, by reviewing 50 tools for CNV calling, of which 11 were included for a benchmark (CLC Genomics Workbench (WGS and WES), cn.MOPS (WGS and WES), CNVkit (WES), CNVnator (WGS), CODEX2 (WGS), Control-FREEC (WGS), DELLY (WGS), ExomeDepth (WES), GATK gCNV (WES and WGS), Lumpy (WGS), and Manta (WES and WGS)), we conclude that CNV identification from NGS data remains challenging. For the best reliability of CNV calling from NGS data, we observed that even if the tools were developed for WES data or allowed it as input, they did not perform well. We suggest WGS as the only NGS-based option for broad calling of CNVs. Furthermore, low precision in all tools leads us to recommend a hypothesis-based approach for finding causative CNVs by NGS in the clinic, and further validation of these candidates by manual inspection, MLPA or array-based approaches. If multiple samples are available from the same protocol, we suggest using these to filter by commonly called CNVs. If only the WGS data is available for the sample, for a higher precision of CNV calls, multiple CNV calling tools should be used. We suggest combining tools which have the best recall (GATK gCNV, Lumpy, DELLY, and cn.MOPS) using consensus callers (e.g., [53]) and prioritize the CNV calls made by such tools.

European Journal of
Human Genetics

Article | Open access | Published: 08 November 2021

Combining callers improves the detection of copy number variants from whole-genome sequencing

Marie Coutelier, Manuel Holtgrewe, Marten Jäger, Ricarda Flöttman, Martin A. Mensah, Malte Spielmann, Peter Krawitz, Denise Horn, Dieter Beule & Stefan Mundlos ✉

European Journal of Human Genetics **30**, 178–186 (2022) | Cite this article

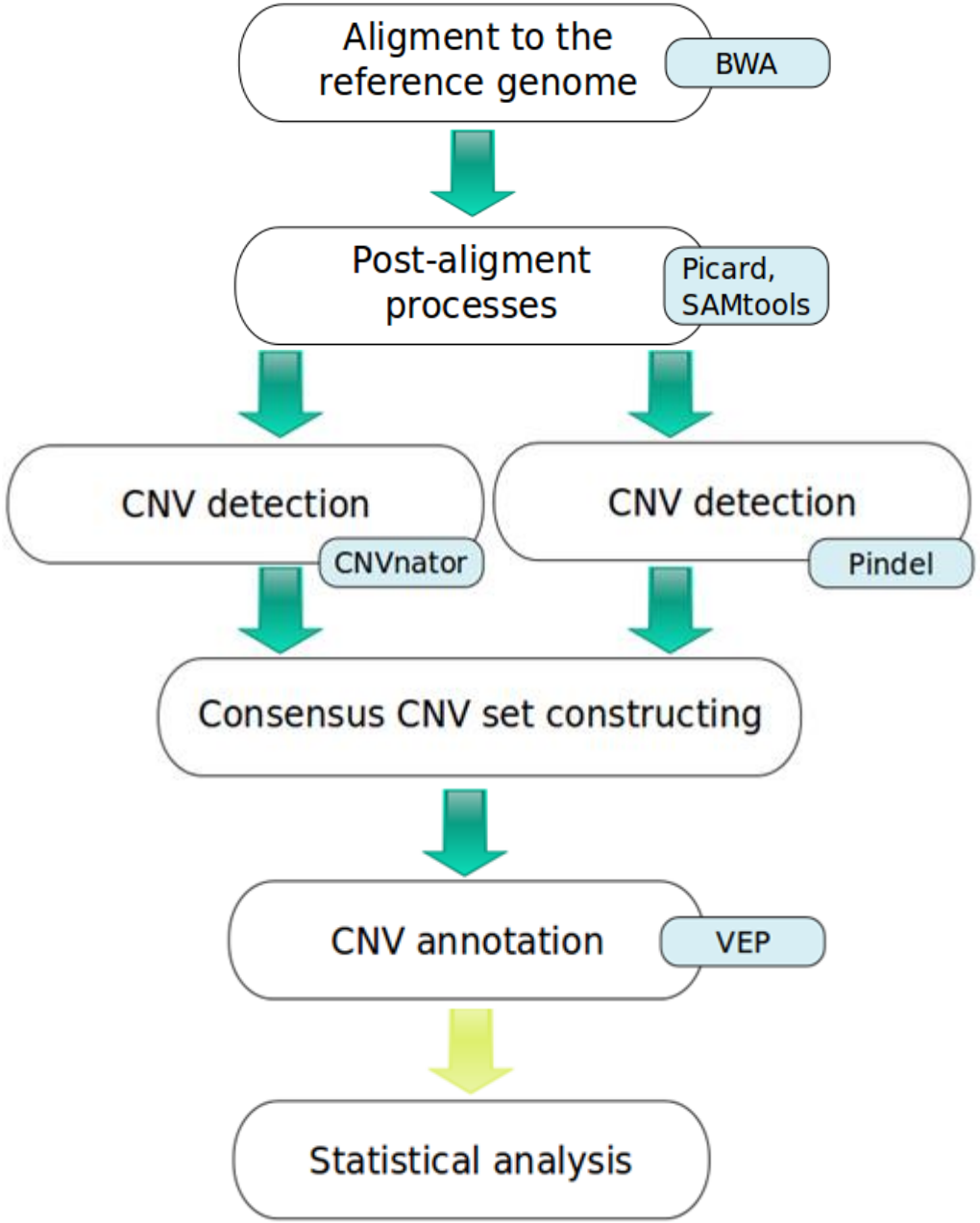
Abstract

Copy Number Variants (CNVs) are deletions, duplications or insertions larger than 50 base pairs. They account for a large percentage of the normal genome variation and play major roles in human pathology. While array-based approaches have long been used to detect them in clinical practice, whole-genome sequencing (WGS) bears the promise to allow concomitant exploration of CNVs and smaller variants. However, accurately calling CNVs from WGS remains a difficult computational task, for which a consensus is still lacking. In this paper, we explore practical calling options to reach the best compromise between sensitivity and sensibility. We show that callers based on different signal (paired-end reads, split reads, coverage depth) yield complementary results. We suggest approaches combining four selected callers (Manta, Delly, ERDS, CNVnator) and a regenotyping tool (SV2), and show that this is applicable in everyday practice in terms of computation time and further interpretation. We demonstrate the superiority of these approaches over array-based Comparative Genomic Hybridization (aCGH), specifically regarding the lack of resolution in breakpoint definition and the detection of potentially relevant CNVs. Finally, we confirm our results on the NA12878 benchmark genome, as well as one clinically validated sample. In conclusion, we suggest that WGS constitutes a timely and economically valid alternative to the combination of aCGH and whole-exome sequencing.

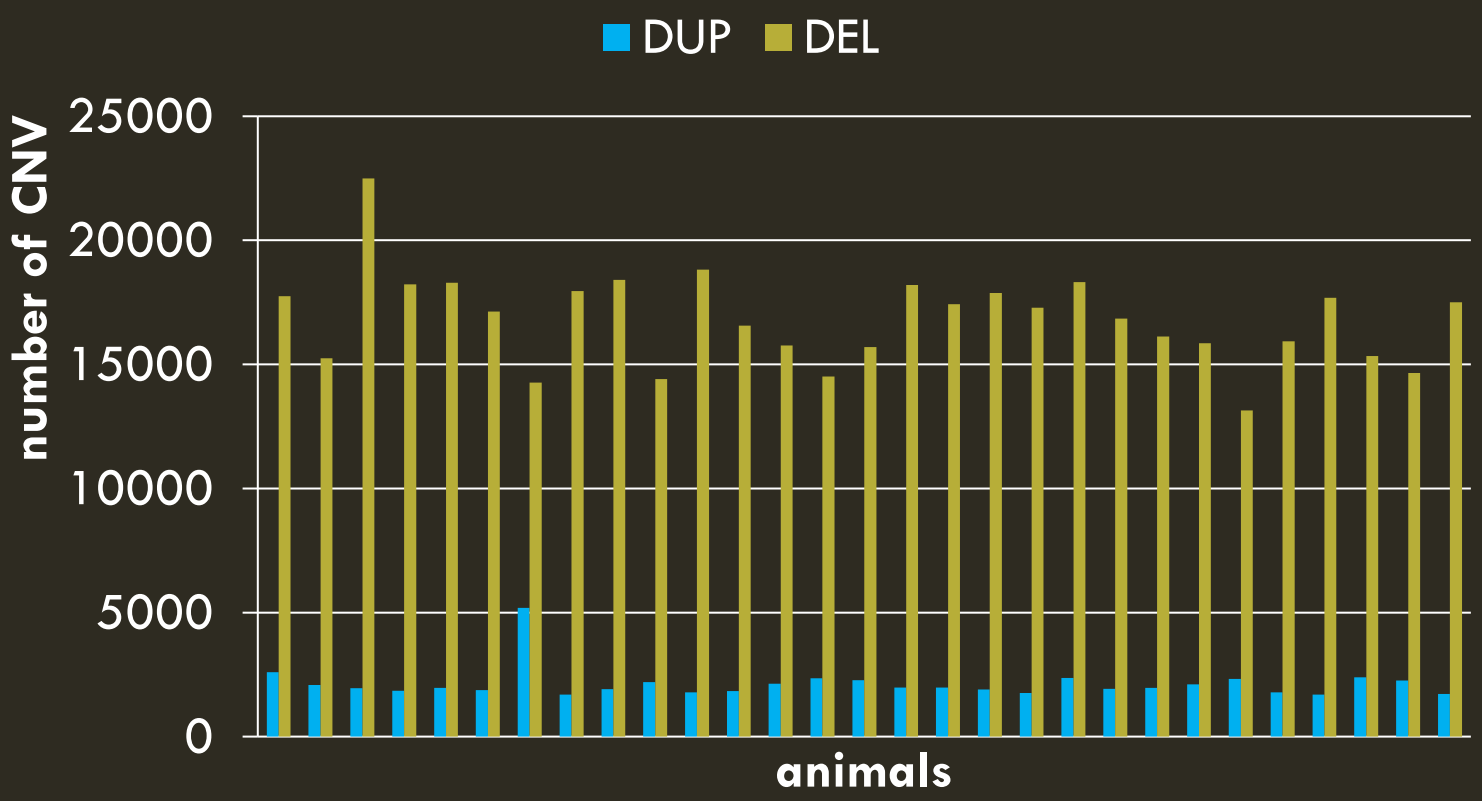
Mielczarek et al. 2019

CNV PIPELINE

→ PRZYKŁAD OD THETA



CNV

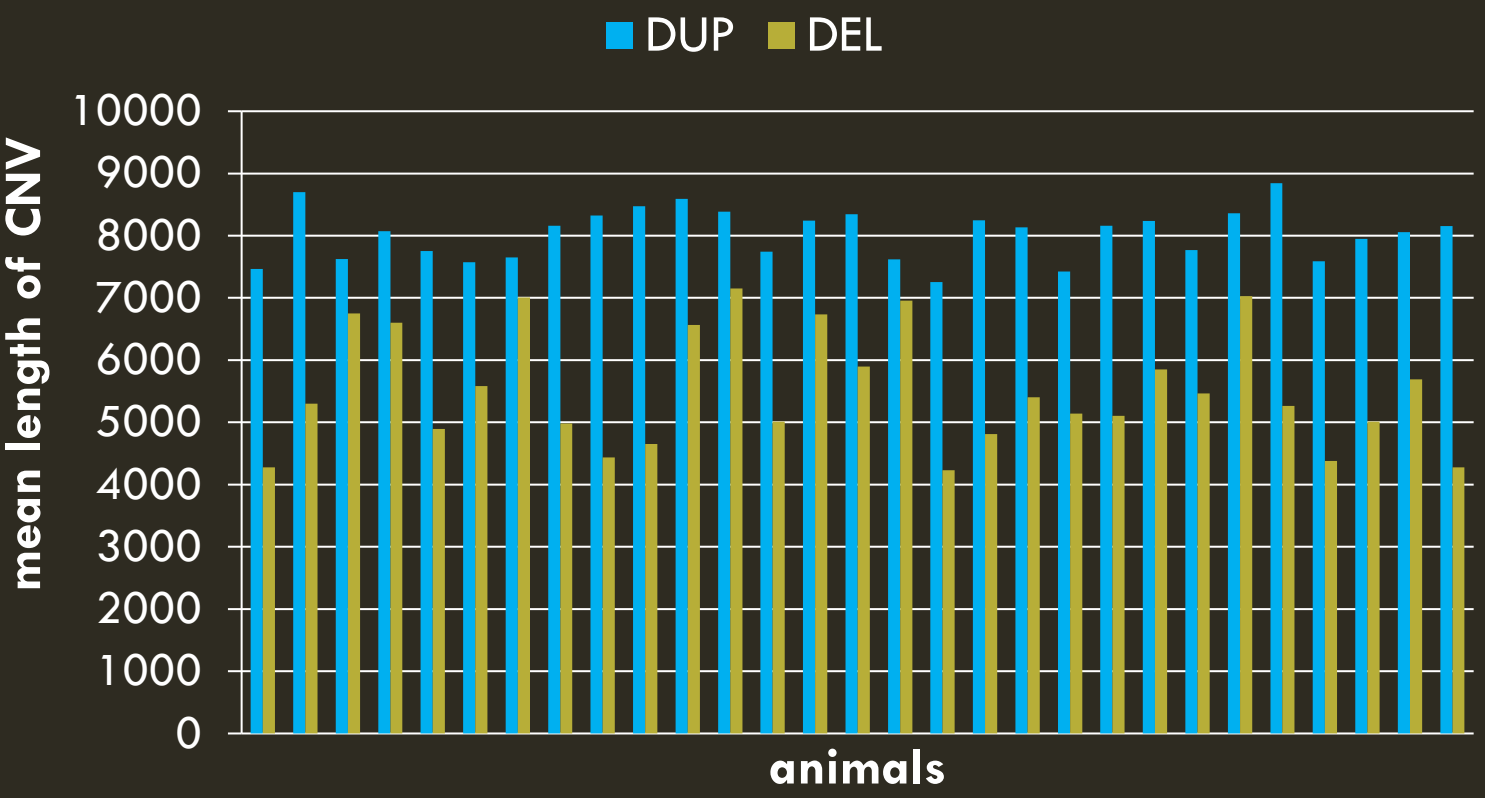


The min and max number
Duplications: 1 694 - 5 187
Deletions: 13 149 - 22 496

The occupancy (%)
Duplications:
min 0.1 (BTA20, BTA22)
max 2.9 (BTA27)
Deletions:
min 1.2 (BTA24)
max 15.8 (BTA25)

WGS

CNV LENGTH

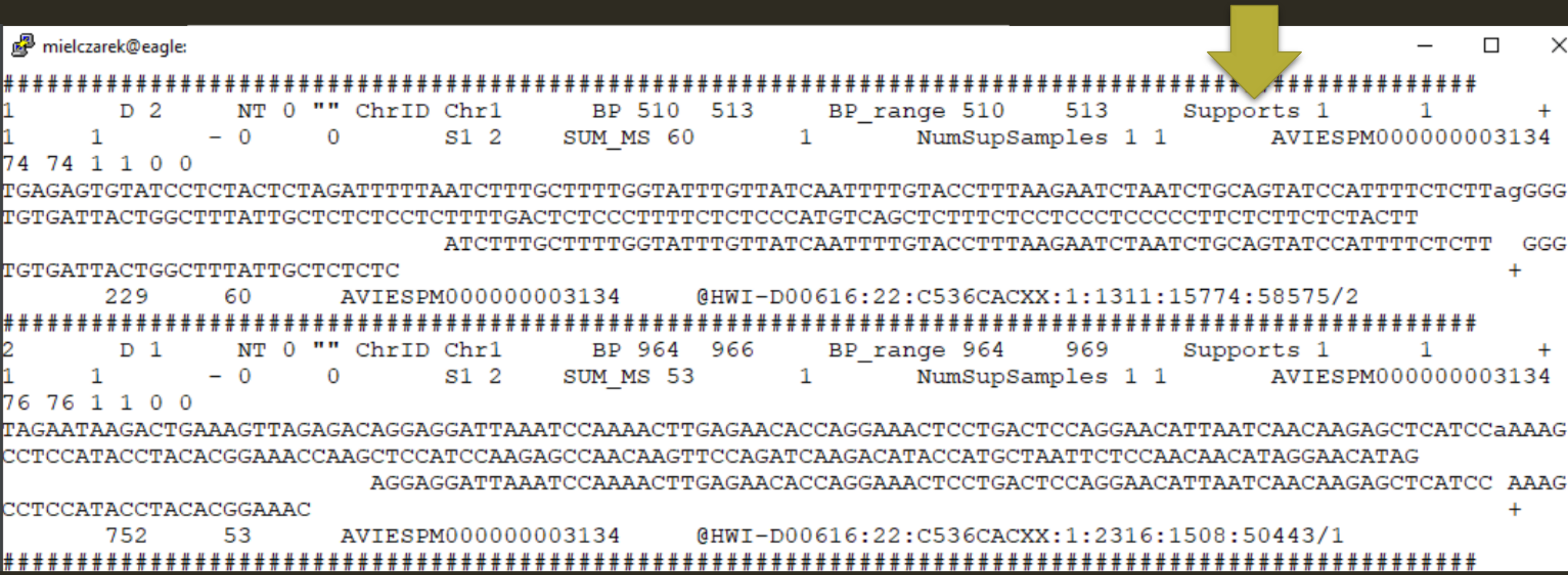


The average length (bp)
Duplications: 7 254 - 8 843
Deletions: 4 233 - 7 154

The min and max lengths (bp)
Duplications: 200 - 439 300
Deletions: 200 - 724 000

FORMATY DANYCH WYJŚCIOWYCH

FORMATY DANYCH WYJŚCIOWYCH: PINDEL



FORMATY DANYCH WYJŚCIOWYCH: CNVNATOR

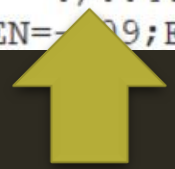
mielczarek@eagle:

```
duplication     chr1:3701-16300 12600   1.80148 0       107163  0       540628  1
deletion        chr1:16301-20400        4100    0.415123        1.82307e-07     341293  0.0312203       2.8036e+07      1
duplication     chr1:24801-32200        7400    1.75999 0       179.448 0       15885.1 1
deletion        chr1:42901-43500        600     0       6.8306  8.82325e-07     1       1       1
duplication     chr1:43501-62600        19100   2.01278 0       1.7469e+09      0       1.84018e+09     1
duplication     chr1:65201-68800        3600    1.93774 4.25108e-06     0.0166806       2.0414  29129.6 1
deletion        chr1:92801-95400        2600    0.290522        4.78528e-06     40.7744 35768.3 4.43648e+07     1
deletion        chr1:215101-217900      2800    0.205357        1.4816e-07      836.124 0.0484782       5.95425e-12     1
deletion        chr1:319701-320500      800     0       0.0484782       5.95425e-12     1       1       1
deletion        chr1:510901-513400      2500    0.402345        0.0109776       5.41966e+07     4133.34 86491.8 1
deletion        chr1:531201-532000      800     0.0372511       0.0374773       0.000423115     1       1       1
deletion        chr1:541901-542800      900     0.0220357       7.31419e-05     2.18887e-08     1       1       1
deletion        chr1:626501-627200      700     0.0143633       0.000832449     7.79728e-07     1       1       1
deletion        chr1:1044501-1045400    900     0       0.00426218      1.54677e-14     1       1       1
deletion        chr1:1372901-1375900    3000    0.60156 0.00383479      54.8515 1706.18 7.67512e+06     1
deletion        chr1:1416201-1424100    7900    0.530027        2.01737e-11     4.44582e-05     2.70123e-11     0.139372        1
deletion        chr1:1739701-1740600    900     0.067292        0.0433481       0.00214204      1       1       1
deletion        chr1:2803901-2814900    11000   0.342297        1.44884e-11     7.53649e+08     1.77081e-11     9.61127e+08     1
deletion        chr1:2847401-2847900    500     0       86.4674 0.000339648     1       1       1
deletion        chr1:2918101-2919200    1100    0.0105824       4.34652e-10     3.91917e-15     1       1       1
deletion        chr1:3219401-3220400    1000    0       0.00038232      4.01814e-17     1       1       1
deletion        chr1:3269901-3271000    1100    0.0830775       0.0130583       0.00761158      1       1       1
deletion        chr1:3489201-3489700    500     0       86.4674 0.000339648     1       1       1
```

FORMATY DANYCH WYJŚCIOWYCH: LUMPY (VCF)

magda@penguin:~/WGS_CNV_bees/Lumpy_files/Lumpy_out

```
#CHROM  POS     ID      REF     ALT     QUAL    FILTER  INFO    FORMAT  S7_CATCAA
LG1     786100  1       N       <DUP>   .       .       SVTYPE=DUP;STRANDS=-+:6;SVLEN=330;END=786430;CIPOS
=-288,9;CIEND=-10,273;CIPOS95=-68,2;CIEND95=-2,89;IMPRECISE;SU=6;PE=6;SR=0    GT:SU:PE:SR     ./.:6:6:0
LG1     357821  2       N       <DEL>   .       .       SVTYPE=DEL;STRANDS=+-:5;SVLEN=-36;END=357857;CIPOS
=0,7;CIEND=-9,0;CIPOS95=0,1;CIEND95=-9,0;IMPRECISE;SU=5;PE=0;SR=5     GT:SU:PE:SR     ./.:5:0:5
LG1     411115  3       N       <DUP>   .       .       SVTYPE=DUP;STRANDS=-+:7;SVLEN=659;END=411774;CIPOS
=-190,9;CIEND=-10,36;CIPOS95=-64,2;CIEND95=-4,11;IMPRECISE;SU=7;PE=7;SR=0     GT:SU:PE:SR     ./.:7:7:0
LG1     506922  4       N       <DEL>   .       .       SVTYPE=DEL;STRANDS=+-:5;SVLEN=-664;END=507586;CIPO
S=-10,8;CIEND=-10,9;CIPOS95=-1,1;CIEND95=-1,1;IMPRECISE;SU=5;PE=3;SR=2        GT:SU:PE:SR     ./.:5:3:2
LG1     507175  5       N       <DEL>   .       .       SVTYPE=DEL;STRANDS=+-:5;SVLEN=-1220;END=508395;CIP
OS=-10,300;CIEND=-297,9;CIPOS95=-1,94;CIEND95=-97,1;IMPRECISE;SU=5;PE=5;SR=0  GT:SU:PE:SR     ./.:5:5:0
LG1     584355  6       N       <DUP>   .       .       SVTYPE=DUP;STRANDS=-+:5;SVLEN=636;END=584991;CIPOS
=-131,9;CIEND=-10,107;CIPOS95=-50,2;CIEND95=-3,28;IMPRECISE;SU=5;PE=5;SR=0    GT:SU:PE:SR     ./.:5:5:0
LG1     584536  7       N       <DEL>   .       .       SVTYPE=DEL;STRANDS=+-:4;SVLEN=-464;END=585000;CIPO
S=-10,311;CIEND=-301,9;CIPOS95=0,120;CIEND95=-103,1;IMPRECISE;SU=4;PE=4;SR=0  GT:SU:PE:SR     ./.:4:4:0
LG1     796763  8       N       <DEL>   .       .       SVTYPE=DEL;STRANDS=+-:4;SVLEN=-23;END=796786;CIPOS
=-10,7;CIEND=-8,9;CIPOS95=-1,0;CIEND95=0,1;IMPRECISE;SU=4;PE=0;SR=4    GT:SU:PE:SR     ./.:4:0:4
LG1     905653  9       N       <DEL>   .       .       SVTYPE=DEL;STRANDS=+-:6;SVLEN=-[illegible]9;END=906162;CIPO
```



LONG READS

LONG READS

Evaluating nanopore sequencing data processing pipelines for structural variation identification

Anbo Zhou[1], Timothy Lin[1] and Jinchuan Xing[1,2*]

The majority of known SVs are ***poorly assayed*** using currently dominant ***short-read sequencing technologies*** → evidence supporting an SV event are indirect (e.g. read depth, mismatch read pairs).

SVs can be detected using ***long-read sequencing technologies*** from Pacific Biosciences (PacBio) and Oxford Nanopore Technologies (ONT) → direct evidence.

Long-reads:

- better mappability in repetitive regions, further extending the part of the genome in which variation can be called reliably (+)
- poor quality (-).

CONCLUSIONS

Evaluating nanopore sequencing data processing pipelines for structural variation identification

Anbo Zhou[1], Timothy Lin[1] and Jinchuan Xing[1,2*]

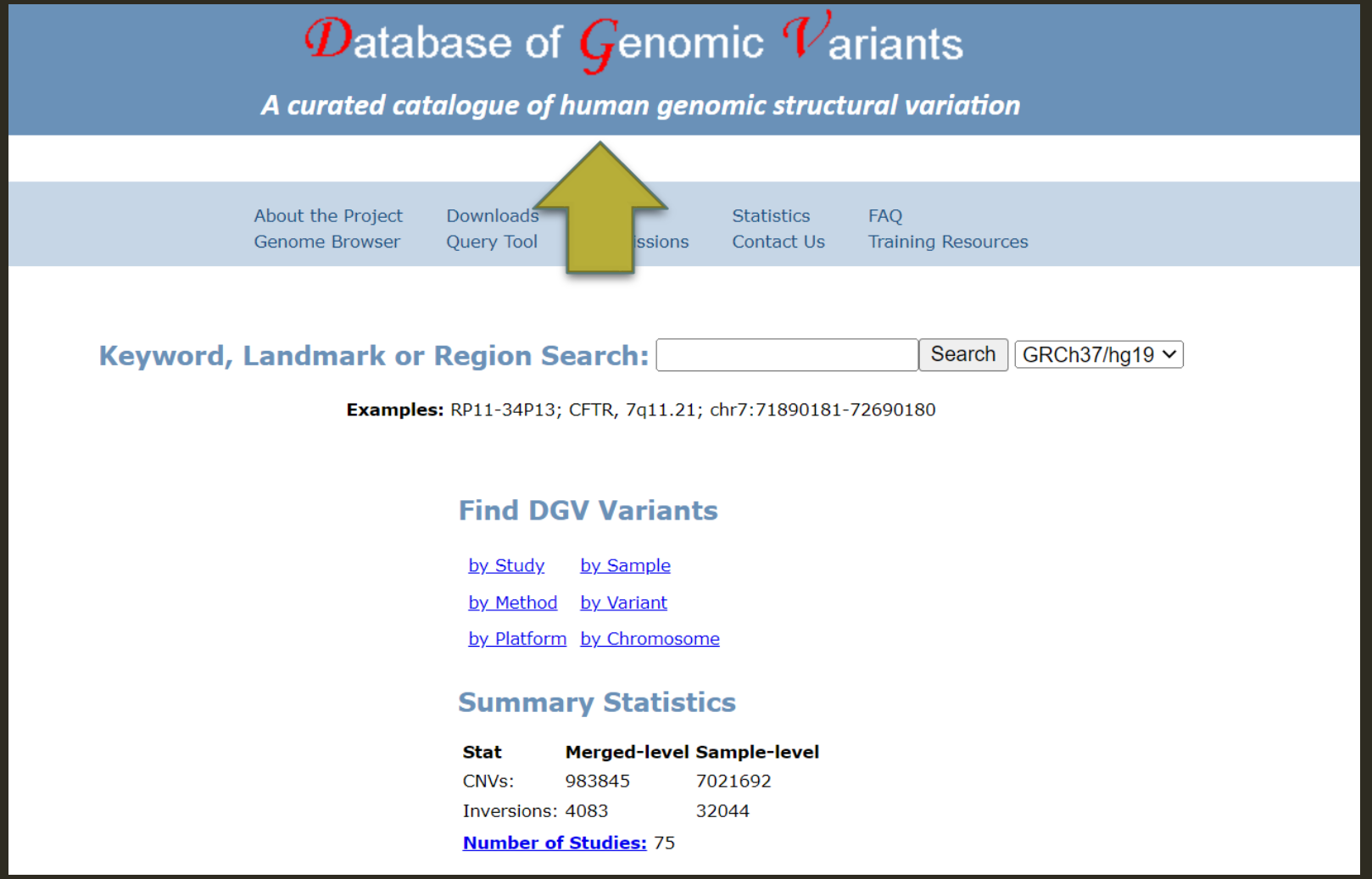
Nanopore sequencing is a rapidly developing technology in terms of both sequencing technology and data analysis.

SV callers' performance diverges between SV types → **recommendations** are tailored to the specific applications:

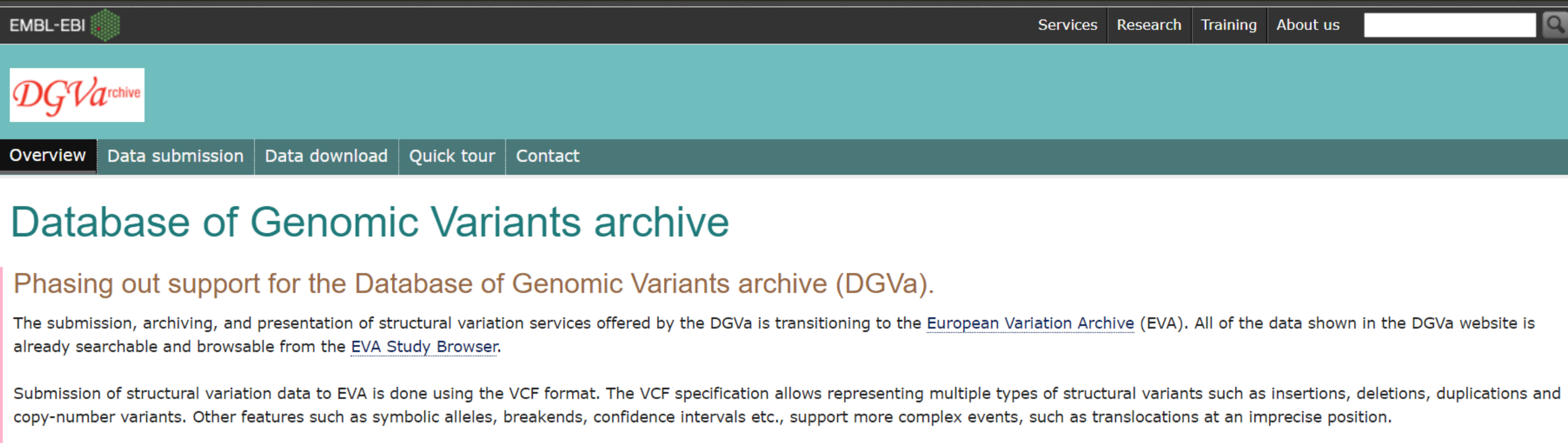
- For an initial analysis, we recommend ***minimap2*** + ***Sniffles*** → high speed and relatively balanced performance calling both insertions and deletions.
- For more detailed analysis, ***multiple tools*** and integrating their results for the best performance.
- When a high-quality true set can be defined, a ***machine learning approach*** can be used to further improve the call set.

BAZY DANYCH CNV PRZYKŁADY

BAZA DANYCH: DGV



BAZA DANYCH: DGVA



BAZA DANYCH: DGVA

Overview | Data submission | Data download | Quick tour | Contact

DGVa > Data download

Data download

Genomic structural variant study data can be downloaded via ftp by following the appropriate link.

DGVa studies

Study	Reference	Organism	Variants	Link
estd233	luo_et_al_2017b	Homo sapiens	1026	Download via FTP
estd232	Blanco-Kelly_et_al_-2017	Homo sapiens	8	Download via FTP
estd231	WONG_et_al_2016	Homo sapiens	7063	Download via FTP
estd229	Fakhro_et_al_2015	Homo sapiens	16676	Download via FTP
estd228	Ansari_et_al_2016	Homo sapiens	12	Download via FTP
estd226	Zlotina_et_al_2016	Homo sapiens	4	Download via FTP
estd225	Magnusson_et_al_2016	Homo sapiens	1917	Download via FTP
estd224	Suktitipat_et_al_2014	Homo sapiens	3576	Download via FTP
estd223	Boussaha_et_al_2015	Bos taurus	6426	Download via FTP

BAZA DANYCH: EVA