Genomics and Transcriptomics

Class 08 - Variant Calling

INSTRUCTOR:
Aureliano Bombarely
Department of Bioscience
Universita degli Studi di Milano
aureliano.bombarely@unimi.it

Outline of Topics

1. Basics about variants in genomics.

2. Manipulating SAM/BAMs and coverage.

3. Simple variants: SNVs, InDels and MNVs.

4. Copy Number Variation (CNV).

5. Structural Variants (SV).

6. Annotating variants and assessing its impact.

Outline of Topics

1. Basics about variants in genomics.

2. Manipulating SAM/BAMs and coverage.

3. Simple variants: SNVs, InDels and MNVs.

4. Copy Number Variation (CNV).

5. Structural Variants (SV).

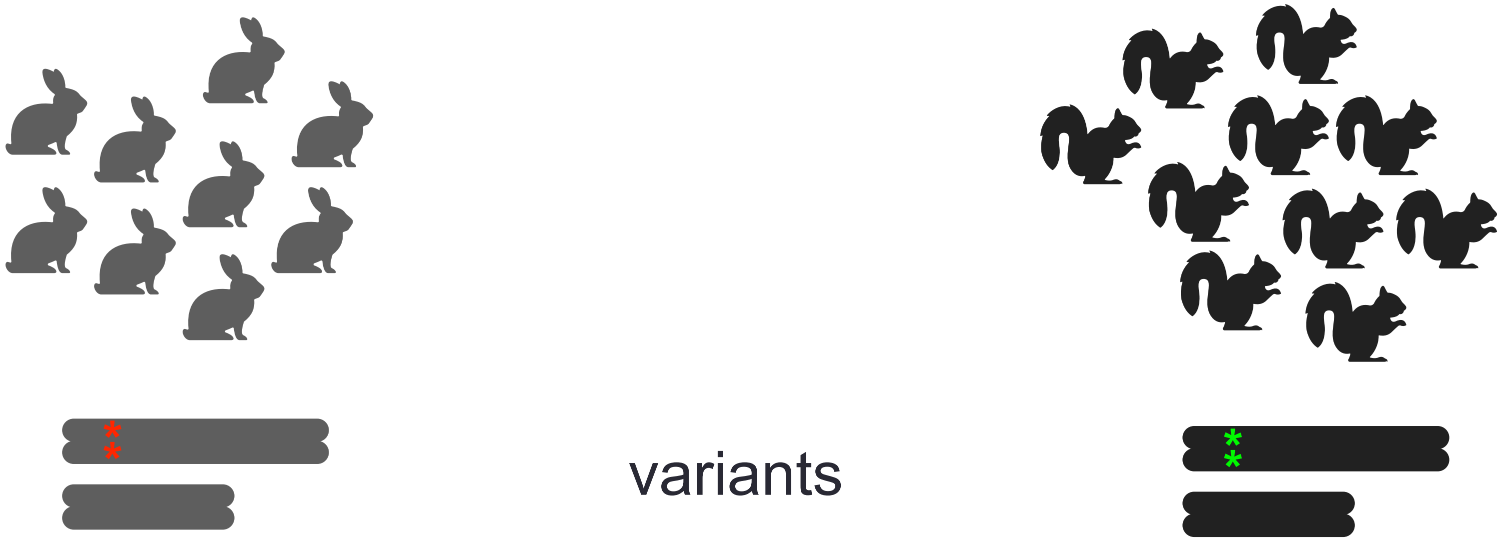
6. Annotating variants and assessing its impact.

1. Basics about variants in genomics.

Genetic variation can be defined as **different forms of a genetic region** of **different individuals** of different populations (polymorphisms) or species (variants).

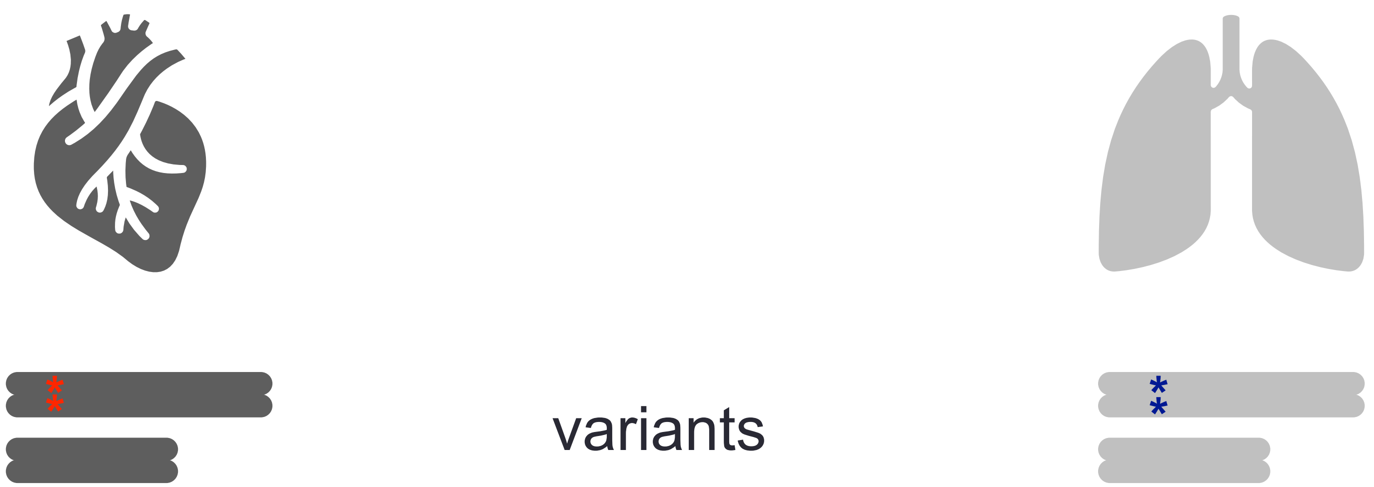
1. Basics about variants in genomics.

Genetic variation can be defined as **different forms of a genetic region** of **different individuals** of different populations (polymorphisms) or species (variants).



1. Basics about variants in genomics.

Genetic variation can be defined as **different forms of a genetic region** of **different individuals** of different populations (polymorphisms) or species/cells (variants).



1. Basics about variants in genomics.

Genetic variation can be defined as **different forms of a genetic region** of **different individuals** of different populations (polymorphisms) or species (variants).

Genetic variation

- Large size changes - Chromosomal reorganisations (Structural Variants)

- Medium size changes - Changes in the gene copy number

- Discrete changes - Single Nucleotide Variants, Insertions/deletions

1. Basics about variants in genomics.

Genetic variation can be defined as **different forms of a genetic region** of **different individuals** of different populations (polymorphisms) or species (variants).

There are different approaches for study genetic variants

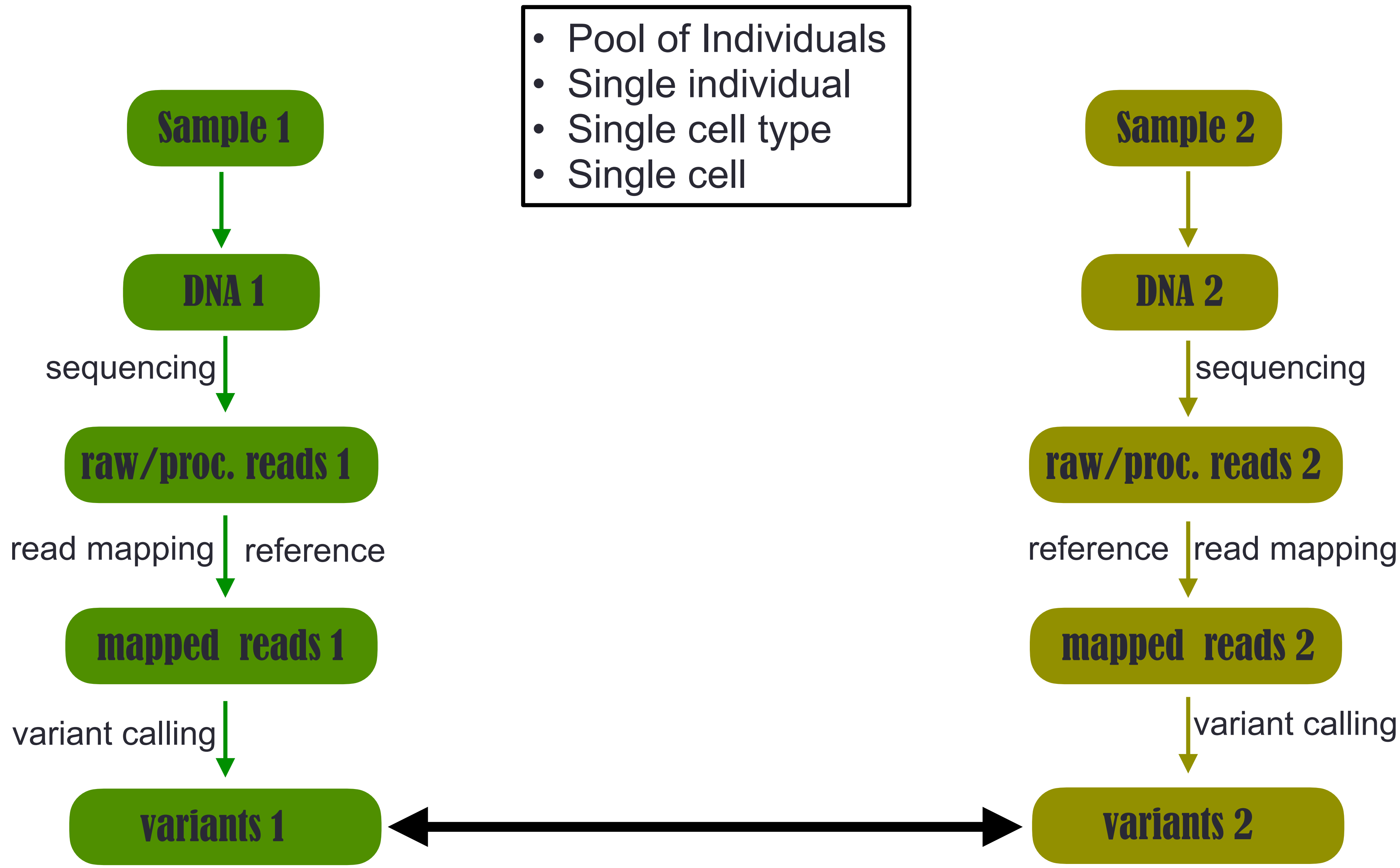
- Cellular techniques (e.g. flow cytometry, FISH…).

- Molecular techniques (e.g. PCR…).

- Genetic/Genomic techniques (e.g. WGS).

1. Basics about variants in genomics.

Variant calling is the process by which identify variants between **different individuals or cell types**.



Outline of Topics

1. Basics about variants in genomics.

2. Manipulating SAM/BAMs and coverage.

3. Simple variants: SNVs, InDels and MNVs.

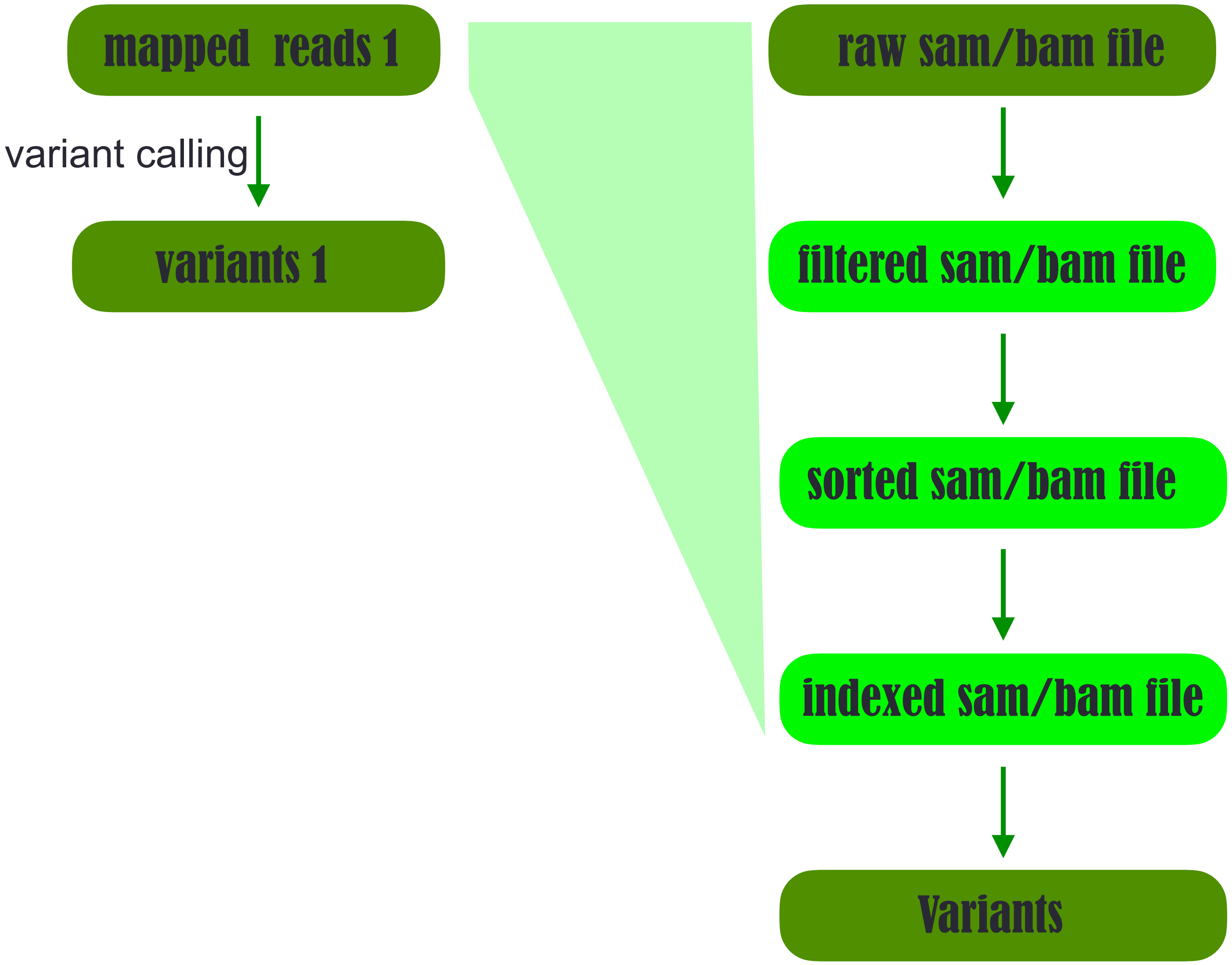
4. Copy Number Variation (CNV).

5. Structural Variants (SV).

6. Annotating variants and assessing its impact.

2. Manipulating SAM/BAMs and coverage.

Before the variant calling it is essential to perform several steps



2. Manipulating SAM/BAMs and coverage.

Before the variant calling it is essential to perform several steps

Oxford Journals › Science & Mathematics › Bioinformatics › Volume 25, Issue 16 › Pp. 2078-2079.

The Sequence Alignment/Map format and SAMtools

Heng Li[1,†], Bob Handsaker[2,†], Alec Wysoker[2], Tim Fennell[2], Jue Ruan[3], Nils Homer[4], Gabor Marth[5], Goncalo Abecasis[6], Richard Durbin[1,*] and 1000 Genome Project Data Processing Subgroup[7]

+ Author Affiliations

« Previous | Next Article »
Table of Contents

This Article

Bioinformatics (2009) 25 (16): 2078-2079.
doi: 10.1093/bioinformatics/btp352
First published online: June 8, 2009

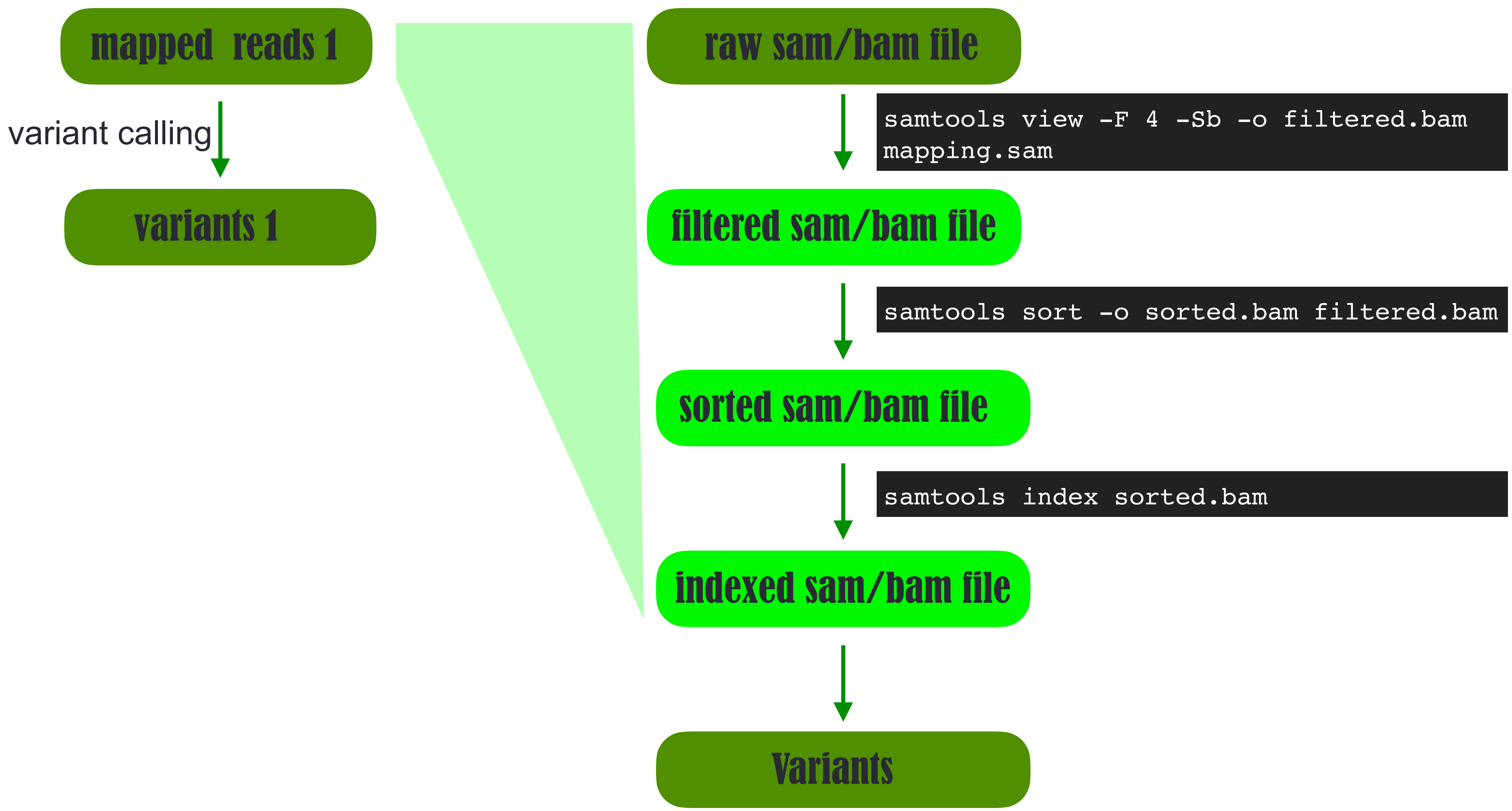
http://samtools.sourceforge.net/

```
Usage:   samtools <command> [options]

Command: view        SAM<->BAM conversion
         sort        sort alignment file
         mpileup     multi-way pileup
         depth       compute the depth
         faidx       index/extract FASTA
         tview       text alignment viewer
         index       index alignment
         idxstats    BAM index stats (r595 or later)
         fixmate     fix mate information
         flagstat    simple stats
         calmd       recalculate MD/NM tags and '=' bases
         merge       merge sorted alignments
         rmdup       remove PCR duplicates
         reheader    replace BAM header
         cat         concatenate BAMs
         bedcov      read depth per BED region
         targetcut   cut fosmid regions (for fosmid pool only)
         phase       phase heterozygotes
         bamshuf     shuffle and group alignments by name
```

2. Manipulating SAM/BAMs and coverage.

Before the variant calling it is essential to perform several steps



2. Manipulating SAM/BAMs and coverage.

Good practices before perform the variant calling:

1. Retrieve information about your mapping.

   1.1. How many reads were mapped?

   1.2. How many regions have coverage of 0?

   1.3. How many regions have a coverage of < 5?

   1.4. What it is the average coverage?

   1.5. What it is the maximum coverage?

2. Know the limitations of your technique and filter your reads accordingly (e.g. for WGS it is worthy to filter PCR duplications).

3. Realign reads if the variant caller does not have this process integrated

2. Manipulating SAM/BAMs and coverage.

Good practices before perform the variant calling:

https://bedtools.readthedocs.io/en/latest/

1. Retrieve information about your mapping.

    1.1. How many reads were mapped?

    1.2. How many regions have coverage of 0?

    1.3. How many regions have a coverage of < 5?

    1.4. What it is the average coverage?

    1.5. What it is the maximum coverage?

2. Know the limitations of your technique and filter your reads accordingly (e.g. for WGS it is worthy to filter PCR duplications).

3. Realign reads if the variant caller does not have this process integrated

2. Manipulating SAM/BAMs and coverage.

1. Retrieve information about your mapping

    - Samtools

    - Bedtools https://bedtools.readthedocs.io/en/latest/

    - Picards tools https://broadinstitute.github.io/picard/

2. Manipulating SAM/BAMs and coverage.

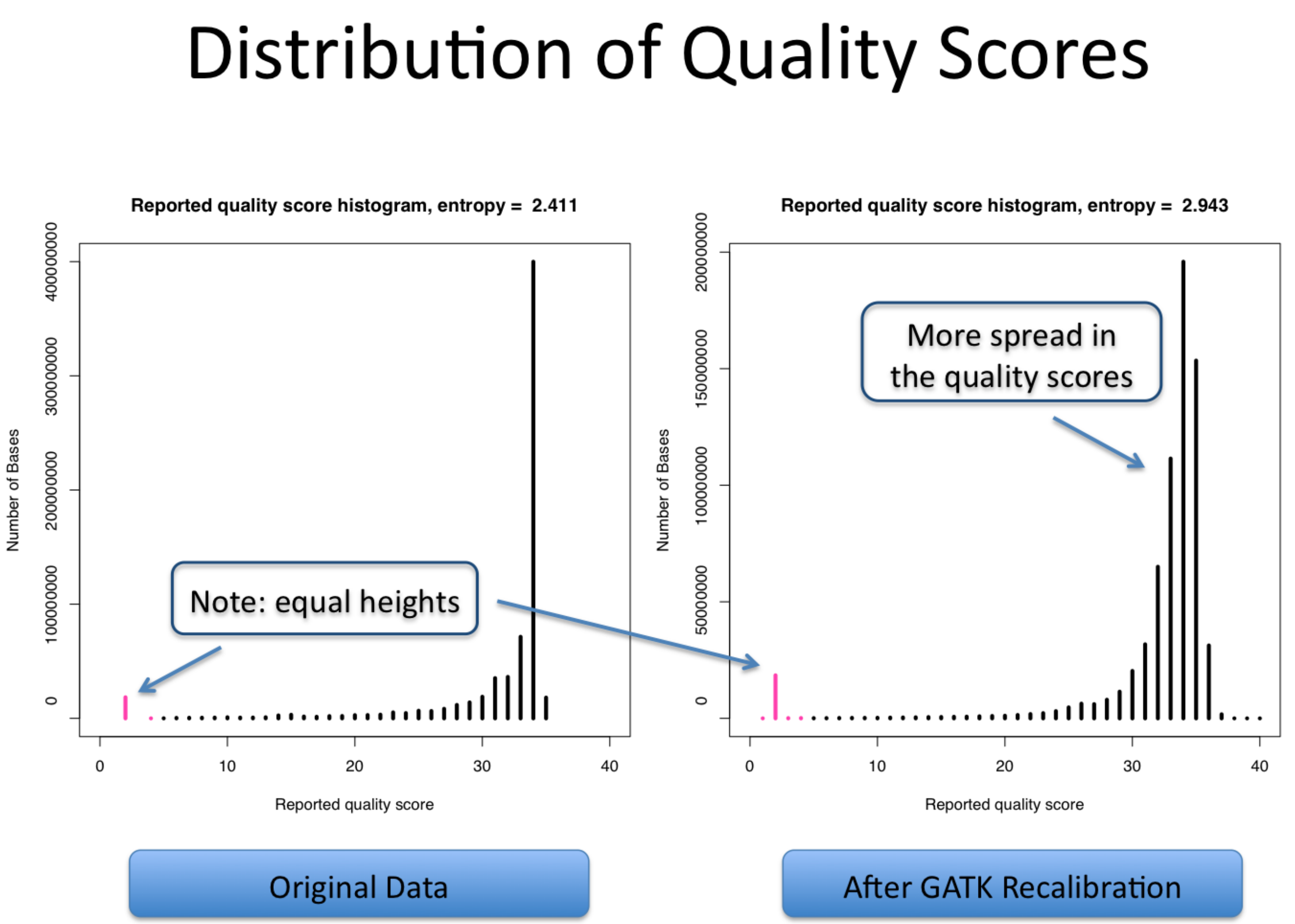
Library preparation problems

Sequencing errors produce biases in the variant call.

Library preparation problems

Sequencing errors - Solutions:

- High coverage (< 20 X) to minimize sequencing errors.
- Recalibrate bases (Base Score Quality Recalibration - BSQR) using tools such as BaseRecalibrator.



2. Manipulating SAM/BAMs and coverage.

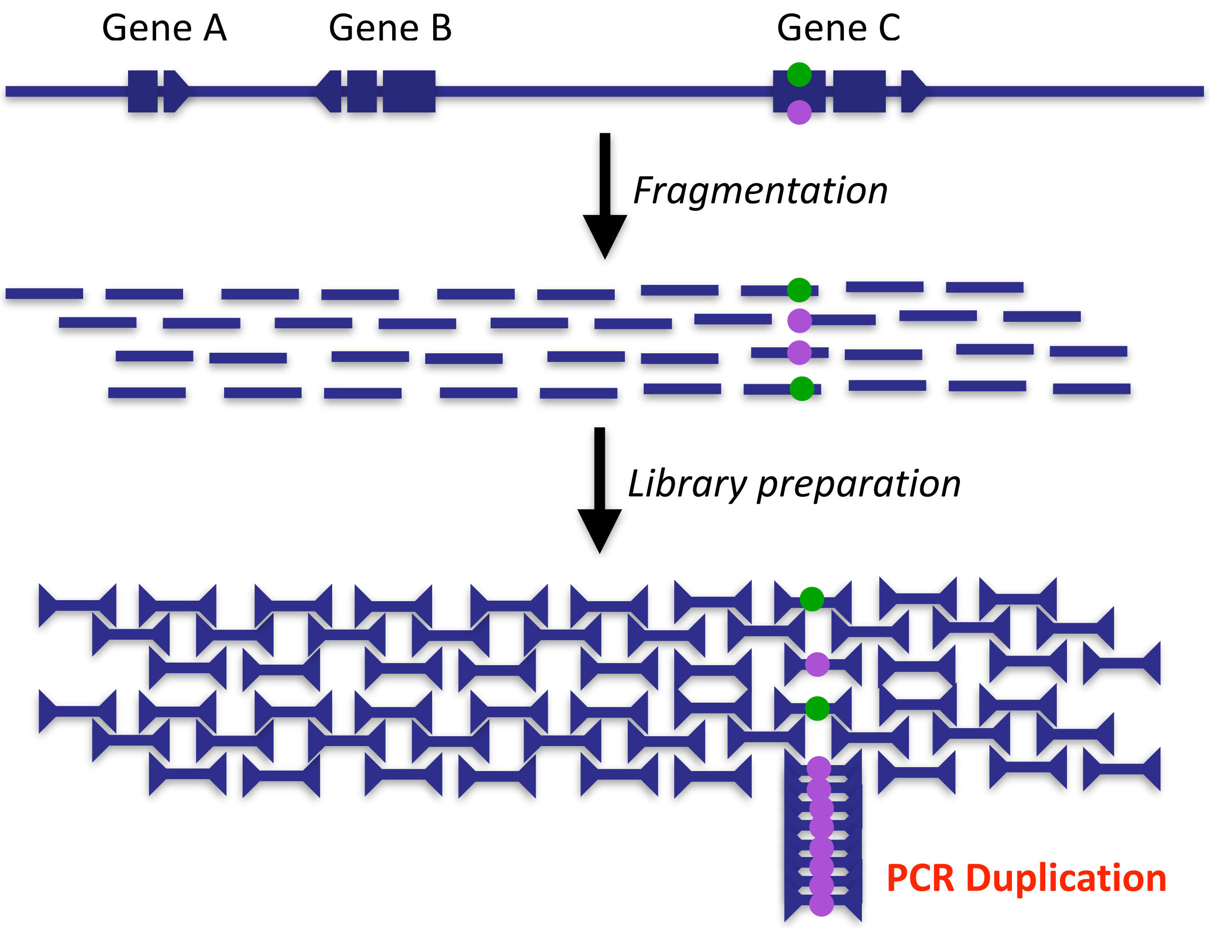
Good practices before perform the variant calling:

1. Retrieve information about your mapping.
   1.1. How many reads were mapped?
   1.2. How many regions have coverage of 0?
   1.3. How many regions have a coverage of < 5?
   1.4. What it is the average coverage?
   1.5. What it is the maximum coverage?
2. Know the limitations of your technique and filter your reads accordingly (e.g. for WGS it is worthy to filter PCR duplications).
3. Realign reads if the variant caller does not have this process integrated

Library preparation problems

PCR duplications produce biases in the variant call (e.g. het.)

- **Library specific problem for Whole Genome Sequencing.**



2. Manipulating SAM/BAMs and coverage.

Library preparation problems

PCR duplications - Solutions:

- **Mark duplicates** with tools such as **samtools rmdup**

↓

CAREFUL*: Some* ***reduced representations*** *techniques with unequal ratios of site amplication* ***WILL PRODUCE THOUSANDS PCR DUPLICATION***

↓

SKIP PCR DUPLICATION MARKING STEP FOR GBS, RAD-SEQ…

2. Manipulating SAM/BAMs and coverage.

Good practices before perform the variant calling:

1. Retrieve information about your mapping.
   1.1. How many reads were mapped?
   1.2. How many regions have coverage of 0?
   1.3. How many regions have a coverage of < 5?
   1.4. What it is the average coverage?
   1.5. What it is the maximum coverage?
2. Know the limitations of your technique and filter your reads accordingly (e.g. for WGS it is worthy to filter PCR duplications).
3. Realign reads if the variant caller does not have this process integrated

2. Manipulating SAM/BAMs and coverage.

Alignment problems

Aligners calculate the alignment correctness and give it a score depending of:

- Length of the alignment.
- Number of mismatches and gaps.
- Uniqueness of the alignment (number of hits).

Good alignment

Misaligned bases

```
coordinates     12345678901234   5678901234567890123456
reference       aggttttttataac---aattaagtctacagagcaacta
sample          aggttttttataacAATaattaagtctacagagcaacta
read1           aggttttttataac***aaAtaa
read2            ggttttttataac***aaAtaaTt
read3                ttttataacAATaattaagtctaca
read4                     CaaT***aattaagtctacagagcaac
read5                      aaT***aattaagtctacagagcaact
read6                        T***aattaagtctacagagcaacta
```

Misaligned bases - Solutions:

- **Read realignment** (IndelRealigner for GATK (obsolete), now it is integrated in the HaplotypeCaller).
- Mark **alignment quality per base (BAQ)** and do not use for variant calling.

Outline of Topics

1. Basics about variants in genomics.

2. Manipulating SAM/BAMs and coverage.

3. Simple variants: SNVs, InDels and MNVs.

4. Copy Number Variation (CNV).

5. Structural Variants (SV).

6. Annotating variants and assessing its impact.

3. Simple variants: SNVs, InDels and MNVs.

Single Nucleotide Variant/Polymorphism (SNV/SNP) is a **substitution** of a single nucleotide at a specific position

```
GACGTGC   Sample 1
|  |||||
GCCGTGC   Sample 2
```

SNVs/SNPs

Insertion/Deletion (InDel/DIV/DIP) is a **insertion or a deletion** of several nucleotides at a specific position

```
GACGTGC   Sample 1
|  |||||
G-CGTGC   Sample 2
```

INDELs/DIVs/DIPs

Multiple Nucleotide Variant/Polymorphism (MNV/MNP) is the **substitution** of several nucleotide at a specific position

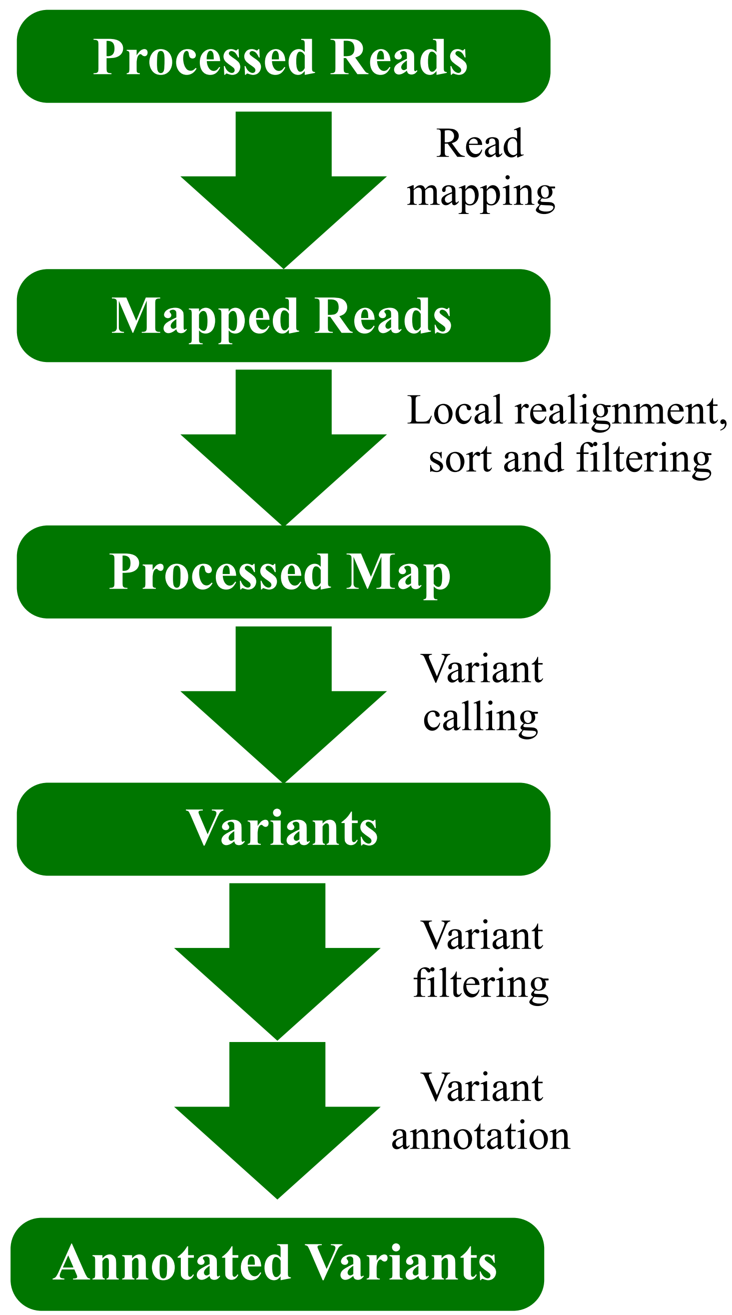
```
GACGTGC   Sample 1
|   ||||
GCTGTGC   Sample 2
```

MNVs/MNPs

3. Simple variants: SNVs, InDels and MNVs.

Considerations

3. Simple variants: SNVs, InDels and MNVs.



Variant calling:

- *Heuristic methods* (read depth)
  - SamTools
  - VarScan
- *Probabilistic methods* (bayesian)
  - GATK
  - FreeBayes
  - SOAPsnp/SOAPindel

Variant calling popular tools

Name	Type	Strength	Weaknesses
SamTools	Heuristic	• Assumes errors are non-independent (matches data) • Good accuracy with low coverage data • Reasonably quick	• Increase false positives at high coverage • Lower quality indel calling
GATK	Probabilistic	• Trains with real data • Excellent accuracy with high coverage data • Low false positive rate	• Assumes errors are independent • High level of preprocessing • Very slow
FreeBayes	Probabilistic	• Combined bam population estimate • Good accuracy with low coverage data • Very very quick	• No training, population level estimate only • Lower quality indel calling

3. Simple variants: SNVs, InDels and MNVs.

Choosing the right tool

Briefings in Bioinformatics Advance Access published January 21, 2013
BRIEFINGS IN BIOINFORMATICS. page 1 of 23 doi:10.1093/bib/bbs086

A survey of tools for variant analysis of next-generation genome sequencing data

Stephan Pabinger, Andreas Dander, Maria Fischer, Rene Snajder, Michael Sperk, Mirjana Efremova, Birgit Krabichler, Michael R. Speicher, Johannes Zschocke and Zlatko Trajanoski

Submitted: 20th August 2012; Received (in revised form): 4th December 2012

Bioinformatics Advance Access published June 27, 2014

Towards Better Understanding of Artifacts in Variant Calling from High-Coverage Samples

Heng Li
Broad Institute of Harvard and MIT, 7 Cambridge Center, Cambridge, MA 02142, USA

Associate Editor: Dr. Jonathan Wren

3. Simple variants: SNVs, InDels and MNVs.

Methods for Variant Evaluation

- Validation by Sanger Sequencing of specific candidates (~5 - 500) using other datasets (e.g. transcriptome) if it is possible.

- Comparison with other method (e.g. genotyping chip).

- Different mapping and variant calling tools comparison (with a “truth set” or a “gold standard” if it is possible).

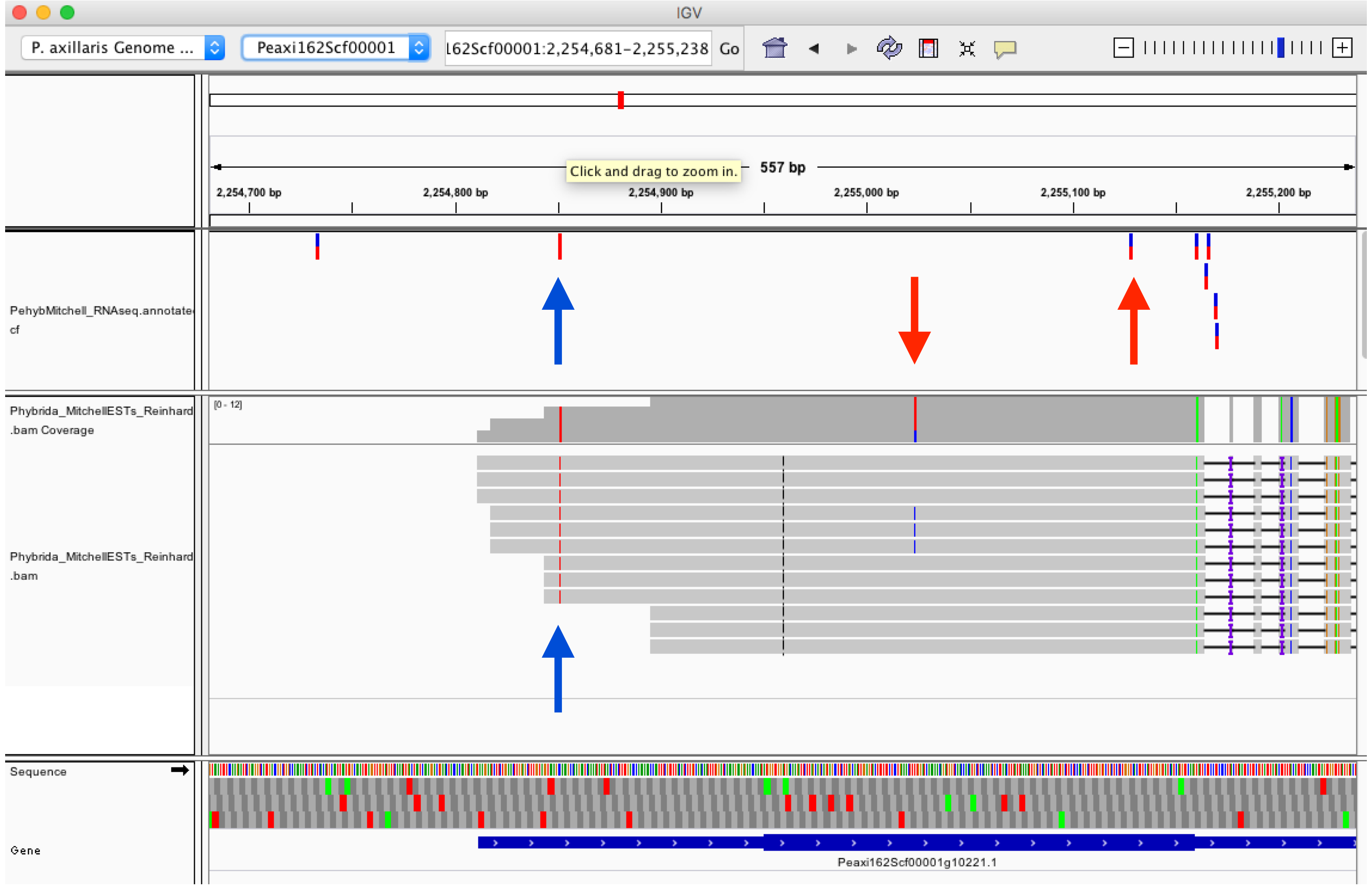
https://gatkforums.broadinstitute.org/gatk/discussion/6308/evaluating-the-quality-of-a-variant-callset

3. Simple variants: SNVs, InDels and MNVs.

- Validation by Sanger Sequencing of specific candidates (~5 - 500) using other datasets (e.g. transcriptome) if it is possible.

Variants from RNASeq (Illumina)

Variants from ESTs (Sanger)



3. Simple variants: SNVs, InDels and MNVs.

- Different mapping, variant calling tools and datasets comparison (with a “truth set” or a “gold standard” if it is possible).

Assumptions:

1. The content of the **truth set** has been **validated.**

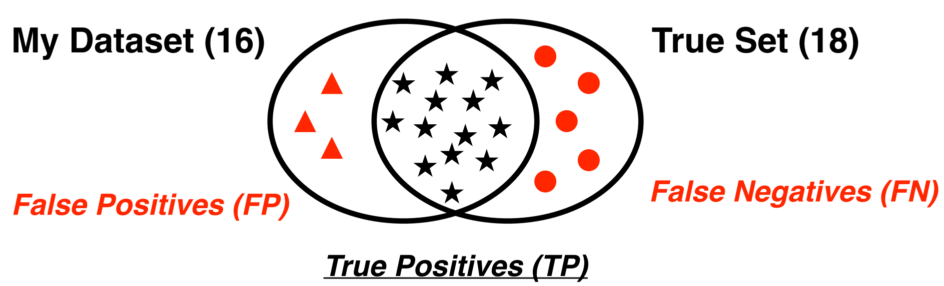
2. Your samples are expected to have similar genomic content as the population of samples that was used to produce the truth set

3. Simple variants: SNVs, InDels and MNVs.

- Different mapping, variant calling tools and datasets comparison (with a “truth set” or a “gold standard” if it is possible).

Metrics:

1. ***Variant level concordance***: Percentage of **variants in your samples** that **match** (are concordant with) variants in your **truth set.**



% SENSITIVITY:
TP * 100 / (TP + FN) = 13 * 100 / (13 + 5) = 72%

% FALSE DISCOVERY RATE:
FP * 100 / (TP + FP) = 3 * 100 / (13 + 3) = 20%

2. ***Genotype concordance***: Percentage of **variants in your genotype** that **match** (are concordant with) variants in your **truth set.**

True Set (9)	A	*	T	C	T	C	C	*	C	A	C
My Dataset (8)	A	T	T	C	*	C	C	T	*	A	*
Matches (6)	1	0	1	1	0	1	1	0	0	1	0

% GT CONCORDANCE:
SumMatches * 100 / TP
6 * 100 / 11 = 54%

3. Simple variants: SNVs, InDels and MNVs.

- Different mapping, variant calling tools and datasets comparison (with a "truth set" or a "gold standard" if it is possible).

Metrics:

3. ***Number of SNPs and INDELs***: Between different datasets should be consistent for the same number of mapped reads.

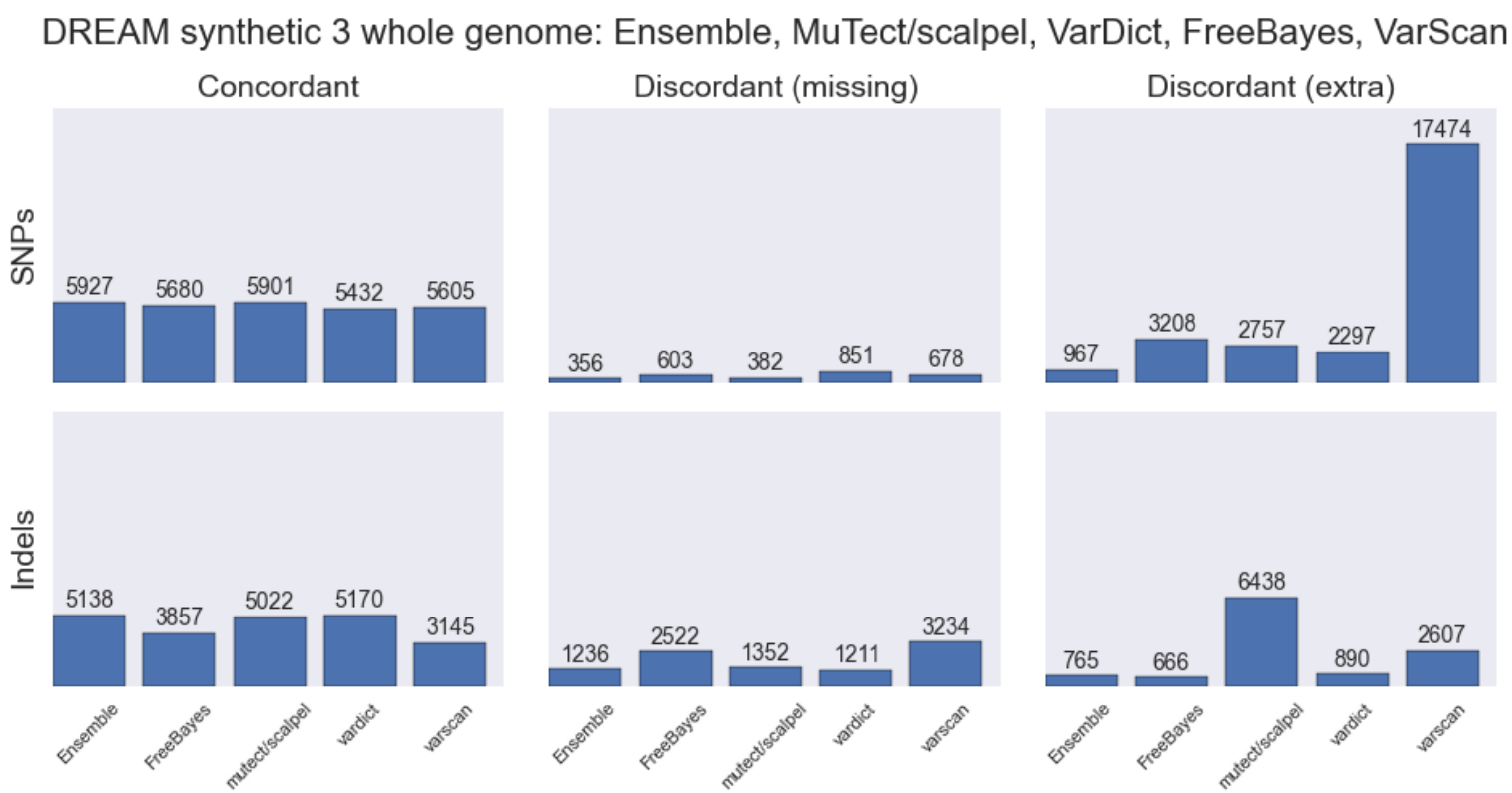
4. ***TiTv Ratio***: Ratio of transition (Ts) to transversion (Tv) SNPs should be random (~0.5). Methylation islands (CpG) and other factors may introduce a bias so expected values will range from 0.5 - 3.0.

5. ***Ratio Insertions/Deletions***: It should be close to 1, except in rare alleles that it could be 0.2 - 0.5.

3. Simple variants: SNVs, InDels and MNVs.

- Different mapping, variant calling tools and datasets comparison (with a “truth set” or a “gold standard” if it is possible).

Comparison between different tools:



https://bcbio.wordpress.com/

3. Simple variants: SNVs, InDels and MNVs.

- Different mapping, variant calling tools and datasets comparison (with a "truth set" or a "gold standard" if it is possible).

Tools:

Name	URL
VariantEvaluation (GATK)	https://software.broadinstitute.org/gatk/documentation/tooldocs/current/org_broadinstitute_gatk_tools_walkers_varianteval_VariantEval.php
GenotypeConcordance (GATK)	https://software.broadinstitute.org/gatk/documentation/tooldocs/current/org_broadinstitute_gatk_tools_walkers_variantutils_GenotypeConcordance.php
VCFTools	http://vcftools.sourceforge.net/
VCFStats	http://lindenb.github.io/jvarkit/
PicardTools	https://broadinstitute.github.io/picard/index.html

Outline of Topics

1. Basics about variants in genomics.

2. Manipulating SAM/BAMs and coverage.

3. Simple variants: SNVs, InDels and MNVs.

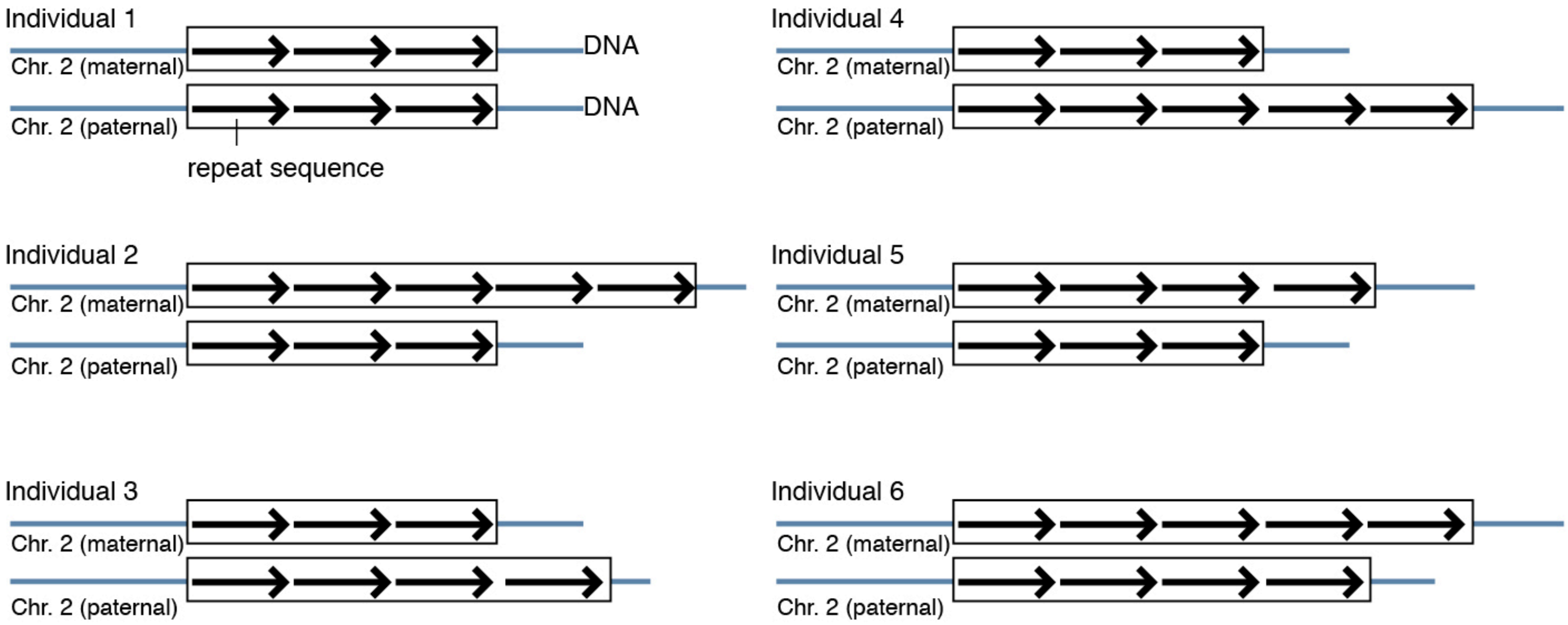
4. Copy Number Variation (CNV).

5. Structural Variants (SV).

6. Annotating variants and assessing its impact.

4. Copy Number Variation (CNV).

A **copy number variation (CNV)** is when the number of copies of a particular gene varies from one individual to the next.



https://www.genome.gov/genetics-glossary/Copy-Number-Variation

4. Copy Number Variation (CNV).

A **copy number variation (CNV)** is when the number of copies of a particular gene varies from one individual to the next.

PLOS COMPUTATIONAL BIOLOGY

OPEN ACCESS PEER-REVIEWED

RESEARCH ARTICLE

Comprehensively benchmarking applications for detecting copy number variation

Le Zhang, Wanyu Bai, Na Yuan, Zhenglin Du

Version 2 Published: May 28, 2019 • https://doi.org/10.1371/journal.pcbi.1007069

4. Copy Number Variation (CNV).

A **copy number variation (CNV)** is when the number of copies of a particular gene varies from one individual to the next.

Software	Methods	Algorithm detail	Input data	Publish	Latest update	Accessibility	URL	Programing Language	#Citations
#Canvas	RD	Expectation-maximization (EM) clustering	BAM	2011	2018/3	Y	https://github.com/Illumina/canvas	C#	29
#cn.MOPS	RD	Mixture Poisson model	BAM	2012	2018/10	Y	http://www.bioinf.jku.at/software/cnmops/cnmops.html	R	226
CNVeM	RD	Expectation-maximization (EM) algorithm	CSV	2013	NA	Y	https://omictools.com/cnvem-tool	C	14
CNVer	RP	Maximum-likelihood, Graphic flow	BAM	2010	2011/5	N	NA	C	158
#CNVnator	RD	Mean shift algorithm	BAM	2011	2016/11	Y	https://github.com/abyzovlab/CNVnator	C++	640
CNVrd2	RD	Expectation-maximization (EM) algorithm	BAM/SAM	2014	2015/11	Y	https://bioconductor.org/packages/release/bioc/html/CNVrd2.html	R	13
#Control-FREEC	RD	LASSO regression	BAM/SAM	2011	2018/8	Y	http://boevalab.com/FREEC/	C++	190
#GROM-RD	RD	Quantile normalization	BAM	2015	2017/5	Y	http://grigoriev.rutgers.edu/software/	C	7
#iCopyDAV	RD	DoC approaches	BAM	2018	2018/3	Y	https://github.com/vogetihrsh/icopydav	R,C++	1
JointSLM	RD	Population-based approach	SAM/BAM	2011	NA	N	NA	R	49
#LUMPY	RD, PEM	A probabilistic framework	BAM/CRAM	2014	2016/3	Y	https://github.com/arq5x/lumpy-sv	C++	157
mrCaNaVAR	RD	mrFAST	SAM	2009	2013/9	Y	http://mrcanavar.sourceforge.net/	C	685
#RDXplorer	RD	Event-wise testing algorithm	BAM	2009	2013/4	Y	https://sourceforge.net/projects/rdxplorer/	Python	496
#ReadDepth	RD	Circular binary segmentation algorithm	Bed Files	2011	2014/8	Y	https://github.com/chrisamiller/readDepth	R	150
#RSICNV	RD	Negative binomial transformations	BAM	2017	2017/7	Y	https://github.com/yhwu/rsicnv	C++	2

Note:
# indicates the software used in this study.

https://doi.org/10.1371/journal.pcbi.1007069.t001

Zhang, Le, et al. "Comprehensively benchmarking applications for detecting copy number variation." *PLoS computational biology* 15.5 (2019): e1007069.

Outline of Topics

1. Basics about variants in genomics.

2. Manipulating SAM/BAMs and coverage.

3. Simple variants: SNVs, InDels and MNVs.

4. Copy Number Variation (CNV).

5. Structural Variants (SV).

6. Annotating variants and assessing its impact.

5. Structural Variants (SV).

Structural variation (SV) is generally defined as a **region of DNA approximately 1 kb and larger** in size and can include **inversions and balanced translocations or genomic imbalances** (insertions and deletions), commonly referred to as copy number variants (CNVs). These CNVs often overlap with segmental duplications, regions of DNA >1 kb present more than once in the genome, copies of which are >90% identical. If present at >1% in a population a CNV may be referred to as copy number polymorphism (CNP).

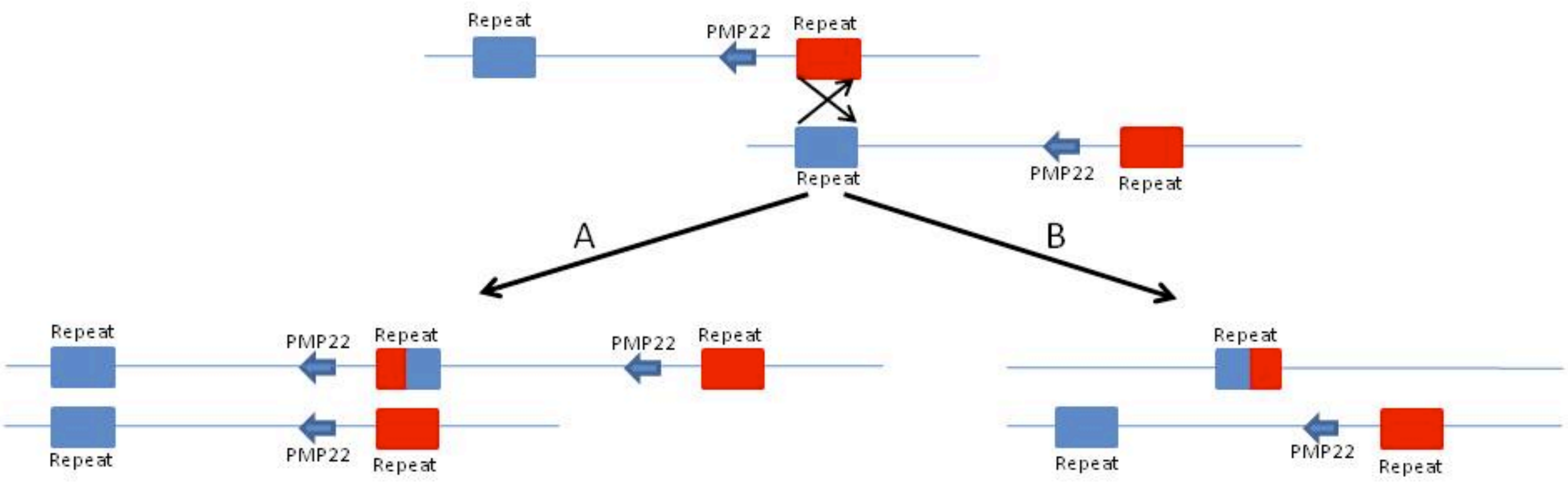


Figure 1: Charcot-Marie Tooth (CMT) disease. Unequal crossing over between two highly homologous repeats on chromosome 17p12 can result in (A) 3 copies of the PMP22 gene with the CMT1A phenotype or the reciprocal (B) and 1 copy of the PMP22 gene with the HNPP phenotype.

https://www.ncbi.nlm.nih.gov/dbvar/content/overview/

5. Structural Variants (SV).

Structural variation (SV) is generally defined as a **region of DNA approximately 1 kb and larger** in size and can include **inversions and balanced translocations or genomic imbalances**

Research | Open Access | Published: 03 June 2019

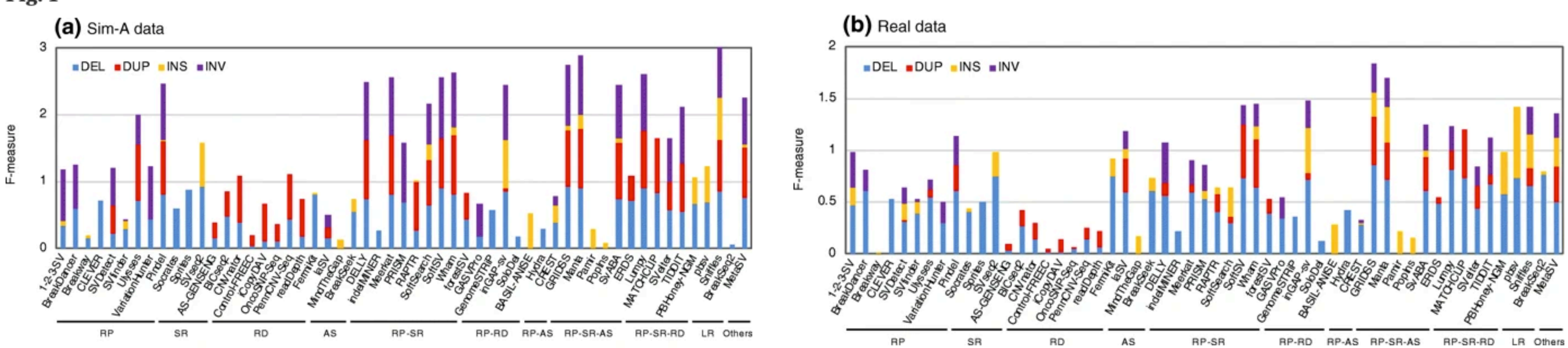
Comprehensive evaluation of structural variation detection algorithms for whole genome sequencing

Shunichi Kosugi, Yukihide Momozawa, Xiaoxi Liu, Chikashi Terao, Michiaki Kubo & Yoichiro Kamatani

Genome Biology **20**, Article number: 117 (2019) | Cite this article

19k Accesses | **10** Citations | **103** Altmetric | Metrics

Fig. 1



5. Structural Variants (SV).

Structural variation (SV) is generally defined as a **region of DNA approximately 1 kb and larger** in size and can include **inversions and balanced translocations or genomic imbalances**

From: Comprehensive evaluation of structural variation detection algorithms for whole genome sequencing

SV type	Tools	Simulated data		Real data		nF*1
		Precision	Recall	Precision	Recall	
DEL	GRIDSS	98.9 (5)	86.6 (2)	87.6 (7)	28.9 (2)	3.57 (1)
	Lumpy	99.1 (4)	81.4 (6)	87.1 (8)	26.1 (4)	3.41 (2)
	SVseq2	96.2 (11)	86.1 (3)	75.7 (17)	24.9 (5)	3.28 (3)
	SoftSV	96.8 (10)	83.6 (4)	80.2 (13)	23.2 (8)	3.25 (7)
	Manta	95.9 (12)	83.1 (5)	74.2 (20)	24.3 (6)	3.21 (5)
	MATCHCLIP	99.4 (2)	71.7 (10)	91.6 (4)	20.9 (11)	3.12 (6)
	inGAP-sv	91.1 (18)	78.6 (7)	78.3 (14)	22.5 (8)	3.10 (7)
DUP	Wham	96.9 (4)	81.7 (4)	57.1 (4)	10.2 (5)	3.92 (1)
	SoftSV	84.2 (14)	67.8 (13)	47.3 (6)	14.3 (3)	3.91 (2)
	MATCHCLIP	87.6 (11)	77.5 (8)	58.0 (3)	9.9 (6)	3.79 (3)
	GRIDSS	91.1 (9)	77.9 (7)	58.4 (2)	9.6 (7)	3.78 (4)
	Manta	99.0 (1)	83.2 (1)	40.4 (9)	6.5 (11)	3.35 (5)
	SvABA	82.6 (15)	69.6 (11)	42.7 (8)	7.2 (9)	3.02 (6)
INS [Unspecified]	pbsv	89.7 (3)	38.2 (5)	72.7 (8)	27.5 (2)	6.68 (1)
	inGAP-sv	99.7 (1)	58.5 (2)	85.5 (2)	11.8 (3)	6.27 (2)
	Sniffles	74.8 (5)	52.5 (3)	65.9 (10)	9.0 (5)	5.08 (3)
	SVseq2	70.4 (8)	64.2 (1)	38.5 (19)	7.1 (9)	4.87 (4)
INS [MEI]	MELT	99.7 (3)	68.9 (3)	88.9 (1)	85.6 *2 (1)	3.21 (1)
	Mobster	100 (1)	67.1 (4)	88.3 (2)	71.9 *2 (2)	3.04 (2)
INV	DELLY	94.7 (8)	81.8 (4)	38.9 (4)	15.6 (2)	3.07 (1)
	TIDDIT	89.2 (14)	77.9 (8)	49.1 (1)	11.7 (5)	2.89 (2)
	1–2-3-SV	70.7 (19)	81.2 (5)	31.8 (9)	14.8 (3)	2.67 (3)
	GRIDSS	96.6 (6)	84.7 (3)	34.2 (8)	10.4 (7)	2.67 (4)

Kosugi, Shunichi, et al. "Comprehensive evaluation of structural variation detection algorithms for whole genome sequencing." *Genome biology* 20.1 (2019): 117.

Outline of Topics

1. Basics about variants in genomics.

2. Manipulating SAM/BAMs and coverage.

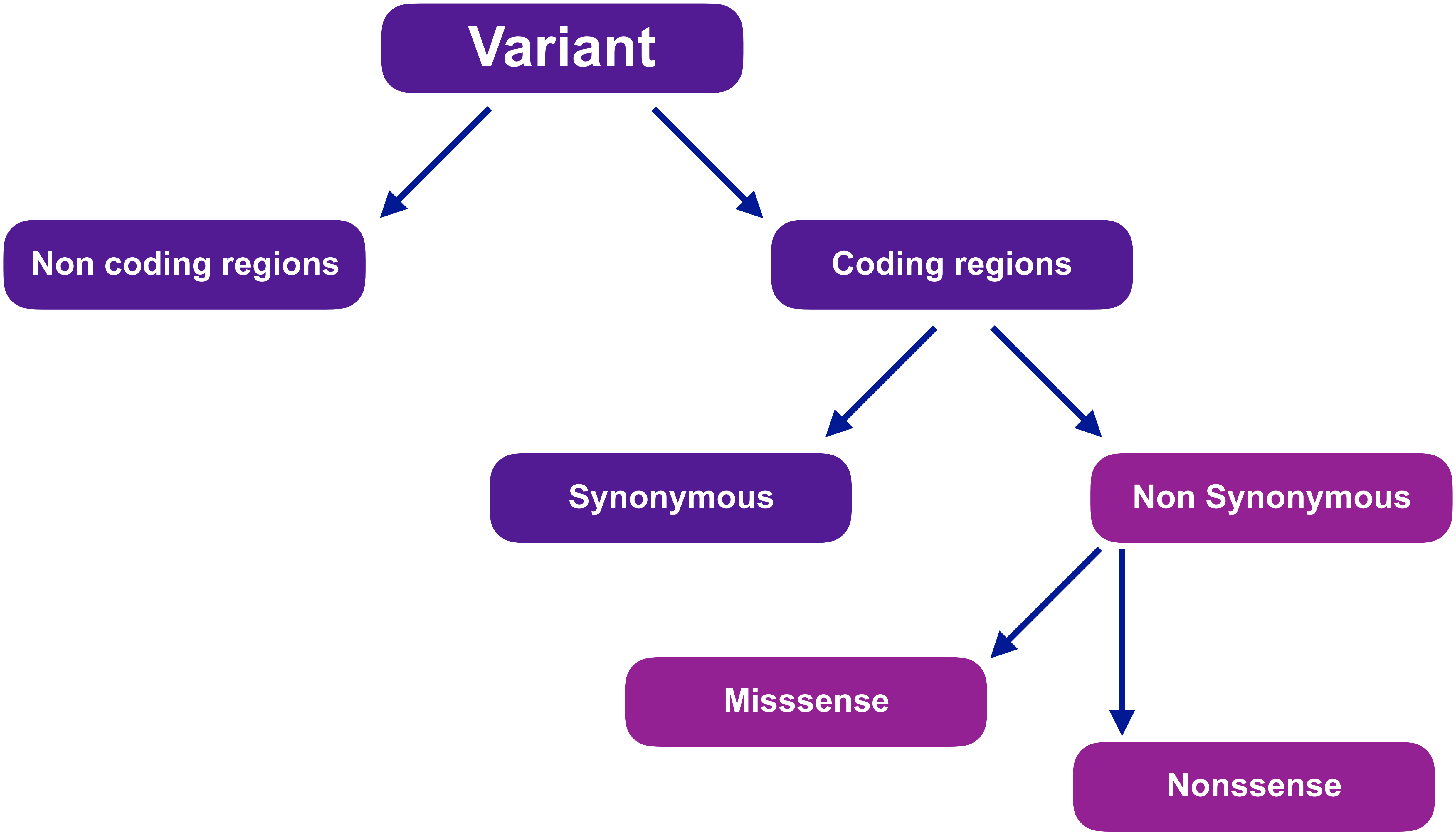
3. Simple variants: SNVs, InDels and MNVs.

4. Copy Number Variation (CNV).

5. Structural Variants (SV).

6. Annotating variants and assessing its impact.

6. Annotating variants and assessing its impact.



6. Annotating variants and assessing its impact.

mRNA is read by groups of **three nucleotides** called **codons**. Each three nucleotides represent an aminoacid that it is carried by a tRNA during the translation.

Amino acids biochemical properties	nonpolar	polar	basic	acidic	Termination: stop codon

Standard genetic code

1st base	2nd base T		2nd base C		2nd base A		2nd base G		3rd base
T	TTT	(Phe/F) Phenylalanine	TCT	(Ser/S) Serine	TAT	(Tyr/Y) Tyrosine	TGT	(Cys/C) Cysteine	T
	TTC		TCC		TAC		TGC		C
	TTA	(Leu/L) Leucine	TCA		TAA	Stop (*Ochre*)[B]	TGA	Stop (*Opal*)[B]	A
	TTG[A]		TCG		TAG	Stop (*Amber*)[B]	TGG	(Trp/W) Tryptophan	G
C	CTT		CCT	(Pro/P) Proline	CAT	(His/H) Histidine	CGT	(Arg/R) Arginine	T
	CTC		CCC		CAC		CGC		C
	CTA		CCA		CAA	(Gln/Q) Glutamine	CGA		A
	CTG[A]		CCG		CAG		CGG		G
A	ATT	(Ile/I) Isoleucine	ACT	(Thr/T) Threonine	AAT	(Asn/N) Asparagine	AGT	(Ser/S) Serine	T
	ATC		ACC		AAC		AGC		C
	ATA		ACA		AAA	(Lys/K) Lysine	AGA	(Arg/R) Arginine	A
	ATG[A]	(Met/M) Methionine	ACG		AAG		AGG		G
G	GTT	(Val/V) Valine	GCT	(Ala/A) Alanine	GAT	(Asp/D) Aspartic acid	GGT	(Gly/G) Glycine	T
	GTC		GCC		GAC		GGC		C
	GTA		GCA		GAA	(Glu/E) Glutamic acid	GGA		A
	GTG		GCG		GAG		GGG		G

https://en.wikipedia.org/wiki/DNA

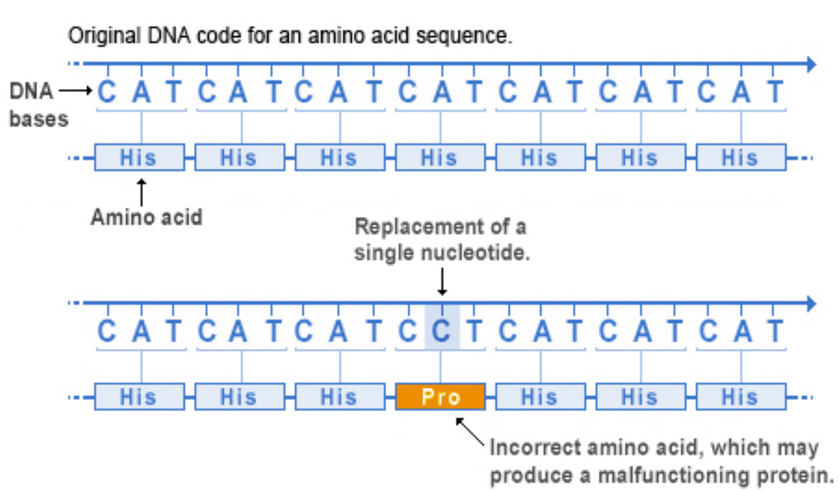
6. Annotating variants and assessing its impact.

A **non-synonymous substitution** is a nucleotide mutation that **alters the amino acid sequence of a protein**. Non-synonymous substitutions differ from synonymous substitutions, which do not alter amino acid sequences and are (sometimes) silent mutations. **As non-synonymous substitutions result in a biological change in the organism, they are subject to natural selection**.

Non Synonymous

Misssense | **Nonssense**

Missense mutation



U.S. National Library of Medicine

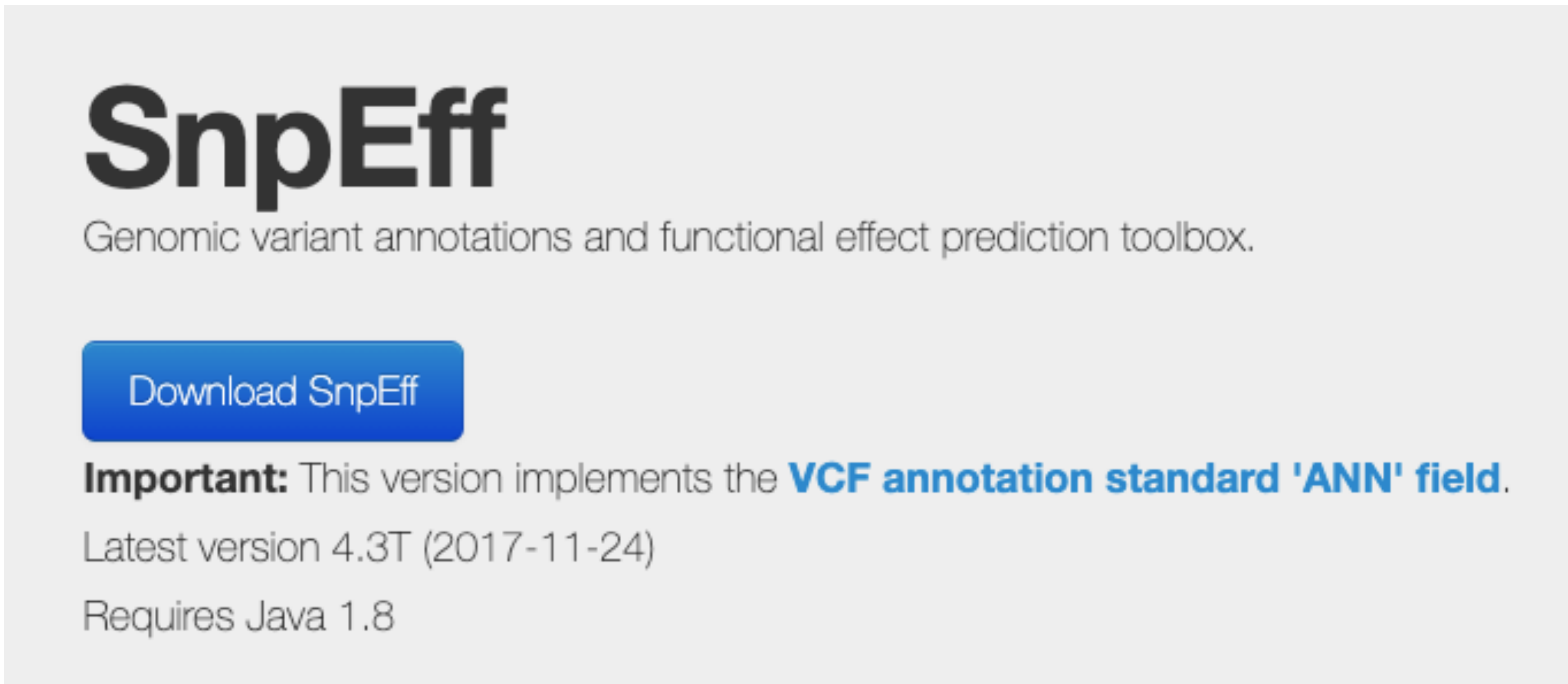
```
     DNA: 5' - ATG ACT CAC TGA GCG CGA AGC TGA - 3'
          3' - TAC TGA GTG ACT CGC GCT TCG ACT - 5'
    mRNA: 5' - AUG ACU CAC UGA GCG CGU AGC UGA - 3'
Protein:       Met Thr His Stop
```

https://en.wikipedia.org/wiki/Nonsynonymous_substitution

6. Annotating variants and assessing its impact.

Program used to annotate variants

http://snpeff.sourceforge.net/



6. Annotating variants and assessing its impact.

Program used to annotate variants http://snpeff.sourceforge.net/

Type	What is means	Example
SNP	Single-Nucleotide Polymorphism	Reference = 'A', Sample = 'C'
Ins	Insertion	Reference = 'A', Sample = 'AGT'
Del	Deletion	Reference = 'AC', Sample = 'C'
MNP	Multiple-nucleotide polymorphism	Reference = 'ATA', Sample = 'GTC'
MIXED	Multiple-nucleotide and an InDel	Reference = 'ATA', Sample = 'GTCAGT'

EFF Sub-field	Meaning
Effect	Effect of this variant. See details here.
Effect impact	Effect impact {High, Moderate, Low, Modifier}. See details here.
Functional Class	Functional class {NONE, SILENT, MISSENSE, NONSENSE}.
Codon_Change / Distance	Codon change: old_codon/new_codon OR distance to transcript (in case of upstream / downstream)
Amino_Acid_Change	Amino acid change: old_AA AA_position/new_AA (e.g. 'E30K')
Amino_Acid_Length	Length of protein in amino acids (actually, transcription length divided by 3).
Gene_Name	Gene name
Transcript_BioType	Transcript bioType, if available.
Gene_Coding	[CODING | NON_CODING]. This field is 'CODING' if any transcript of the gene is marked as protein coding.
Transcript_ID	Transcript ID (usually ENSEMBL IDs)
Exon/Intron Rank	Exon rank or Intron rank (e.g. '1' for the first exon, '2' for the second exon, etc.)
Genotype_Number	Genotype number corresponding to this effect (e.g. '2' if the effect corresponds to the second ALT)
Warnings / Errors	Any warnings or errors (not shown if empty).

6. Annotating variants and assessing its impact.

Program used to annotate variants http://snpeff.sourceforge.net/

Effect Seq. Ontology	Effect Classic	Note & Example	Impact
coding_sequence_variant	CDS	The variant hits a CDS.	MODIFIER
chromosome	CHROMOSOME_LARGE DELETION	A large part (over 1% or 1,000,000 bases) of the chromosome was deleted.	HIGH
duplication	CHROMOSOME_LARGE_DUPLICATION	Duplication of a large chromoome segment (over 1% or 1,000,000 bases).	HIGH
inversion	CHROMOSOME_LARGE_INVERSION	Inversion of a large chromoome segment (over 1% or 1,000,000 bases).	HIGH
coding_sequence_variant	CODON_CHANGE	One or many codons are changed e.g.: An MNP of size multiple of 3	LOW
inframe_insertion	CODON_INSERTION	One or many codons are inserted e.g.: An insert multiple of three in a codon boundary	MODERATE
frameshift_variant	FRAME_SHIFT	Insertion or deletion causes a frame shift e.g.: An indel size is not multple of 3	HIGH