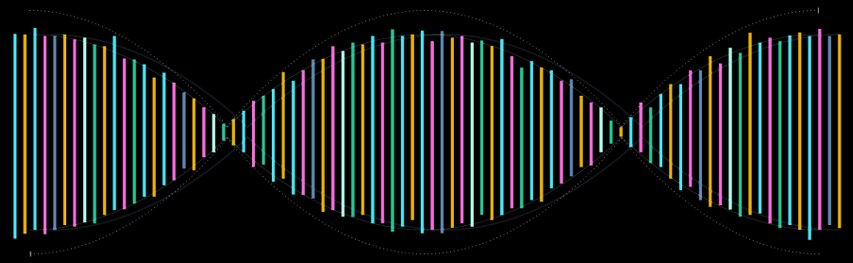
Genotek

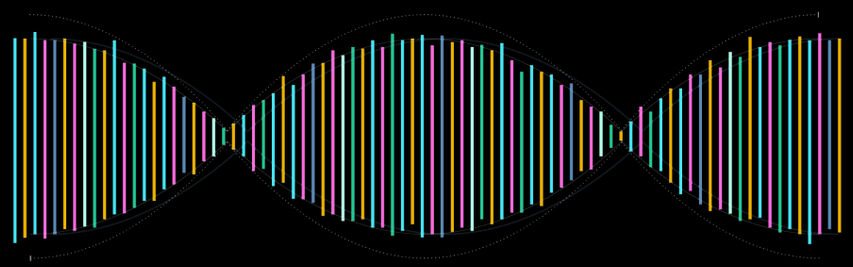
Практическая биоинформатика

Занятие 1

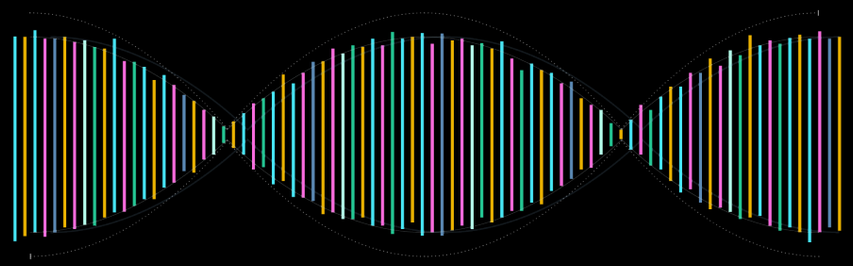
Александр Ракитько



...AATACAGTGCGA...
...AATGCAGTGCGA...



...AATGCAGTGCGA...
...AATGCAGTCCGA...



...AATGCAGTGCGA...
...AATACAGTGCGA...

Технологии анализа генома

ПЦР
(1 мутация)

- Отдельные «горячие» мутации

Микрочиповое генотипирование
(650 тыс. мутаций)

- Риски заболеваний
- Наследственные заболевания
- Фармакогенетика
- Питание
- Происхождение

Экзомное секвенирование
(12-30 млн. мутаций)

- Наследственные заболевания
- Фармакогенетика

Полногеномное секвенирование
(3 млрд. мутаций)

- Риски заболеваний
- Наследственные заболевания
- Фармакогенетика
- Питание
- Происхождение

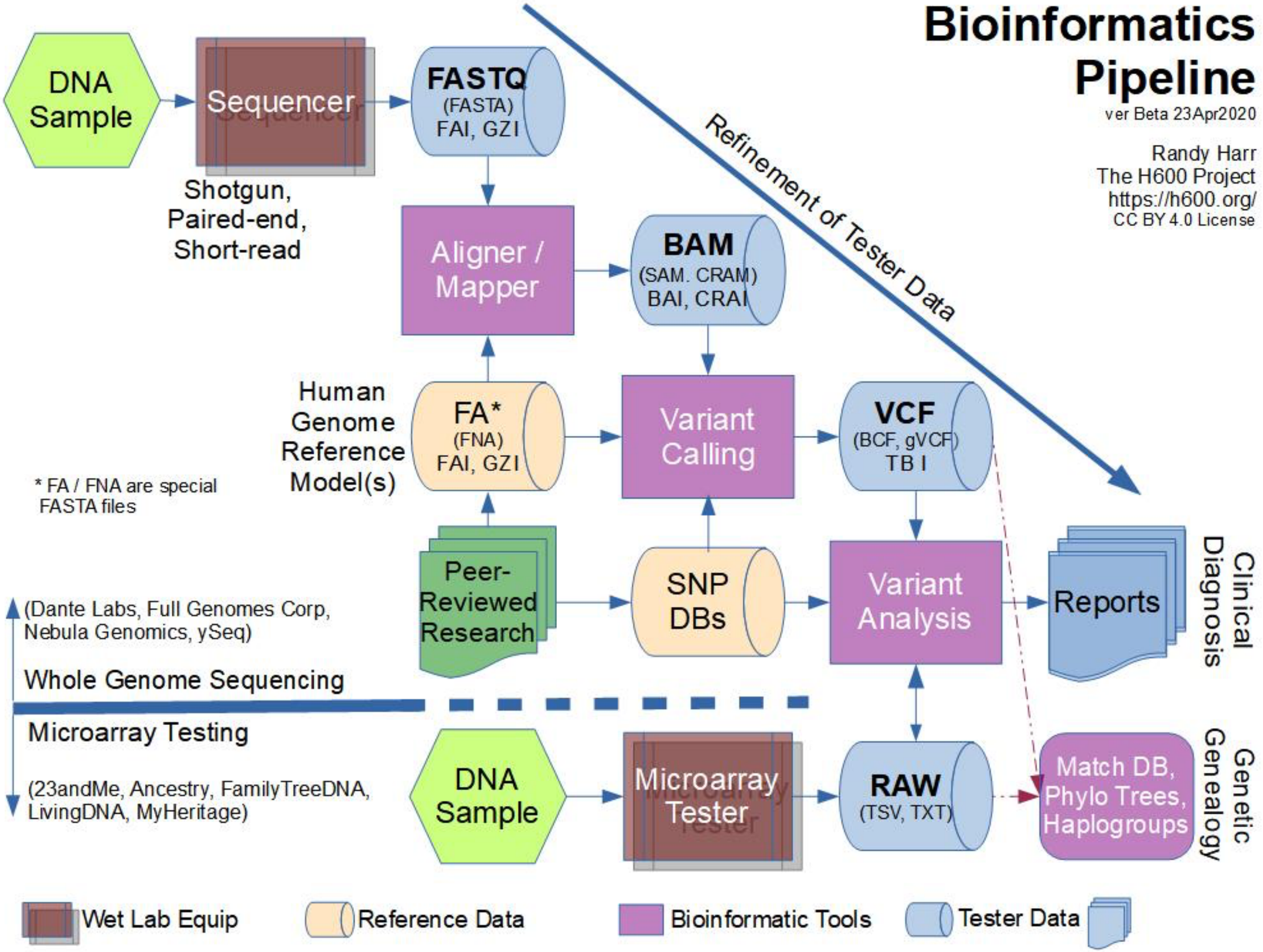
Форматы данных

- Данные разного типа хранятся в разных форматах, универсального не существует
- При возможности стоит использовать стандартные и распространенные форматы данных
- XLSX — последнее, что стоит использовать
- Описание большинства форматов можно найти здесь: https://genome.ucsc.edu/FAQ/FAQformat.html

Bioinformatics Pipeline

ver Beta 23Apr2020

Randy Harr
The H600 Project
https://h600.org/



FASTA file

```
>NG_008679.1:5001-38170 Homo sapiens paired box 6 (PAX6)
ACCCTCTTTTCTTATCATTGACATTTAAACTCTGGGGCAGGTCCTCGCGTAGAACGCGGCTGTCAGATCT
GCCACTTCCCCTGCCGAGCGGCGGTGAGAAGTGTGGGAACCGGCGCTGCCAGGCTCACCTGCCTCCCCGC
CCTCCGCTCCCAGGTAACCGCCCGGGCTCCGGCCCCGGCCCGGCTCGGGGCCCGCGGGGCCTCTCCGCTG
CCAGCGACTGCTGTCCCCAAATCAAAGCCCGCCCCAAGTGGCCCCGGGGCTTGATTTTTGCTTTTAAAAG
GAGGCATACAAAGATGGAAGCGAGTTACTGAGGGAGGGATAGGAAGGGGGGTGGAGGAGGGACTTGTCTT
TGCCGAGTGTGCTCTTCTGCAAAAGTAGCAAAATGTTCCACTCCTAAGAGTGGACTTCCAGTCCGGCCCT
GAGCTGGGAGTAGGGGGCGGGAGTCTGCTGCTGCTGTCTGCTAAAGCCACTCGCGACCGCGAAAAATGCA
GGAGGTGGGGACGCACTTTGCATCCAGACCTCCTCTGCATCGCAGTTCACGACATCCACGCTTGGGAAAG
TCCGTACCCGCGCCTGGAGCGCTTAAAGACACCCTGCCGCGGGTCGGGCGAGGTGCAGCAGAAGTTTCCC
GCGGTTGCAAAGTGCAGATGGCTGGACCGCAACAAAGTCTAGAGATGGGGTTCGTTTCTCAGAAAGACGC
```

FIGURE 7.1: An example fasta file showing the first part of the PAX6 gene.

https://compgenomr.github.io/book/fasta-and-fastq-formats.html

https://www.ncbi.nlm.nih.gov/nuccore/1798174254

FASTQ file

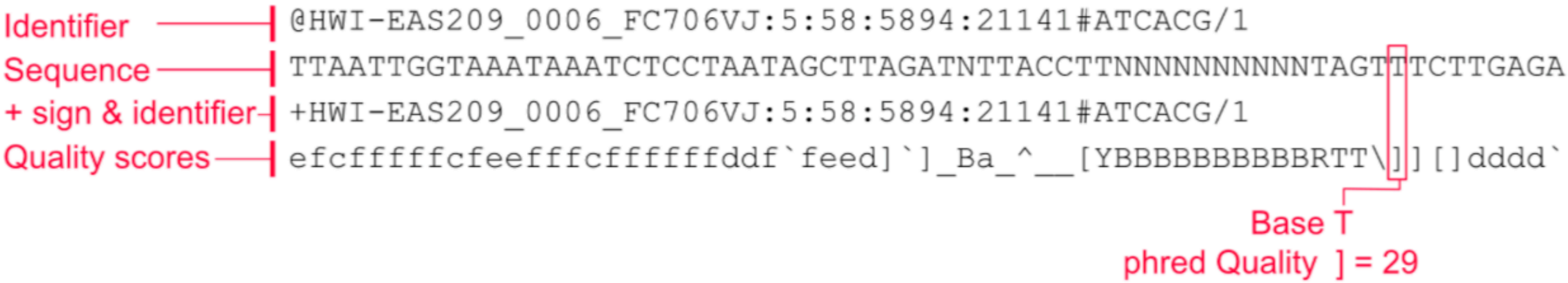


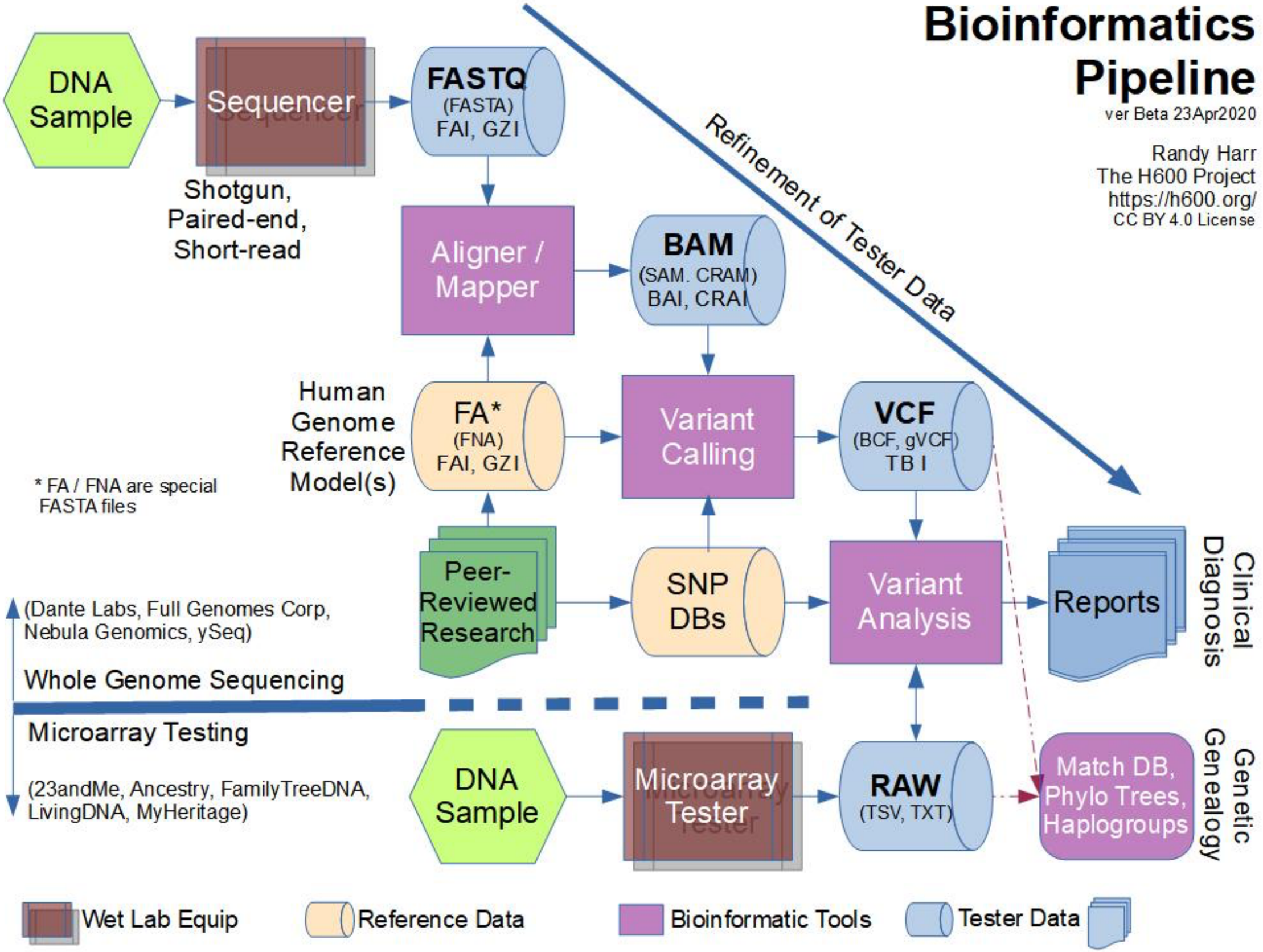
FIGURE 7.2: FASTQ format and a brief explanation of each line in the format.

https://compgenomr.github.io/book/fasta-and-fastq-formats.html

Bioinformatics Pipeline

ver Beta 23Apr2020

Randy Harr
The H600 Project
https://h600.org/



SAM file

http://zyxue.github.io/2017/09/26/sam-format-example.html

A

```
                  10        20        30        40
Coor      12345678901234  5678901234567890123456789012345
ref       AGCATGTTAGATAA**GATAGCTGTGCTAGTAGGCAGTCAGCGCCAT

+r001/1        TTAGATAAAGGATA*CTG
+r002         aaaAGATAA*GGATA
+r003       gcctaAGCTAA
+r004                     ATAGCT..............TCAGC
-r003                           ttagctTAGGC
-r001/2                                        CAGCGGCAT
```

B

```
@HD VN:1.5 SO:coordinate
@SQ SN:ref LN:45
r001    99 ref  7 30 8M2I4M1D3M = 37  39 TTAGATAAAGGATACTG *
r002     0 ref  9 30 3S6M1P1I4M * 0   0 AAAAGATAAGGATA     *
r003     0 ref  9 30 5S6M       * 0   0 GCCTAAGCTAA        * SA:Z:ref,29,-,6H5M,17,0;
r004     0 ref 16 30 6M14N5M    * 0   0 ATAGCTTCAGC        *
r003  2064 ref 29 17 6H5M       * 0   0 TAGGC              * SA:Z:ref,9,+,5S6M,30,1;
r001   147 ref 37 30 9M         = 7 -39 CAGCGGCAT          * NM:i:1
```

Header section

QUAL (read quality; * meaning such information is not available)

Alignment section

QNAME (query template name, aka. read ID)

FLAG (indicates alignment information about the read, e.g. paired, aligned, etc.)

RNAME (reference sequence name, e.g. chromosome /transcript id)

POS (1-based position)

MAPQ (mapping quality)

CIGAR (summary of alignment, e.g. insertion, deletion)

RNEXT (reference sequence name of the primary alignment of the NEXT read; for paired-end sequencing, NEXT read is the paired read; corresponding to the RNAME column)

PNEXT (Position of the primary alignment of the NEXT read in the template; corresponding to the POS column)

TLEN (the number of bases covered by the reads from the same fragment. In this particular case, it's 45 - 7 + 1 = 39 as highlighted in Panel A). Sign: plus for leftmost read, and minus for rightmost read

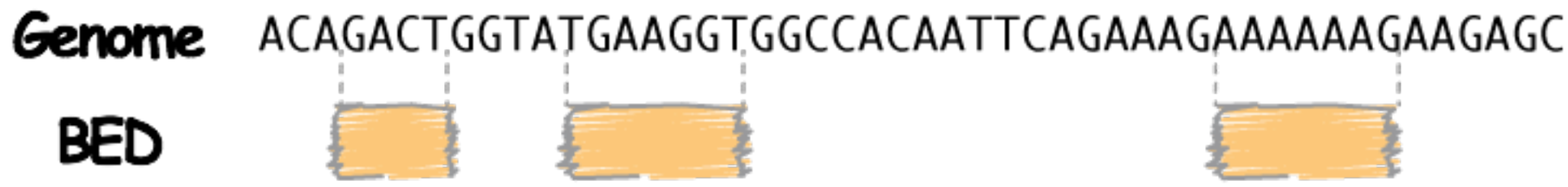
SEQ (read sequence)

Optional fields in the format of TAG:TYPE:VALUE

SAM / BAM / CRAM

- Выравнивания хранятся в SAM формате. SAM можно сжать в BAM или CRAM
- Работать с ними можно с помощью **samtools** (просмотр, объединение, фильтрация)
- Всегда сортируем и индексируем
- Для визуализации можно использовать IGV
- BAM можно конвертировать в FASTA/FASTQ

BED file




```
chr1   213941196   213942363
chr1   213942363   213943530
chr1   213943530   213944697
```

```
chr7   127471196   127472363   Pos1   0   +
chr7   127472363   127473530   Pos2   0   +
chr7   127473530   127474697   Pos3   0   +
```

Chromosome ID →

Start location

End location

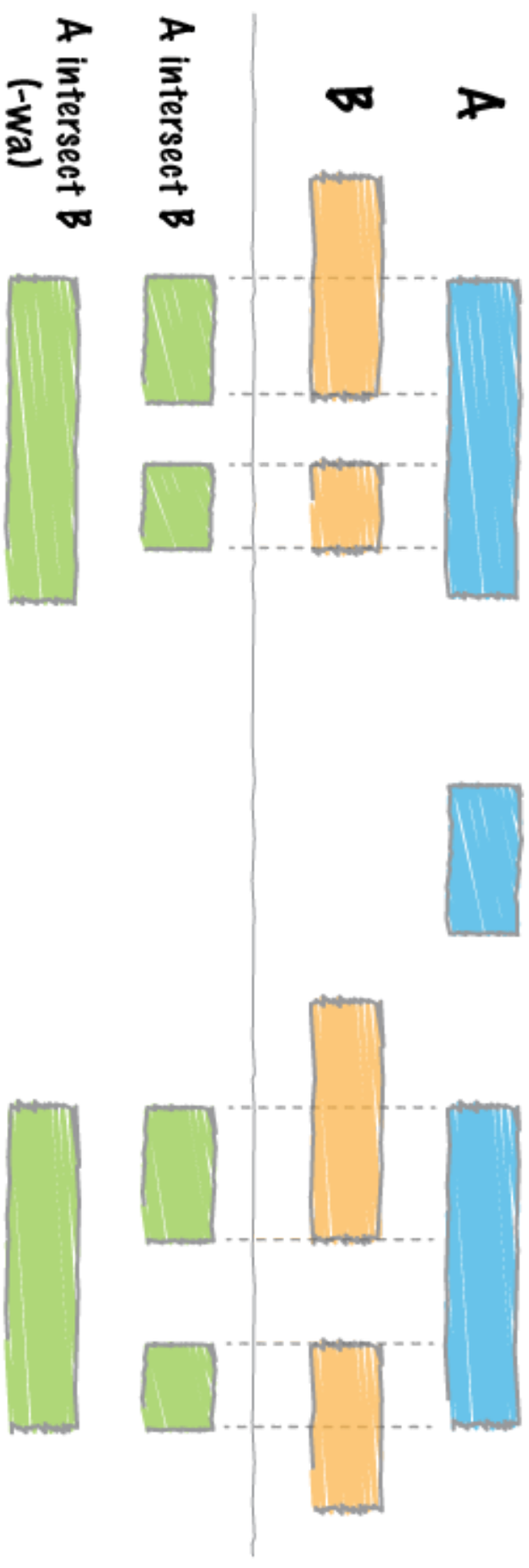
Name

Phase (reading frame)

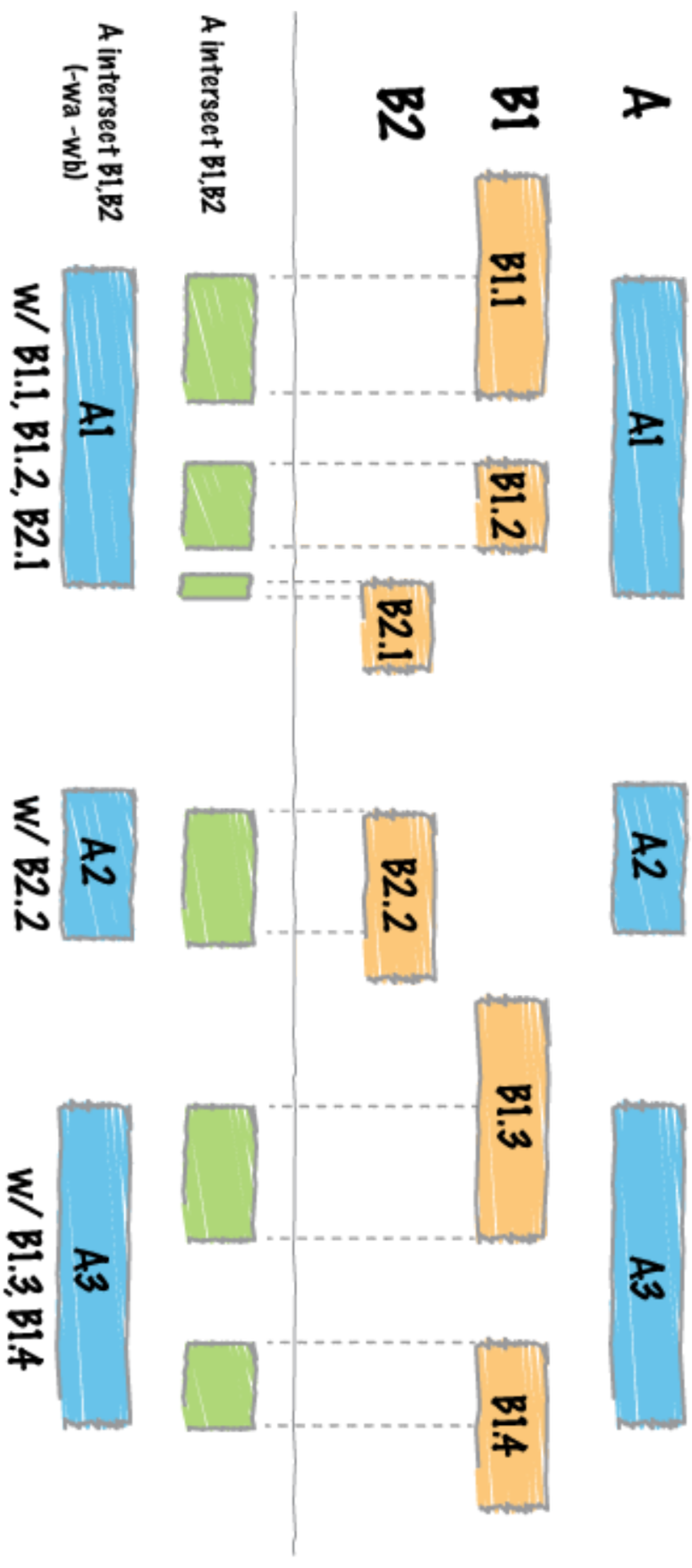
Strand

https://hbctraining.github.io/Intro-to-ChIPseq-flipped/lessons/07_handling_peaks_bedtools.html

Intersect w/ 1 database



Intersect w/ 2 or more databases



BED file

- Используется для хранения и работы с геномными интервалами
- Например, можно использовать для координат генов, экзонов и тд.
- Не содержит header
- Удобный инструмент для работы **bedtools**

VCF format

HEADER

```
##fileformat=VCFv4.1
##fileDate=20090805
##tcgaversion=1.1
##vcfProcessLog=<InputVCF=<file1.vcf>,InputVCFSource=<caller1>,InputVCFVer=<1.0>,InputVCFParam=<a1,b>,InputVCFgeneAnno=<anno1.gaf>>
##reference=ftp://ftp.ncbi.nih.gov/genbank/genomes/Eukaryotes/vertebrates_mammals/Homo_sapiens/GRCh37/special_requests/GRCh37-lite.fa
##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">
```

INFO meta-information

```
##FILTER=<ID=q10,Description="Quality below 10">
##FILTER=<ID=s50,Description="Less than 50% of samples have data">
```

FILTER meta-information

```
##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">
```

FORMAT meta-information

```
##SAMPLE=<ID=NORMAL,Individual=TCGA-01-1000,File=TCGA-01-1000-1.bam,Platform=Illumina,Source=dbGAP,Accession=1234>
##SAMPLE=<ID=TUMOR,Individual=TCGA-01-1000,File=TCGA-01-1000-2.bam,Platform=Illumina,Source=dbGAP,Accession=4567>
##PEDIGREE=<Name_0=TUMOR,Name_1=NORMAL>
```

Fixed fields

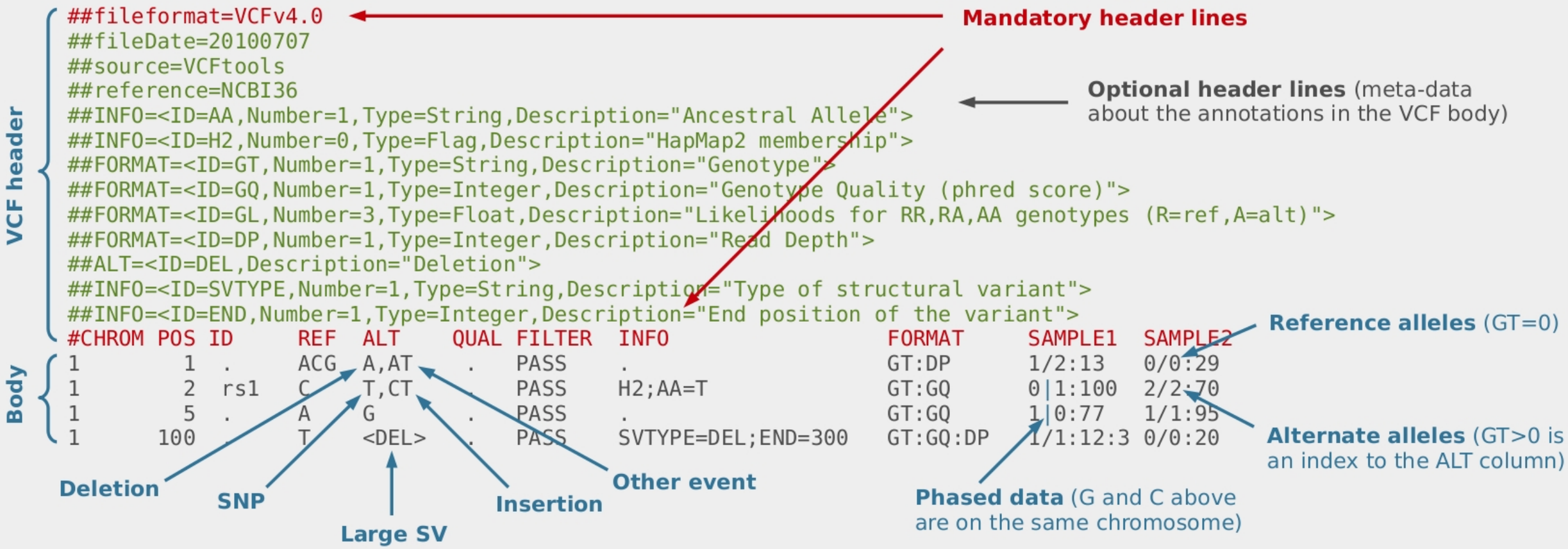
Optional: FORMAT field specifying data type + Per-sample genotype data

BODY

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	NORMAL	TUMOR
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
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
20	1110696	rs6040355	A	G,T	67	PASS	NS=2;DP=10;AF=0.333,0.667;DB	GT:GQ:DP:HQ	1|2:21:6:23,27	2|1:2:0:18,2
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
20	1234567	microsat1	GTC	G,GTCTC	50	PASS	NS=3;DP=9;AA=G	GT:GQ:DP	0/1:35:4	0/2:17:2

https://davetang.github.io/learning_vcf_file/

VCF format


VCF header
##fileformat=VCFv4.0
##fileDate=20100707
##source=VCFtools
##reference=NCBI36
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality (phred score)">
##FORMAT=<ID=GL,Number=3,Type=Float,Description="Likelihoods for RR,RA,AA genotypes (R=ref,A=alt)">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##ALT=<ID=DEL,Description="Deletion">
##INFO=<ID=SVTYPE,Number=1,Type=String,Description="Type of structural variant">
##INFO=<ID=END,Number=1,Type=Integer,Description="End position of the variant">
#CHROM POS ID REF ALT QUAL FILTER INFO FORMAT SAMPLE1 SAMPLE2
Body
1 1 . ACG A,AT . PASS . GT:DP 1/2:13 0/0:29
1 2 rs1 C T,CT . PASS H2;AA=T GT:GQ 0|1:100 2/2:70
1 5 . A G . PASS . GT:GQ 1|0:77 1/1:95
1 100 . T <DEL> . PASS SVTYPE=DEL;END=300 GT:GQ:DP 1/1:12:3 0/0:20
Mandatory header lines
Optional header lines (meta-data about the annotations in the VCF body)
Reference alleles (GT=0)
Alternate alleles (GT>0 is an index to the ALT column)
Deletion
SNP
Large SV
Insertion
Other event
Phased data (G and C above are on the same chromosome)

VCF file

- Содержит данные генотипирования
- Часто содержит только позиции с альтернативными аллелями
- Подходит для хранения больших референсных панелей, фазированных данных
- Удобный инструмент для работы **bcftools / vcftools**

PLINK format

*.ped

FID	IID	PID	MID	Sex	P	rs1	rs2	rs3
1	1	0	0	2	1	CT	AG	AA
2	2	0	0	1	0	CC	AA	AC
3	3	0	0	1	1	CC	AA	AC

*.map

Chr	SNP	GD	BPP
1	rs1	0	870000
1	rs2	0	880000
1	rs3	0	890000

*.fam

FID	IID	PID	MID	Sex	P
1	1	0	0	2	1
2	2	0	0	1	0
3	3	0	0	1	1

*.bed

Contains binary version of the SNP info of the *.ped file. (not in a format readable for humans)

*.bim

Chr	SNP	GD	BPP	Allele 1	Allele 2
1	rs1	0	870000	C	T
1	rs2	0	880000	A	G
1	rs3	0	890000	A	C

Covariate file

FID	IID	C1	C2	C3
1	1	0.00812835	0.00606235	-0.000871105
2	2	-0.0600943	0.0318994	-0.0827743
3	3	-0.0431903	0.00133068	-0.000276131

Legend			
FID	Family ID	rs{x}	Alleles per subject per SNP
IID	Individual ID	Chr	Chromosome
PID	Paternal ID	SNP	SNP name
MID	Maternal ID	GD	Genetic distance (morgans)
Sex	Sex of subject	BPP	Base-pair position (bp units)
P	Phenotype	C{x}	Covariates (e.g., Multidimensional Scaling (MDS) components)

PLINK format

- Содержит данные генотипирования
- Позволяет сохранять родственные связи и фенотипы
- Легче, чем VCF
- «Теряет» фазирование
- Удобный инструмент для работы **plink**

Форматы данных

Формат	Данные	Инструменты
FASTA	Последовательность нуклеотидов	
FASTQ	Последовательности ридов	
SAM/BAM/CRAM	Выровненные риды	samtools
BED	Регионы	bedtools
VCF	Генотипы	bcftools/vcftools
PLINK	Генотипы	plink