Genomic data

Libor Mořkovský, Václav Janoušek

Genomic data

- Genome from the bioinformatic perspective
- Where does the genomic data come from?
- Common genomic data formats
- Specialized tools for genomic data

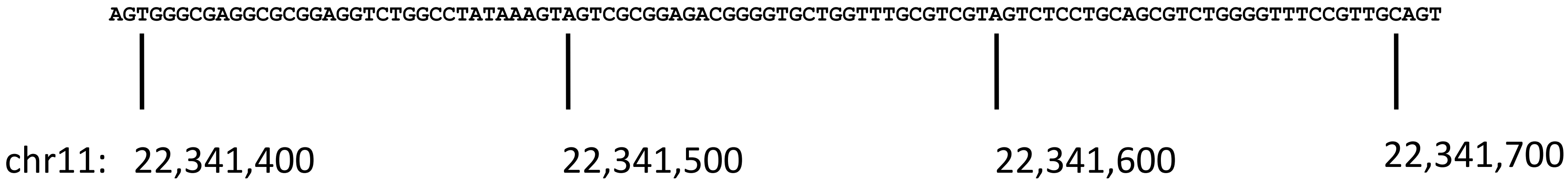
Genome from the bioinformatic perspective

- sequence

```
AGTGGGCGAGGCGCGGAGGTCTGGCCTATAAAGTAGTCGCGGAGACGGGG
TGCTGGTTTGCGTCGTAGTCTCCTGCAGCGTCTGGGGTTTCCGTTGCAGT
CCTCGGAACCAGGACCTCGGCGTGGCCTAGCGAGTTATGGCGACGAAGGC
CGTGTGCGTGCTGAAGGGCGACGGCCCAGTGCAGGGCATCATCAATTTCG
AGCAGAAGGCAAGGGCTGGGACGGAGGCTTGTTTGCGAGGCCGCTCCCAC
CCGCTCGTCCCCCCGCGCACCTTTGCTAGGAGCGGGTCGCCCGCCAGGCC
TCGGGGCCGCCCTGGTCCAGCGCCCGGTCCCGGCCCGTGCCGCCCGGTCG
GTGCCTTCGCCCCCAGCGGTGCGGTGCCCAAGTGCTGAGTCACCGGGCGG
GCCCGGGCGCGGGGCGTGGGACCGAGGCCGCCGCGGGGCTGGGCCTGCGC
GTGGCGGGAGCGCGGGGAGGGATTGCCGCGGGCCGGGGAGGGGCGGGGGC
GGGCGTGCTGCCCTCTGTGGTCCTTGGGCCGCCGCCGCGGGTCTGTCGTG
GTGCCTGGAGCGGCTGTGCTCGTCCCTTGCTTGGCCGTGTTCTCGTTCCT
GAGGGTCCCGCGGACACCGAGTGGCGCAGTGCCAGGCCCAGCCCGGGGAT
GGCGACTGCGCCTGGGCCCGCCTGGTGTCTTCGCATCCCTCTCCGCTTTC
CGGCTTCAGCGCTCTAGGTCAGGGAGTCTTCGCTTTTGTACAGCTCTAAG
GCTAGGAATGGTTTTTATATTTTTAAAAGGCTTTGGAAAACAAAAATACG
CAACAGAGACCGTTTGTGTGACACTTTGCAGGGAAGTTTGCTGGCCTCTG
TTCTAGGTCATGATTGGGCTGCAAGGGCAGAGAAGGTAGCCTTGAACAGA
GGTCCTTTTCCTCCTCCTAAGCTCCGGGAGCCAGAGGTTTAACTGACCCT
```

Genome from the bioinformatic perspective

- physical map

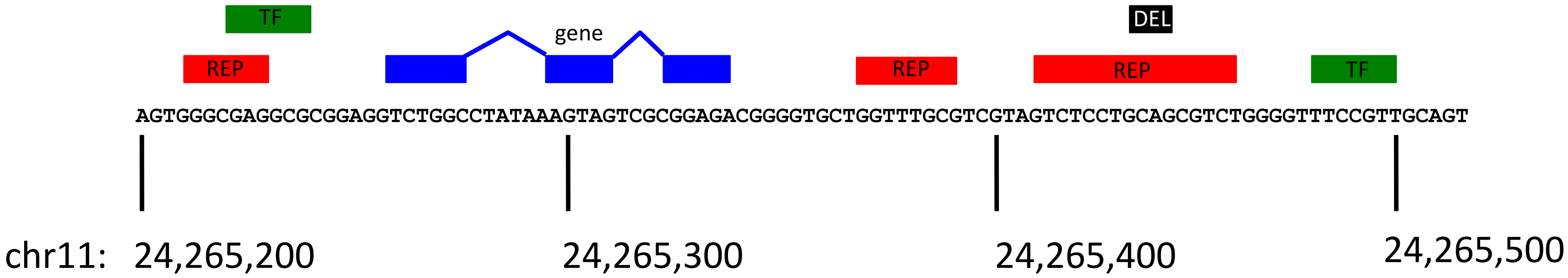


AGTGGGCGAGGCGCGGAGGTCTGGCCTATAAAGTAGTCGCGGAGACGGGGTGCTGGTTTGCGTCGTAGTCTCCTGCAGCGTCTGGGGTTTCCGTTGCAGT

chr11: 22,341,400 22,341,500 22,341,600 22,341,700

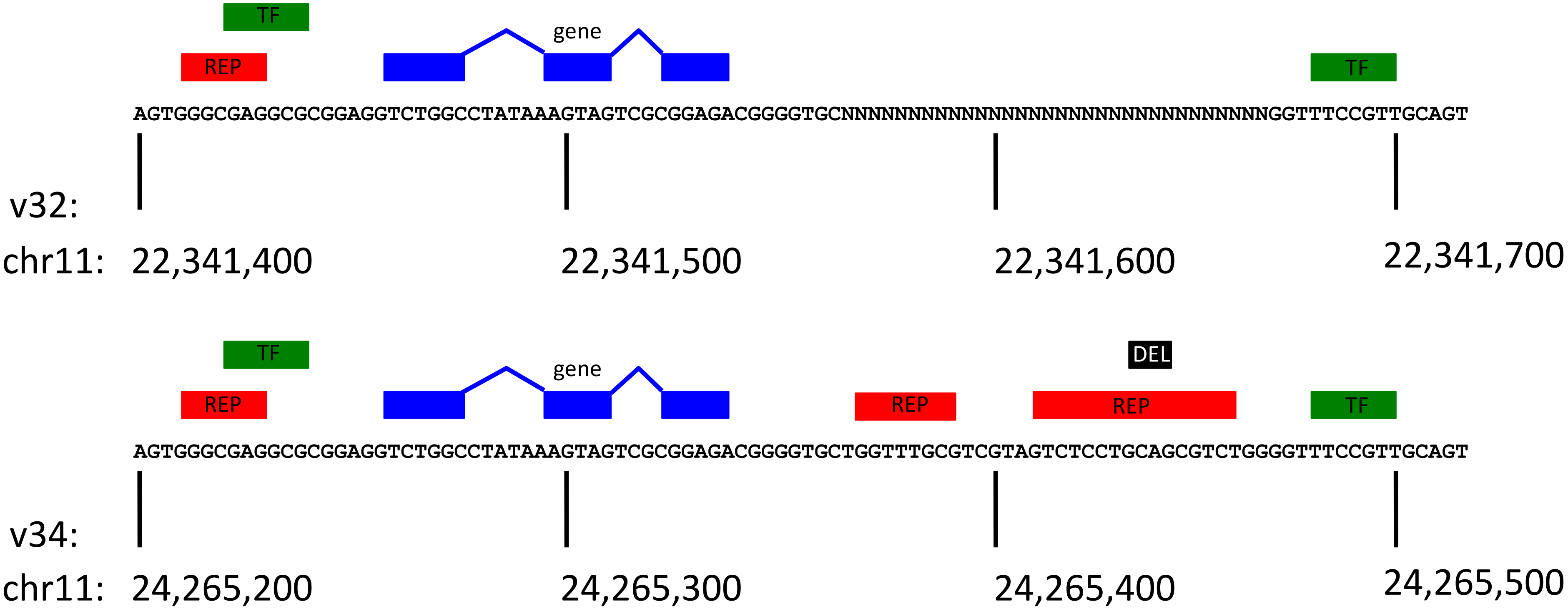
Genome from the bioinformatic perspective

- annotations


TF
REP
gene
DEL
REP
REP
TF
AGTGGGCGAGGCGCGGAGGTCTGGCCTATAAAGTAGTCGCGGAGACGGGGTGCTGGTTTGCGTCGTAGTCTCCTGCAGCGTCTGGGGTTTCCGTTGCAGT
chr11: 24,265,200
24,265,300
24,265,400
24,265,500

Genome from the bioinformatic perspective

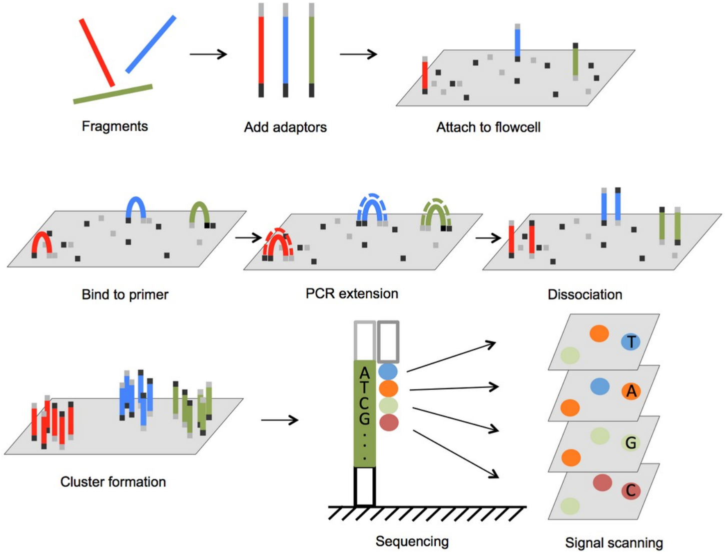
- versioned reference

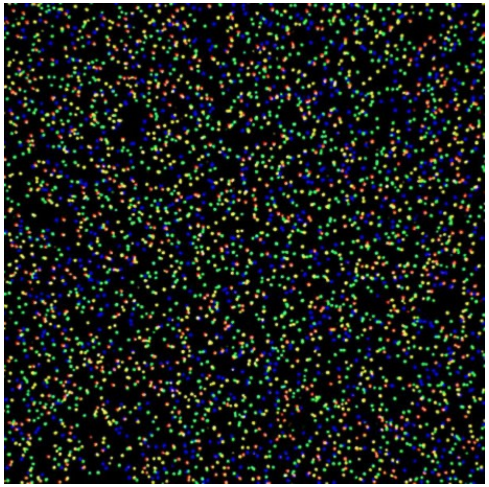

TF
REP
gene
TF
AGTGGGCGAGGCGCGGAGGTCTGGCCTATAAAGTAGTCGCGGAGACGGGGTGCNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNGGTTTCCGTTGCAGT
v32:
chr11: 22,341,400
22,341,500
22,341,600
22,341,700
TF
REP
gene
REP
DEL
REP
TF
AGTGGGCGAGGCGCGGAGGTCTGGCCTATAAAGTAGTCGCGGAGACGGGGTGCTGGTTTGCGTCGTAGTCTCCTGCAGCGTCTGGGGTTTCCGTTGCAGT
v34:
chr11: 24,265,200
24,265,300
24,265,400
24,265,500

Where does the genomic data come from?

Get a sequence

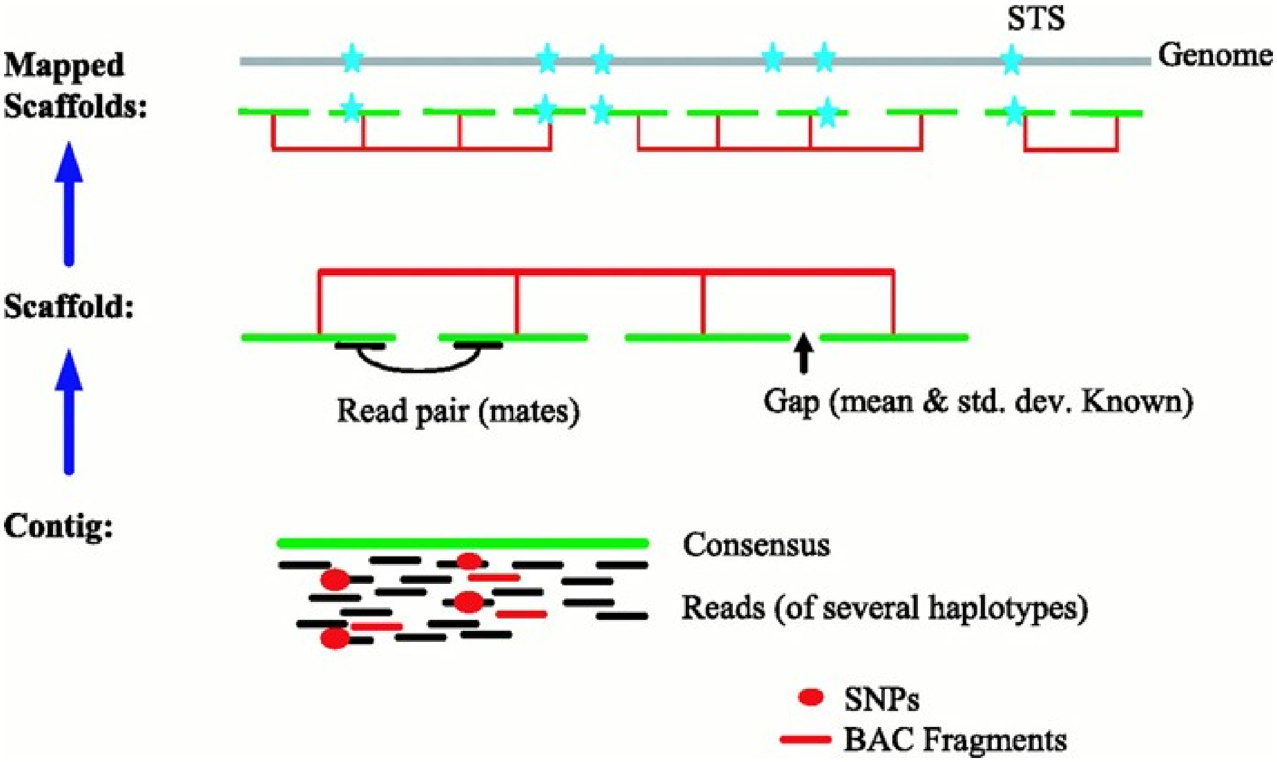
- Various methods:
  - NGS: Illumina, IonTorrent
  - TGS: PacBio, NanoPore
- They all produce short stretches of DNA (reads) of various length (100 bp - 100 kbp)
- Reads can form pairs (i.e. physical distance known between them) which is used for assembly





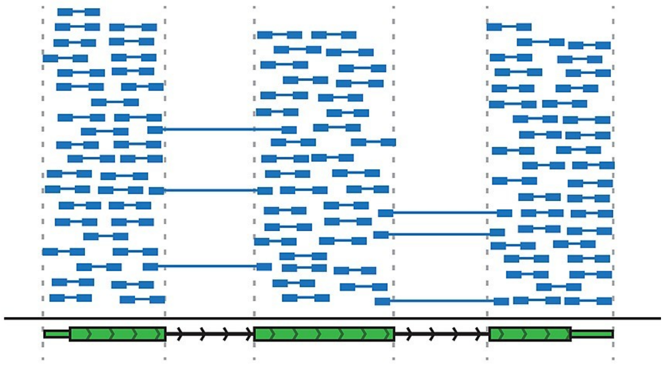
Map the sequence

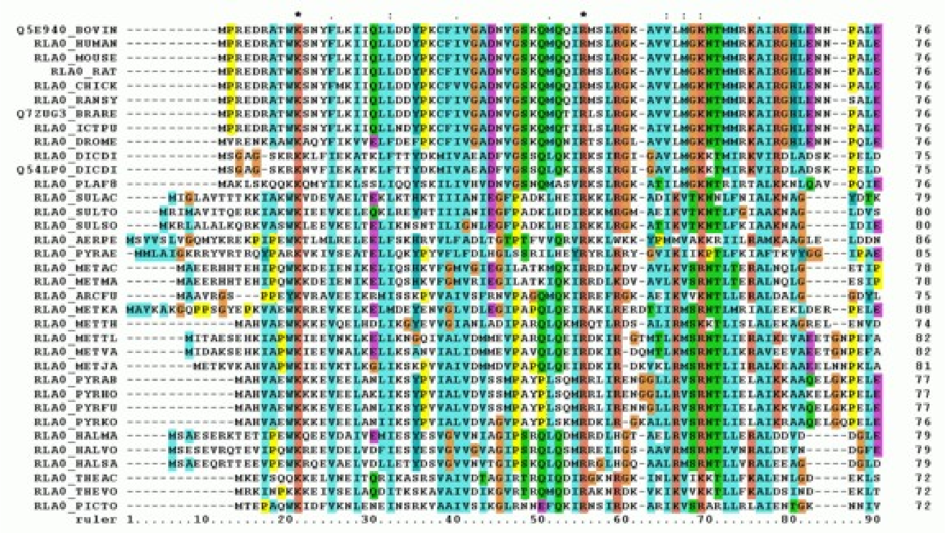
- *Reads* are *assembled* into continuous *contigs*
- *Paired-end reads* help to create a *scaffold* of contings
- Scaffolds are then mapped to *chromosomes*



Sequence annotation

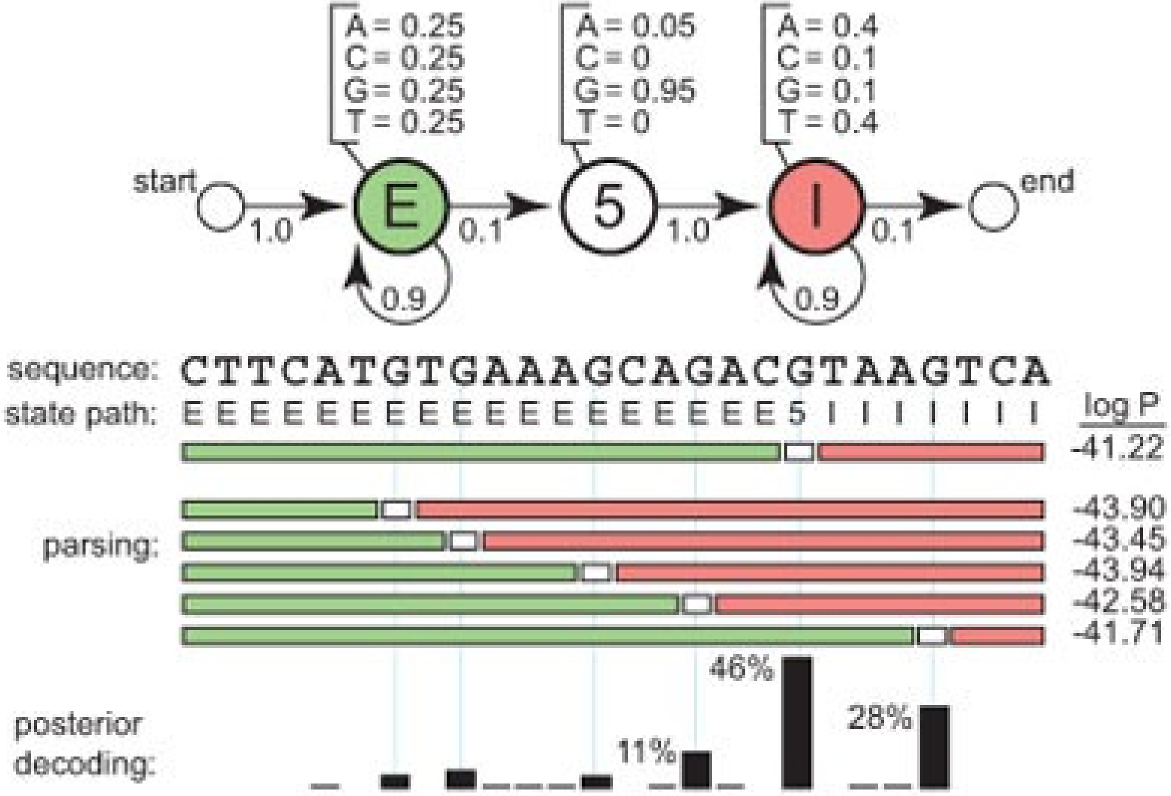
- sequence similarity:
  - to known features (sequence similarity to ESTs, RNA-seq)
  - to homologous features in other organisms (homology – gene/protein families)





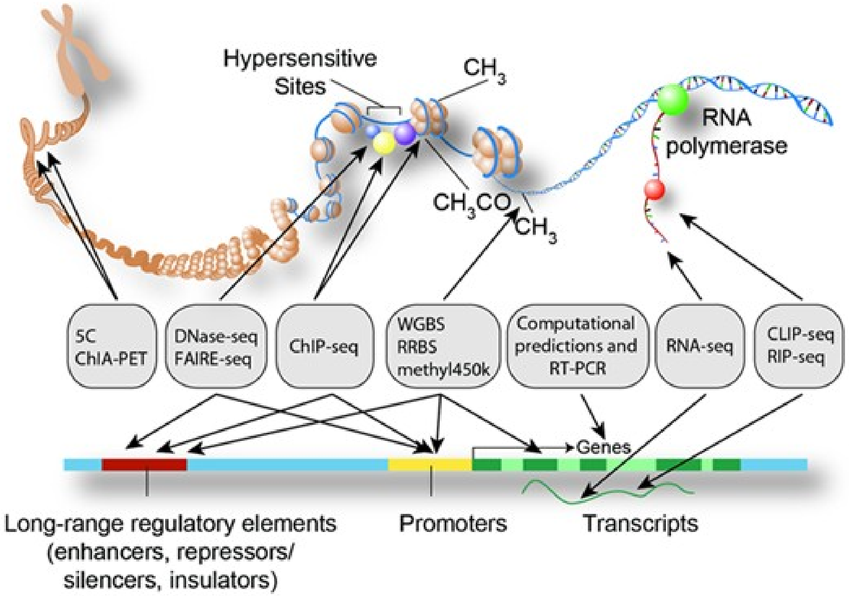
Sequence annotation

- feature prediction using models:
  - using Hidden Markov Models to predict gene structure



Sequence annotation

- Other non-coding functional elements
  - TF binding sites, etc.
  - interspecies sequence conservation
  - ChIP-seq (protein-DNA interaction)
  - DNAseI Hypersensitive Sites (open chromatin sites)



ENCODE

Sequence annotation

- Other features
  - Variation data (SNPs, INDELS)
  - Structural variation data (CNVs)
  - Repeat data (RepeatMasker)
  - Epigenomic data (methylation, histone acetylation)
  - Functional data (Gene Ontology, KEGG, ...)
  - Gene Expression

Where are genomic data stored?

NCBI National Center for Biotechnology Information

e! Ensembl

UCSC Genome Browser

Common genomic data formats

Common genomic data formats

- Regular text files of a specific format
  - easy to open and explore
  - easy to work with
  - .fasta, .fastq, .bed, .gff, .gtf, .vcf, ...
- Binaries
  - more efficient for large datasets
  - fast retrieval by specific tools
  - .2bit, .gz, .bcf

Storing sequences: FASTA

```
>ID_seq|specific_info
AGTGGGCGAGGCGCGGAGGTCTGGCCTATAAAGTAGTCGCGGAGACGGGG
TGCTGGTTTGCGTCGTAGTCTCCTGCAGCGTCTGGGGTTTCCGTTGCAGT
CCTCGGAACCAGGACCTCGGCGTGGCCTAGCGAGTTATGGCGACGAAGGC
CGTGTGCGTGCTGAAGGGCGACGGCCCAGTGCAGGGCATCATCAATTTCG
AGCAGAAGGCAAGGGCTGGACGGAGGCTTGTTTGCGAGGCCGCTCCCAC
CCGCTCGTCCCCCCGCGCACCTTTGCTAGGAGCGGGTCGCCCGCCAGGCC
TCGGGGCCGCCCTGGTCCAGCGCCCGGTCCCGGCCCGTGCCGCCCGGTCG
GTGCCTTCGCCCCCAGCGGTGCGGTGCCCAAGTGCTGAGTCACCGGGCGG
```

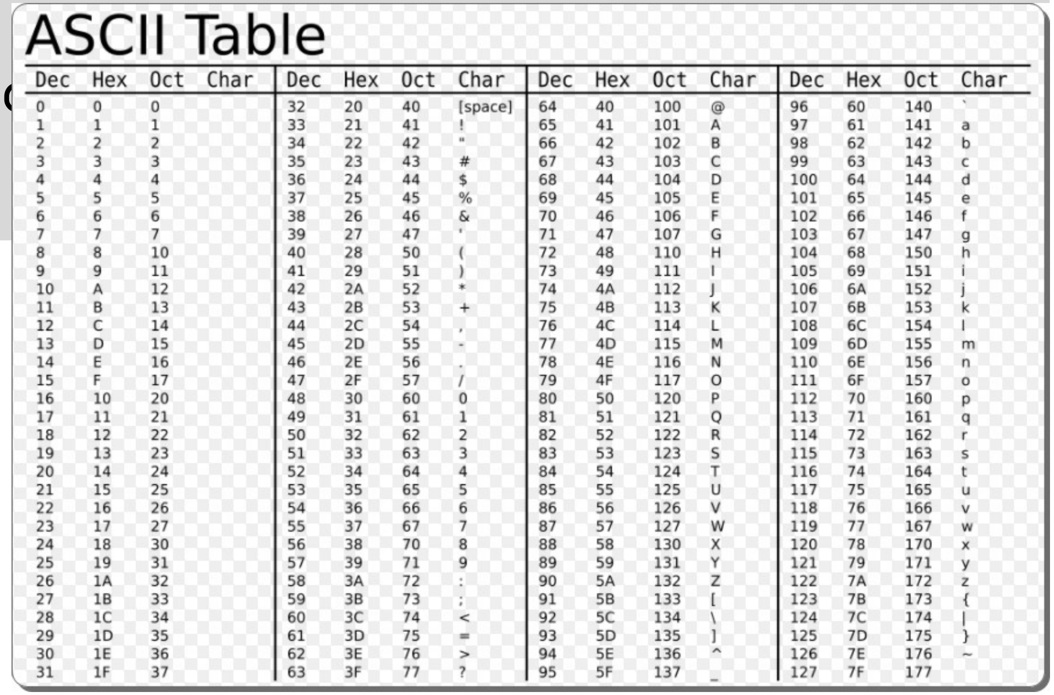
Storing reads: FASTQ

```
@ID_seq1
AGTGGGCGAGGCGCGGAGGTCTGGCCTATAAAGTAGTCGCGGAGACGGGG
+
```

ASCII

```
!''*((((***+))%%%++)(%%%%).1***-+*''))**55CCF>>>>>
@ID_seq2
CCTCGGAACCAGGACCTCGGCGTGGCCTAGCGAGTTATGGCGACGAAGGC
+
')%'* (***+)*''))*%%++5
```

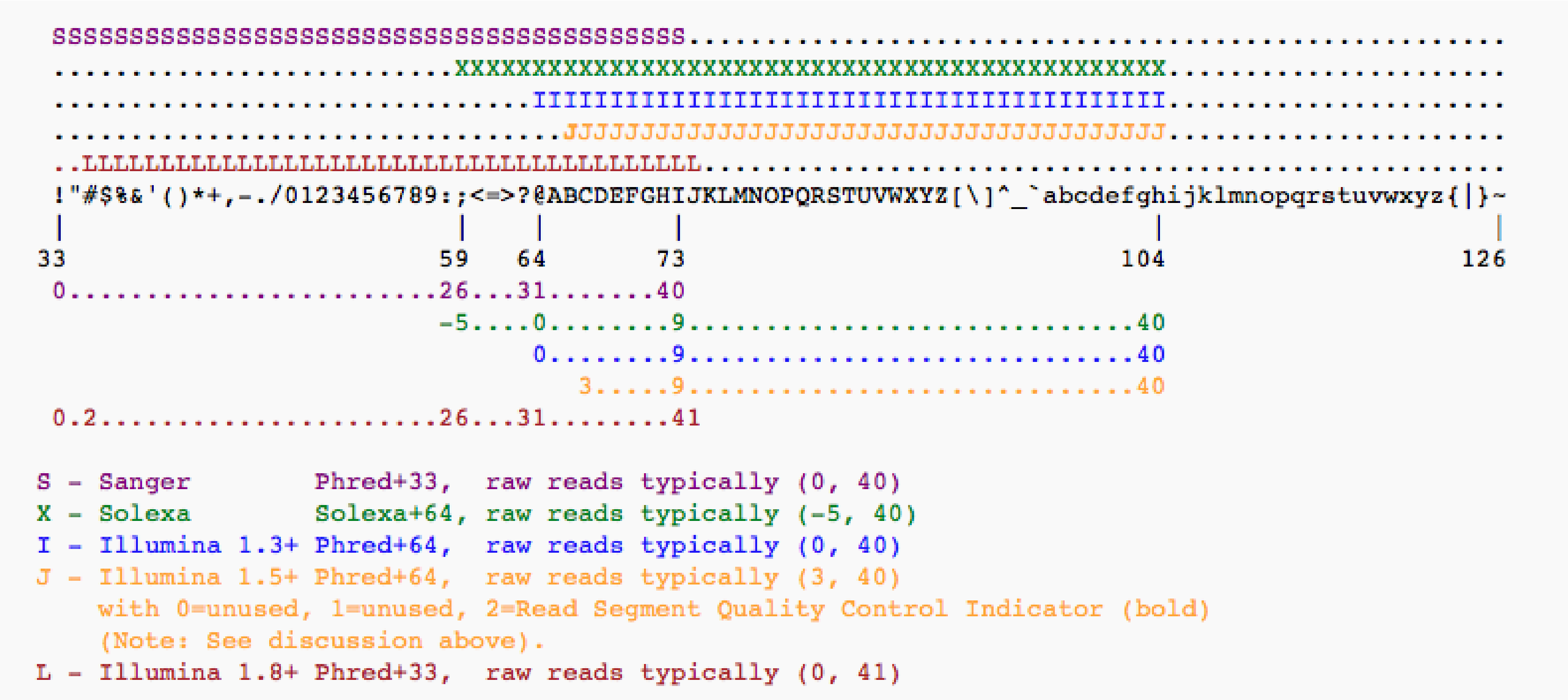
ASCII = American Standard Code for Information Interchange



ASCII Table

Dec	Hex	Oct	Char	Dec	Hex	Oct	Char	Dec	Hex	Oct	Char	Dec	Hex	Oct	Char
0	0	0		32	20	40	[space]	64	40	100	@	96	60	140	`
1	1	1		33	21	41	!	65	41	101	A	97	61	141	a
2	2	2		34	22	42	"	66	42	102	B	98	62	142	b
3	3	3		35	23	43	#	67	43	103	C	99	63	143	c
4	4	4		36	24	44	$	68	44	104	D	100	64	144	d
5	5	5		37	25	45	%	69	45	105	E	101	65	145	e
6	6	6		38	26	46	&	70	46	106	F	102	66	146	f
7	7	7		39	27	47	'	71	47	107	G	103	67	147	g
8	8	10		40	28	50	(	72	48	110	H	104	68	150	h
9	9	11		41	29	51	)	73	49	111	I	105	69	151	i
10	A	12		42	2A	52	*	74	4A	112	J	106	6A	152	j
11	B	13		43	2B	53	+	75	4B	113	K	107	6B	153	k
12	C	14		44	2C	54	,	76	4C	114	L	108	6C	154	l
13	D	15		45	2D	55	-	77	4D	115	M	109	6D	155	m
14	E	16		46	2E	56	.	78	4E	116	N	110	6E	156	n
15	F	17		47	2F	57	/	79	4F	117	O	111	6F	157	o
16	10	20		48	30	60	0	80	50	120	P	112	70	160	p
17	11	21		49	31	61	1	81	51	121	Q	113	71	161	q
18	12	22		50	32	62	2	82	52	122	R	114	72	162	r
19	13	23		51	33	63	3	83	53	123	S	115	73	163	s
20	14	24		52	34	64	4	84	54	124	T	116	74	164	t
21	15	25		53	35	65	5	85	55	125	U	117	75	165	u
22	16	26		54	36	66	6	86	56	126	V	118	76	166	v
23	17	27		55	37	67	7	87	57	127	W	119	77	167	w
24	18	30		56	38	70	8	88	58	130	X	120	78	170	x
25	19	31		57	39	71	9	89	59	131	Y	121	79	171	y
26	1A	32		58	3A	72	:	90	5A	132	Z	122	7A	172	z
27	1B	33		59	3B	73	;	91	5B	133	[	123	7B	173	{
28	1C	34		60	3C	74	<	92	5C	134	\	124	7C	174	|
29	1D	35		61	3D	75	=	93	5D	135	]	125	7D	175	}
30	1E	36		62	3E	76	>	94	5E	136	^	126	7E	176	~
31	1F	37		63	3F	77	?	95	5F	137	_	127	7F	177	

FASTQ: ASCII to PHRED



```
 SSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSS.....................................................
 ..........................XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX......................
 ...............................IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII......................
 .................................JJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJ......................
 ..LLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLL....................................................
 !"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstuvwxyz{|}~
 |                         |    |        |                              |                     |
33                        59   64       73                            104                   126
 0........................26...31.......40
                          -5....0........9.............................40
                                0........9.............................40
                                  3.....9.............................40
 0.2......................26...31........41

S - Sanger        Phred+33,  raw reads typically (0, 40)
X - Solexa        Solexa+64, raw reads typically (-5, 40)
I - Illumina 1.3+ Phred+64,  raw reads typically (0, 40)
J - Illumina 1.5+ Phred+64,  raw reads typically (3, 40)
    with 0=unused, 1=unused, 2=Read Segment Quality Control Indicator (bold)
    (Note: See discussion above).
L - Illumina 1.8+ Phred+33,  raw reads typically (0, 41)
```

PHRED: quality scores

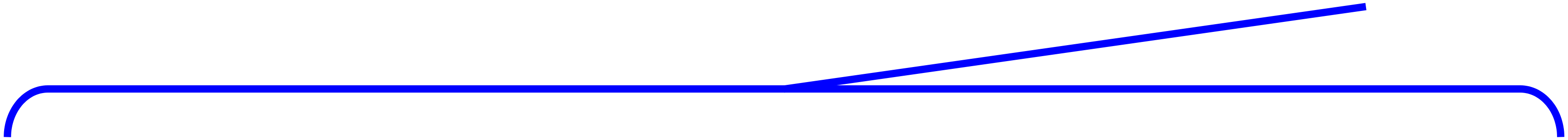
Phred Quality Score	Probability of incorrect base call	Base call accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1000	99.9%
40	1 in 10,000	99.99%

$$\text{Phred} = -10 \log_{10} P$$

Storing annotations: GFF/GTF

- GFF
  - General Feature Format (any kind of annotation/feature)
- GTF
  - Gene Transfer Format (specific form of GFF used to store gene annotation)
- 9 TAB separated fields
- actual content of individual fields depends on the database and type of data

seqname	source	feature	start	end	score	strand	frame	attribute
2	protein_coding	CDS	2419108	2419128	.	+	0	gene_id "ENSG00000223972";
X	protein_coding	CDS	1186934	1440976	.	-	0	gene_id "ENSG00000123546";



```
gene_id "ENSG00000223972"; gene_name "DDX11L1"; gene_source "havana"; gene_biotype "protein_coding";
```

tag "value";

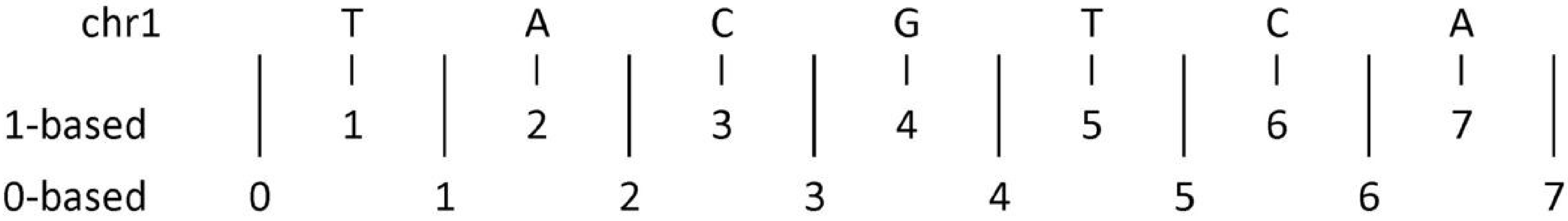
Storing annotations: BED

- 3/4/6/12 columns
- used by UCSC Genome Browser to visualize various features

chrom	chromStart	chromEnd	name	score	strand
2	2419108	2419128	ENSG00000223972	.	+
X	1186934	1440976	ENSG00000123546	.	-

Storing annotations: BED

- 0-based vs. 1-based coordinate system



chr1 T A C G T C A

1-based 1 2 3 4 5 6 7

0-based 0 1 2 3 4 5 6 7

	1-based	0-based
Indicate a single nucleotide	chr1:4-4 G	chr1:3-4 G
Indicate a range of nucleotides	chr1:2-4 ACG	chr1:1-4 ACG
Indicate a single nucleotide variant	chr1:5-5 T/A	chr1:4-5 T/A

Storing variation data: VCF

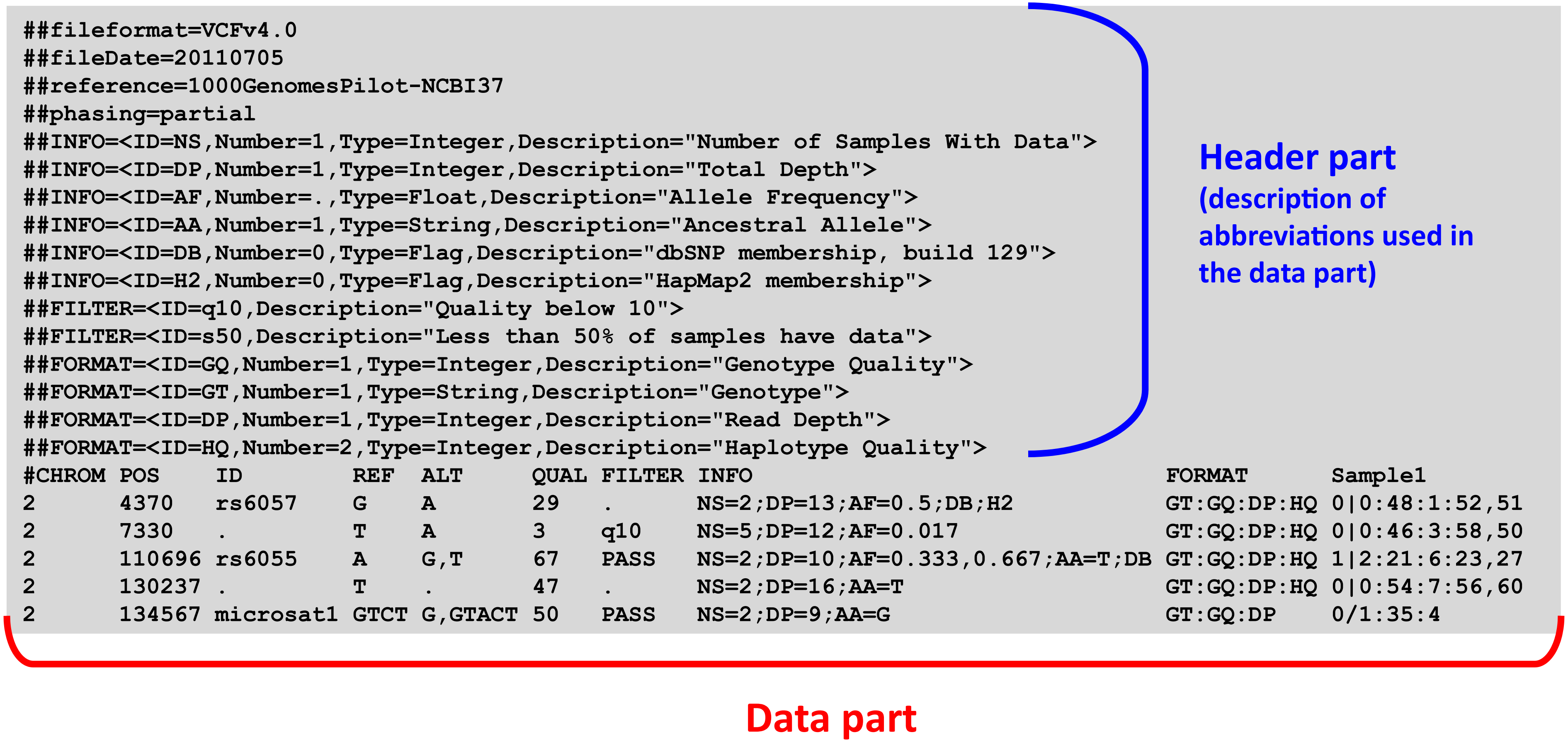
- Variant Call Format

```
##fileformat=VCFv4.0
##fileDate=20110705
##reference=1000GenomesPilot-NCBI37
##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=.,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=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##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      Sample1
2      4370    rs6057     G    A       29   .      NS=2;DP=13;AF=0.5;DB;H2           GT:GQ:DP:HQ 0|0:48:1:52,51
2      7330    .          T    A       3    q10    NS=5;DP=12;AF=0.017               GT:GQ:DP:HQ 0|0:46:3:58,50
2      110696  rs6055     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      130237  .          T    .       47   .      NS=2;DP=16;AA=T                   GT:GQ:DP:HQ 0|0:54:7:56,60
2      134567  microsat1  GTCT G,GTACT 50   PASS   NS=2;DP=9;AA=G                    GT:GQ:DP    0/1:35:4
```

```
< /data-shared/vcf_examples/luscinia_vars_flags.vcf.gz zcat |  less -S
```

Storing variation data: VCF

- Variant Call Format



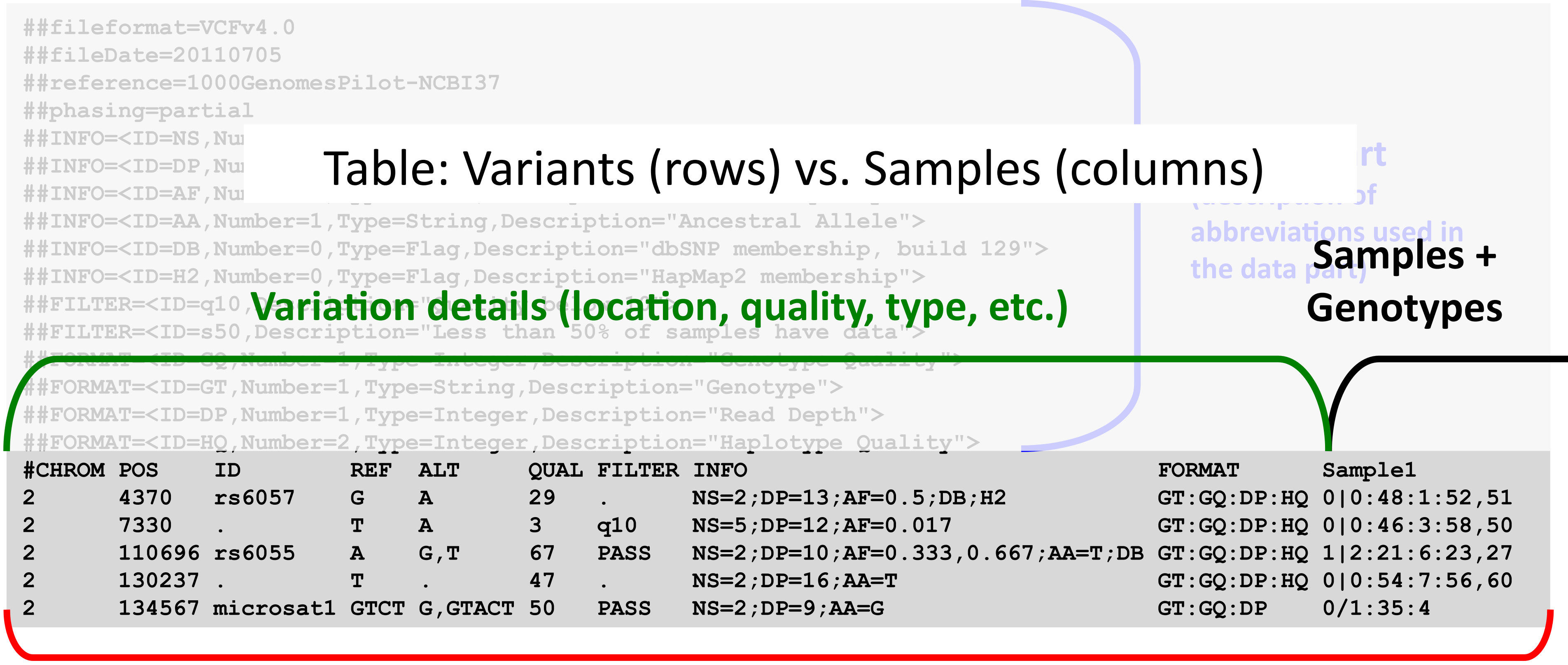
```
##fileformat=VCFv4.0
##fileDate=20110705
##reference=1000GenomesPilot-NCBI37
##phasing=partial
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of Samples With Data">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth">
##INFO=<ID=AF,Number=.,Type=Float,Description="Allele Frequency">
##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=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##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      Sample1
2      4370   rs6057    G    A       29   .      NS=2;DP=13;AF=0.5;DB;H2           GT:GQ:DP:HQ 0|0:48:1:52,51
2      7330   .         T    A       3    q10    NS=5;DP=12;AF=0.017               GT:GQ:DP:HQ 0|0:46:3:58,50
2      110696 rs6055    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      130237 .         T    .       47   .      NS=2;DP=16;AA=T                   GT:GQ:DP:HQ 0|0:54:7:56,60
2      134567 microsat1 GTCT G,GTACT 50   PASS   NS=2;DP=9;AA=G                    GT:GQ:DP    0/1:35:4
```

Header part
(description of abbreviations used in the data part)

Data part

Storing variation data: VCF

- Variant Call Format



```
##fileformat=VCFv4.0
##fileDate=20110705
##reference=1000GenomesPilot-NCBI37
##phasing=partial
##INFO=<ID=NS,Nu
##INFO=<ID=DP,Nu
##INFO=<ID=AF,Nu
##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=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##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      Sample1
2      4370    rs6057     G    A       29   .      NS=2;DP=13;AF=0.5;DB;H2           GT:GQ:DP:HQ 0|0:48:1:52,51
2      7330    .          T    A       3    q10    NS=5;DP=12;AF=0.017               GT:GQ:DP:HQ 0|0:46:3:58,50
2      110696  rs6055     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      130237  .          T    .       47   .      NS=2;DP=16;AA=T                   GT:GQ:DP:HQ 0|0:54:7:56,60
2      134567  microsat1  GTCT G,GTACT 50   PASS   NS=2;DP=9;AA=G                    GT:GQ:DP    0/1:35:4
```

Table: Variants (rows) vs. Samples (columns)

rt
(description of
abbreviations used in
the data part)

Variation details (location, quality, type, etc.)

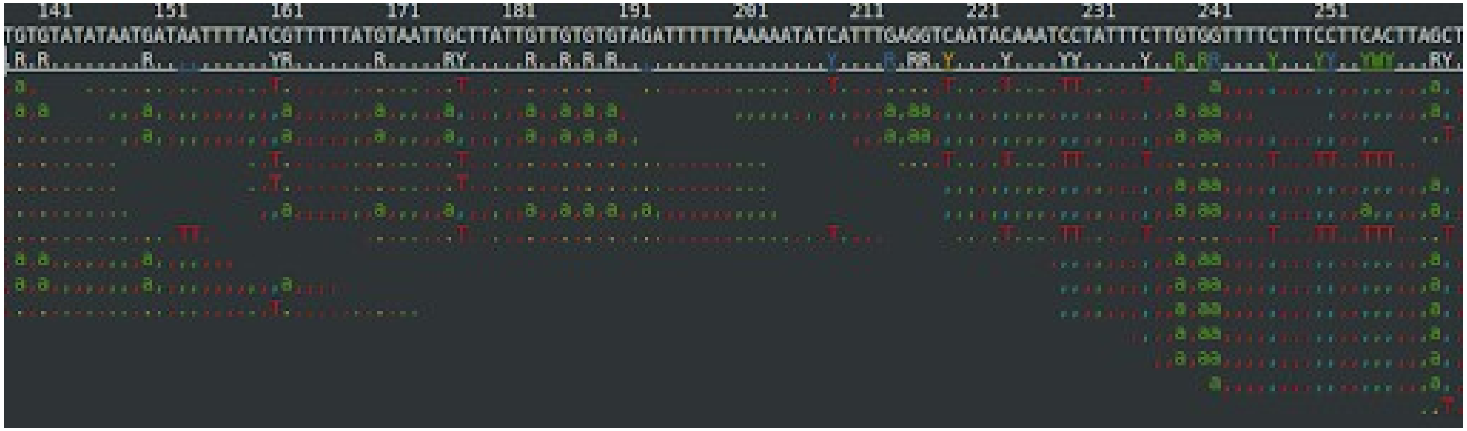
Samples + Genotypes

Data part

Specialized tools for genomic data

samtools

- Working with SAM/BAM files (i.e read alignment data)
- Manipulation with SAM/BAM (sorting, merging, subsetting)
- Summary statistics (read depth by position)
- Viewing read alignment in command line:



```
   141       151       161       171       181       191       201       211       221       231       241       251
TGTGTATATAATGATAATTTTTATCGTTTTTATGTAATTGCTTATTGTTGTGTGTAGATTTTTTAAAAATATCATTTGAGGTCAATACAAATCCTATTTCTTGTGGTTTTCTTTCCTTCACTTAGCT
```

https://samtools.github.io

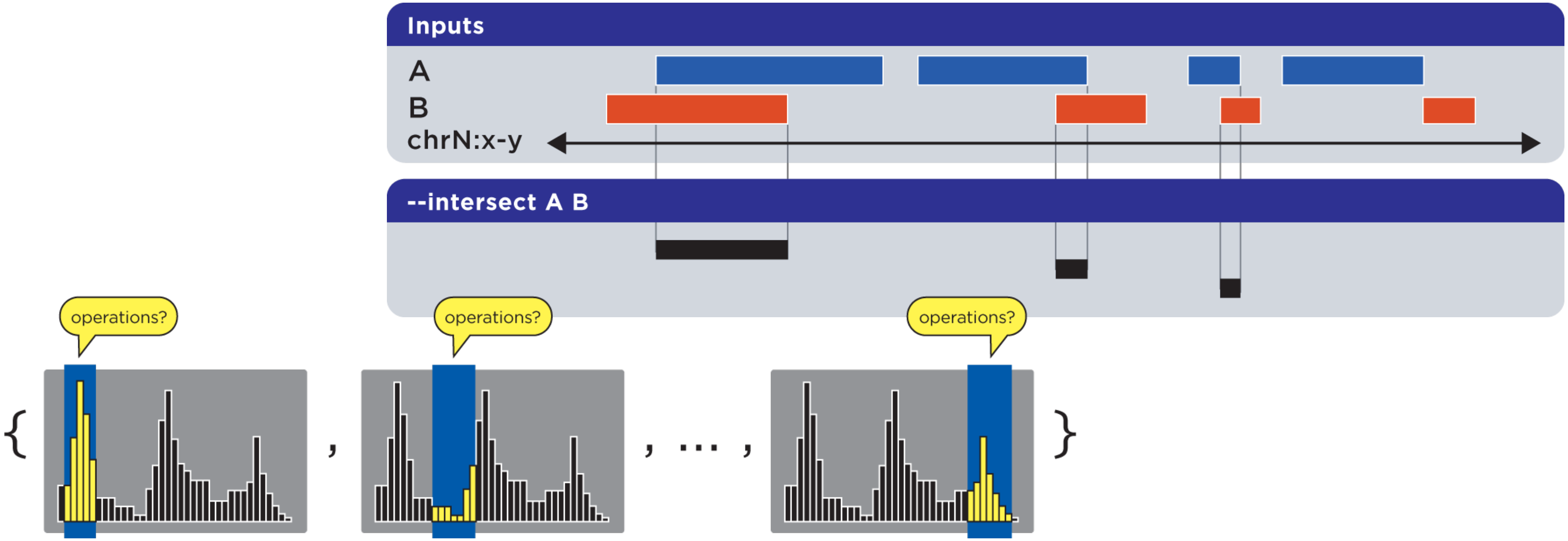
bcftools/vcftools

- variant call files (vcf/bcf)
- bcftools:
  - annotation, concatenation, merging, converting to different formats, filtering based on various criteria, variant calling
- vcftools:
  - mainly filtering/creating subsets
  - population genetics (allele frequency, Hardy-Weinberg, Fst, Pi, Tajima, linkage disequilibrium,...)

https://vcftools.github.io/index.html
https://samtools.github.io/bcftools/

bedtools/bedops

- Operations with genomic data based on their physical position in genome (chromosome, feature start, feature end, strand)
- Usually intersections, overlaps, summary by specific regions (e.g. coverage), sliding window analysis, randomization



https://bedtools.readthedocs.io/en/latest/
https://github.com/bedops/bedops

What did we learned?

- How does genome look from the bioinformatic perspective
- Where does the genomic data come from?
- Common genomic data formats
- Specialized tools for genomic data