

Introduction to Genomics

Libor Mořkovský, Václav Janoušek,

Anastassiya Zidkova, Anna Přistoupilová, Filip Sedlák

<http://ngs-course.readthedocs.org/en/praha-january-2017/>

Genome

The genome is the genetic material of an organism including both the genes and the non-coding sequences.

Bioinformatic perspective

- sequence
- physical map
- annotations
- versioned reference

Bioinformatic perspective

- sequence

AGTGGGCGAGGCGCGGAGGTCTGGCCTATAAAGTAGTCGCGGAGACGGGG
TGCTGGTTTGCCTCGTAGTCTCCTGCAGCGTCTGGGGTTTCCGTTGCAGT
CCTCGGAACCAGGACCTCGGCGTGGCCTAGCGAGTTATGGCGACGAAGGC
CGTGTGCGTGCTGAAGGGCGACGGCCAGTGCAGGGCATCATCAATTTTCG
AGCAGAAGGCAAGGGCTGGGACGGAGGCTTGTTTGCAGGGCCGCTCCAC
CCGCTCGTCCCCCGCGCACCTTTGCTAGGAGCGGGTCGCCCCGCCAGGCC
TCGGGGCCGCCCTGGTCCAGCGCCCGGTCCCGGCCCGTGCCGCCCGGTCTG
GTGCCTTCGCCCCCAGCGGTGCGGTGCCCAAGTGCTGAGTCACCGGGCGG
GCCCCGGGCGCGGGGCGTGCGGACCGAGGCCGCCGCGGGGCTGGGCCTGCGC
GTGGCGGGAGCGCGGGGAGGGATTGCCGCGGGCCGGGGAGGGGCGGGGGC
GGGCGTGCTGCCCTCTGTGGTCCTTGGGCCGCCGCCGCGGGTCTGTCTGTG
GTGCCTGGAGCGGCTGTGCTCGTCCCTTGCTTGGCCGTGTTCTCGTTCCT
GAGGGTCCCGCGGACACCGAGTGGCGCAGTGCCAGGCCAGCCCCGGGGAT
GGCGACTGCGCCTGGGCCCGCCTGGTGTCTTCGCATCCCTCTCCGCTTTC
CGGCTTCAGCGCTCTAGGTCAGGGAGTCTTCGCTTTTGTACAGCTCTAAG
GCTAGGAATGGTTTTTATATTTTAAAAGGCTTTGGAAAACAAAAATACG
CAACAGAGACCGTTTGTGTGACACTTTGCAGGGAAGTTTGCTGGCCTCTG
TTCTAGGTCATGATTGGGCTGCAAGGGCAGAGAAGGTAGCCTTGAACAGA
GGTCCTTTTCCTCCTCCTAAGCTCCGGGAGCCAGAGGTTTAACTGACCCT

Bioinformatic perspective

- physical map

AGTGGGCGAGGCGCGGAGGTCTGGCCTATAAAGTAGTCGCGGAGACGGGGTGCTGGTTTGCGTCGTAGTCTCCTGCAGCGTCTGGGGTTTCCGTTGCAGT

chr11: 22,341,400

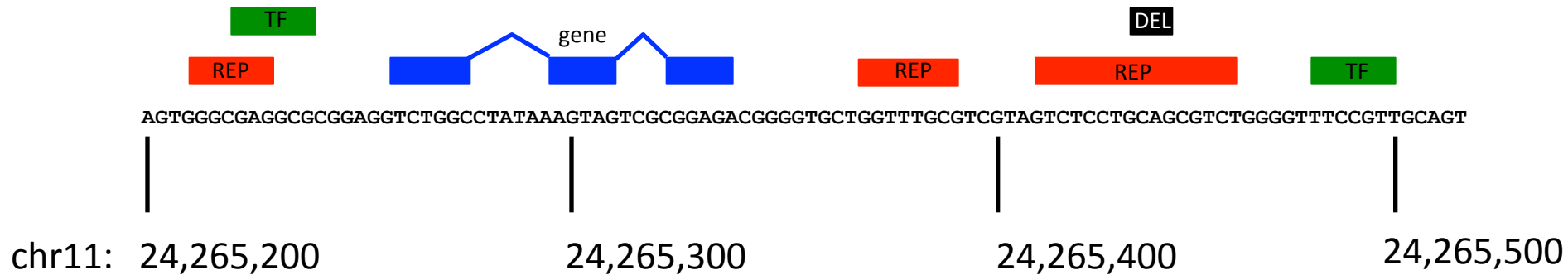
22,341,500

22,341,600

22,341,700

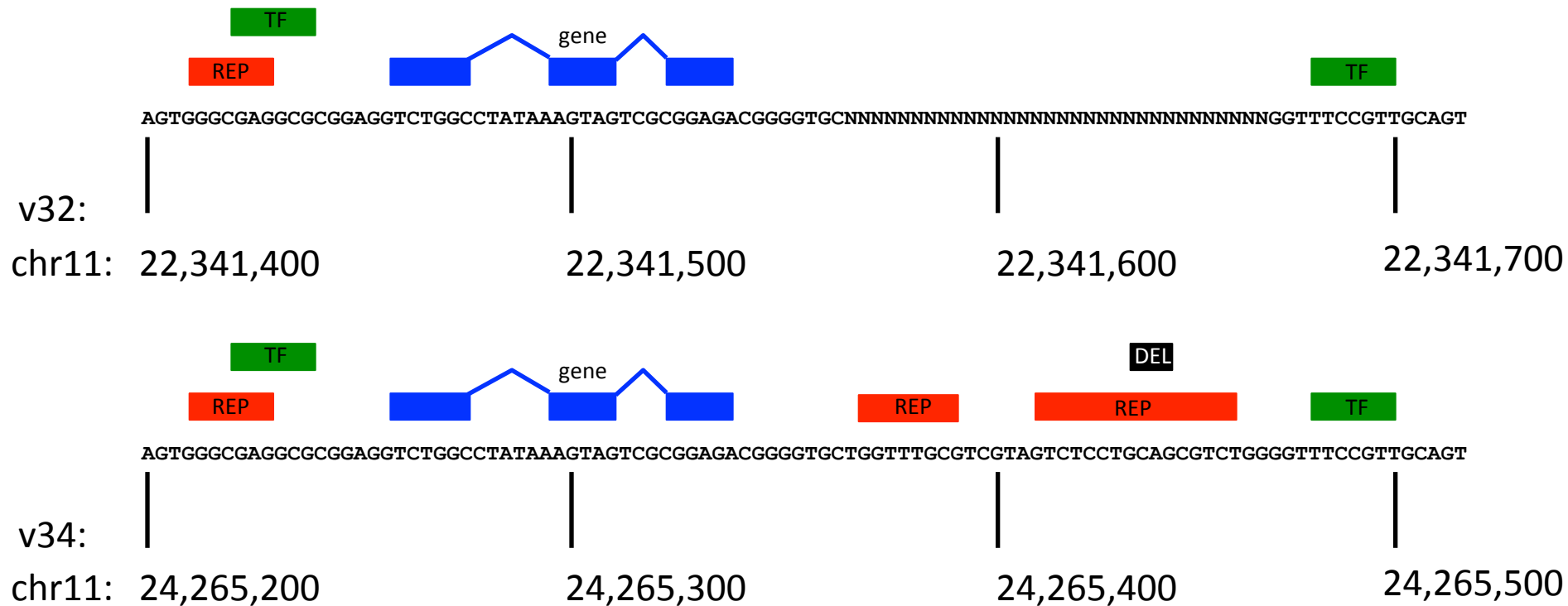
Bioinformatic perspective

- annotations



Bioinformatic perspective

- versioned reference



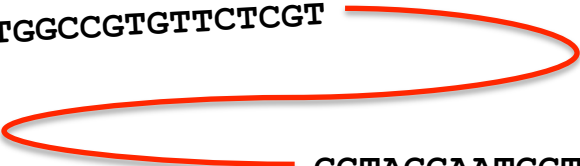
How to get a genome?

- get a sequence
- map the sequence
- annotate the sequence
- refine the sequence

Get a sequence

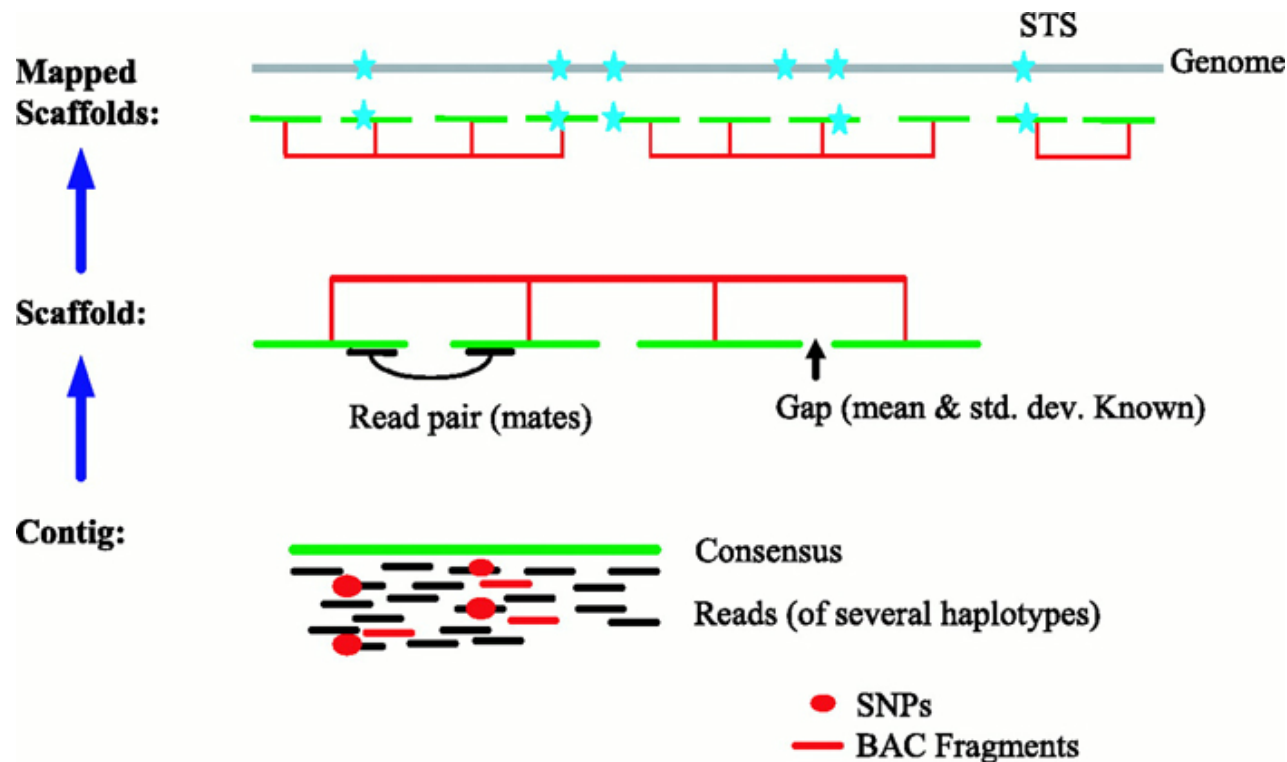
- Old ways (Sanger) or new ways (NGS)...
... *all produce "reads"*
... *or "pairs of reads" ...*

CGTGGGACCGAGGCCGCCGCGGGGCTGGGCCT GGCGACGGCCCAGTGCAGGGCATCATC
GGCGACGGCCCAGTGCAGGGCATCATC
CTGGTGTCTTCGCATCCCTCTCCGCTTTC
TGCAAGGGCAGAGAAGGTAGCCTTGAACAGA TGCAAGGGCAGAGAAGGTAGCCTTGAACAGA
GCTGTGCTCGTCCCTTGCTTGGCCGTGTTCTCGT GCTAGGAATGGTTTTTATATTTTAAAAGGC



Map the sequence

- Reads are assembled into continuous contigs
- Pair reads help to create a scaffold of contigs
- Scaffolds are then mapped to chromosomes

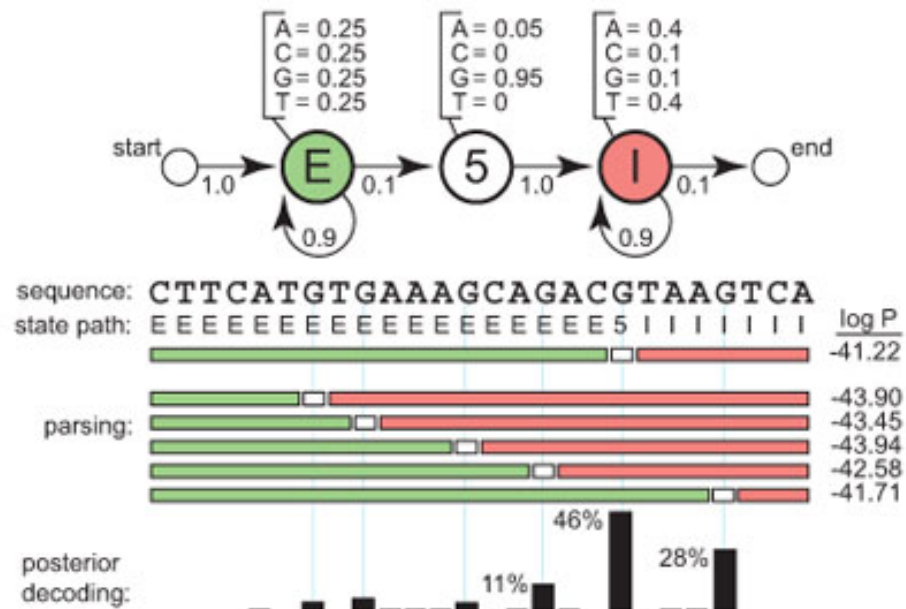


Annotate the sequence

- Annotation approaches
 - sequence similarity
 - to known features
 - to homologous features in other organisms
 - feature prediction using models

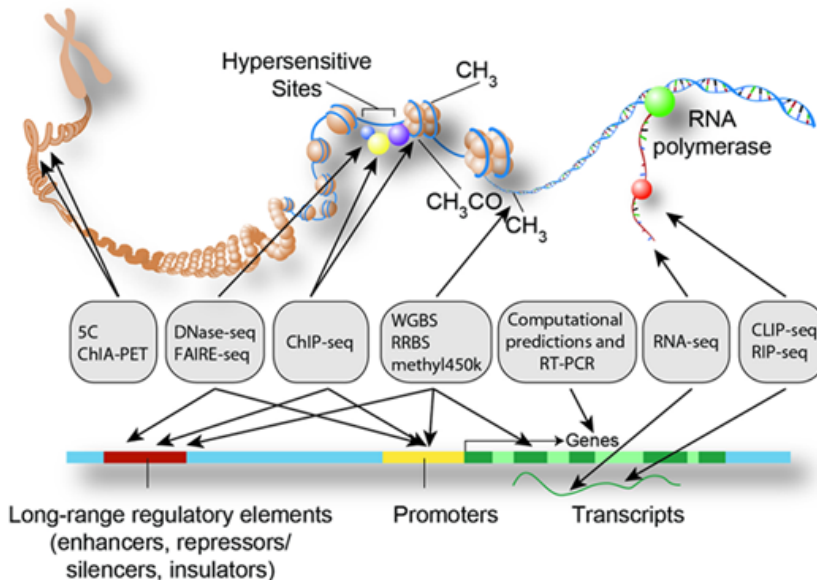
Annotate the sequence

- Gene prediction
 - sequence similarity to ESTs, RNA-seq
 - homology – gene/protein families
 - using Hidden Markov Models to predict gene structure



Annotate the sequence

- Other non-coding functional elements
 - TF binding sites, etc.
 - interspecies sequence conservation
 - ChIP-seq, DNaseI Hypersensitive Sites, etc.



Annotate the sequence

- 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

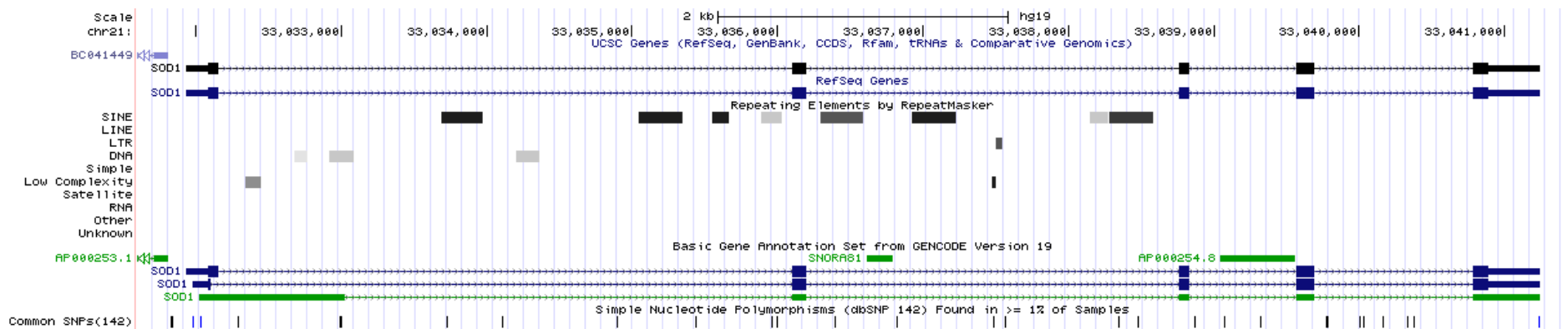
Where to find genomic data?



UCSC Genome Bioinformatics

Where to find genomic data?

UCSC Genome Bioinformatics



The way the genomic data are stored

- 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  
TGCTGGTTTTCGTCGTAGTCTCCTGCAGCGTCTGGGGTTTCCGTTGCAGT  
CCTCGGAACCAGGACCTCGGCGTGGCCTAGCGAGTTATGGCGACGAAGGC  
CGTGTGCGTGCTGAAGGGCGACGGCCCAGTGCAGGGCATCATCAATTTTCG  
AGCAGAAGGCAAGGGCTGGGACGGAGGCTTGTTTTCGAGAGCCGCTCCAC  
CCGCTCGTCCCCCGCGCACCTTTGCTAGGAGCGGGTCGCCCCGCCAGGCC  
TCGGGGCCGCCCTGGTCCAGCGCCCCGGTCCCGGCCCGTGCCGCCCGGTTCG  
GTGCCTTCGCCCCCAGCGGTGCGGTGCCCAAGTGCTGAGTCACCGGGCGG
```

Storing reads: FASTQ

```
@ID_seq1
```

```
AGTGGGCGAGGCGCGGAGGTCTGGCCTATAAAGTAGTCGCGGAGACGGGG
```

```
+
```

```
! ' ' * ( ( ( ( * * * + ) ) % % % + + ) ( % % % % ) . 1 * * * - + * ' ' ) ) * * 5 5 C C F > > > >
```

```
@ID_seq2
```

```
CCTCGGAACCAGGACCTCGGCGTGGCCTAGCGAGTTATGGCGACGAAGGC
```

```
+
```

```
' ) % ' * ( * * * + ) * ' ' ) ) * % % + + 5 C ) ( % % % ( ! ( ( % ) . 1 * * * - + * 5 C F > > > >
```

FASTQ: quality scores (PHRED)

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%

FASTQ: PHRED to ASCII

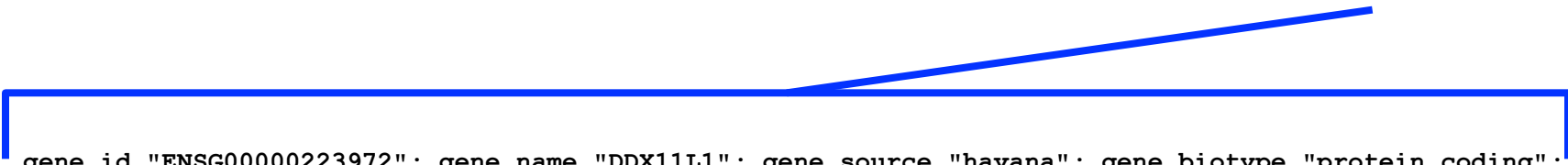


S - Sanger Phred+33, raw reads typically (0, 40)
X - Solexa Solexa+64, raw reads typically (-5, 40)
I - Illumina 1.3+ Phred+64, raw reads typically (0, 40)
J - Illumina 1.5+ Phred+64, raw reads typically (3, 40)
with 0=unused, 1=unused, 2=Read Segment Quality Control Indicator (bold)
(Note: See discussion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

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";
```

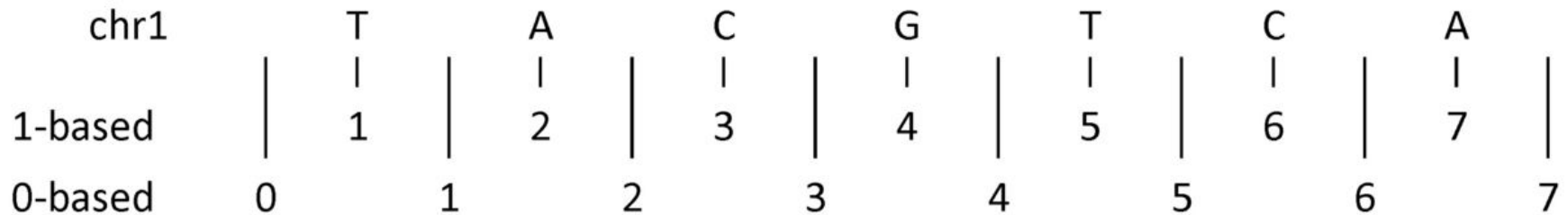
Storing annotations: BED

- 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



	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
```

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
2 4370 rs6057 G A 29 . NS=2;DP=13;AF=0.5;DB;H2
2 7330 . T A 3 q10 NS=5;DP=12;AF=0.017
2 110696 rs6055 A G,T 67 PASS NS=2;DP=10;AF=0.333,0.667;AA=T;DB
2 130237 . T . 47 . NS=2;DP=16;AA=T
2 134567 microsat1 GTCT G,GTACT 50 PASS NS=2;DP=9;AA=G
```

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

FORMAT	Sample1
GT:GQ:DP:HQ	0 0:48:1:52,51
GT:GQ:DP:HQ	0 0:46:3:58,50
GT:GQ:DP:HQ	1 2:21:6:23,27
GT:GQ:DP:HQ	0 0:54:7:56,60
GT:GQ:DP	0/1:35:4

Data part

Storing variation data: VCF

- Variant Call Format

```
##fileformat=VCFv4.0
##fileDate=20110705
##reference=1000GenomesPilot-NCBI37
##phasing=partial
##INFO=<ID=NS,Num
```

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

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

Samples +
Genotypes

```
##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,Number=1,Type=Flag,Description="Quality score < 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 part

Storing annotations: VCF

- Explore VCF file

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