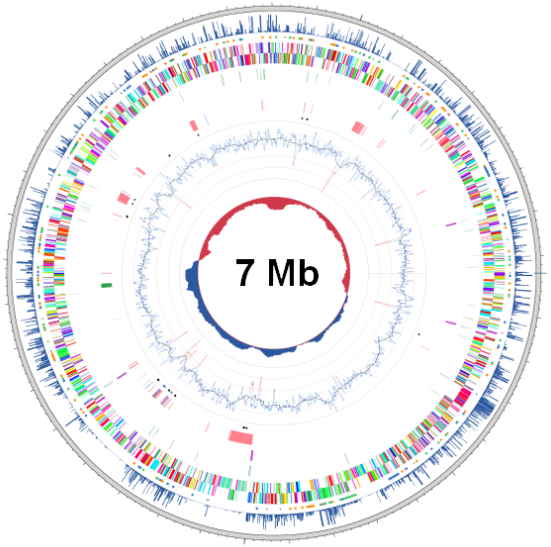
Задачи геномики ("ДНК биоинформатика")

Ярослав Баранов
МНЛ«Компьютерные технологии», Университет ИТМО

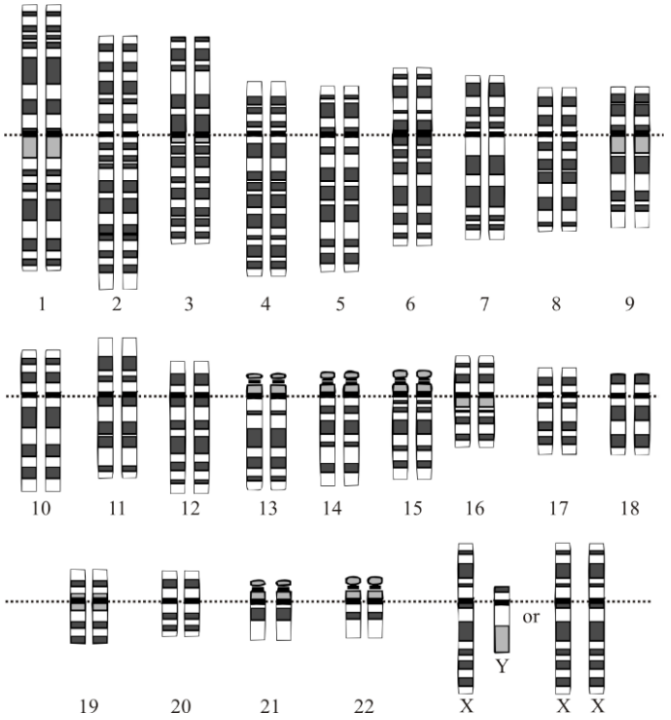
Геномика

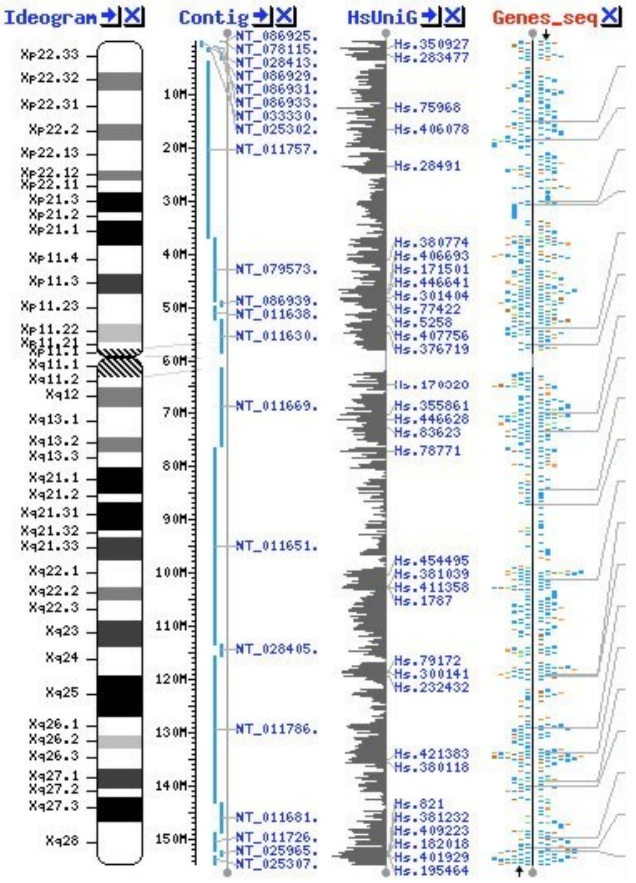
- Area of genetics that concerns the sequencing and analysis of an organism's genetic information
- DNA sequencing + bioinformatics **=>** sequence, assemble and analyze the function and structure of genomes (the complete set of DNA within a single cell of an organism)

Bacterial genome

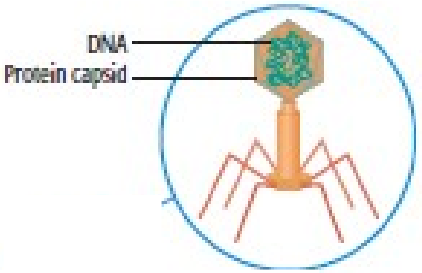


Human genome



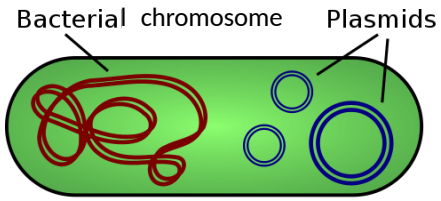


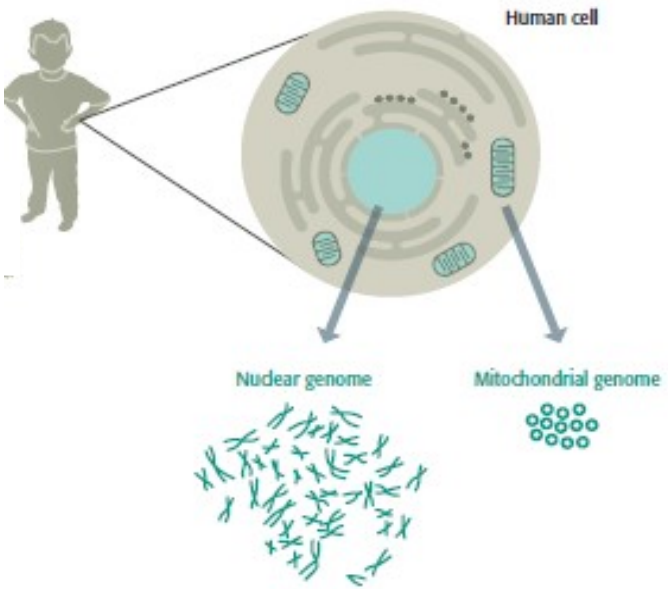
Секвенирование и сборка геномов De novo



viral

prokaryotic





eukaryotic

2000 – draft human genome sequence
2003 – completed (kind of)

3300 books of 1000 pages with 1000 bp per page

Ensembl genomes:

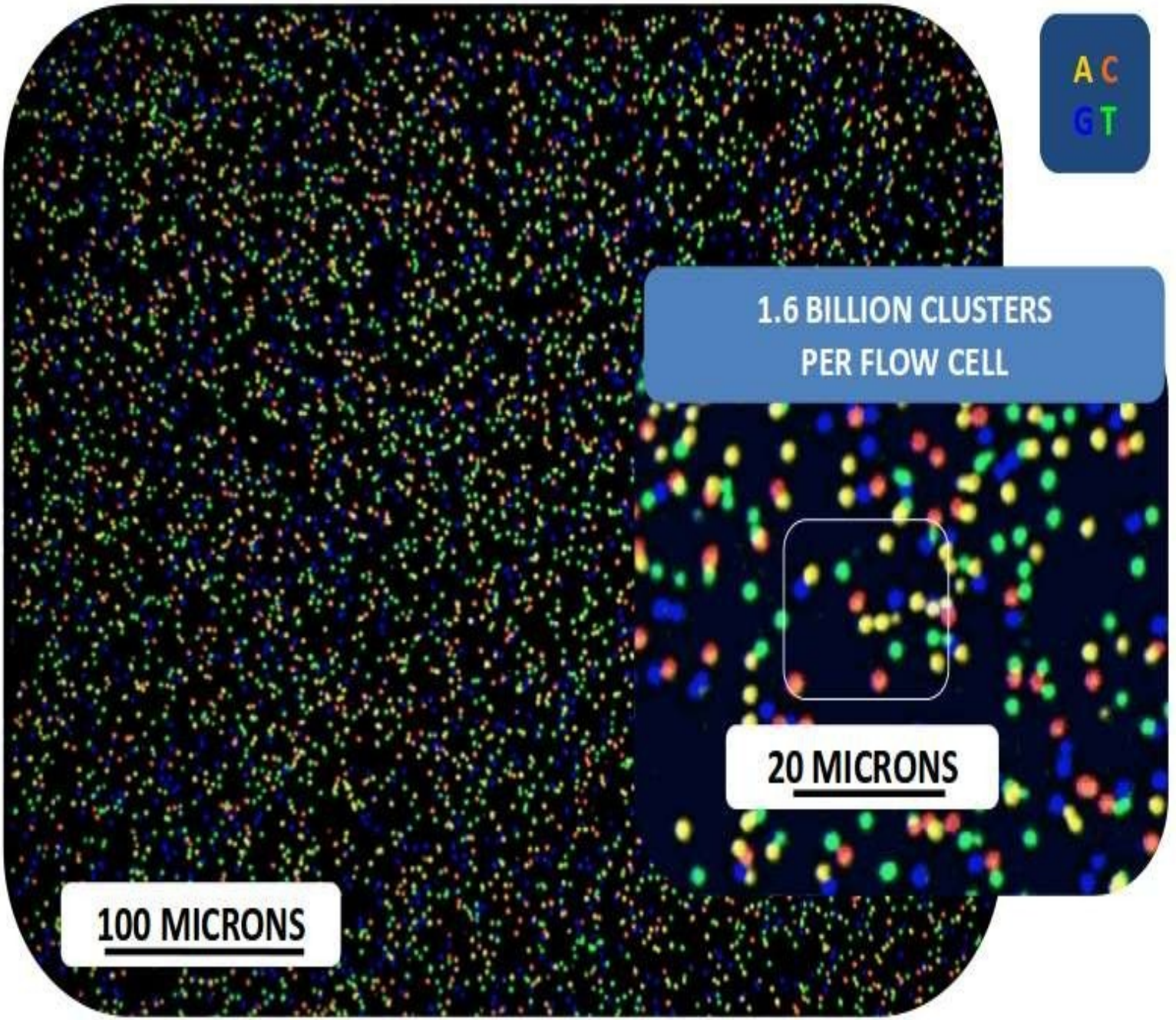
- 69 высшие животные + модельные организмы
- 55 насекомых
- 39 растений
- 563 грибов
- Более 200 протистов
- Более 20 000 бактерий

+ регулярные пополнения

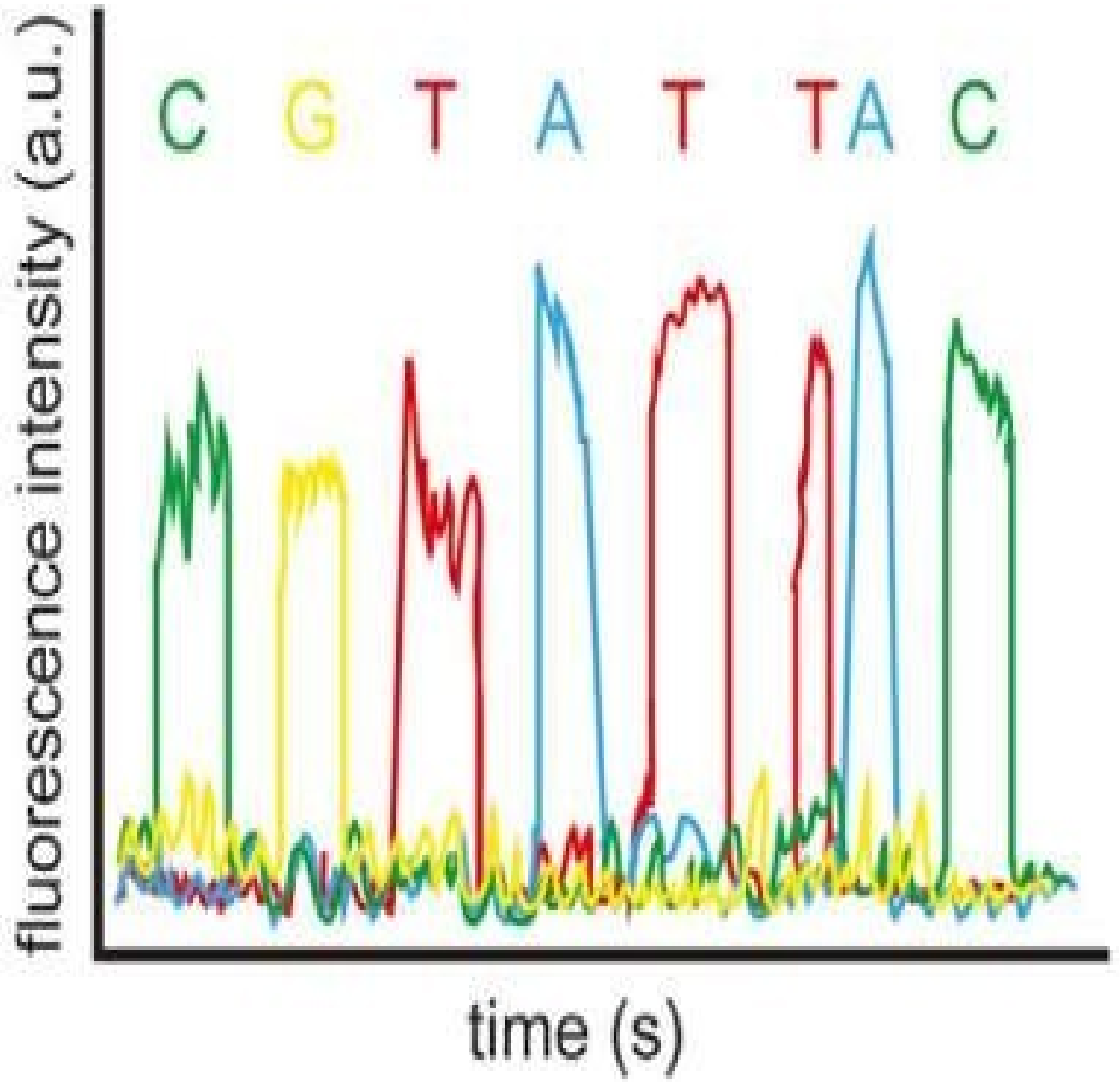
Секвенаторы

- Sanger — First developed in 1986
- Illumina Genome Analyzer (HiSeq/MiSeq/NextSeq)
- Pacific Biosciences (PACBIO RSII)
- Oxford Nanopore

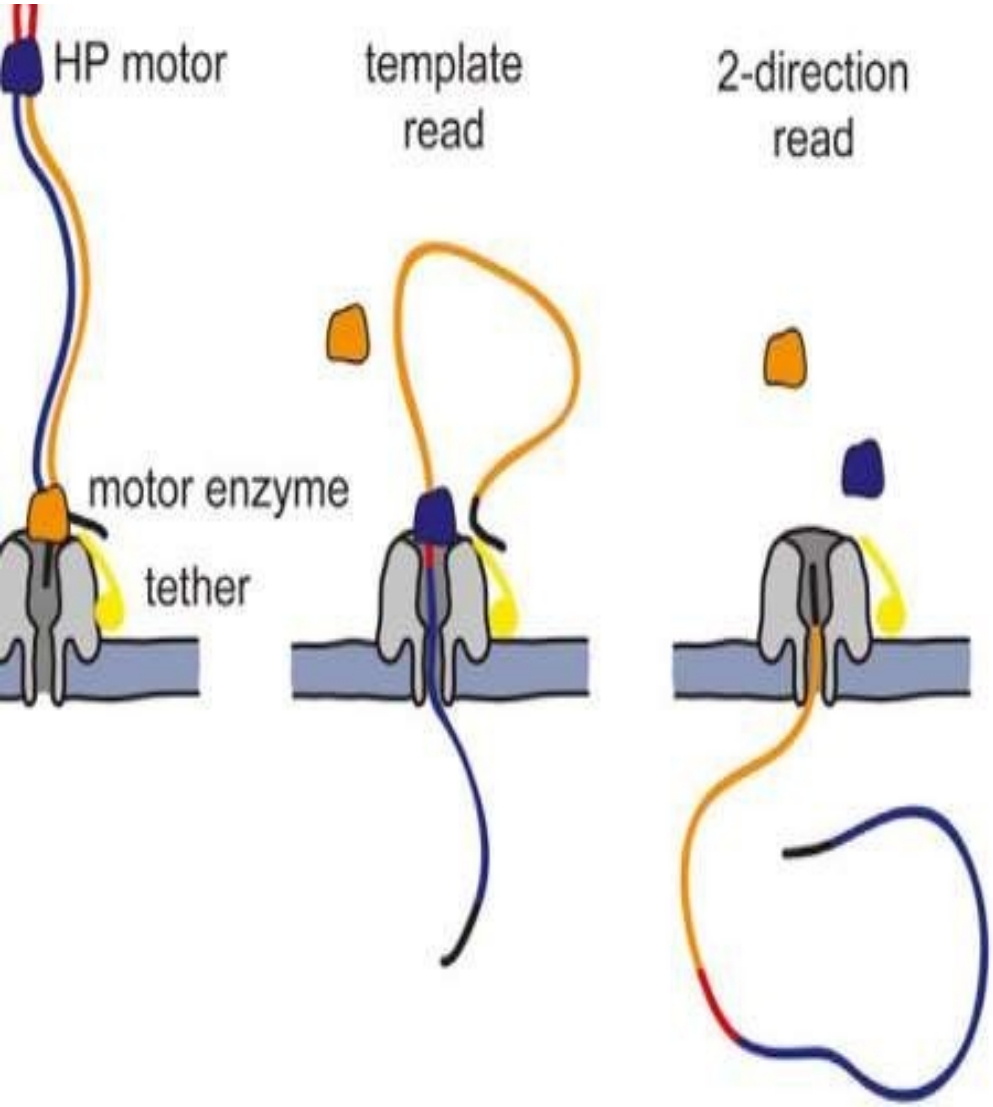
Исходный сигнал (Illumina)

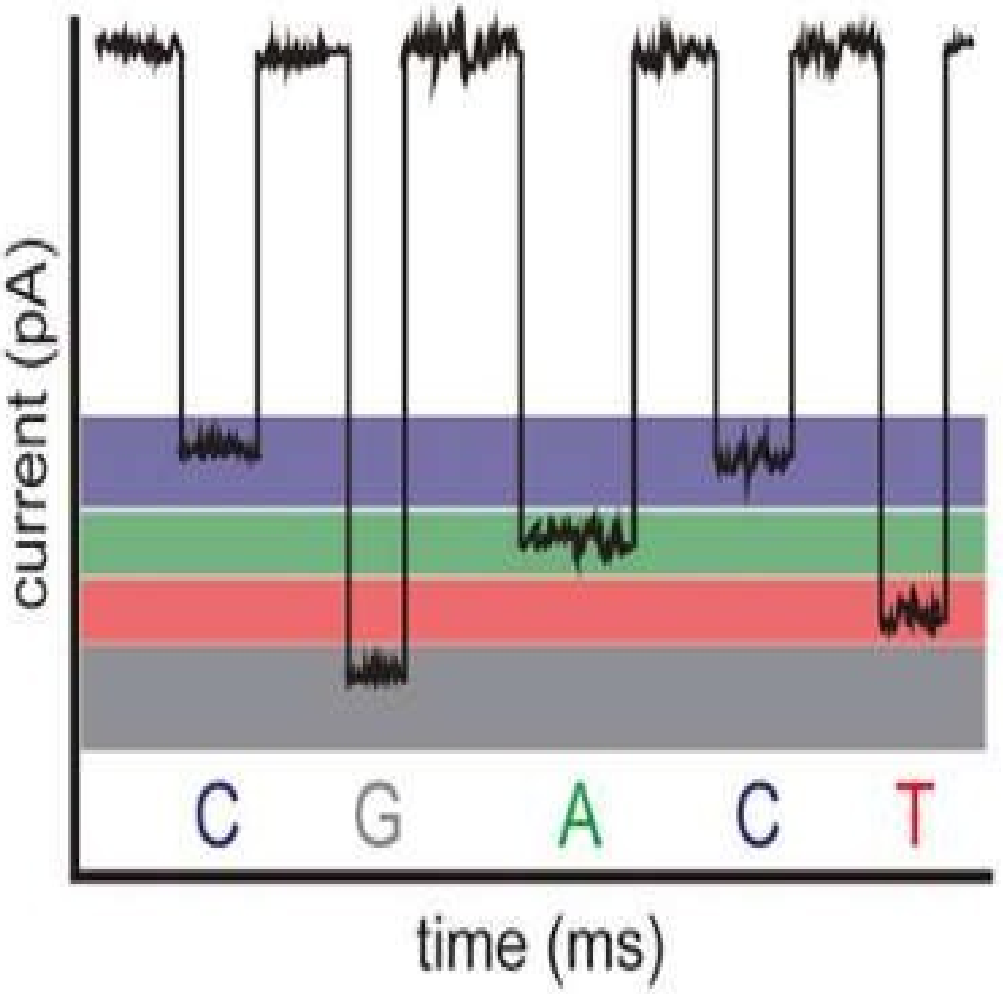


Исходный сигнал (Pacific bioisciences)



Исходный сигнал (Oxford Nanopore)





Геном

Физические методы (Сонификация)

Genomic Fragments
(200 b or 400 b or 1kb)

Сжатие и хранение данных

~ 10^8 записей

или

~ 10-30 Гбайт

Зачем секвенировать геномы людей?

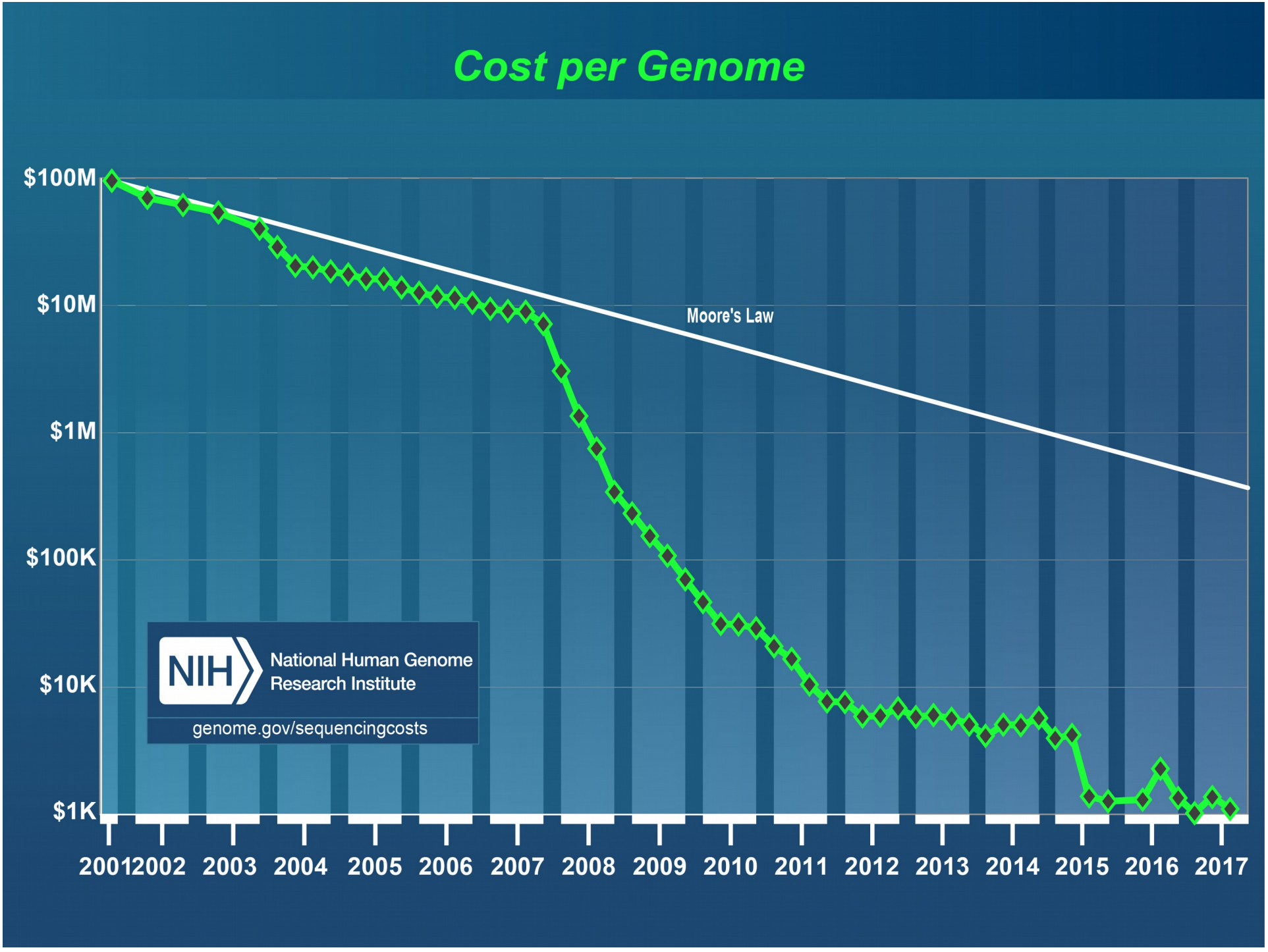
- Диагностика заболеваний (в т.ч. пренатальная)
- Подбор индивидуального лечения
- Оценка риска развития заболеваний в будущем
- Оценка предрасположенностей
- Оценка риска развития заболеваний у детей

Зачем секвенировать другие геномы?

- Подбор лечения для конкретного варианта бактерии или вируса

- Более «осмысленная» селекция и биотехнология сельско-хозяйственных организмов

Стоимость секвенирования генома человека




PHASE TWO • INTERPRETATION
SHERMAN Hollow Ledger
I THINK I FOUND A CORNER PIECE.
3 BILLION PIECES
GENOME

Формат FASTQ

@SRR566546.970 HWUSI-EAS1673_11067_FC7070M:4:1:2299:1109 length=50
TTGCCTGCCTATCATTTTAGTGCCTGTGAGGTGGAGATGTGAGGATCAGT
+SRR566546.970 HWUSI-EAS1673_11067_FC7070M:4:1:2299:1109 length=50
hhhhhhhhhhghhghhhhhfhhhhhfffffe`ee[`X]b[d[ed`[Y[^Y

@SRR566546.971 HWUSI-EAS1673_11067_FC7070M:4:1:2374:1108 length=50
GATTTGTATGAAAGTATACAACTAAAACTGCAGGTGGATCAGAGTAAGTC
+SRR566546.971 HWUSI-EAS1673_11067_FC7070M:4:1:2374:1108 length=50
hhhhgfhhcghghggfcffdhfehhhhcehdchhdhahehffffde`bVd

@.....

@.....

left-to-right increasing order of quality (ASCII):
!"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstuvwxyz{|}~

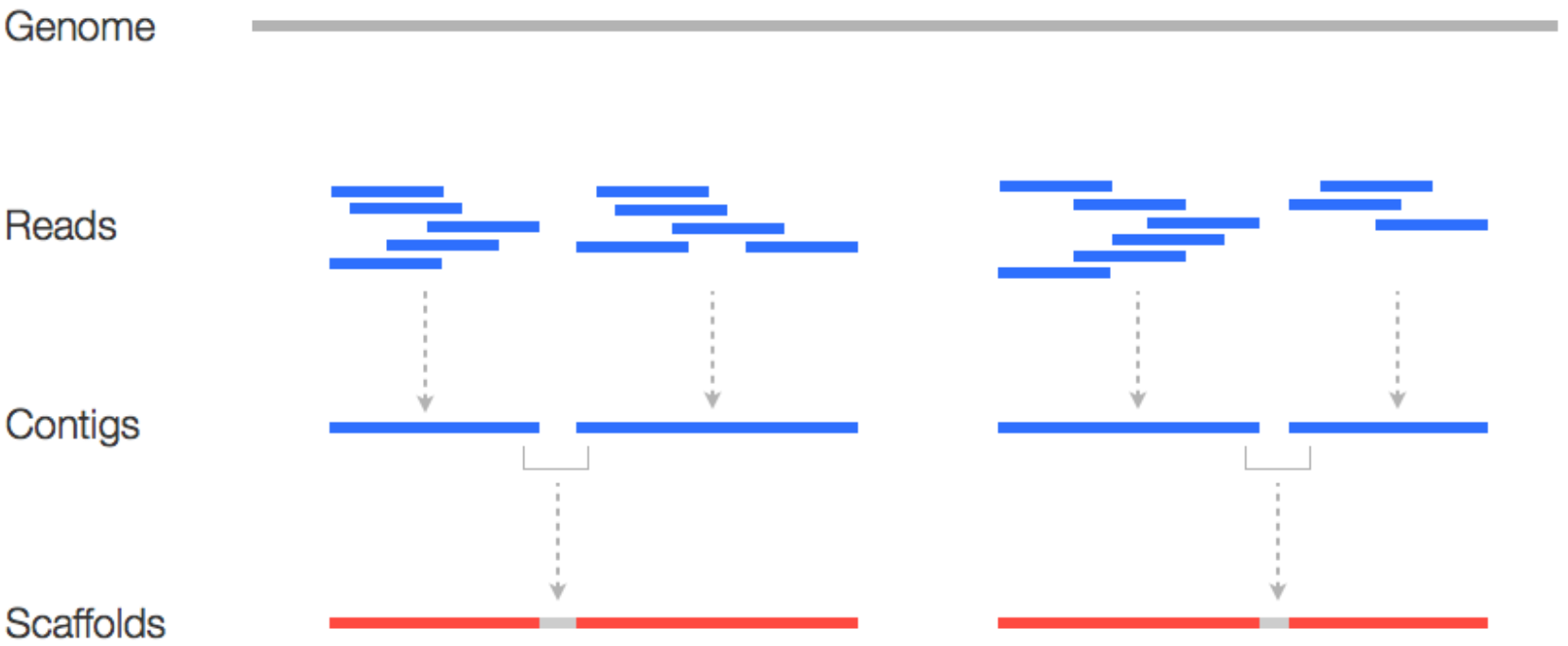
Формат FASTA

;LCBO - Prolactin precursor - Bovine
; a sample sequence in FASTA format
MDSKGSSQKGSRLLLLLVVSNLLLCQGVVSTPVCPNGPGNCQVSLRDLFDRAVMVSHYIHDLSS
EMFNEFDKRYAQGKGFITMALNSCHTSSLPTPEDKEQAQQTHHEVLMSLILGLLRSWNDPLYHL
VTEVRGMKGAPDAILSRAIEIEEENKRLLEGMEMIFGQVIPGAKETEPYPVWSGLPSLQTKDED
ARYSAFYNLLHCLRRDSSKIDTYLKLLNCRIIYNNNC*

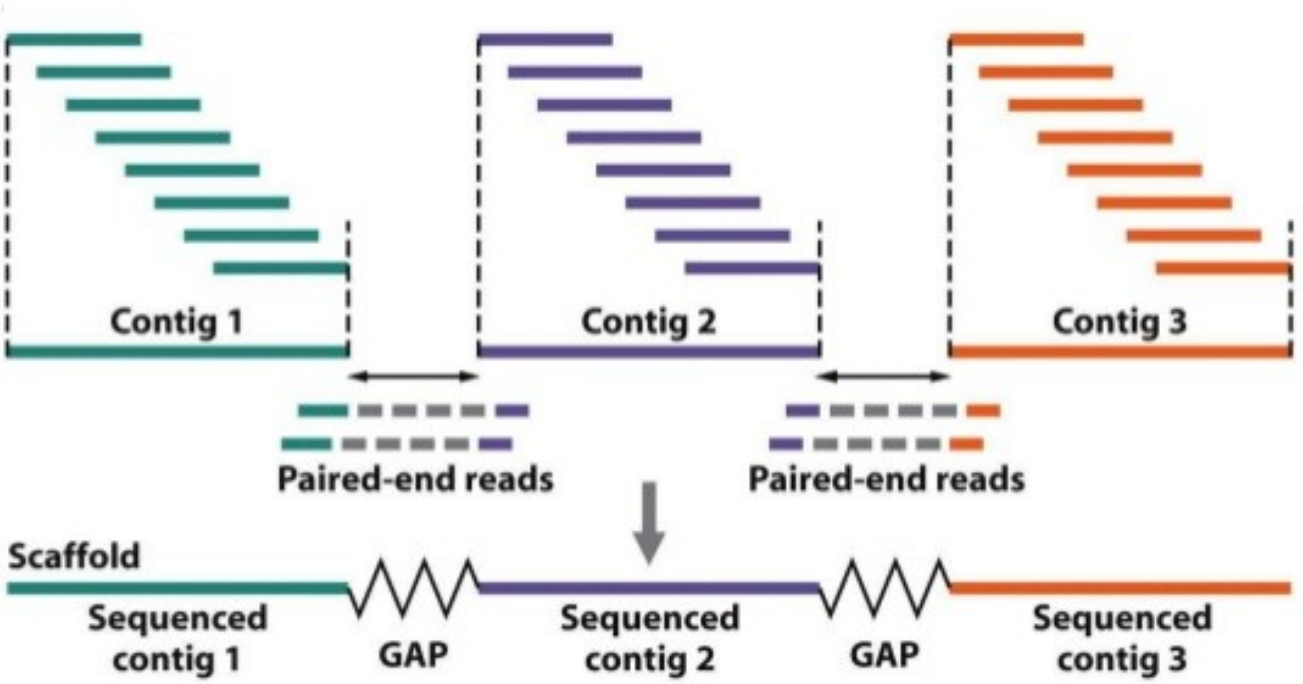
>MCHU - Calmodulin - Human, rabbit, bovine, rat, and chicken
ADQLTEEQIAEFKEAFSLFDKDGDGTITTKELGTVMRSLGQNPTEAELQDMINEVDADGNGTID
FPEFLTMMARKMKDTDSEEEIREAFRVFDKDGNGYISAAELRHVMTNLGEKLTDEEVDEMIREA
DIDGDGQVNYEEFVQMMTAK*

>gi|5524211|gb|AAD44166.1| cytochrome b [Elephas maximus maximus]
LCLYTHIGRNIYYGSYLYSETWNTGIMLLLITMATAFMGYVLPWGQMSFWGATVITNLFSAIPYIGTNLV
EWIWGGFSVDKATLNRFFAFHFILPFTMVALAGVHLTFLHETGSNNPLGLTSDSDKIPFHPYYTIKDFLG
LLILILLLLLLALLSPDMLGDPDNHMPADPLNTPLHIKPEWYFLFAYAILRSVPNKLGGVLALFLSIVIL
GLMPFLHTSKHRSMMLRPLSQALFWTLTMDLLTLTWIGSQPVEYPYTIIGQMASILYFSIILAFLPIAGX
IENY

Genome assembly



- Read - continuous sequence produced by sequencer
- Coverage - the number of short reads that overlap each other within a specific genomic region (how many times the particular base or region is read)
- Contig - set of overlapping segments (reads) of DNA sequences forming continuous consensus sequence
- Scaffold - set of linked non-contiguous series of genomic sequences, consisting of contigs separated by gaps of roughly known length


Contig 1
Contig 2
Contig 3
Paired-end reads
Paired-end reads
Scaffold
Sequenced contig 1
GAP
Sequenced contig 2
GAP
Sequenced contig 3

Сборка генома De novo

CTGCATCGACTAC

CGACTACGACTAG

ACGCCGCTGCA

CGGACTGACTG

TGCATCGACTA

GCATCGGACTG

ACTAGCGAGCT

GCGACGCCG

ACGACTAGCGAGCT

TGACTGCATCGA

AAGCTGCGA

GCCGCTGCATC

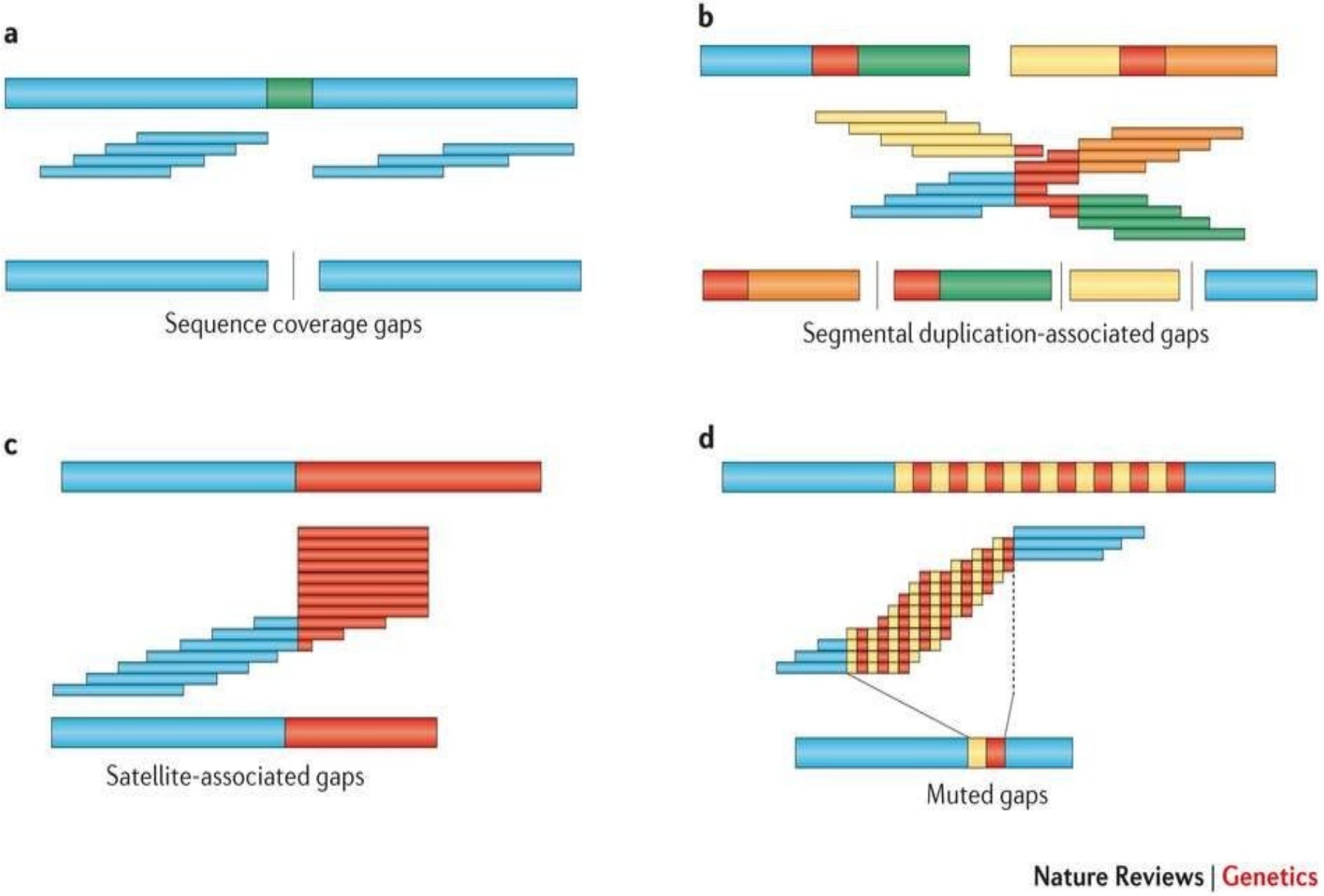
AAGCTGCGACGCCGCTGCATCGGACTGACTGCATCGACTACGACTAGCGAGCT

Сборка генома De novo

CTGCATCGACTAC
ACGCCGCTGCA CGGACTGACTG CGACTACGACTAG GCGACGCCG
GCATCGGACTG TGCATCGACTA ACTAGCGAGCT
AAGCTGCGA GCCGCTGCATC TGACTGCATCGA ACGACTAGCGAGCT

AAGCTGCGACGCCGCTGCATCGGACTGACTGCATCGACTACGACTAGCGAGCT

Трудности сборки De novo



Chaisson et al., 2015

Ресеквенирование

CTGCATCGGCTAC

CGGACTGACTG CGGCTACGACTAG

ACGCCGGCA

GCGACGCCG GCATCGGACTG TGCATCGGCTA ACTAGCGAGCT

AAGCTGCGA GCCGGCATC TGACTGCATCGG ACGACTAGCGAGCT

AAGCTGCGACGCCGCTGCATCGGACTGACTGCATCGACTACGACTAGCGAGCT

Ресеквенирование

CTGCATCGGCTAC

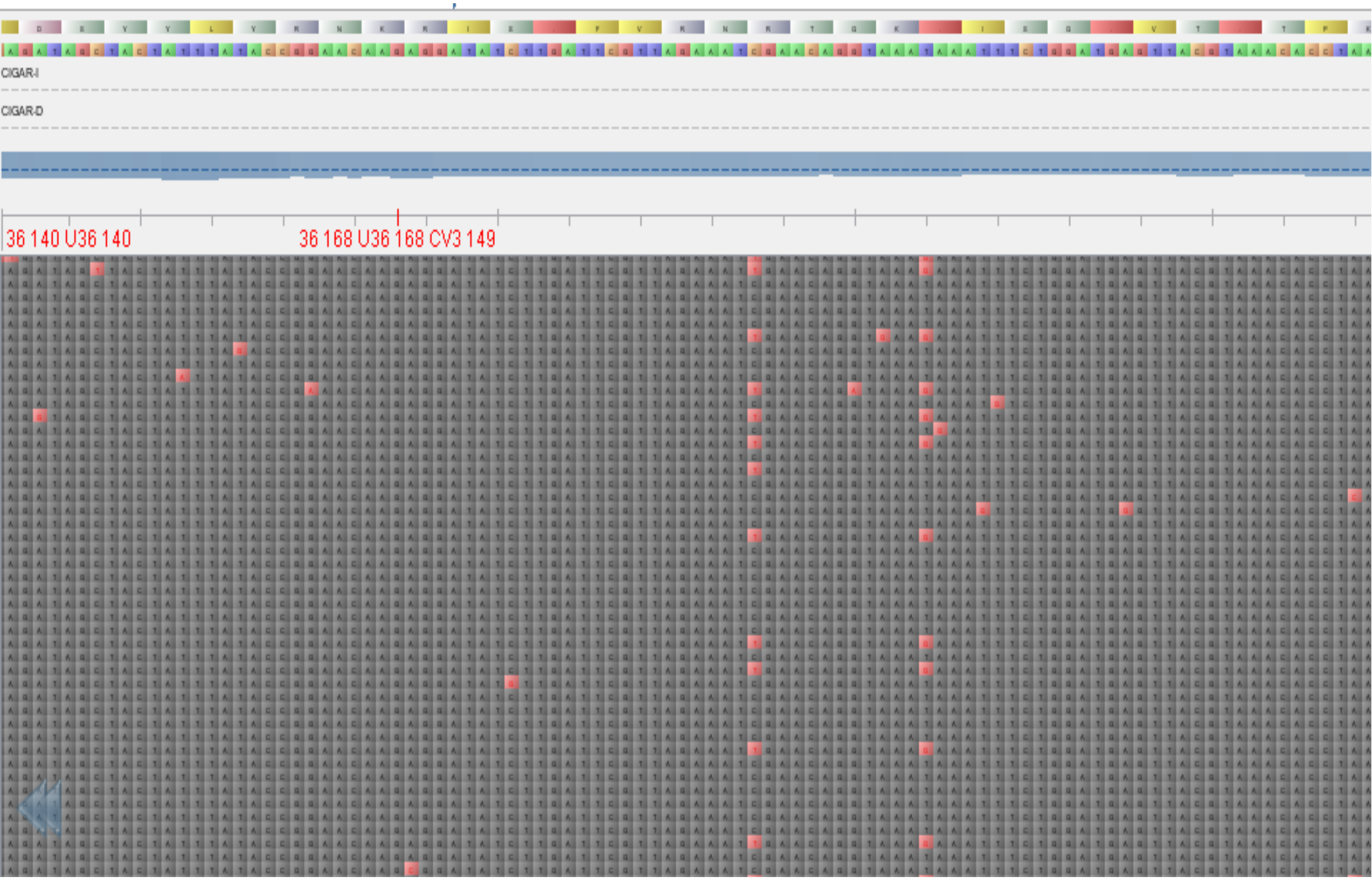
ACGCCG--GCA CGGACTGACTG CGGCTACGACTAG GCGACGCCG

GCATCGGACTG TGCATCGGCTA ACTAGCGAGCT

AAGCTGCGA GCCG--GCATC TGACTGCATCGG ACGACTAGCGAGCT

AAGCTGCGACGCCGCTGCATCGGACTGACTGCATCGACTACGACTAGCGAGCT

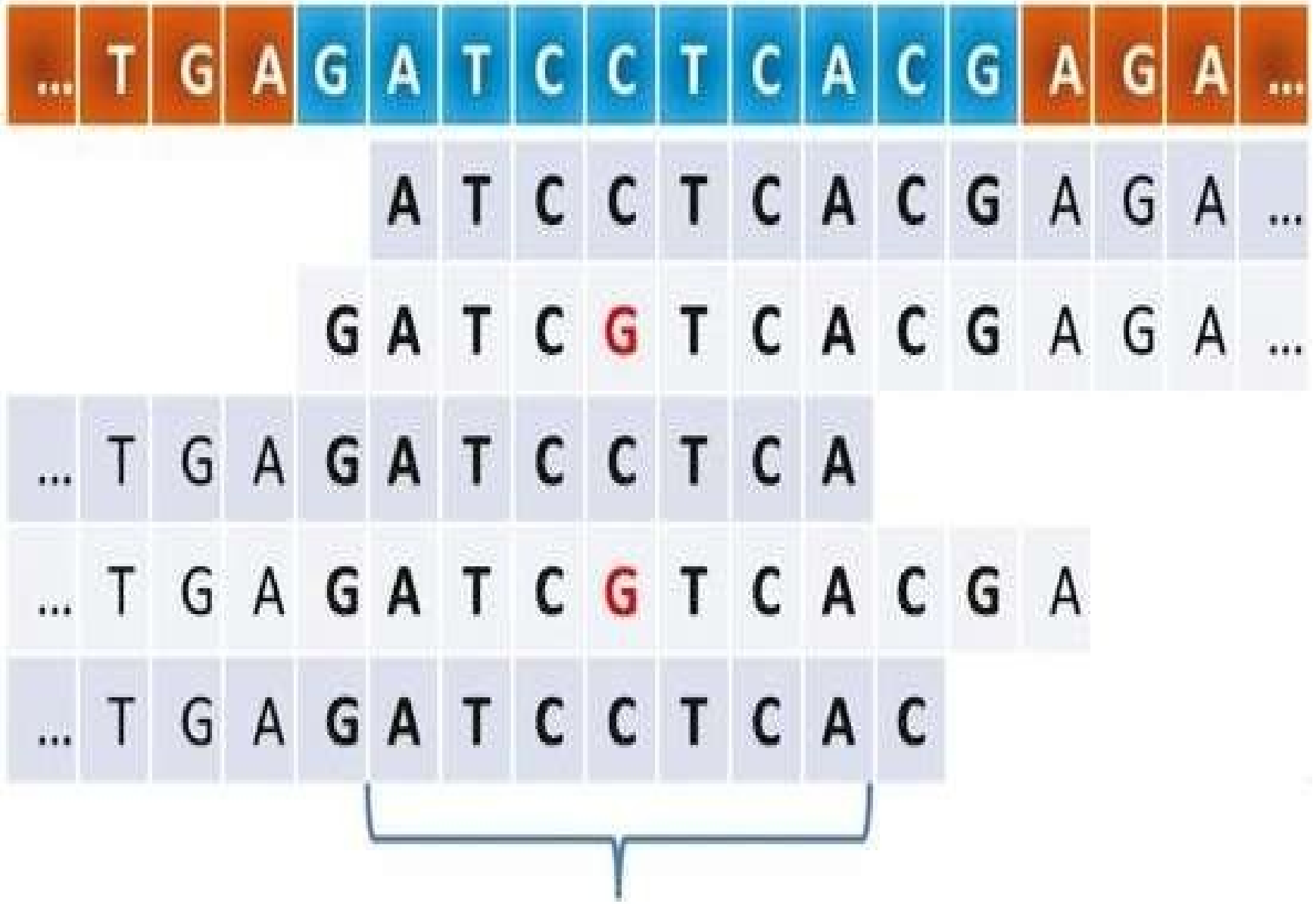
Поиск отличий (мутаций)



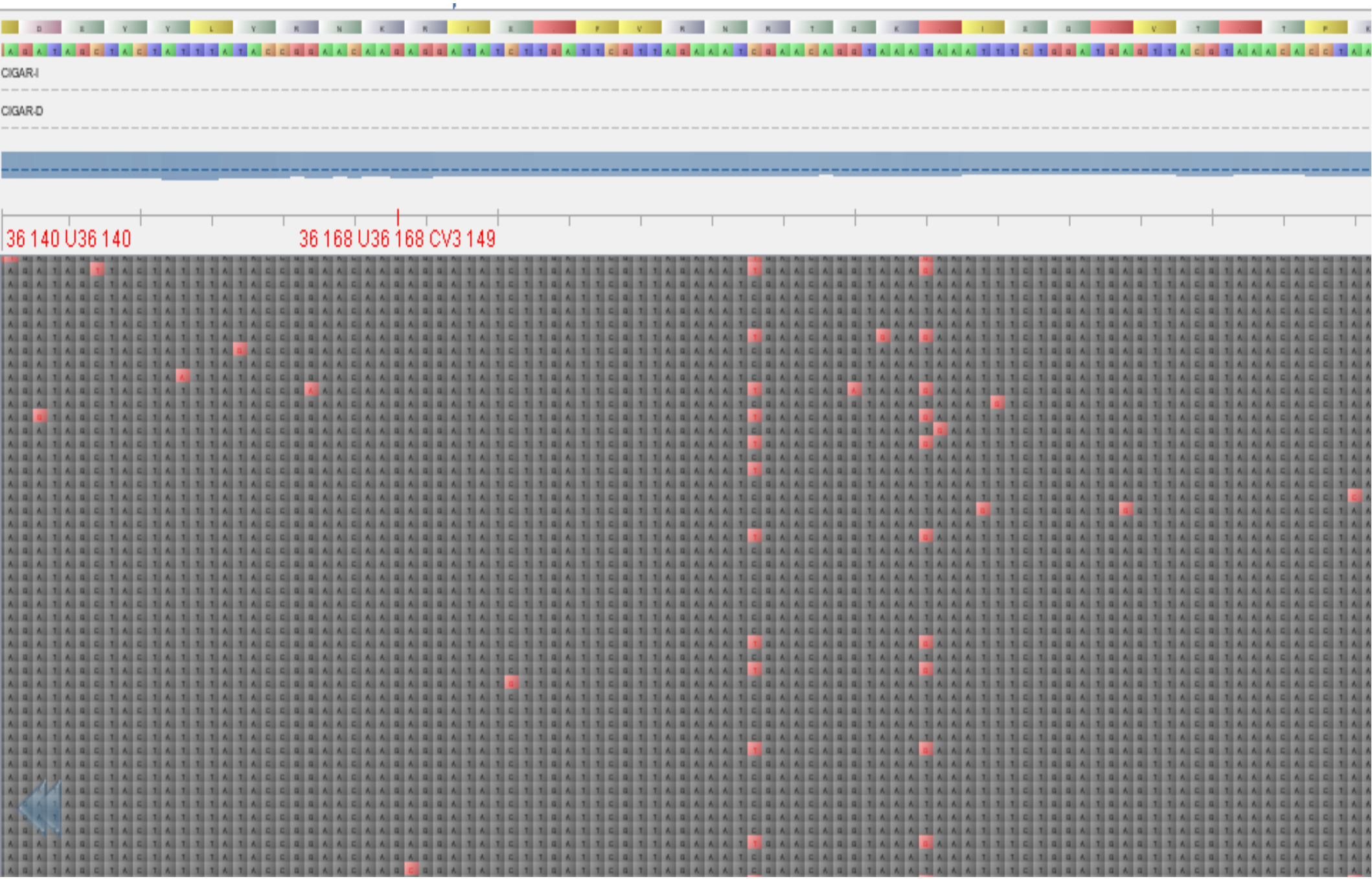
Формат VCF

```
#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|0:4
8:1:51,51 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|0:4
9:3:58,50 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|2:2
1:6:23,27 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|0:5
4:7:56,60 0|0:48:4:51,51 0/0:61:2
20     1234567 microsat1 GTCT  G,GTACT 50   PASS   NS=3;DP=9;AA=G                    GT:GQ:DP    0/1:3
5:4       0/2:17:2       1/1:40:3
```

Оценка покрытия



Поиск отличий (мутаций)

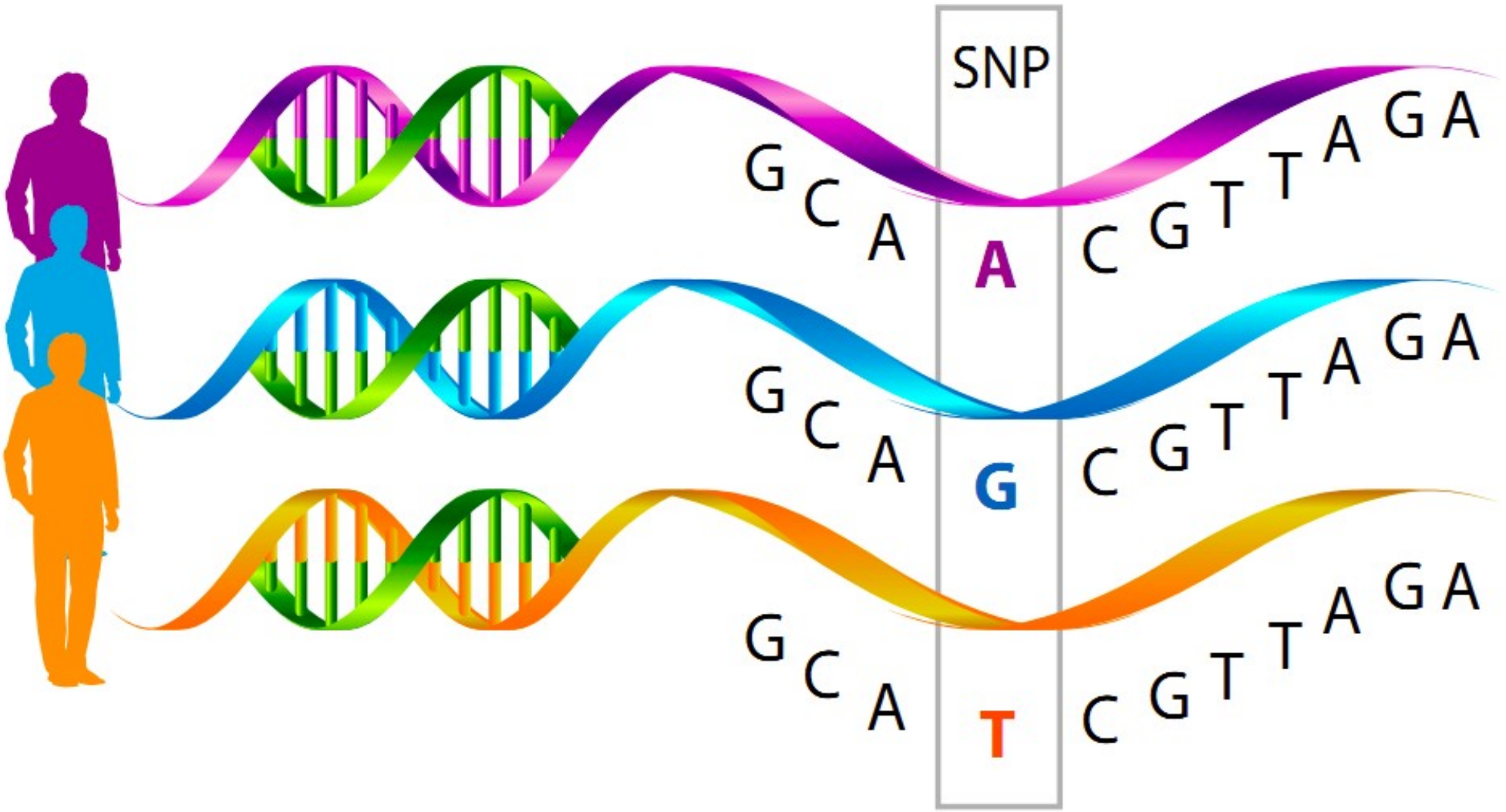


Indel examples

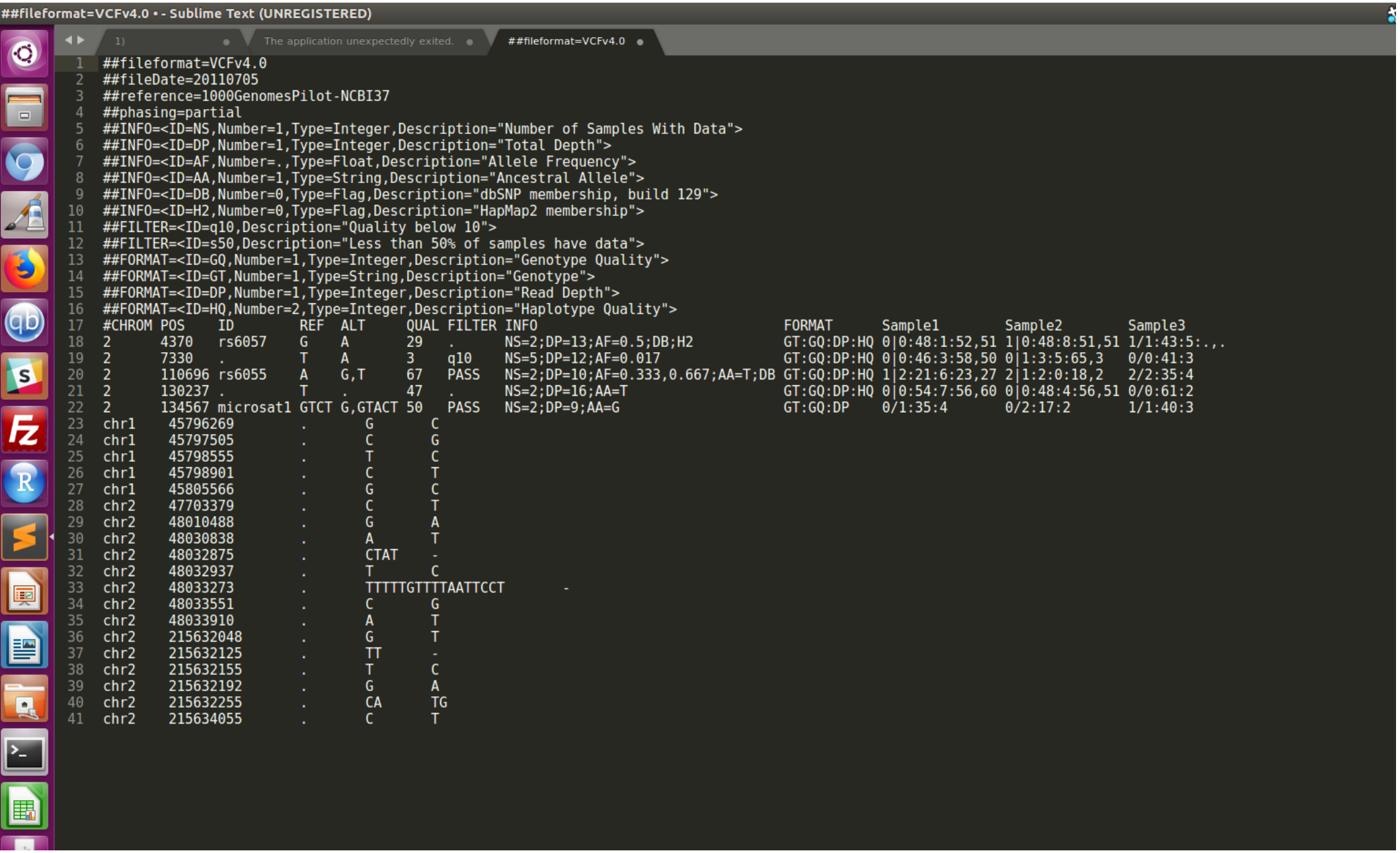
wild-type sequence

3 bp deletion

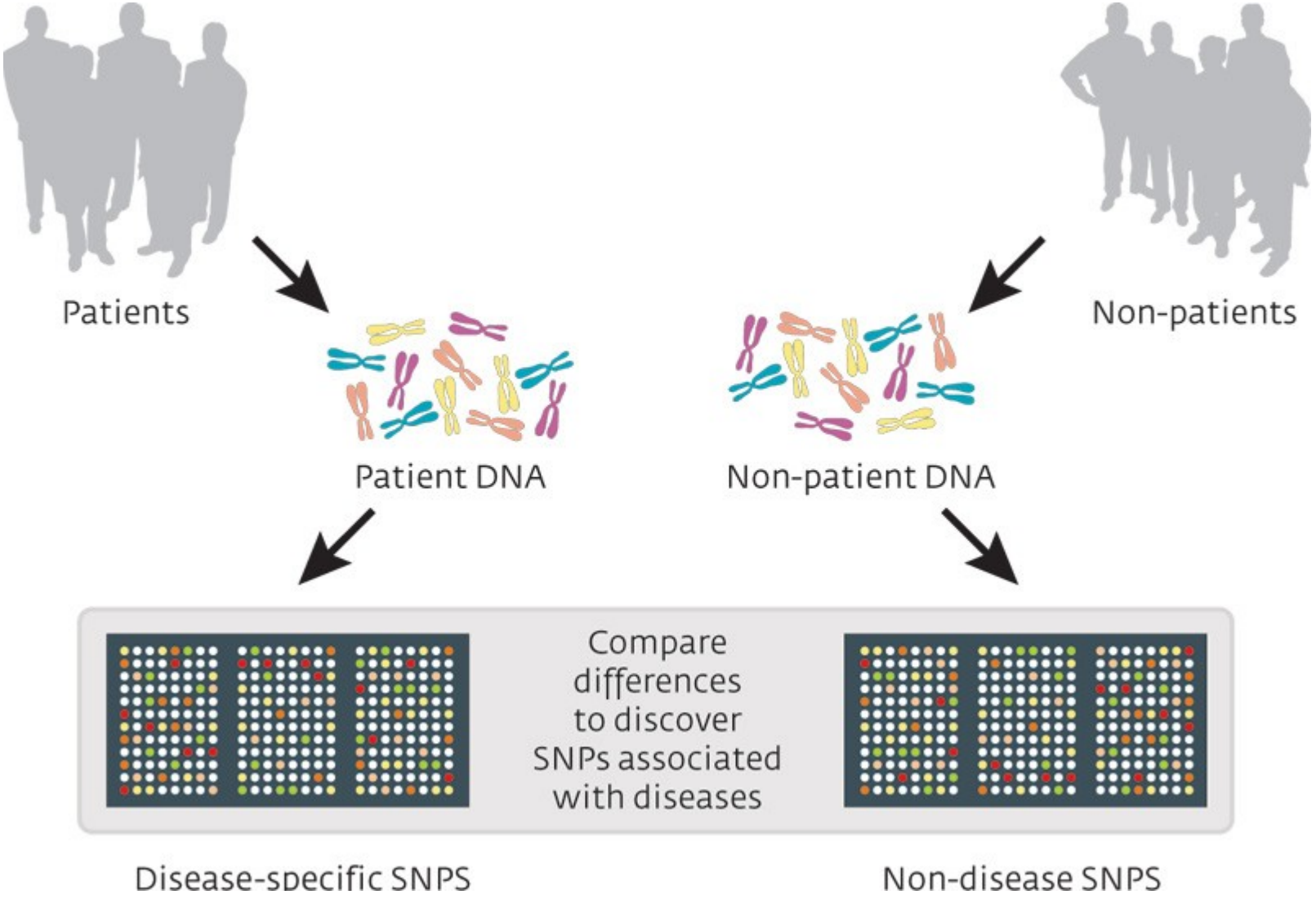
4 bp insertion (orange)

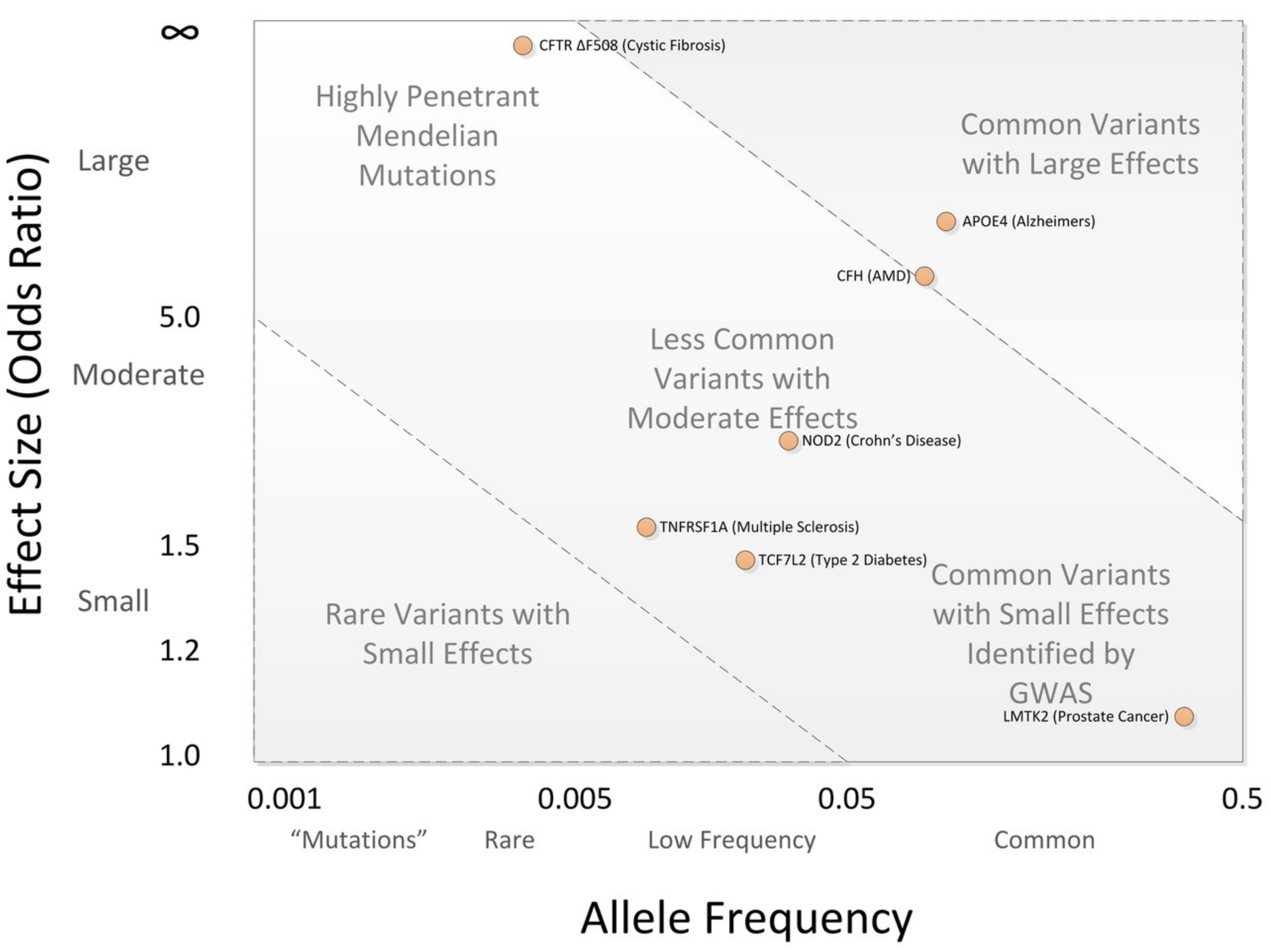

SNP
G C A A C G T T A G A
G C A G C G T T A G A
G C A T C G T T A G A

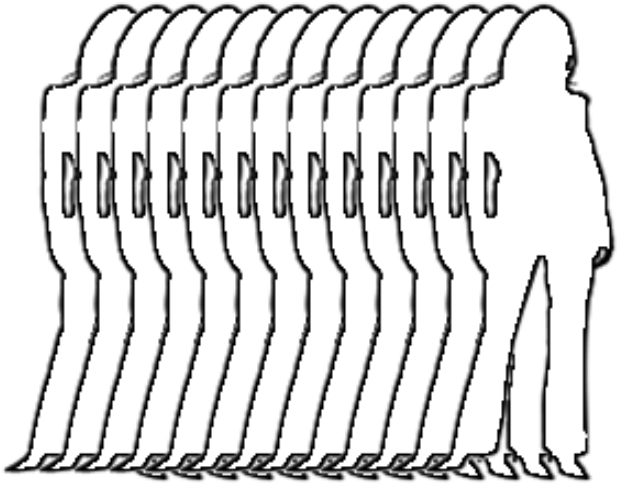
Формат VCF


##fileformat=VCFv4.0 • - Sublime Text (UNREGISTERED)
1)
The application unexpectedly exited.
##fileformat=VCFv4.0
##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 Sample2 Sample3
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 1|0:48:8:51,51 1/1:43:5:.,.
2 7330 . T A 3 q10 NS=5;DP=12;AF=0.017 GT:GQ:DP:HQ 0|0:46:3:58,50 0|1:3:5:65,3 0/0:41:3
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|1:2:0:18,2 2/2:35:4
2 130237 . T . 47 . NS=2;DP=16;AA=T GT:GQ:DP:HQ 0|0:54:7:56,60 0|0:48:4:56,51 0/0:61:2
2 134567 microsat1 GTCT G,GTACT 50 PASS NS=2;DP=9;AA=G GT:GQ:DP 0/1:35:4 0/2:17:2 1/1:40:3
chr1 45796269 . G C
chr1 45797505 . C G
chr1 45798555 . T C
chr1 45798901 . C T
chr1 45805566 . G C
chr2 47703379 . C T
chr2 48010488 . G A
chr2 48030838 . A T
chr2 48032875 . CTAT -
chr2 48032937 . T C
chr2 48033273 . TTTTGTTTTAATTCCT -
chr2 48033551 . C G
chr2 48033910 . A T
chr2 215632048 . G T
chr2 215632125 . TT -
chr2 215632155 . T C
chr2 215632192 . G A
chr2 215632255 . CA TG
chr2 215634055 . C T

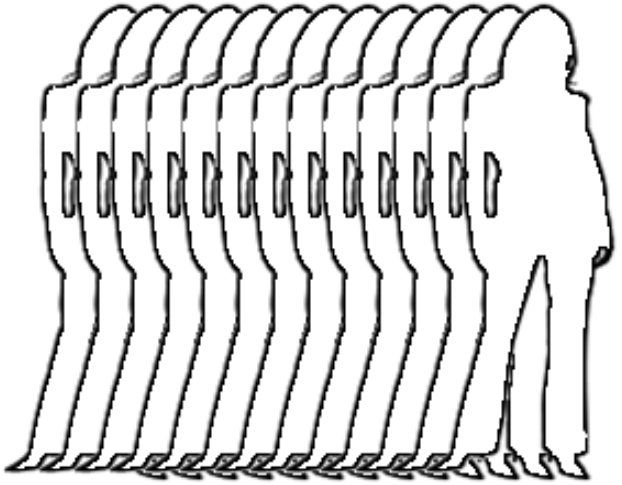
GWAS (Genome-wide association study)




Effect Size (Odds Ratio)
∞
Large
5.0
Moderate
1.5
Small
1.2
1.0
CFTR ΔF508 (Cystic Fibrosis)
Highly Penetrant Mendelian Mutations
Common Variants with Large Effects
APOE4 (Alzheimers)
CFH (AMD)
Less Common Variants with Moderate Effects
NOD2 (Crohn's Disease)
TNFRSF1A (Multiple Sclerosis)
TCF7L2 (Type 2 Diabetes)
Rare Variants with Small Effects
Common Variants with Small Effects Identified by GWAS
LMTK2 (Prostate Cancer)
0.001
0.005
0.05
0.5
"Mutations"
Rare
Low Frequency
Common
Allele Frequency



GC CC GG GC CC GC GC
GG CC GC GG GC GG



GC CC GC GC GG CC CC
CC GC GC GG GC GG

SNP1

Cases
Count of G:
2104 of 4000

Frequency of G:
52.6%

Controls
Count of G:
2676 of 6000

Frequency of G:
44.6%

P-value:
$5.0 \cdot 10^{-15}$

SNP2

Cases
Count of G:
1648 of 4000

Frequency of G:
41.2%

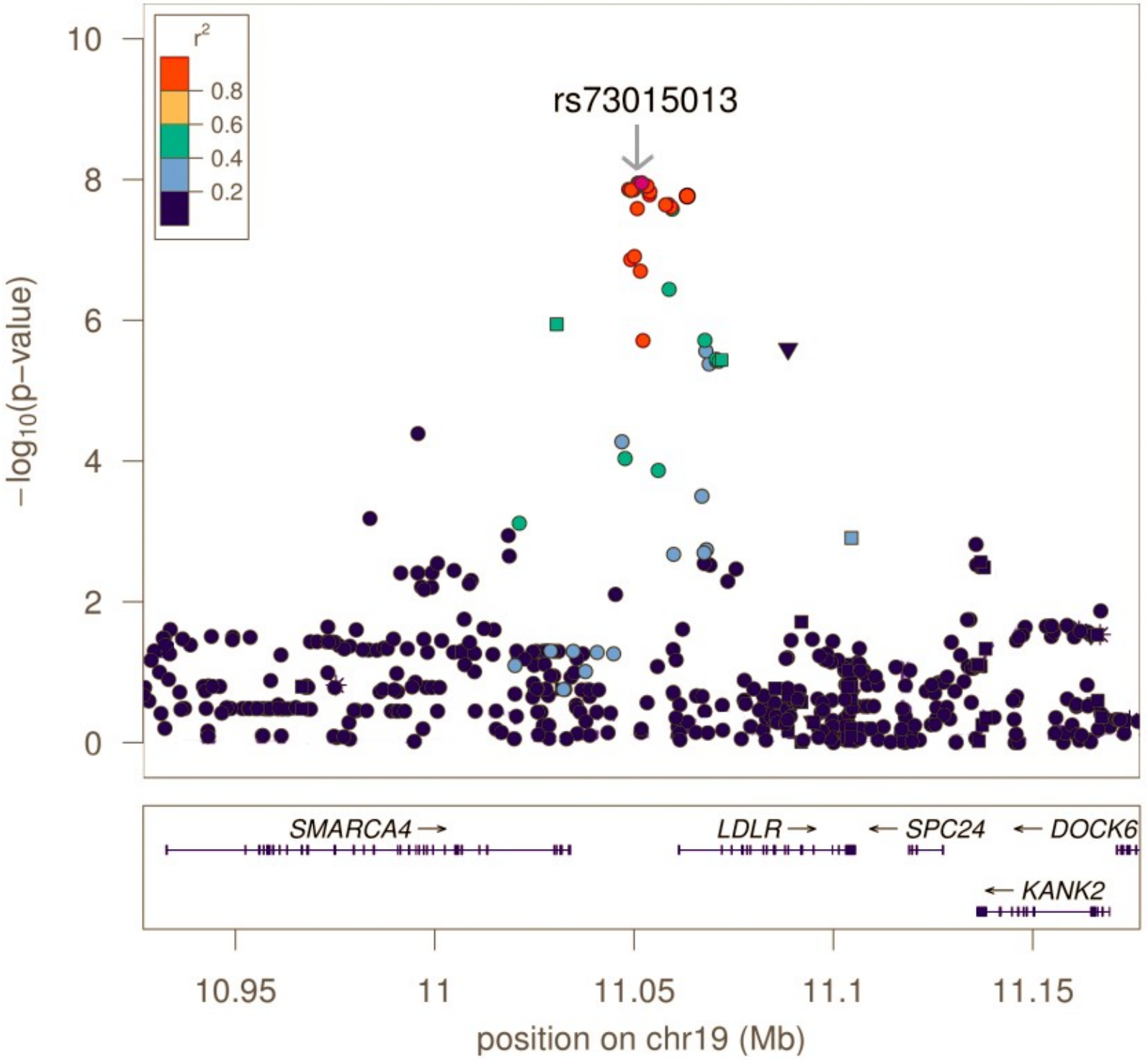
Controls
Count of G:
2532 of 6000

Frequency of G:
42.2%

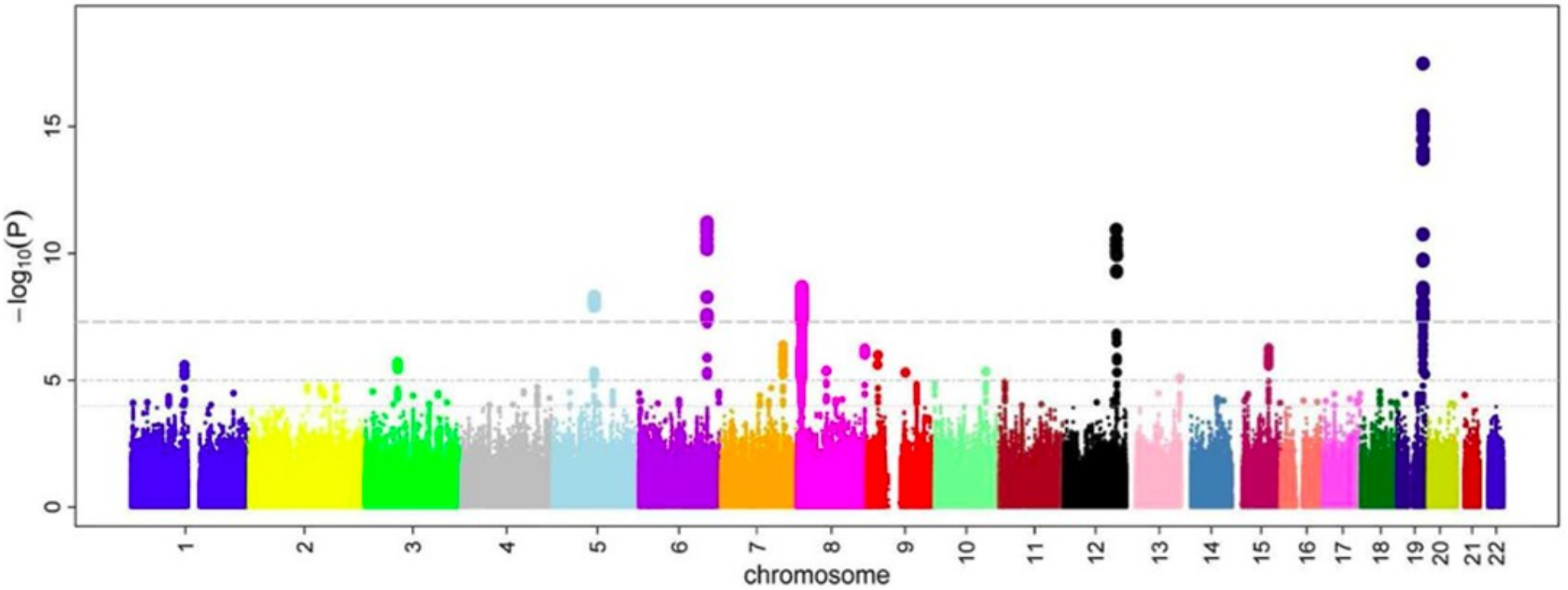
P-value:
0.33

SNP …

Repeat for all SNPs


r²
0.8
0.6
0.4
0.2
rs73015013
10
8
6
4
2
0
−log10(p−value)
SMARCA4 →
LDLR →
← SPC24
← DOCK6
← KANK2
10.95
11
11.05
11.1
11.15
position on chr19 (Mb)

Manhattan plot



Можно почитать дома

- A field guide to whole-genome sequencing, assembly and annotation

  https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4231593/

- Genome-wide association study

  https://en.wikipedia.org/wiki/Genome-wide_association_study

Спасибо за внимание!