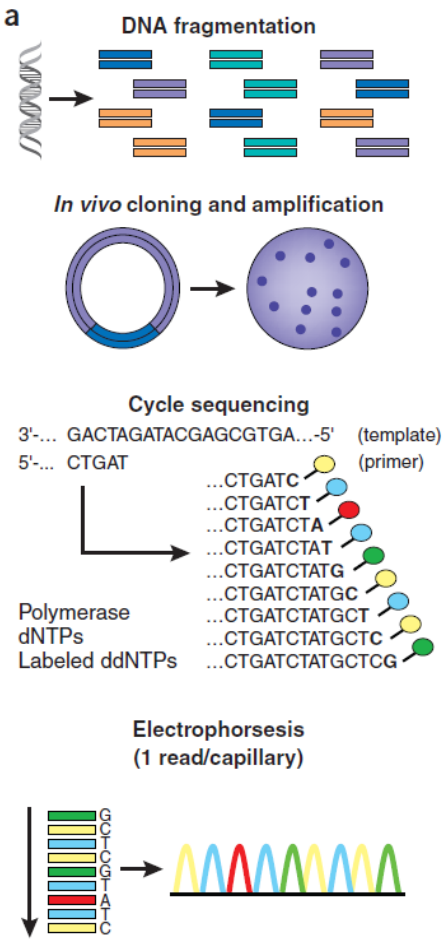
Sequenzierung

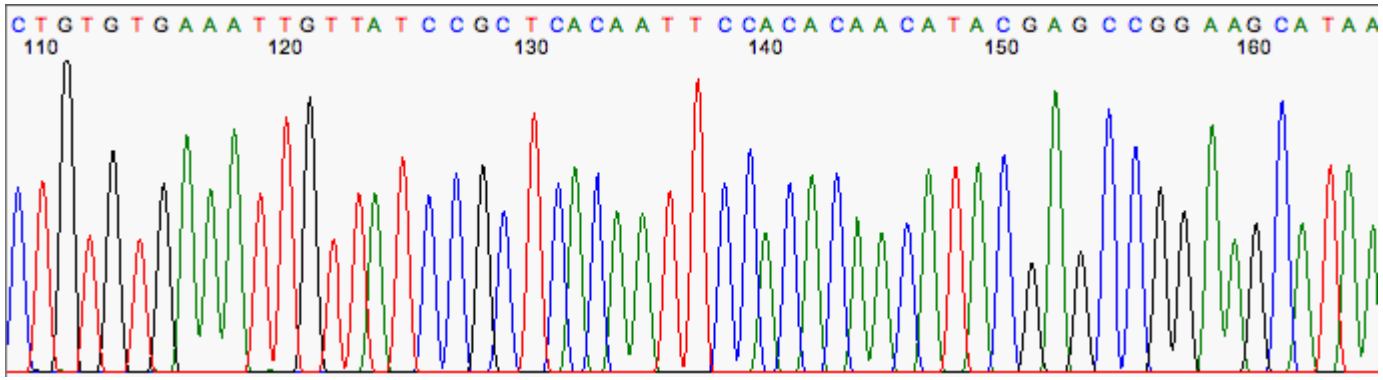
Aufgabenstellung

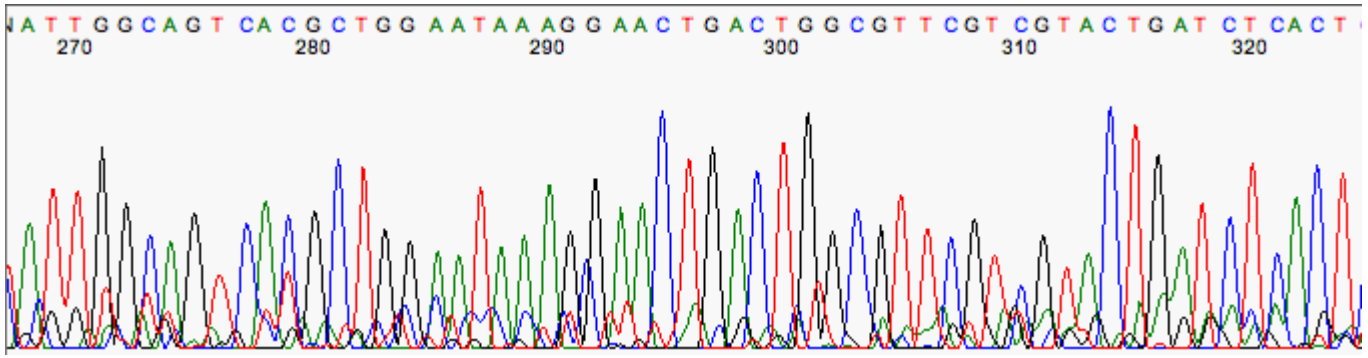
- Genomische Sequenzierung (DNA)
  - Re-Sequenzierung eines bekannten Genoms (z.B. Mensch)
  - De novo Sequenzierung
- Sequenzierung von RNA mittels Umschreiben in cDNA – Sequenzieren von Transkripten oder kleinen RNAs (z.B. miRNA)

Sanger sequencing



- DNA is fragmented
- Cloned to a plasmid vector
- Cyclic sequencing reaction
- Separation by electrophoresis
- Readout with fluorescent tags





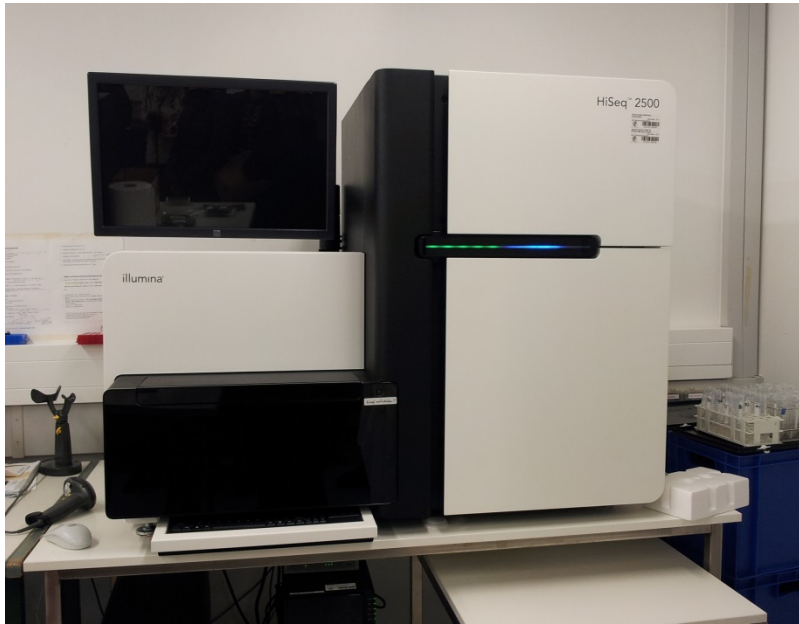
http://seqcore.brcf.med.umich.edu/doc/dnaseq/interpret.html

Historie

- 1977: Sequenzierung von Bakteriophagen PhiX175. 5386 Nukleotide (Fred Sanger, Nobelpreis 1980 – sein zweiter nach 1958 für die Sequenzierung von Insulin)
- 1990er: Sequenzierung kompletter Genome mittels Sanger Sequenzierung
  - Hefe (S. cervisiae), Wurm (C. elegans), Fliege (Drosophila melanogaster), Maus, Mensch, …
- 1990er Jahre: EST Sequenzierung: EST = expressed sequence tag, Sequenzierung von Bruchstücken der mRNAs (cDNA)

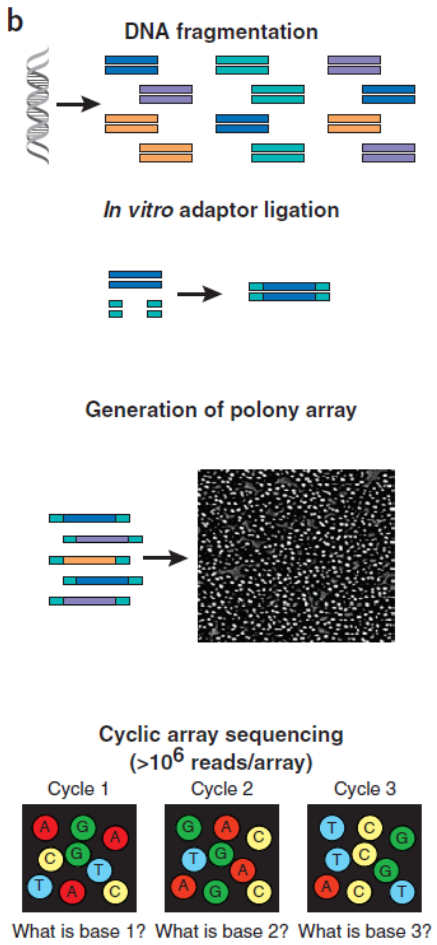
Next Generation Sequencing (NGS)

- Ab ca. 2005
- Sequenzieren vieler kleiner (heute etwa 100 bp) Fragmente: „reads“
- In einem Experiment kann man mehrere 100 Mio reads bestimmen
- Gerät: Illumina HiSeq Serie

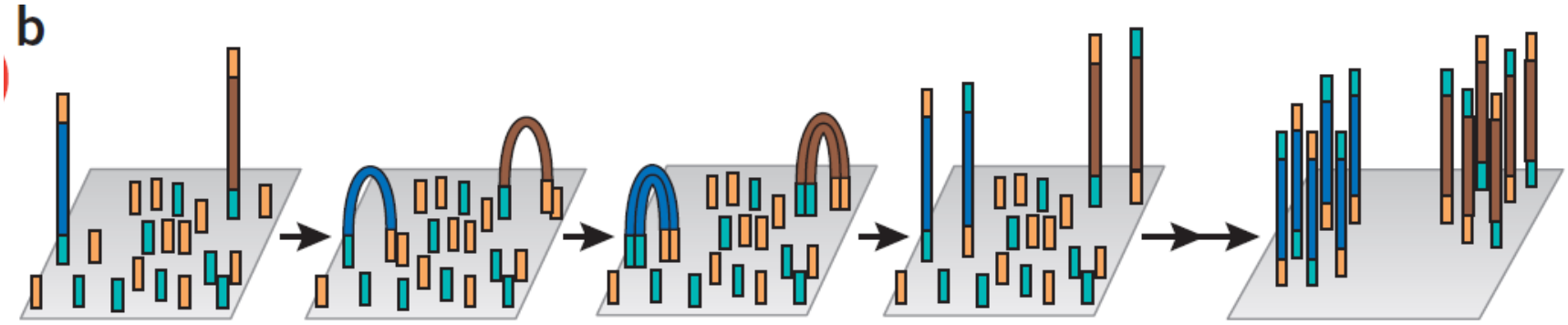


Cyclic-array methods

- DNA is fragmented
- Adaptors ligated to fragments
- Several possible protocols yield array of PCR colonies.
- Enyzmatic extension with fluorescently tagged nucleotides.
- Cyclic readout by imaging the array.

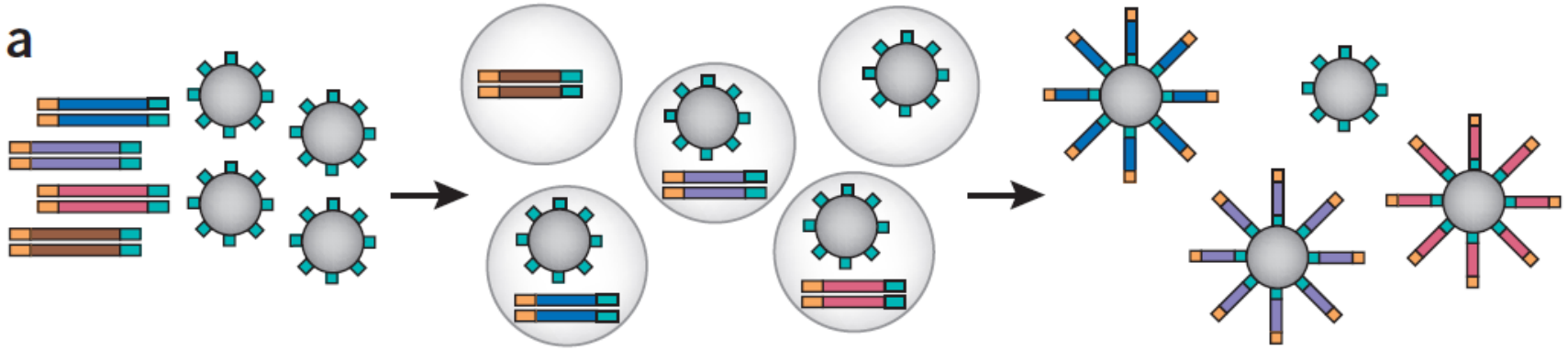


Bridge PCR

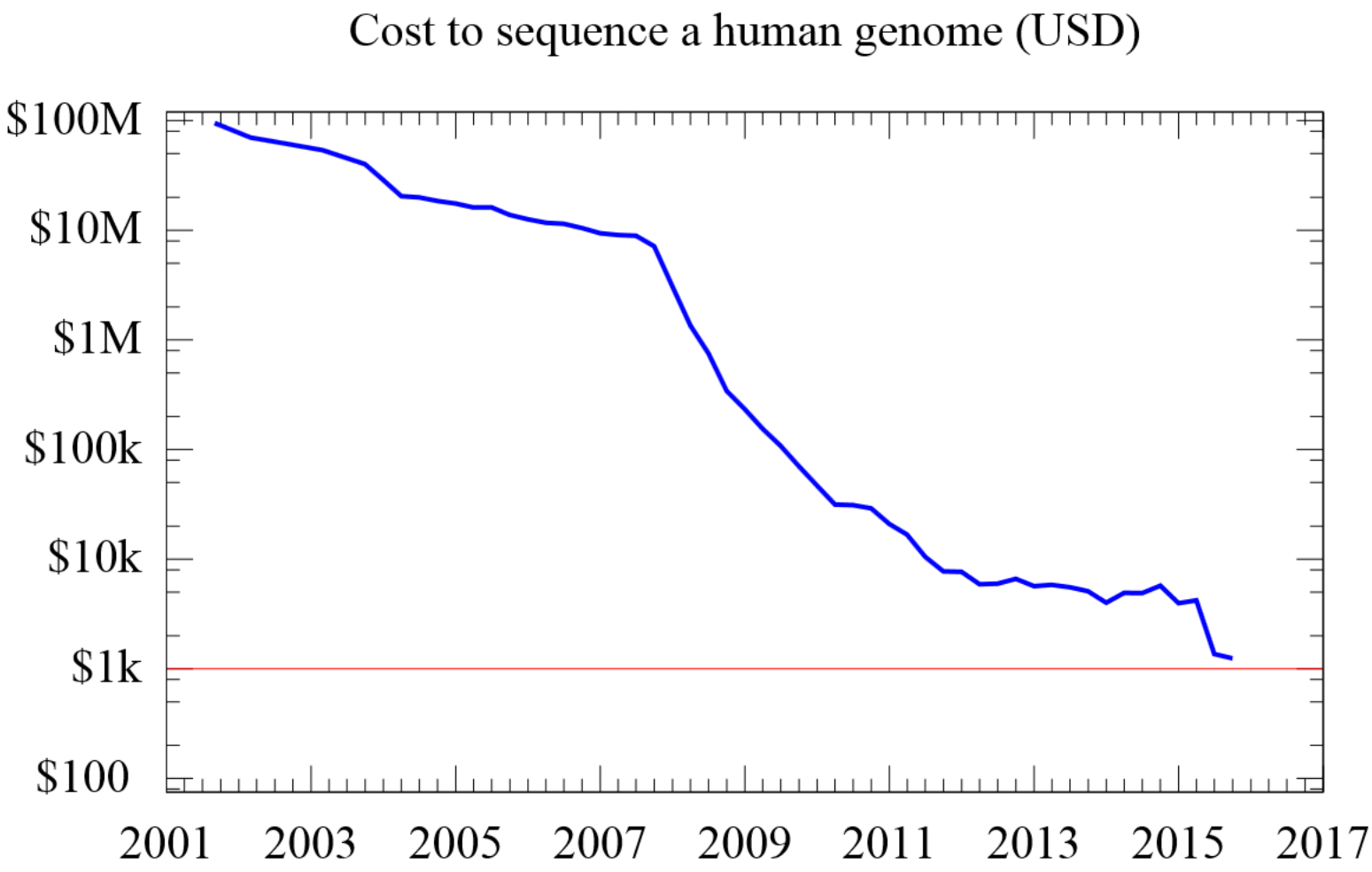


- DNA fragments are flanked with adaptors.
- A flat surface coated with two types of primers, corresponding to the adaptors.
- Amplification proceeds in cycles, with one end of each bridge tethered to the surface.
- Used by Solexa/Illumina.

Emulsion PCR



- Fragments, with adaptors, are PCR amplified within a water drop in oil.
- One primer is attached to the surface of a bead.
- Used by 454, Polonator and SOLiD.


Cost to sequence a human genome (USD)
$100M
$10M
$1M
$100k
$10k
$1k
$100
2001
2003
2005
2007
2009
2011
2013
2015
2017

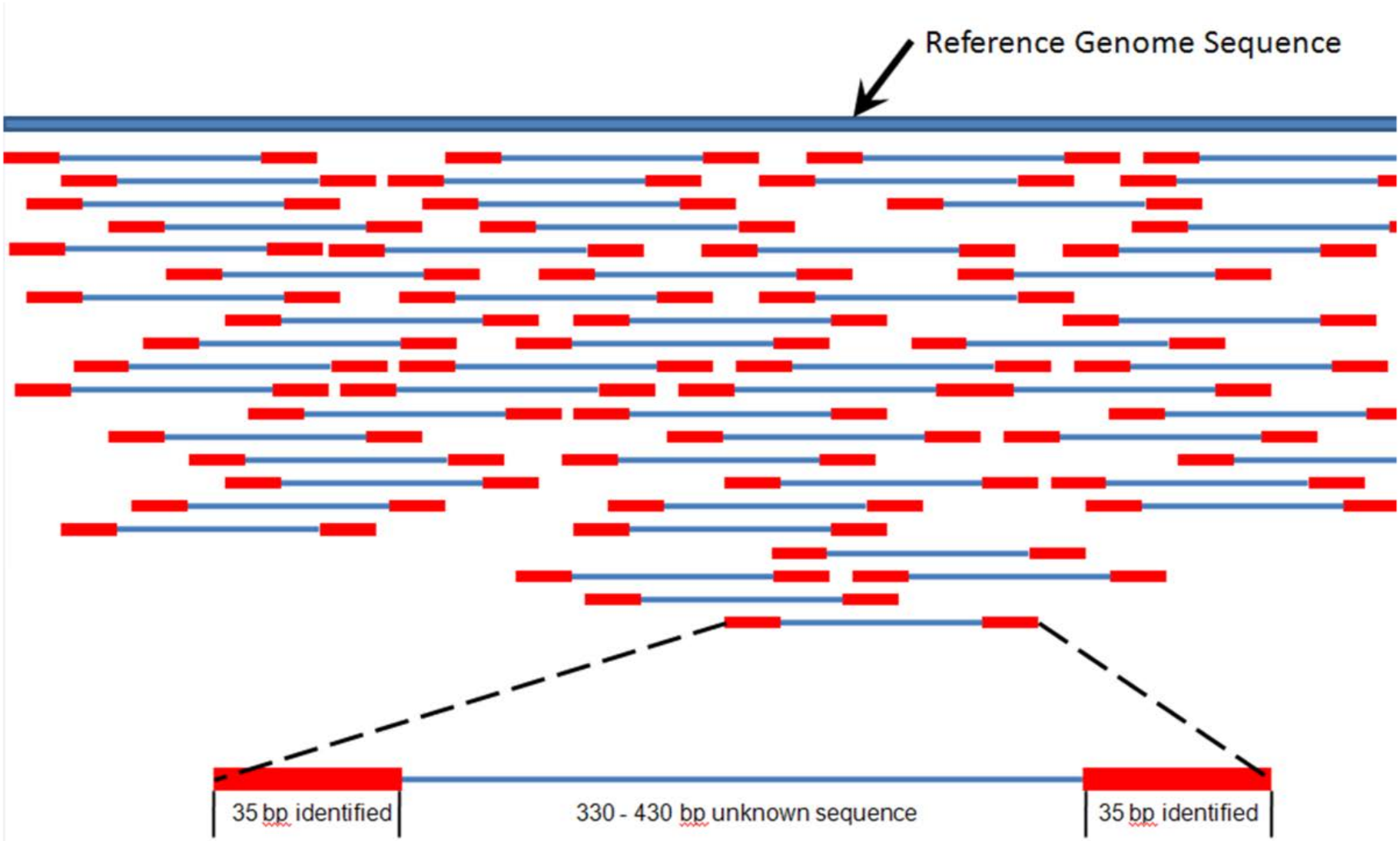
Qualität

- Zuverlässigkeit der read-Sequenzen ist unterschiedlich und von der Technologie abhängig
- Enden können schlechter sein (daher: clipping)
- Homopolymer-runs sind problematisch
- GC bias

- Meist wird zu der Sequenz eine Qualitätswert ausgegeben

Man bestimmt viele Bruchstücke

- Ausgangsmaterial: DNA von vielen Zellen. „Grundsequenz“
- Resultat der Sequenzierung: 100e Millionen reads von Stücken der DNA, die sich wiederholen
- Man weiss nicht welches Stück von wo in der Grundsequenz kam


Reference Genome Sequence
35 bp identified
330 - 430 bp unknown sequence
35 bp identified

Shotgun sequencing & Assembly

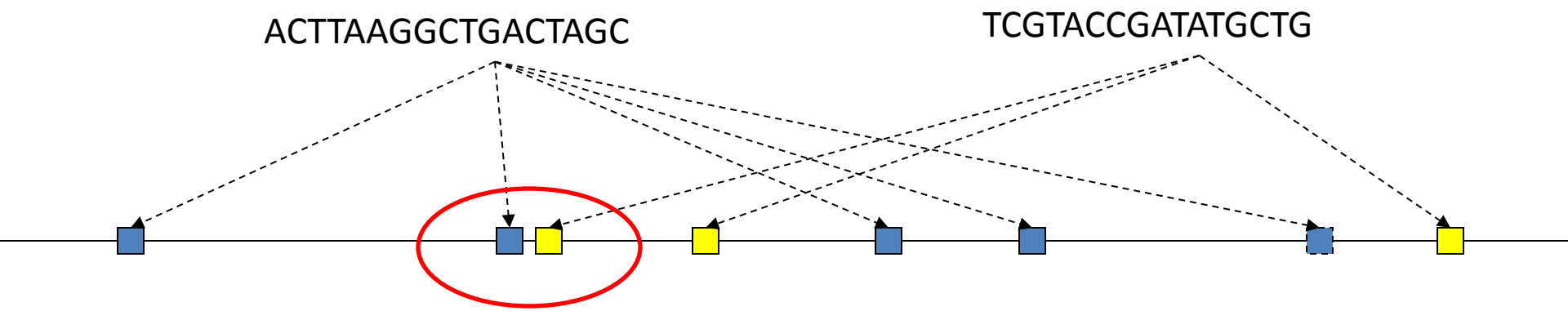
- Sanger sequence reads ca 500-800 Basen lang
- Große DNA Stücke, z.B. BACs, Bacterial artifical chromosome. Länge 100-300 kb.
- Zerlegen und klonieren: Clone. Insert einige 1000 bp. Von einer oder von beiden Seiten ansequenzieren.
- Heute 100 bp-reads, entweder einzeln oder paired-end
- Mehr Fehler als bei Sanger Sequenzierung – kompensiert durch höhere Abdeckung

- Wikipedia: „Shotgun sequencing“, „DNA sequencing theory“

Re-Sequenzierung

- Annahme: Das Genom des Organismus ist im Prinzip bekannt, z.B. Humangenom, oder mus musculus
- Aufgabe: Sequenziere eine weiteres Individuum, oder einen nah verwandte strain oder Spezies
- Versuche die neu bestimmten reads auf das bekannte Genom zu „mappen“, um deren Reihenfolge (und die korrekte Sequenz) zu bestimmen

Read length and pairing



- Short reads are problematic, because short sequences do not map uniquely to the genome.
- Solution #1: Get longer reads.
- Solution #2: Get paired reads.

Mapping Software

- BLAST zu langsam (Vorverarbeitung der query)
- Hashing: k-mer index for seeds.
- Suffix trees, suffix arrays: Vorverarbeitung des Textes. Speicherbedarf ist ein Mehrfaches des Genoms.
  - Suffix tree: 10-20fach; suffix array: 8fach
  - Beispiel: Humangenom 3 GB, Suffix tree mehr als 30GB, suffix array 24GB.
  - Wieviel RAM hat Ihr Computer?

Reminder: Secondary Storage Data Structures

- Data structure resides on disk
- B-trees (1972), string B-tree (1996)
- Suffix arrays were designed to reside on disk (not any more)

- Secondary Storage Data Structures sind nicht schnell genug für read mapping!
  Datenstruktur muss in RAM passen.

Software

- Erste Generation: eland (hashing), vmatch, …
- SOAP, MAQ (hashing)
- Bowtie, SOAP2, BWA … Burrows-Wheeler transform

- Bowtie uses as little as 1.3GB of RAM for the index of the human genome (according to the authors, see Table 5)
- See: “Ultrafast and memory-efficient alignment of short DNA sequences to the human genome, by Ben Langmead, Cole Trapnell, Mihai Pop and Steven L Salzberg. Genome Biology 2009

Burrows-Wheeler transform & FM index

- BW Transform is a string (of equal length to the text).
  - BWT can be transformed back into the text
  - BWT can be compressed efficiently
- FM Index: Allows counting and searching of strings in the BWT. By Ferragina and Manzini (2000), but FM stands for „Full text index in Minute space“
- See Intro be Ben Langmead: „Introduction to the Burrows-Wheeler Transform and FM Index”, bwt_fm.pdf

Bowtie uses a different and novel indexing strategy to create an ultrafast, memory-efficient short read aligner geared toward mammalian re-sequencing. In our experiments using reads from the 1,000 Genomes project, Bowtie aligns 35-base pair (bp) reads at a rate of more than 25 million reads per CPU-hour, which is more than 35 times faster than Maq and 300 times faster than SOAP under the same conditions. Bowtie employs a Burrows-Wheeler index based on the full-text minute-space (FM) index, which has a memory footprint of only about 1.3 gigabytes (GB) for the human genome. The small footprint allows Bowtie to run on a typical desktop computer with 2 GB of RAM. The index is small enough to be distributed over the internet and to be stored on disk and re-used. Multiple processor cores can be used simultaneously to achieve even greater alignment speed. We have used Bowtie to align 14.3× coverage worth of human Illumina reads from the 1,000 Genomes project in about 14 hours on a single desktop computer with four processor cores.

NGS File formats

FastQ format

```
@seq1
TGGATCCTTAATAAACAAGGATGTTTCTGCATCATT
+
IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIBIII
@seq2
CACTTGTTCGGTGTTGTGGGGAAATGATGCAGAAAA
+
IIIIIIIIIIIIIIIIIIIIIIIIIIIIII3IIIII'
etc.
```

NGS File formats

SAM-Format

seq1 0 1 255048 0 36M * 0 0 TGGATCCTTAATA AACAAGGATGTTTCTGCATCATT IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIBIII NM:i:0 MD:Z: 36 AS:i:36 XS:i:36 NH:i:2
seq2 16 1 255069 0 1M1D35M * 0 0 TTTTCTGCATCATTTCCCCACAACACCGAACAAGTG 'IIIII3III IIIIIIIIIIIIIIIIIIIIIIIII NM:i:1 MD:Z:1^G35 AS:i:35 XS:i:35 NH:i:2

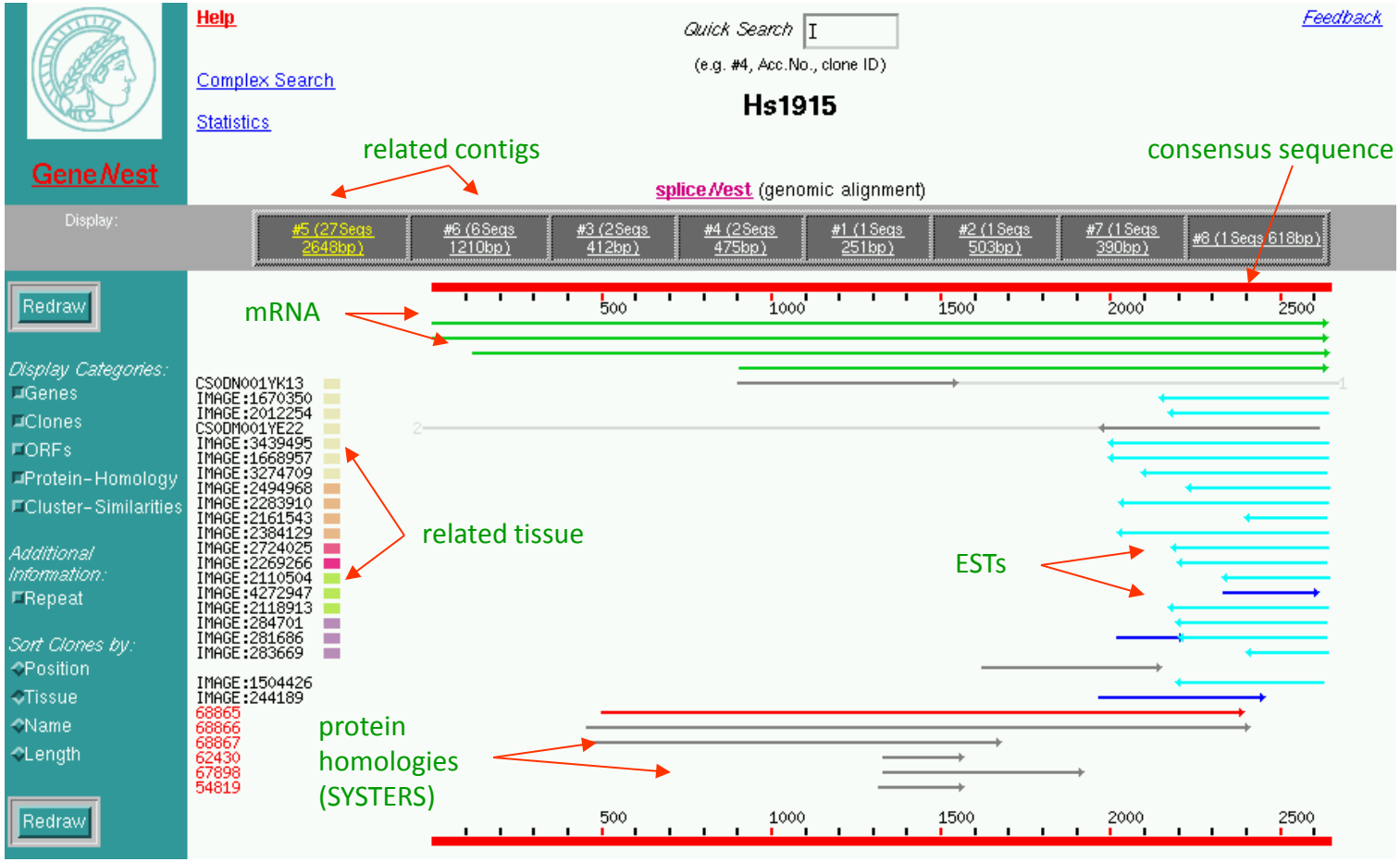
- http://samtools.sourceforge.net/samtools.shtml

Transkriptsequenzierung

- Cancer genomics, fusion genes
- Genetics, SNPs, variant calling (homozygous, heterozygous, frequencies vs sequencing errors)

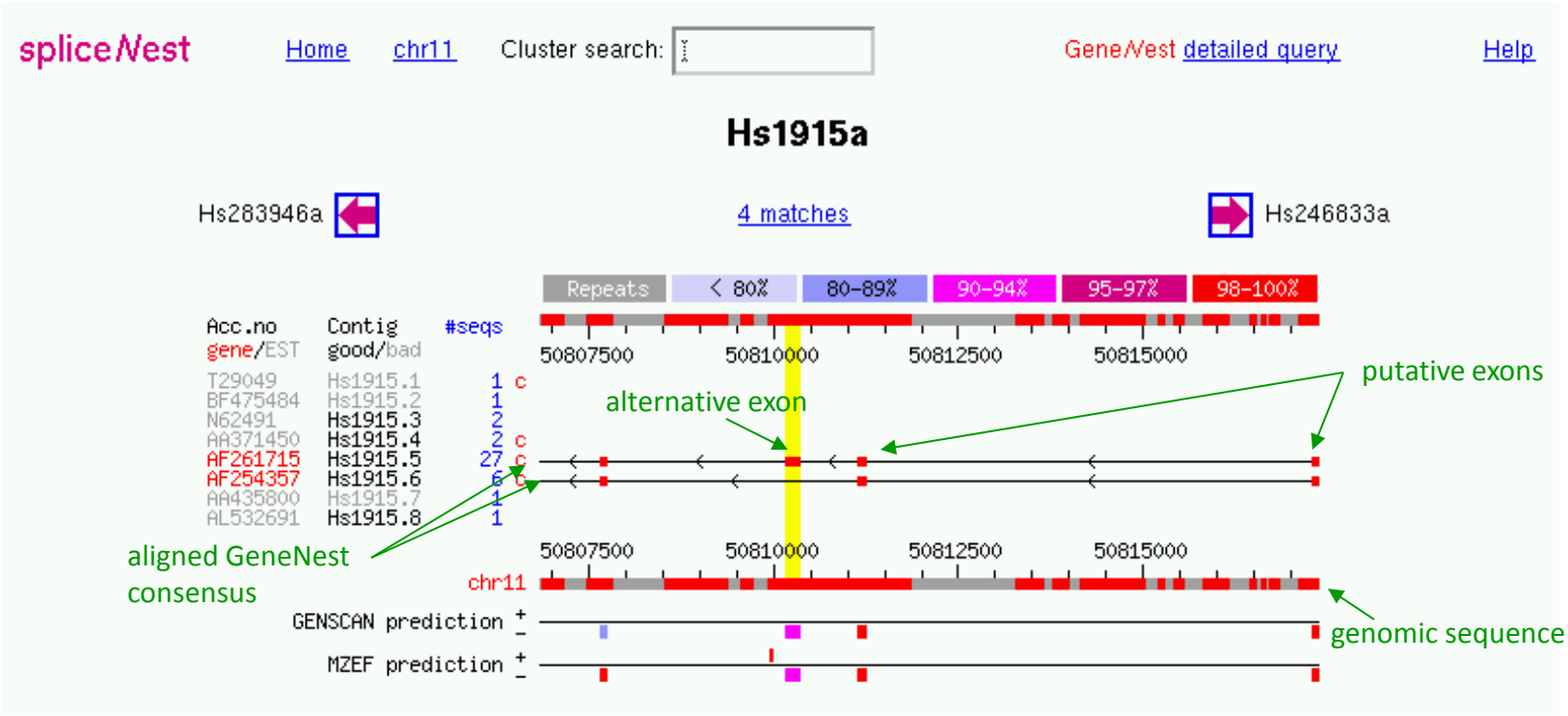
GeneNest visualization

(http://GeneNest.molgen.mpg.de)

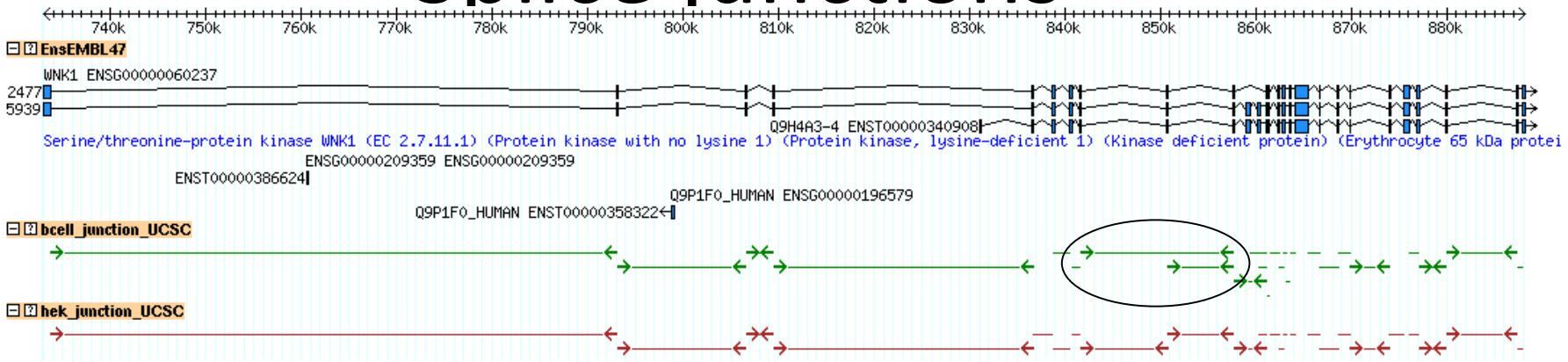


SpliceNest

(http://SpliceNest.molgen.mpg.de)

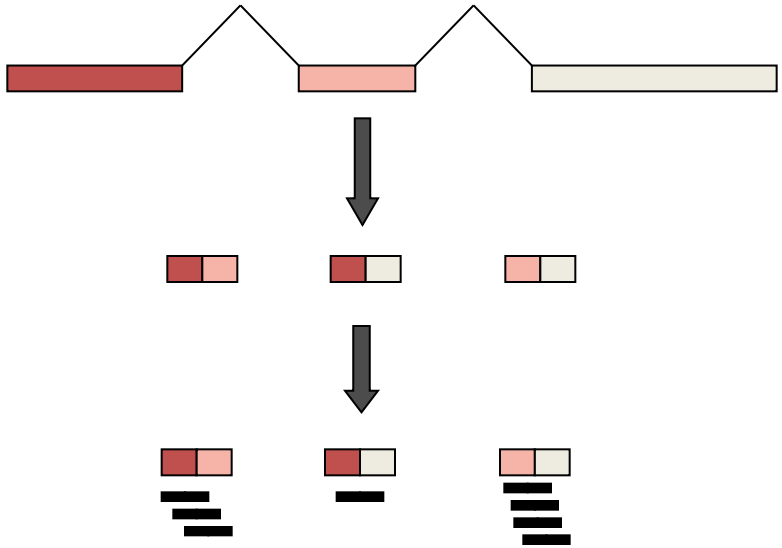


Splice junctions

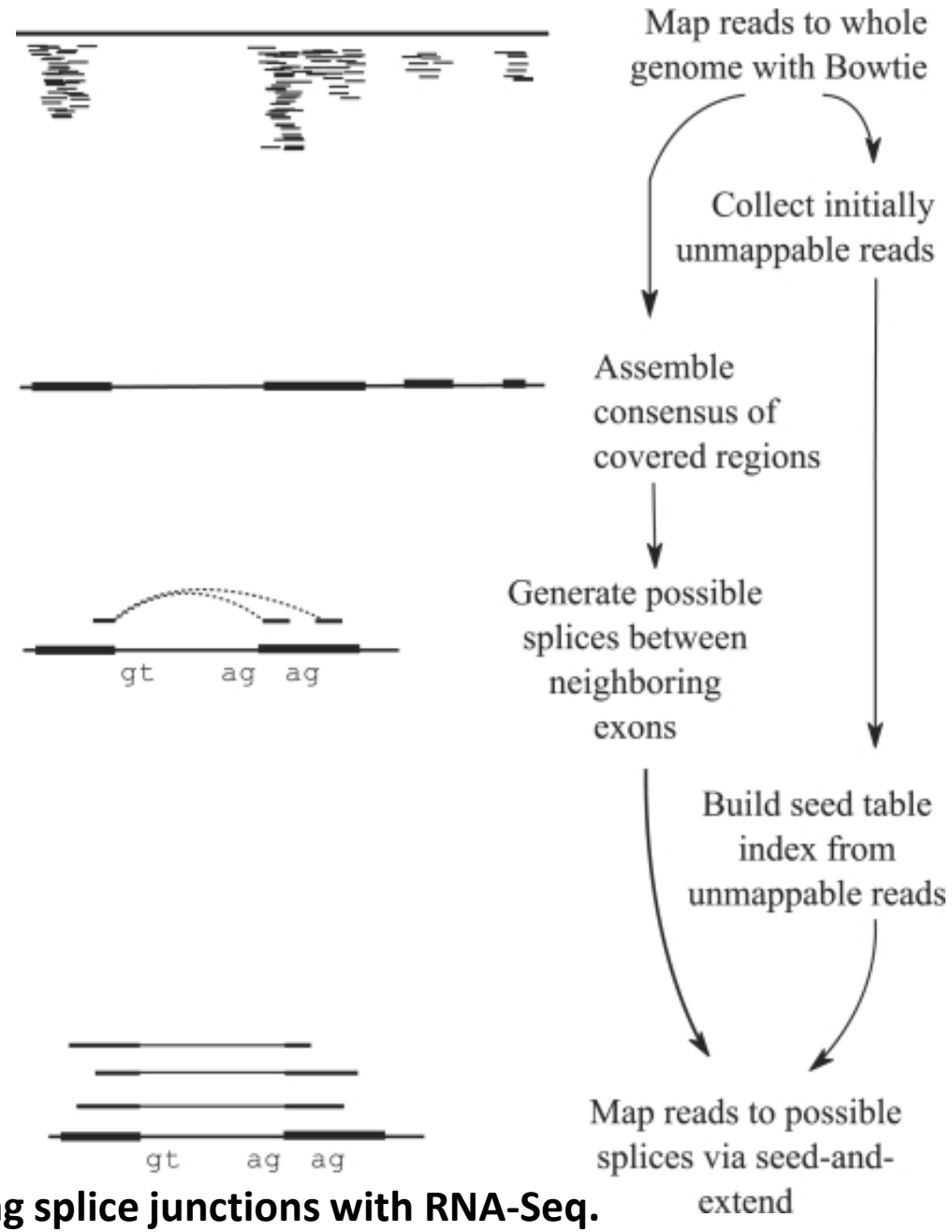


Align unmatched reads to artificial junctions

(~ 2,8 x 10^6 artificial junctions)



Sultan et al. (2008) A Global View of Gene Activity and Alternative Splicing by Deep Sequencing of the Human Transcriptome. *Science*, 321(5891):956-960



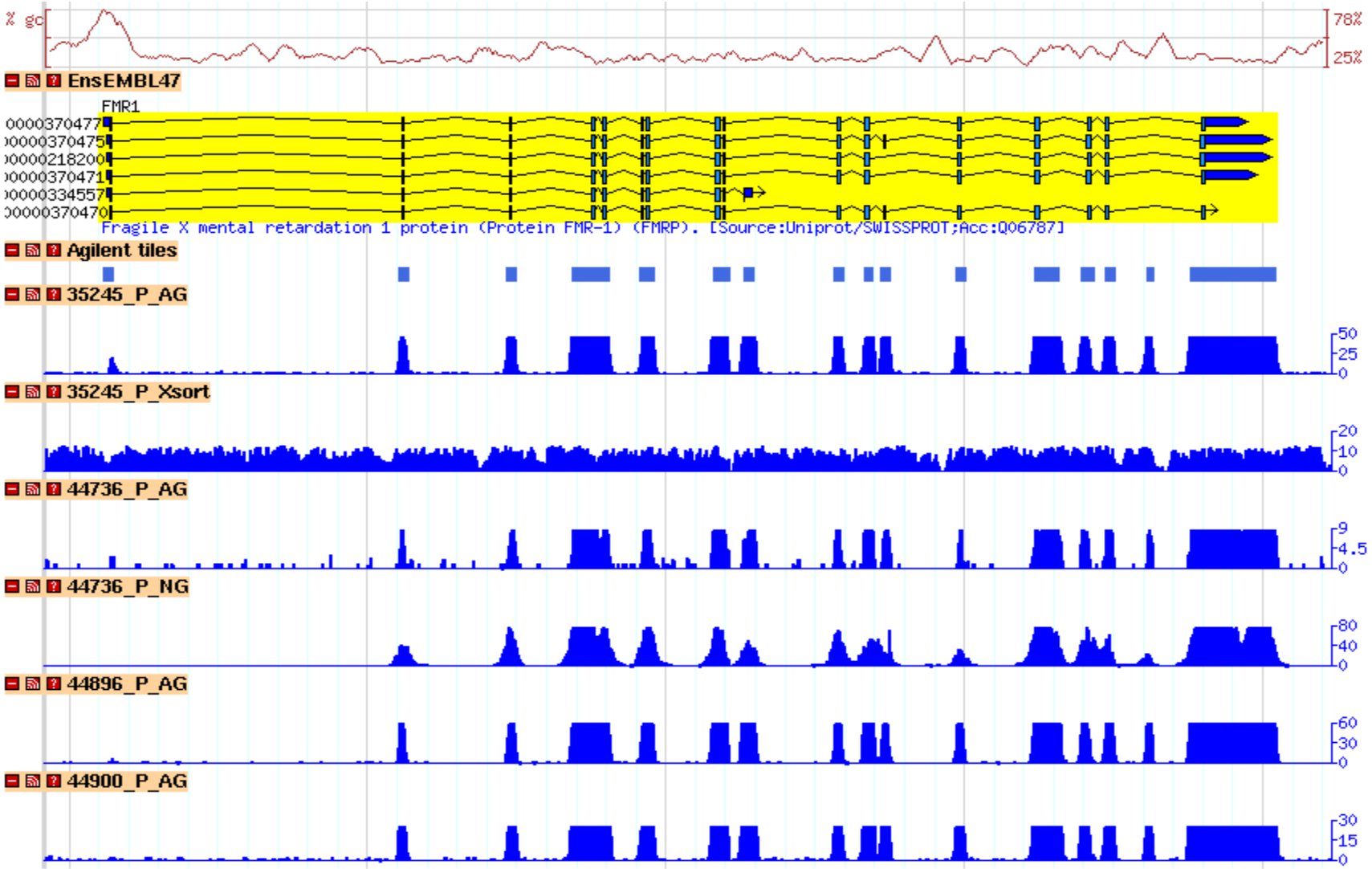
TopHat: discovering splice junctions with RNA-Seq.

Trapnell C[1], Pachter L, Salzberg SL.

Quantifizierung und Sampling

- Angenommen, es sind ca 1/3 aller Gene in einer Zelle exprimiert. Manche häufig (viele mRNA Moleküle), andere gering (wenige mRNA Moleküle)
- ESTs: ca 100K reads aus einer cDNA Bibliothek
- RNA-seq: 100 Mio reads

FMR1


% gc
78%
25%
EnsEMBL47
FMR1
00000370477
00000370475
00000218200
00000370471
00000334557
00000370470
Fragile X mental retardation 1 protein (Protein FMR-1) (FMRP). [Source:Uniprot/SWISSPROT;Acc:Q06787]
Agilent tiles
35245_P_AG
50
25
0
35245_P_Xsort
20
10
0
44736_P_AG
9
4.5
0
44736_P_NG
80
40
0
44896_P_AG
60
30
0
44900_P_AG
30
15
0

Andere Anwendungen

Genetics, SNPs, variant calling (homozygous, heterozygous, frequencies vs sequencing errors)

Variant calling

Homozygous (and hemi) mutations vs heterozygous

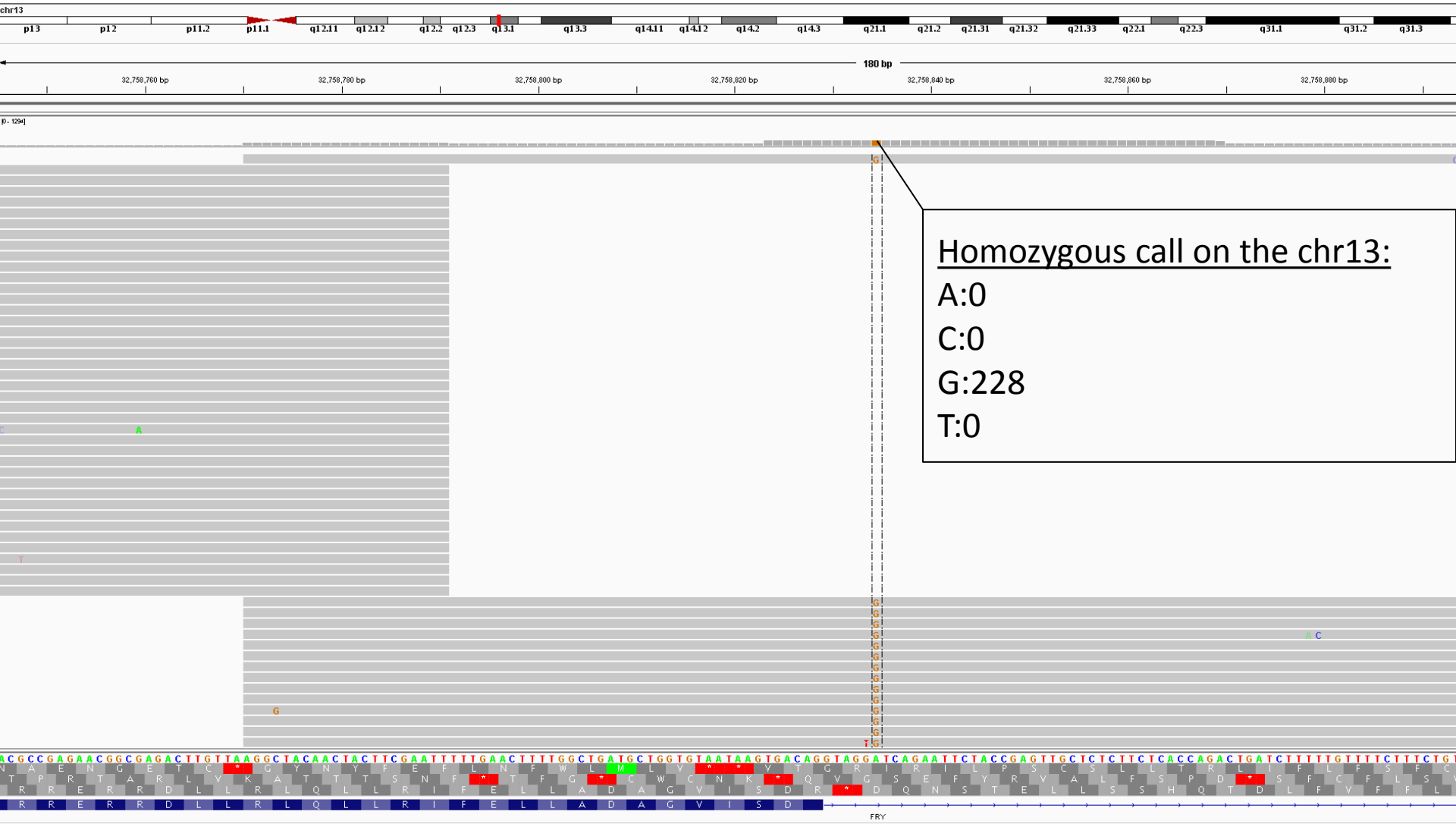
```
Homozygous mutations on chr1-22 and XX:
Reference:                  AAAAAAAAA
1st_copy Sequenced_chr1: AACAAAAAA
2nd_copy Sequenced_chr2: AACAAAAAA
```

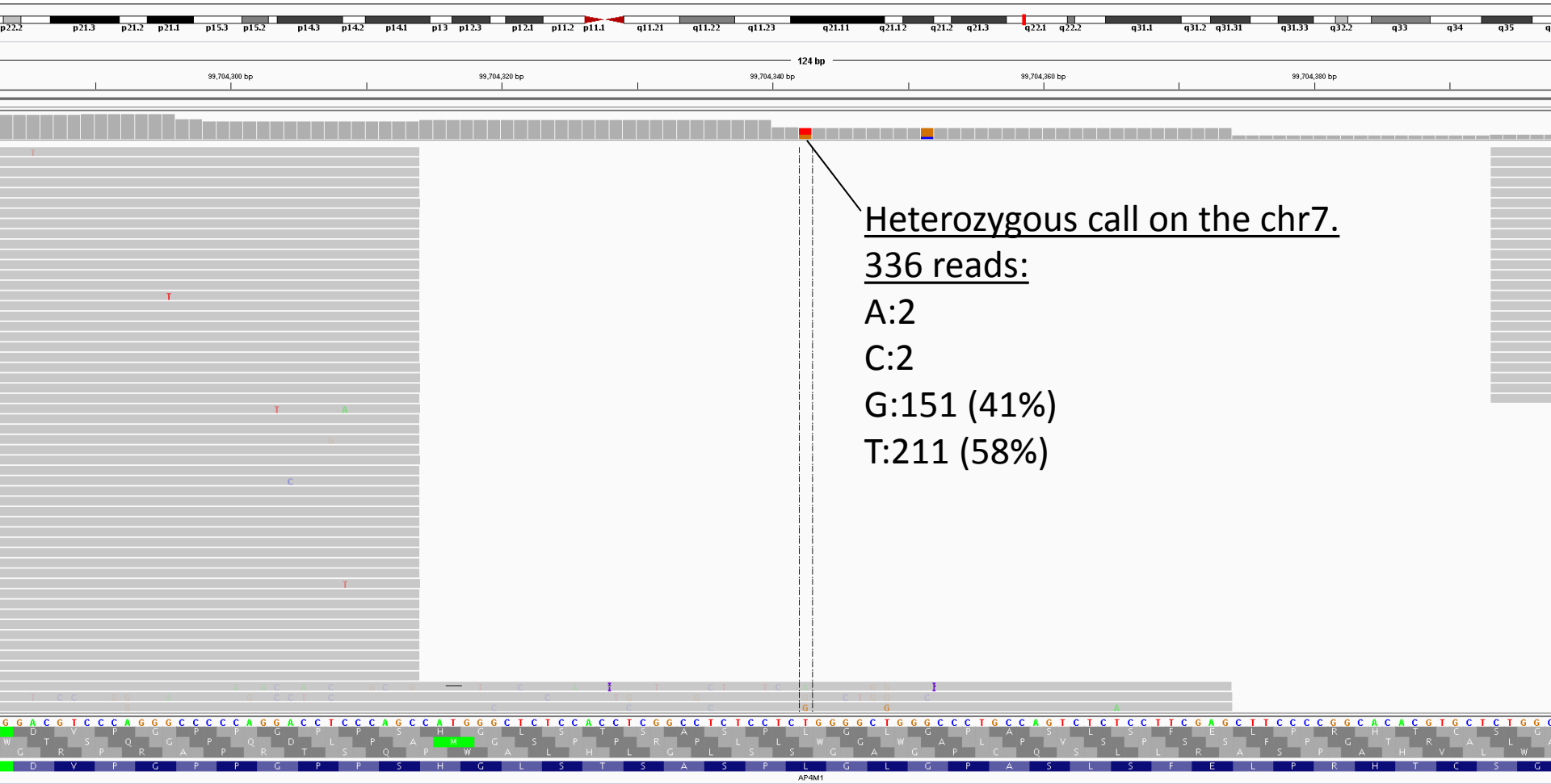
```
Heterozygous mutations on chr1-22 and XX:
Reference:                  AAAAAAAAA
1st_copy Sequenced_chr1: AACAAAAAA
2nd_copy Sequenced_chr2: AAAAAAAAA
```

```
Hemizygous mutations on XY (male: X and Y):
Reference:                  AAAAAAAAA
1st_copy Sequenced_chr1: AACAAAAAA
2nd_copy Sequenced_chr2: AACAAAAAA
```

Hemizygous call on the X from a male person

chrX:24,932,759-24,932,833 (hg18)


chr13
p13 p12 p11.2 p11.1 q12.11 q12.12 q12.2 q12.3 q13.1 q13.3 q14.11 q14.12 q14.2 q14.3 q21.1 q21.2 q21.31 q21.32 q21.33 q22.1 q22.3 q31.1 q31.2 q31.3
180 bp
32,758,760 bp
32,758,780 bp
32,758,800 bp
32,758,820 bp
32,758,840 bp
32,758,860 bp
32,758,880 bp
Homozygous call on the chr13:
A:0
C:0
G:228
T:0



Heterozygous call on the chr7.

336 reads:

A:2

C:2

G:151 (41%)

T:211 (58%)

Indel vs base exchange