Genomics and Transcriptomics

High-throughput methods

Eszter Ari
Dept. of Genetics
Eötvös Loránd Univ.
Budapest, Hungary
arieszter@gmail.com

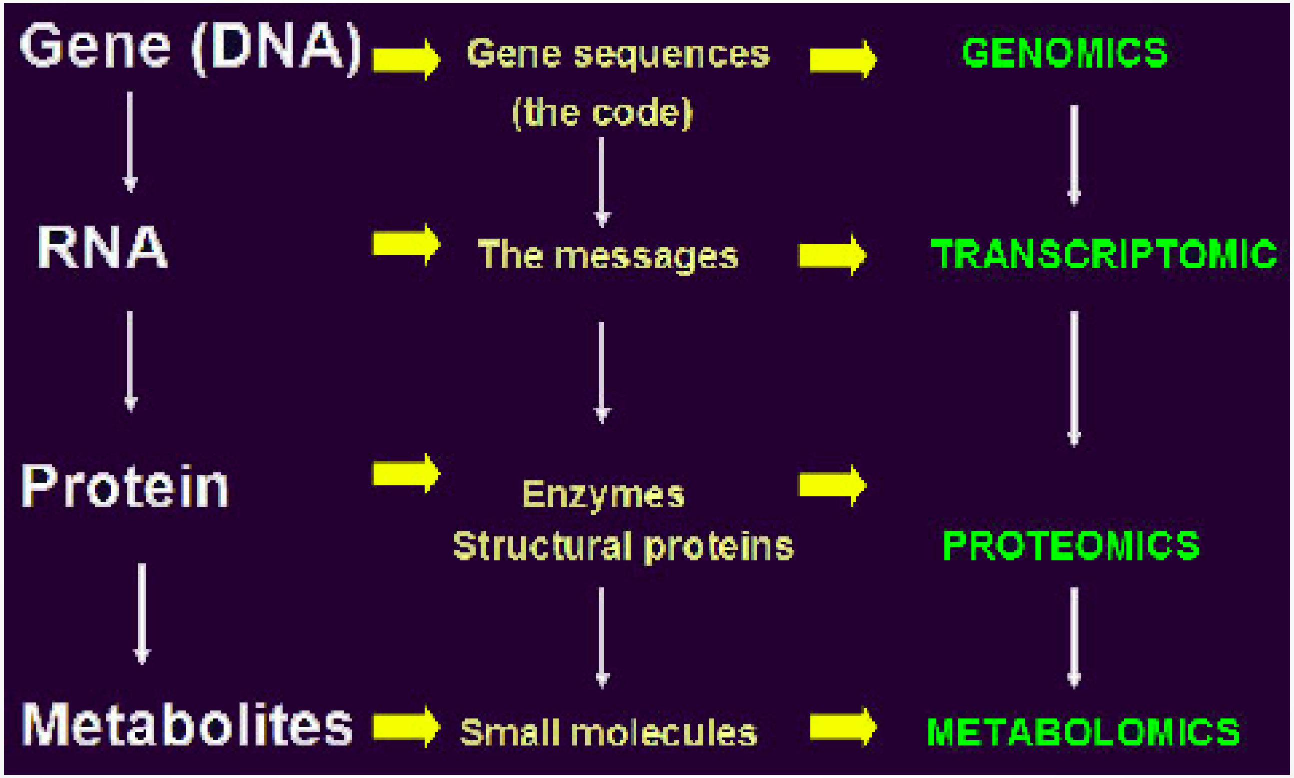
Thematics

- Genomics
  - Genomes, projects
  - Applications
  - Genome sequencing
    - de novo sequencing
    - re-sequencing
  - SNP analysis

Genomics

- Genom: complete set of genetic material within an organism
  - It is coded with DNA (or RNA in some viruses)
  - Genes and non-coding sequences
- Genomics investigates of
  - whole genomes
  - intercations between genes and non-coding regions
  - genome structures
  - gene locations
  - differences between genomes
- In contrast: genetics usually investigate functions of a single gene.
- Bioinformatics is massively needed to investigate genomes.

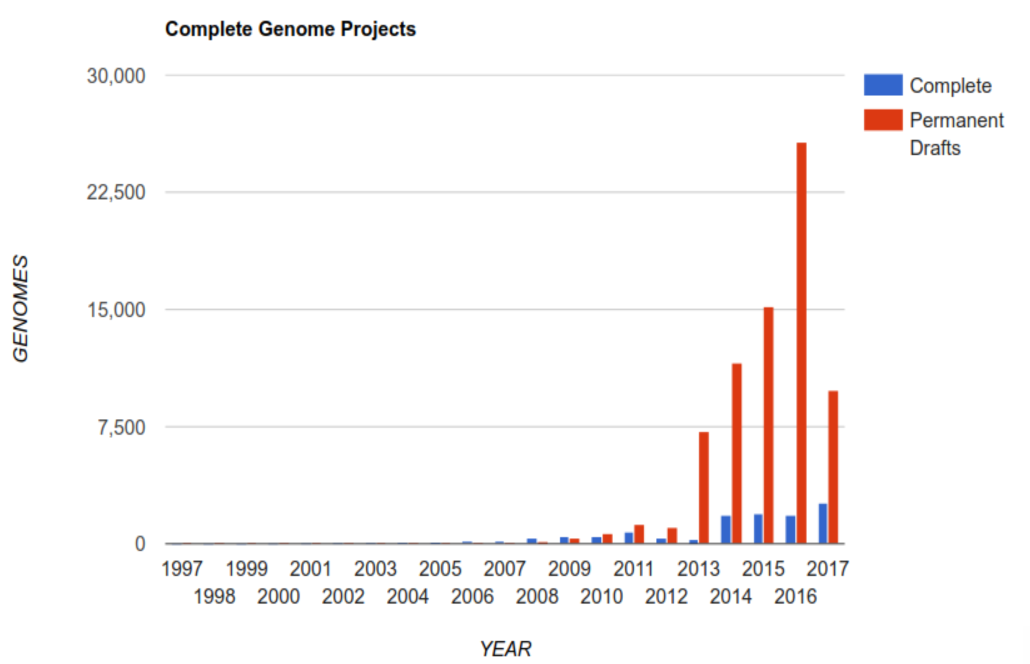
Omics

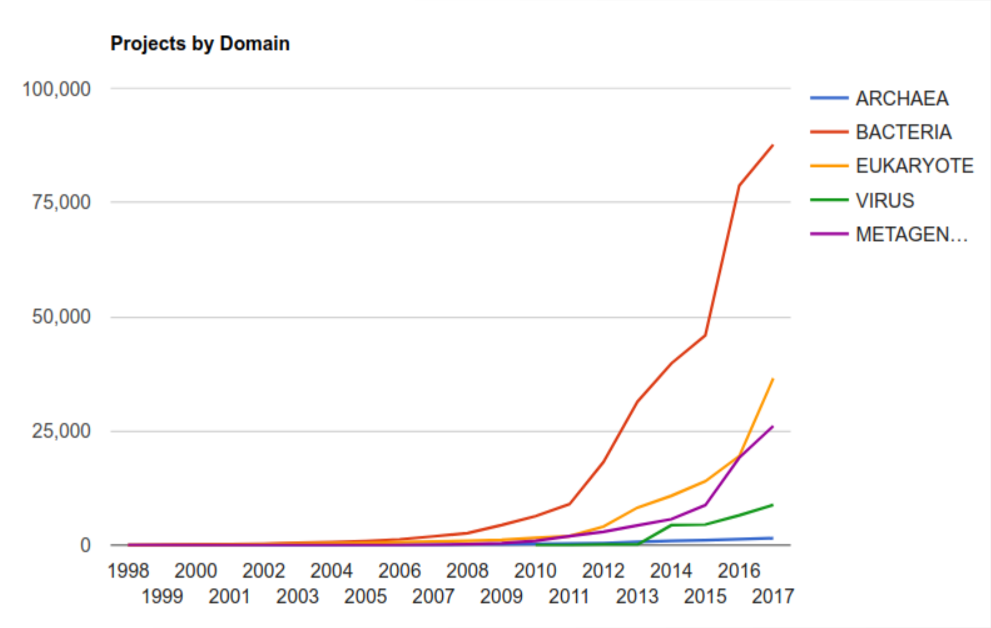


Gene (DNA)	Gene sequences (the code)	GENOMICS
RNA	The messages	TRANSCRIPTOMIC
Protein	Enzymes Structural proteins	PROTEOMICS
Metabolites	Small molecules	METABOLOMICS

Genome programs

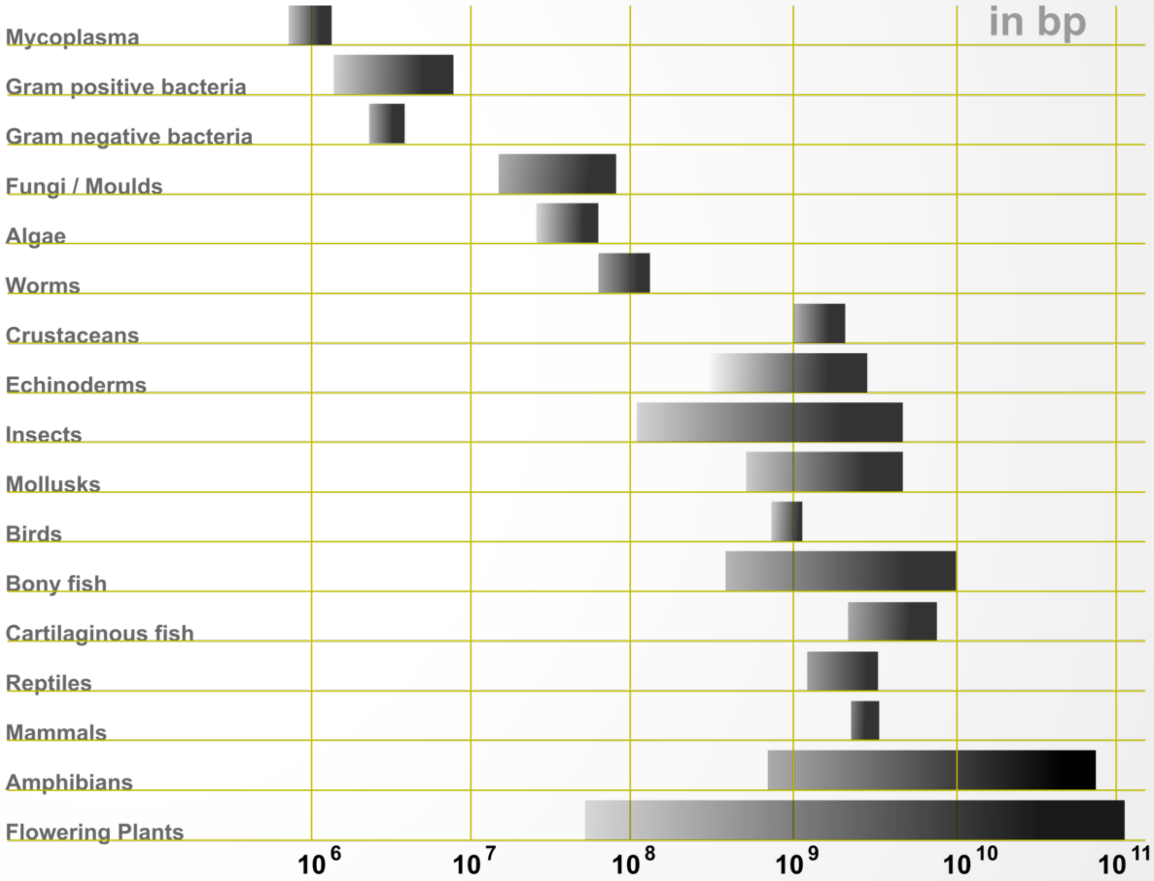
- GOLD, Genomes online database: https://gold.jgi.doe.gov/



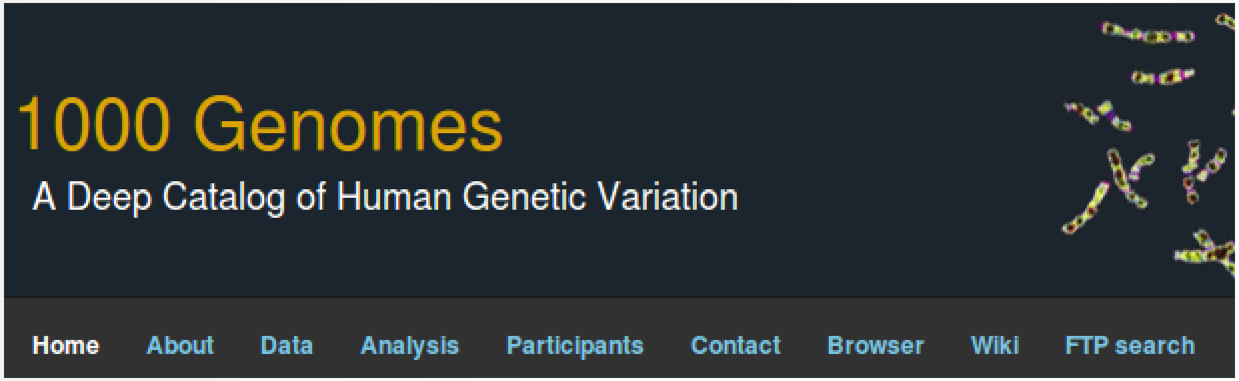


Genome size

- Virus
  (2 kb - 700 kb, kilobase = 1000 nt)
  - 1-2 stranded DNA or RNA
    - First sequenced genome:
      Phi-X174 phage, Fred Sanger, 1977
- Bacteria (139 kb - 13.000 kb)
  Archea (500 kb - 5.700 kb)
  - 2 stranded haploid chromosomes
    - plasmides
- Eucarya (8,2 Mb - 220.000 Mb,
  megabase = 1.000.000 nt)
  - diploid chromosomes - nuclear
  - Organelles with genome:
    mitochondria (16,6 kb)
    chloroplast (120 kb - 170 kb)
    - Human genome:
      June 2000 – Feb 2001



Genome programs



http://www.1000genomes.org/

http://www.uk10k.org/

- Aims of Beijing genomics Institute (BGI, China) sequencing center: million human genomes, million microbe genomes, million plant and animal genomes
  - The Million Human Genome Project
- 100,000 foodborne pathogen genome project
- Up to 100,000 NHS patients - human
- 50,000 Faroe Islanders Project - human
- 20,000 Global pneumococcal project - human
- 10,000 Genome 10k vertebrate sequencing project
- 10,000 autism genome projekt - human
- 5,000 arthropod genome sequencing project
- ...

Applications

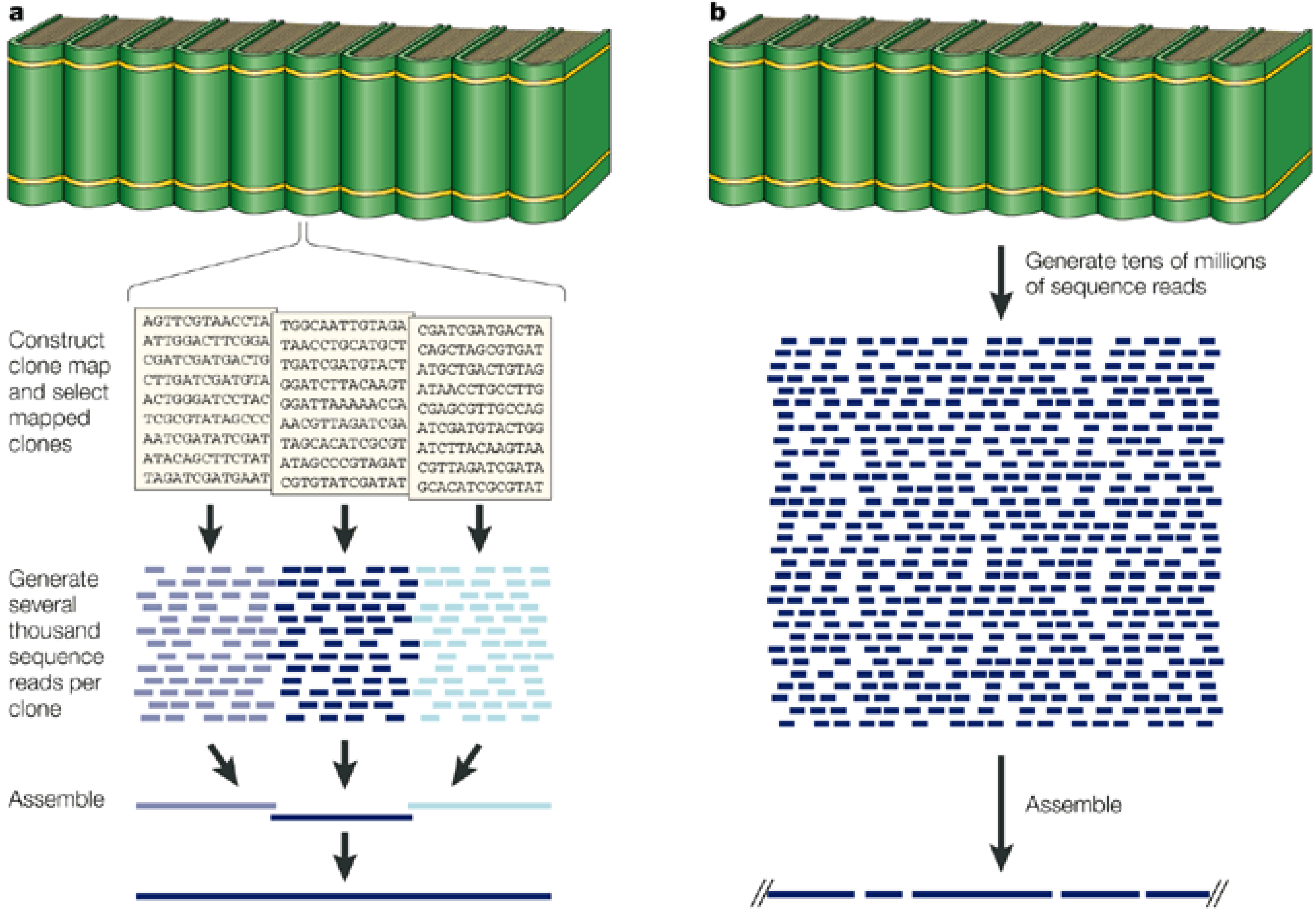
- Genetics
  - ie.: gene locations, environment, regulation, recombination hot-spots
- Populationgenetics
  - ie: explore the history of a population using SNP frequencies
- Evolutiongenetics
  - ie: investigate which part of the genome is under selection
  - phylogenomics
- Paleontology
- Medicine
  - diagnostics
  - Personal therapy, ie: genetherapy
  - ie. cancaer research
- Drug developement
- Agriculture (GMO)
- Food industry
- Forencinc science

How do we get the data?

Genome sequencing: in the past and today

- Different strategies for genome sequencing:
  - In the past:
    - Clone based hierarchical sequencing (BAC – bacterial artificial chromosome – libraries)
    - Whole genom shutgun sequencing
  - Today:
    - Massively paralell Next Generation Sequencing (NGS)

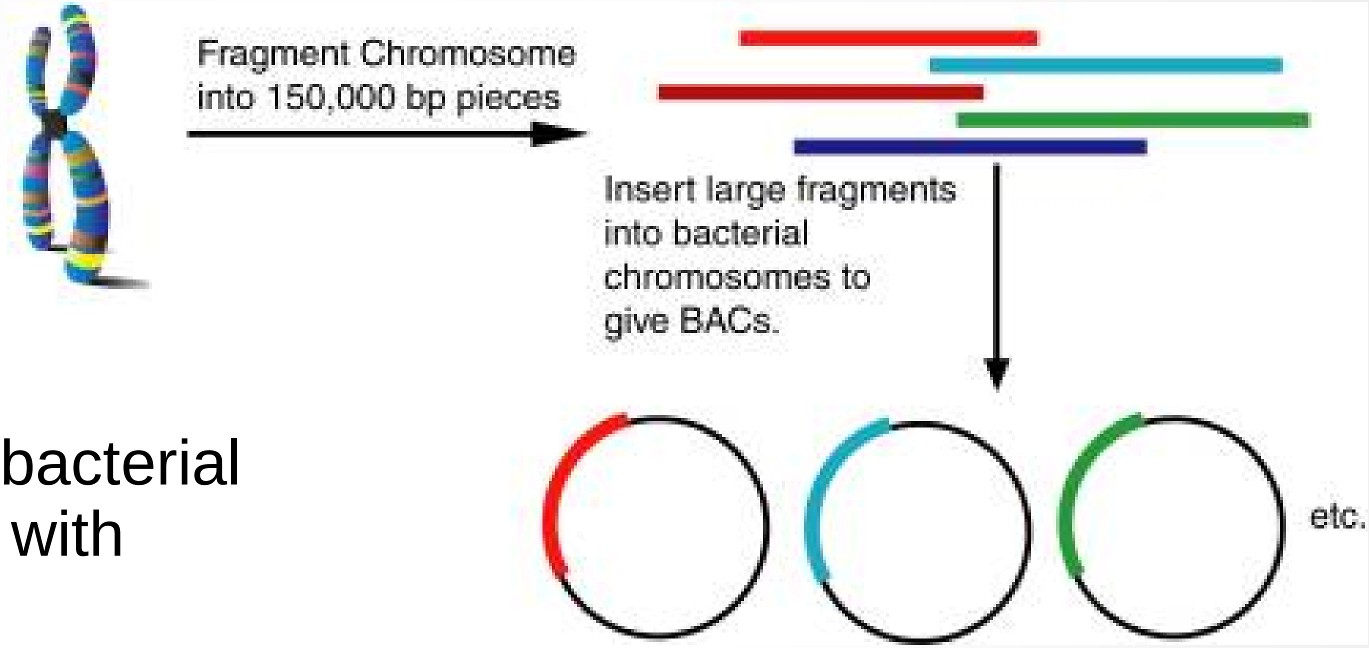
Clone-by-clone vs. whole genome shotgun



Nature Reviews | Genetics

Green ED (2001) Strategies for the systematic sequencing of complex genomes. Nat Rev Genet 2: 573–583

Clone based hierarchical sequencing (BAC to BAC)

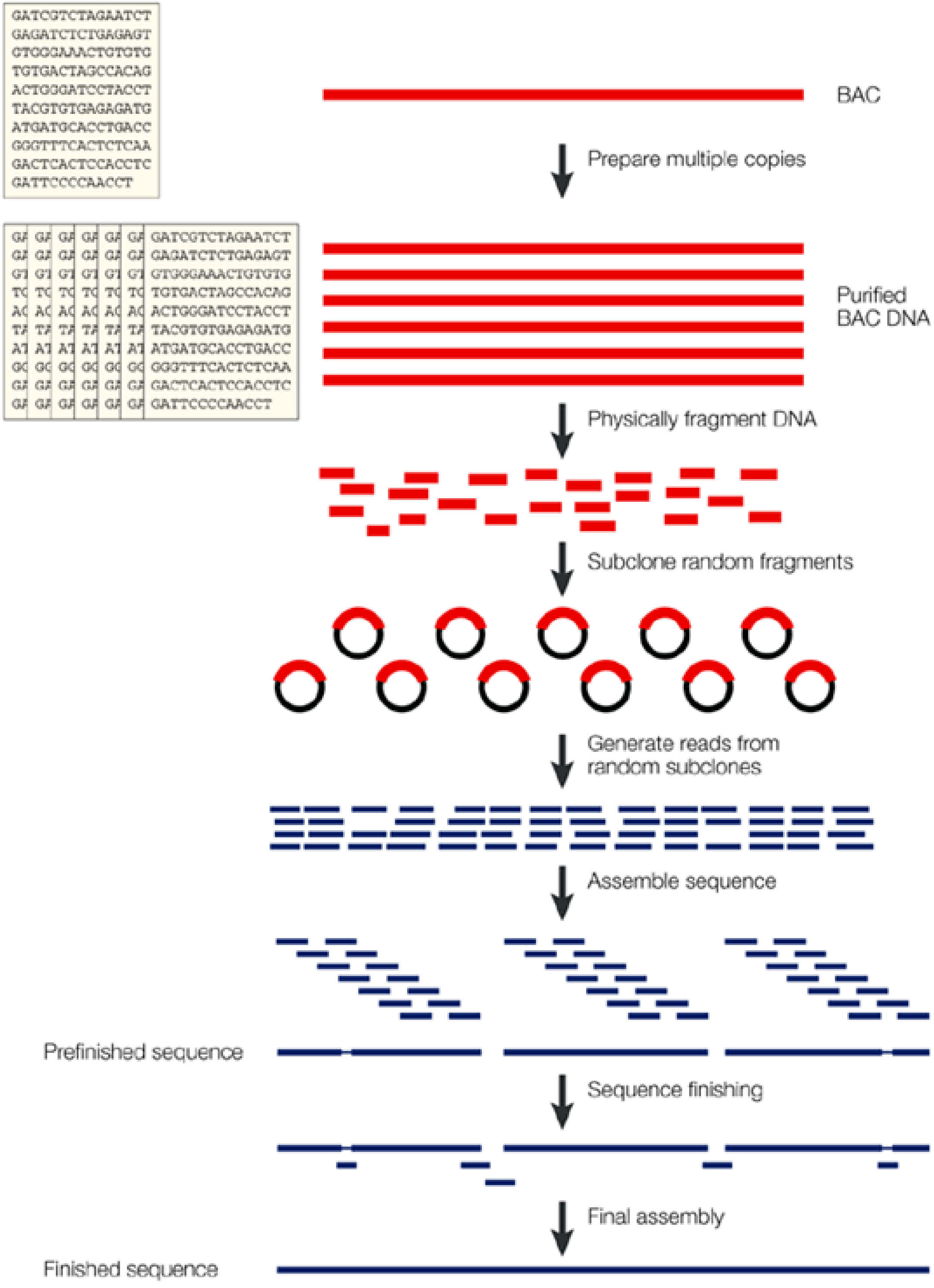

Fragment Chromosome into 150,000 bp pieces
Insert large fragments into bacterial chromosomes to give BACs.
etc.

Sequencing viral and bacterial genomes was started with clone based method.

- The whole genomes were cutted to ~40 - 150 kb overlapping pieces
  - Genomic location of each piece were determined (ie. Using unique STS sties or FISH)
- Cloning – amplification (*E. coli*, BAC - Bacterial Artificial Chromosome - contigs)
  - BAC library: contains the whole genome of a species

Clone based hierarchical sequencing

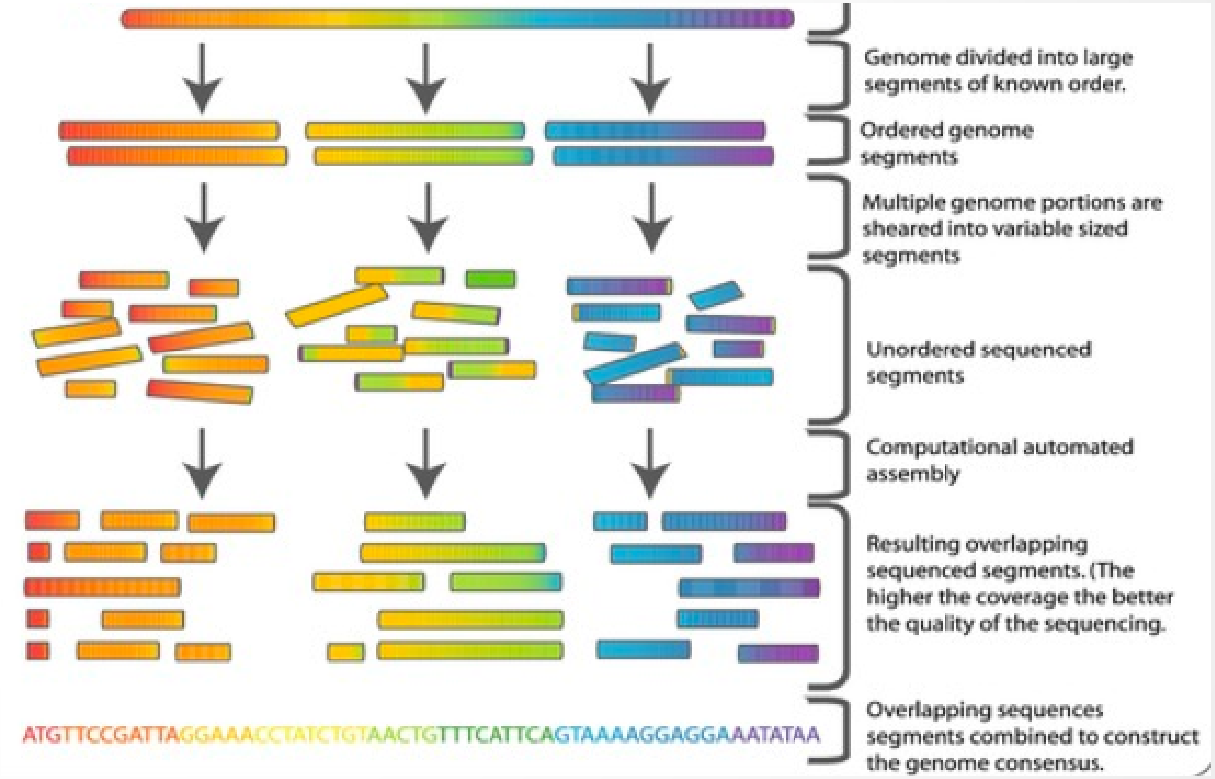
- Amplification
- Fragmentation
- Amplification: subclone libraries
- Reads from subclones



Nature Reviews | Genetics

Clone based hierarchical sequencing

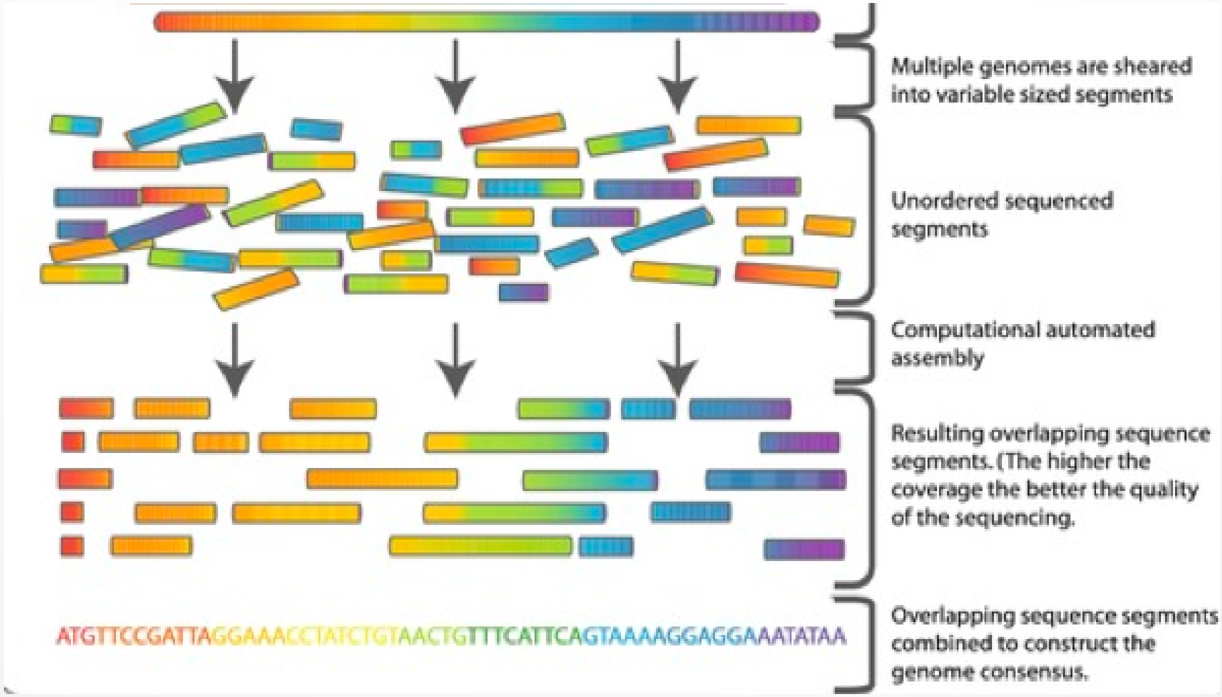
- – Sanger sequencing
- – Base calling:
  - Quality scores: PHRED
- – Bioinformatics: genome assembly
  - PHARP software
  - Assembly the order of nucleotides of the BAC contigs based on the reads
  - Assembly the whole genome based on BAC contigs


Genome divided into large segments of known order.
Ordered genome segments
Multiple genome portions are sheared into variable sized segments
Unordered sequenced segments
Computational automated assembly
Resulting overlapping sequenced segments. (The higher the coverage the better the quality of the sequencing.
Overlapping sequences segments combined to construct the genome consensus.
ATGTTCCGATTAGGAAACCTATCTGTAACTGTTTCATTCAGTAAAAGGAGGAAATATAA

http://www.phrap.org/phredphrapconsed.html

Whole genome shotgun sequencing - recent

- „Shutgun” breaking-up the whole genome (i.e pass through in a capillar)
  - 2 - 10 kilobase
  - Sequencing the pieces
- Assembly using computer
  - TIGR Assembler – first whole genome assembler software


Multiple genomes are sheared into variable sized segments
Unordered sequenced segments
Computational automated assembly
Resulting overlapping sequence segments. (The higher the coverage the better the quality of the sequencing.
ATGTTCCGATTAGGAAACCTATCTGTAACTGTTTCATTCAGTAAAAGGAGGAAATATAA
Overlapping sequence segments combined to construct the genome consensus.

Celera
Craig Venter
1996

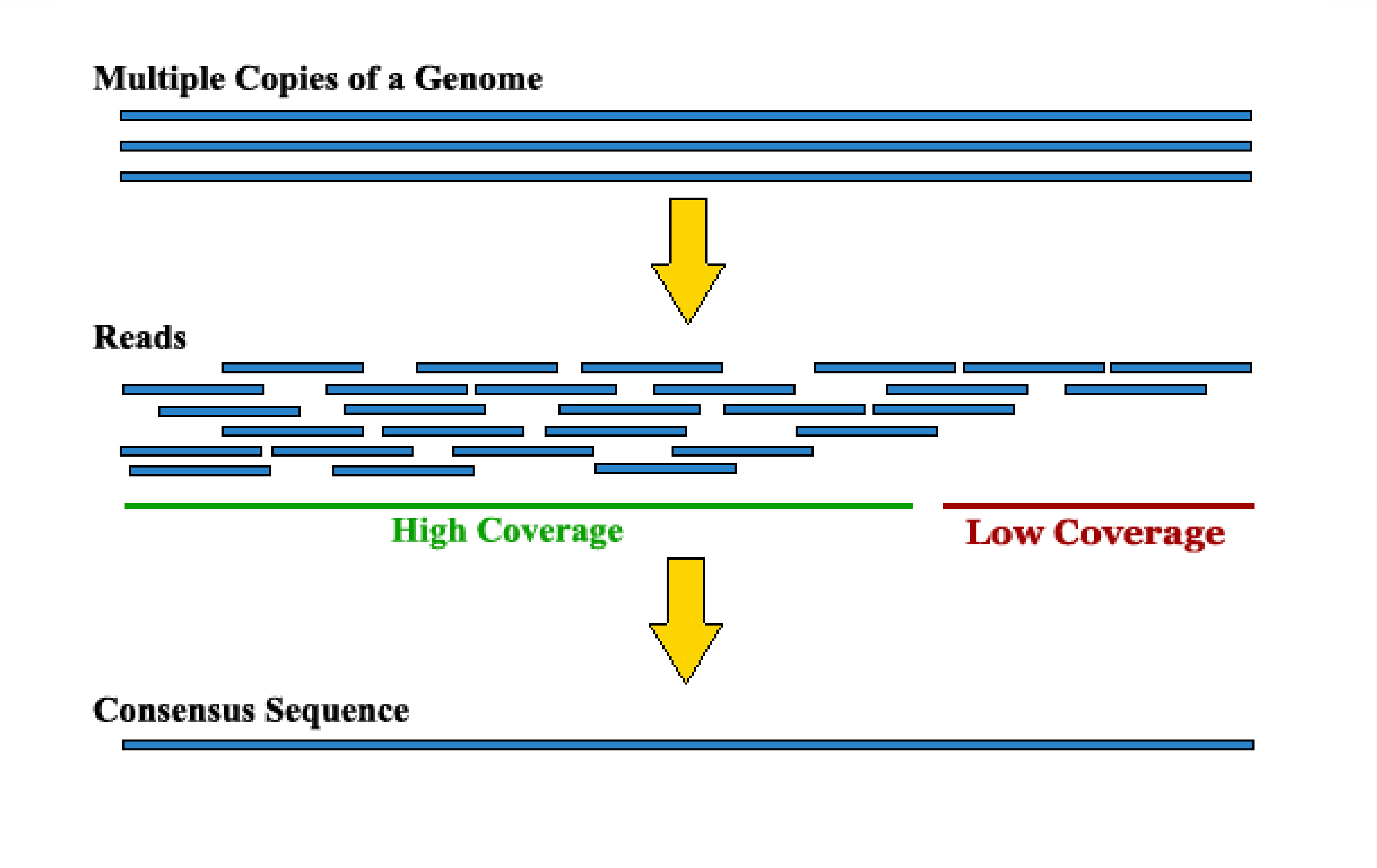
Comparison

- Clone based sequencing
  - Less chance to make errors during assembly
  - We know the place of the comtigs for sure
  - Time consuming
  - Expensive
  - Needs less computations: dealing with 100-200 Kb data at the same time

- Whole genom shutgun
  - More chance to make errors during assembly
  - We do not know the place of the comtigs
  - Fast
  - Less expensive
  - Computationally intensive: dealing with more Gb data at the same time

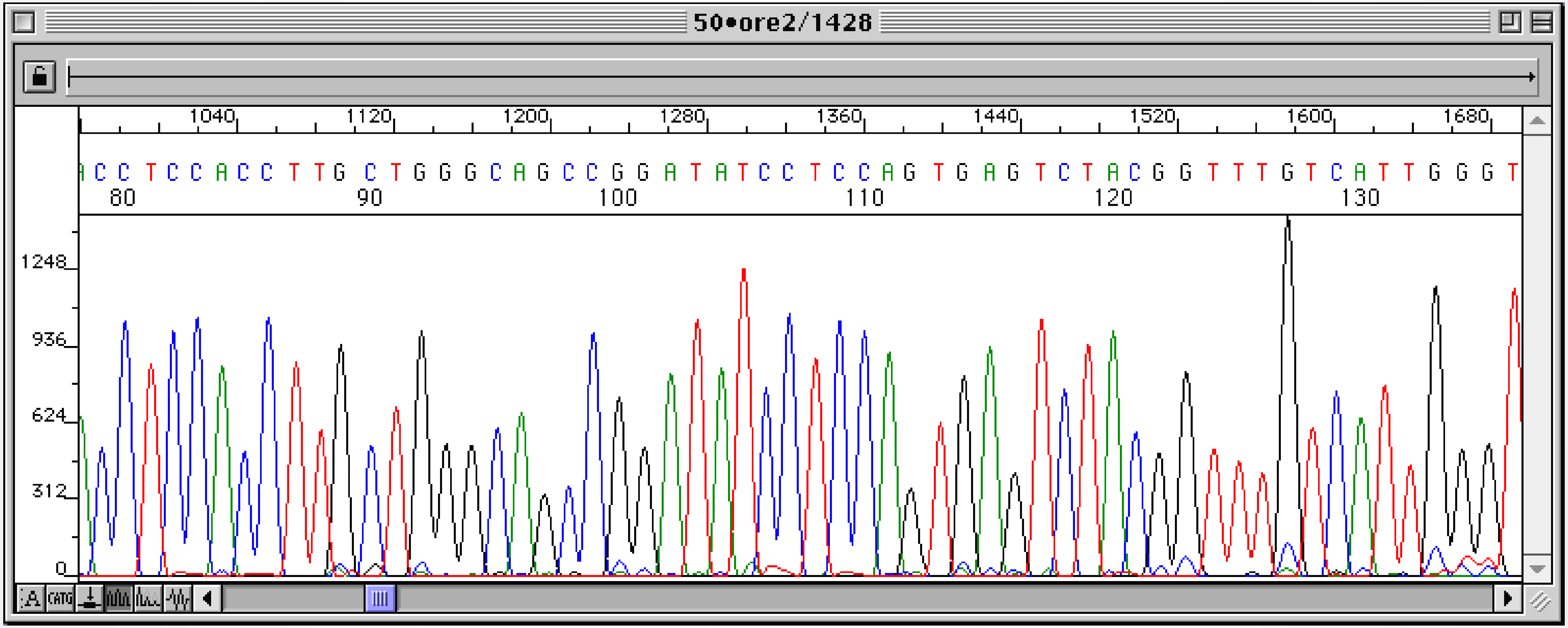
High coverage is needed

Coverage


Multiple Copies of a Genome
Reads
High Coverage
Low Coverage
Consensus Sequence

Chain-terminating Sanger sequencing

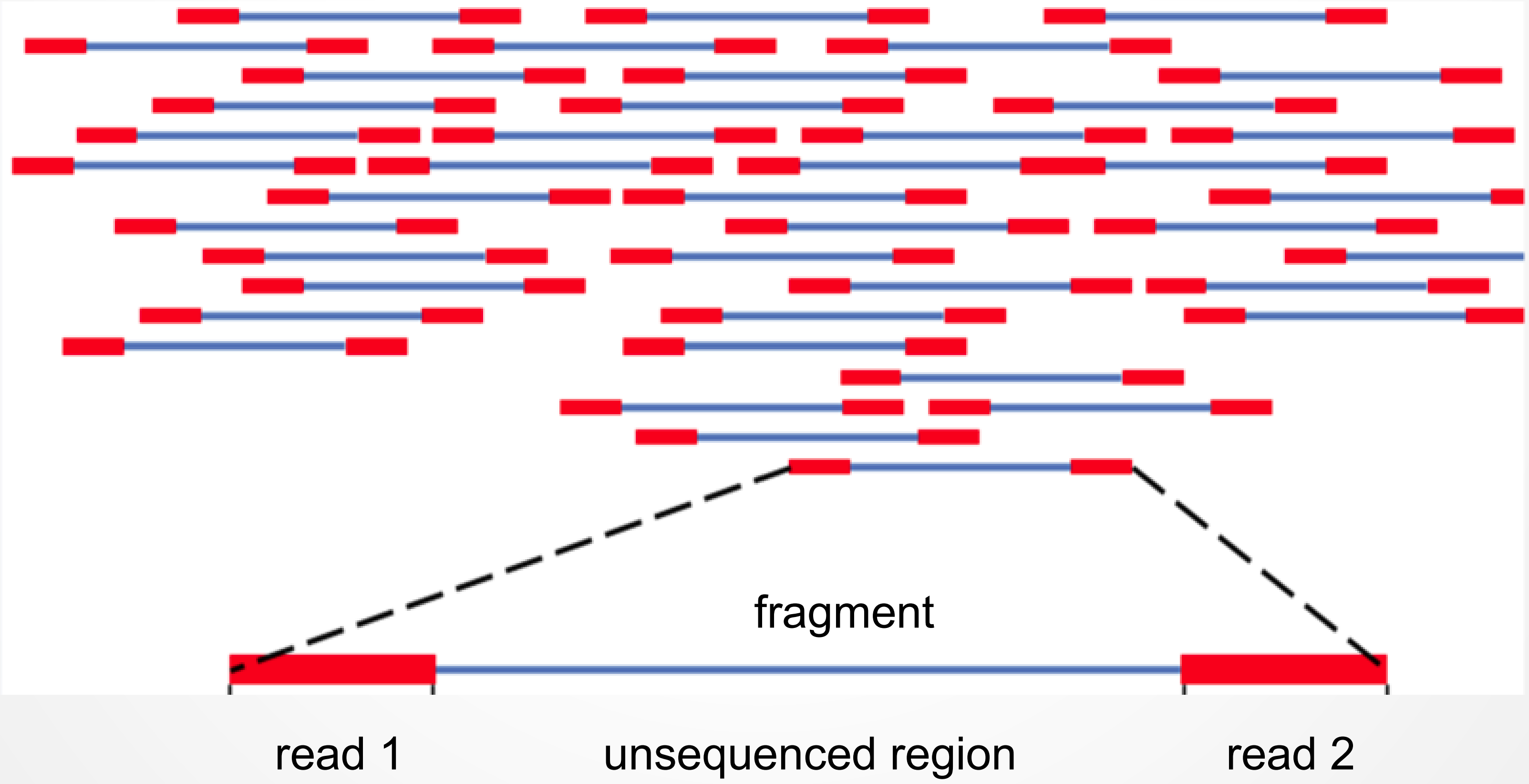
The dideoxynucleotides are fluorescently labeled for detection in automated sequencing machines. → Electropherogram


50•ore2/1428
1040
1120
1200
1280
1360
1440
1520
1600
1680
A C C T C C A C C T T G C T G G G C A G C C G G A T A T C C T C C A G T G A G T C T A C G G T T T G T C A T T G G G T
80
90
100
110
120
130
1248
936
624
312
0
A
CATG

Read length: 900-1000 nucleotides

Next Generation Sequencing - NGS

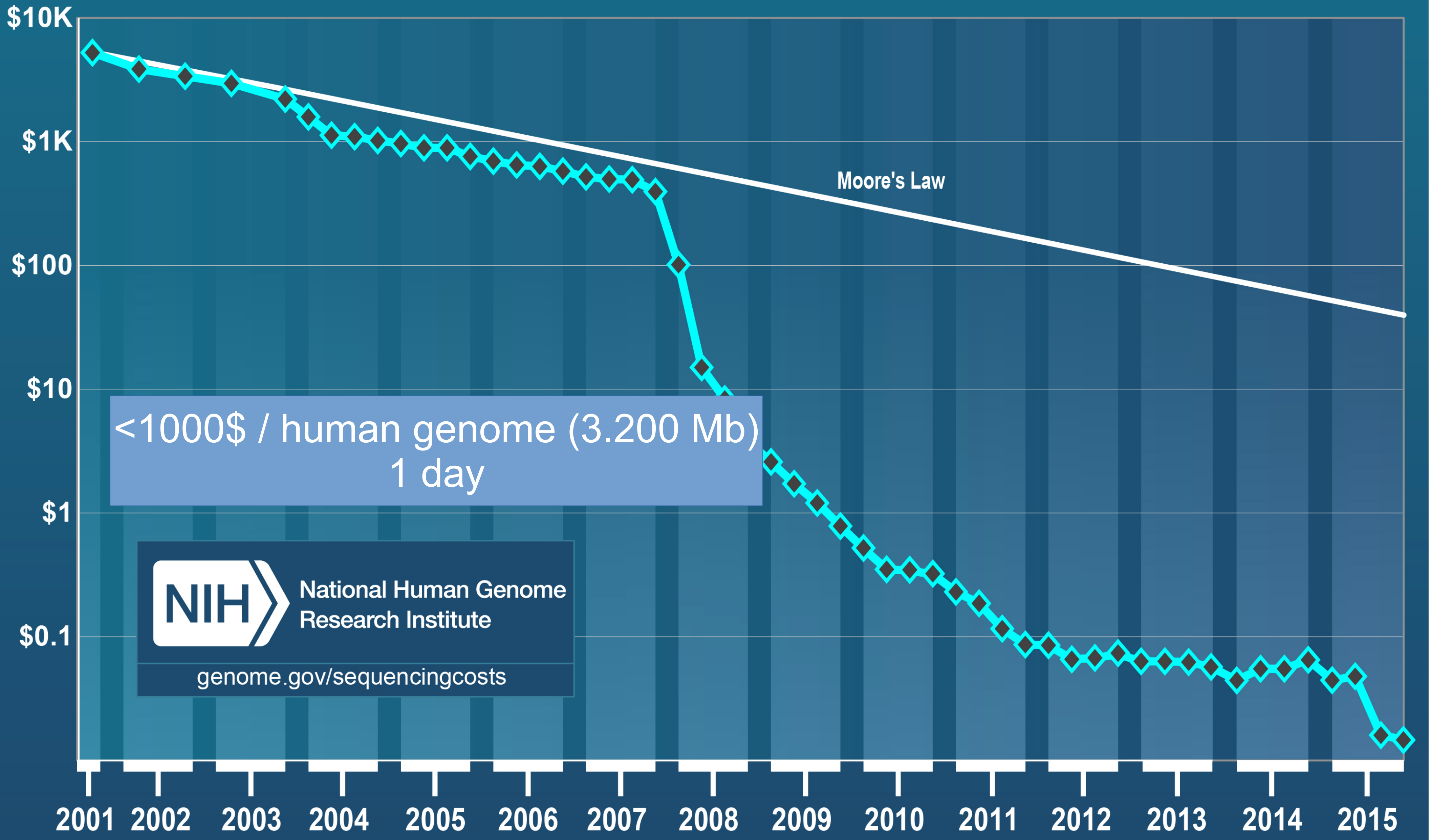
- **High throughput (highly parallel), sequencing a lot of regions at the same time → fast, cheap**
- Sequencing the beginning (single end sequencing), or the beginning and the end (paired end seq.) of fragmnets.
- Sequencing 1 million DNA fragments at the same time



Next Generation Sequencing - NGS

- Could be strand specific (forward, reverse)
- Methods (not based on Sanger sequencing):
  - Illumina (Solexa) sequencing
  - SOLiD sequencing
  - Ion Torrent sequencing
  - Pyrosequencing (454)
  - PacBio
  - Oxford nanopore
  - ...
- Read lengths: 50-700-thousands nt
- Million reads per day
  - Cost: 5 cent ~ 1 $ / 1.000.000 nt
- Sequnecing is fast (Human genome: a day), but the assembly is complicated and computationally intense

Cost per Raw Megabase of DNA Sequence

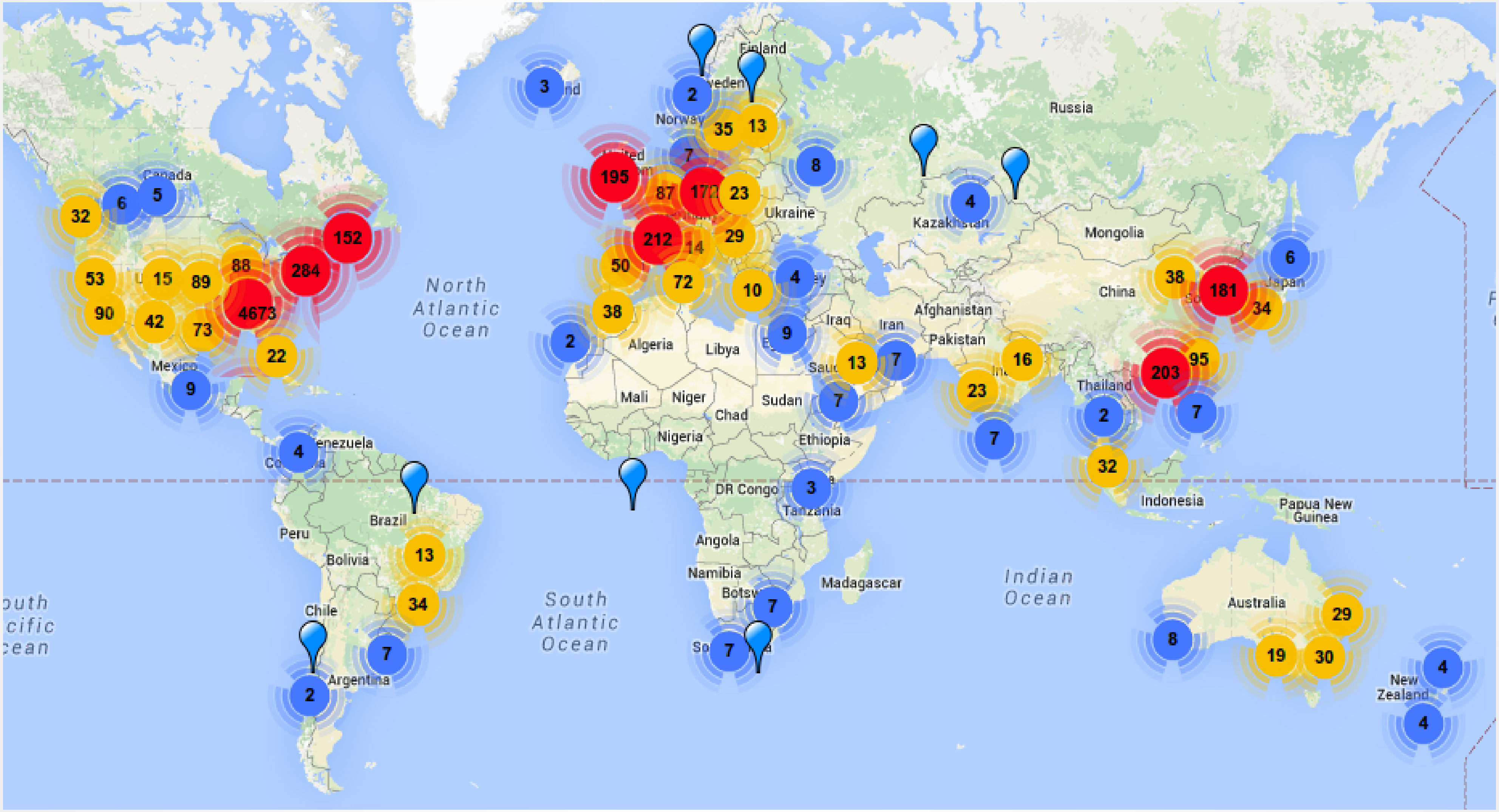


<1000$ / human genome (3.200 Mb)
1 day

NIH National Human Genome Research Institute

genome.gov/sequencingcosts

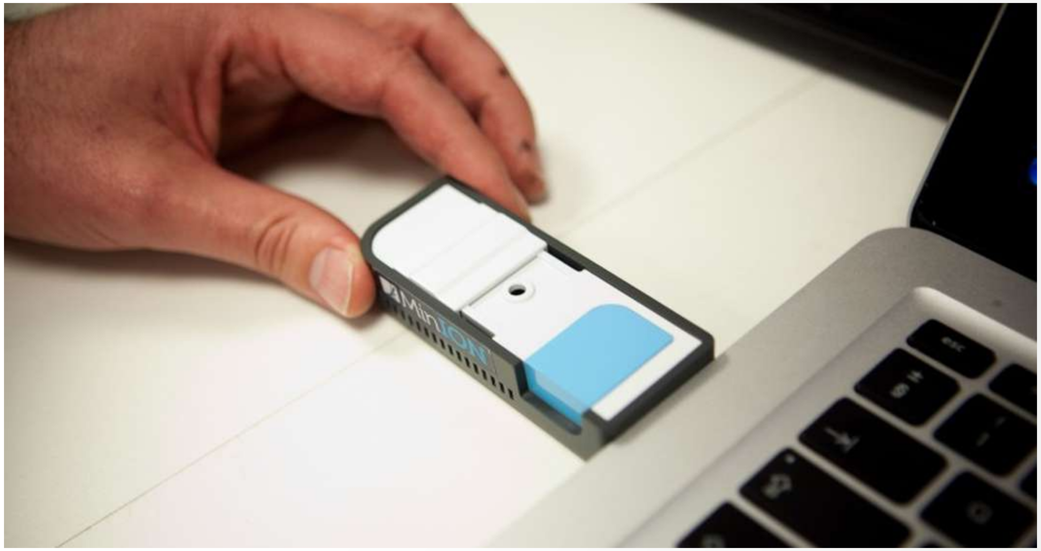
High-throughput sequencing instruments world-wide (2015)



http://omicsmaps.com/

454, HiSeq, Illumina GA2, Ion Torrent, MiSeq, PacBio, Polonator, Proton, SOLiD

NGS instruments



Illumina HiSeq

MinION

Illumina sequencing

Video
https://www.youtube.com/watch?v=HMyCqWhwB8E

Steps of genome analysis

1. Quality checking

2. Trimming: filter out low quality reads (or read parts)

3.a) Newly sequenced genome: *de novo* assembly

3.b) Genome re-sequencing: mapping

4. Unfold genetic diversity: statistical analysis

Steps of genome analysis

1. Quality checking
2. Trimming: filter out low quality reads (or read parts)

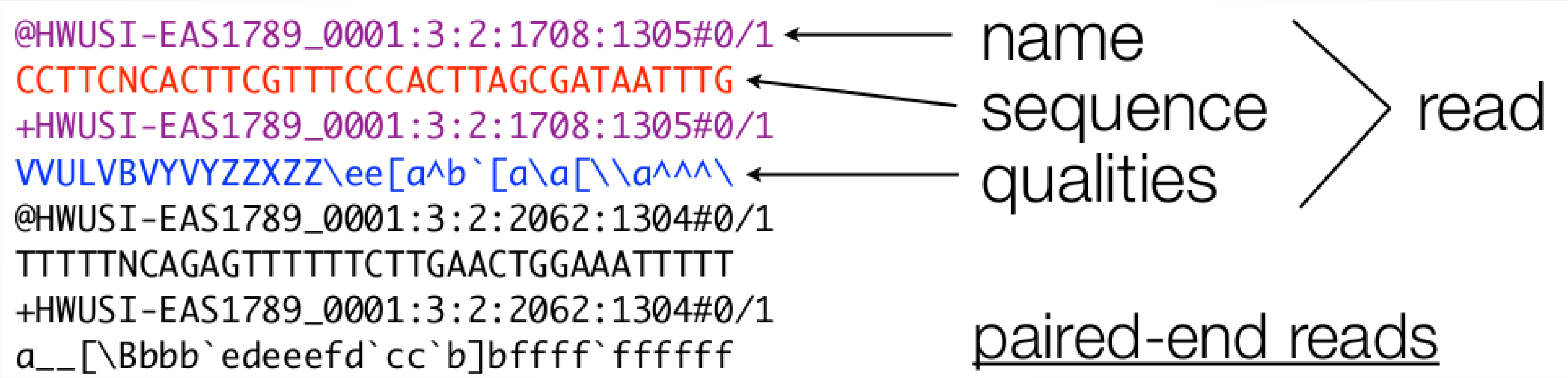
3.a) Newly sequenced genome: *de novo* assembly

3.b) Genome re-sequencing: mapping

4. Unfold genetic diversity: statistical analysis

The reads

- Result of NGS: ie. fastQ file
  - quality checking (ie.: FastQC software)
  - trimming: filter out low quality reads (or read parts)



```
@HWUSI-EAS1789_0001:3:2:1708:1305#0/1
CCTTCNCACTTCGTTTCCCACTTAGCGATAATTTG
+HWUSI-EAS1789_0001:3:2:1708:1305#0/1
VVULVBVYVYZZXZZ\ee[a^b`[a\a[\\a^^^\
@HWUSI-EAS1789_0001:3:2:2062:1304#0/1
TTTTTNCAGAGTTTTTTCTTGAACTGGAAATTTTT
+HWUSI-EAS1789_0001:3:2:2062:1304#0/1
a__[\Bbbb`edeeefd`cc`b]bffff`ffffff
```

Steps of genome analysis

1. Quality checking
2. Trimming: filter out low quality reads (or read parts)

3.a) Newly sequenced genome: *de novo* assembly

3.b) Genome re-sequencing: mapping

4. Unfold genetic diversity: statistical analysis

De-novo genome assembly

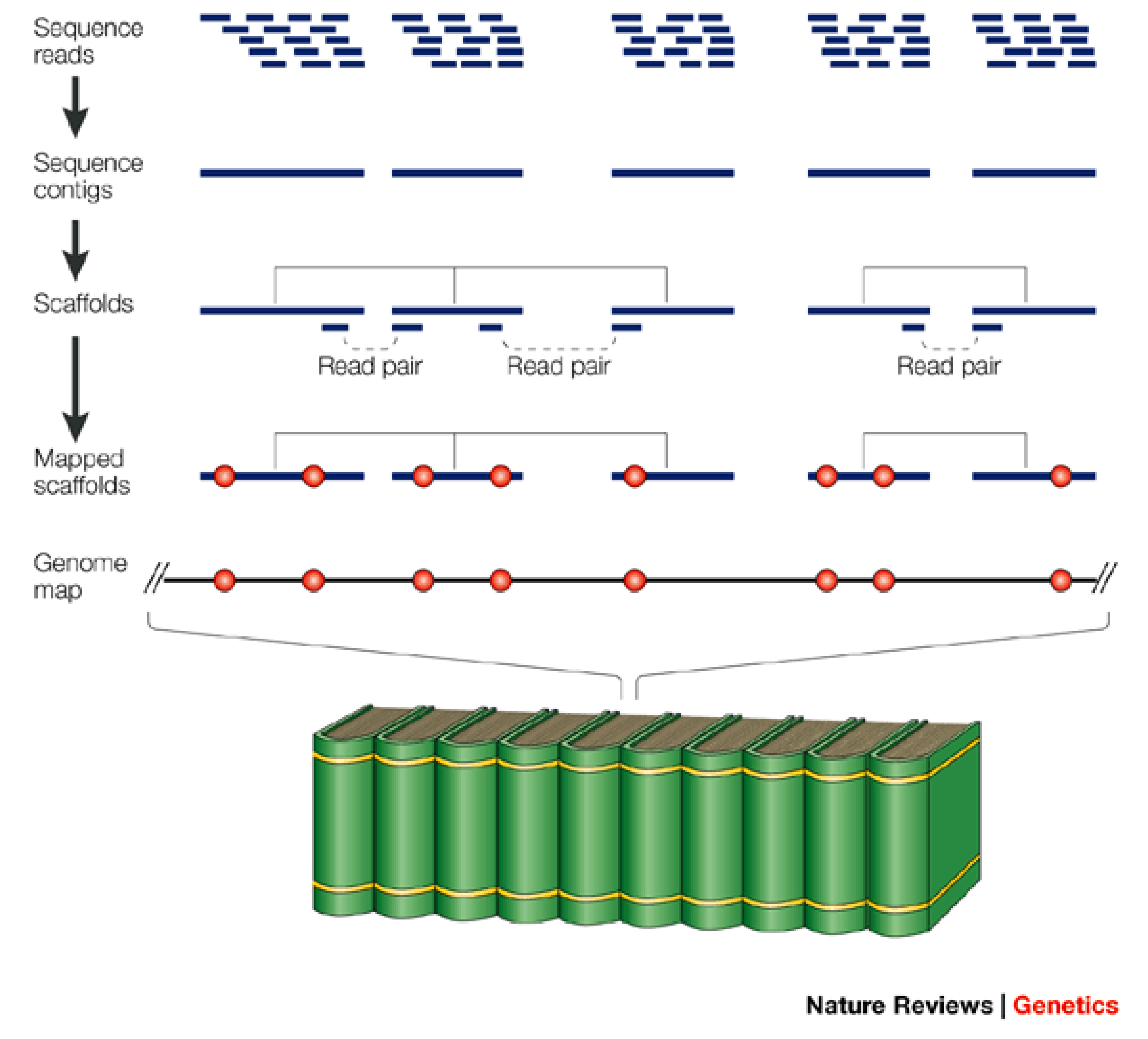
- Construction of the whole genome sequence based on reads
  - Among Eucariotes the fruit fly genome was the first which was assembled by purly this method
  - Human genome: 2-3 billion reads (100X coverage)
- Gready algorithm:
  1. Pairwise alignment of all possible read pairs (based on sequence similarity)
  2. Merging the 2 reads that are the most similary – overlap the most
  3. Repeat step 2 till there are single reads
- Assembler softwares: ABySS, Celera WGA, Edna, Euler, MIRA, Newbler, SOAPdenovo, …
- Problem: we cannot check if the assembly was correcte – if the genome was newly sequenced
  - Causes of an incorrect assembly:
    - Repetitive regions – we should exclude these
    - Reads that aligned to a wrong place and/or in a wrong orientation

Aligning and non-aligning reads

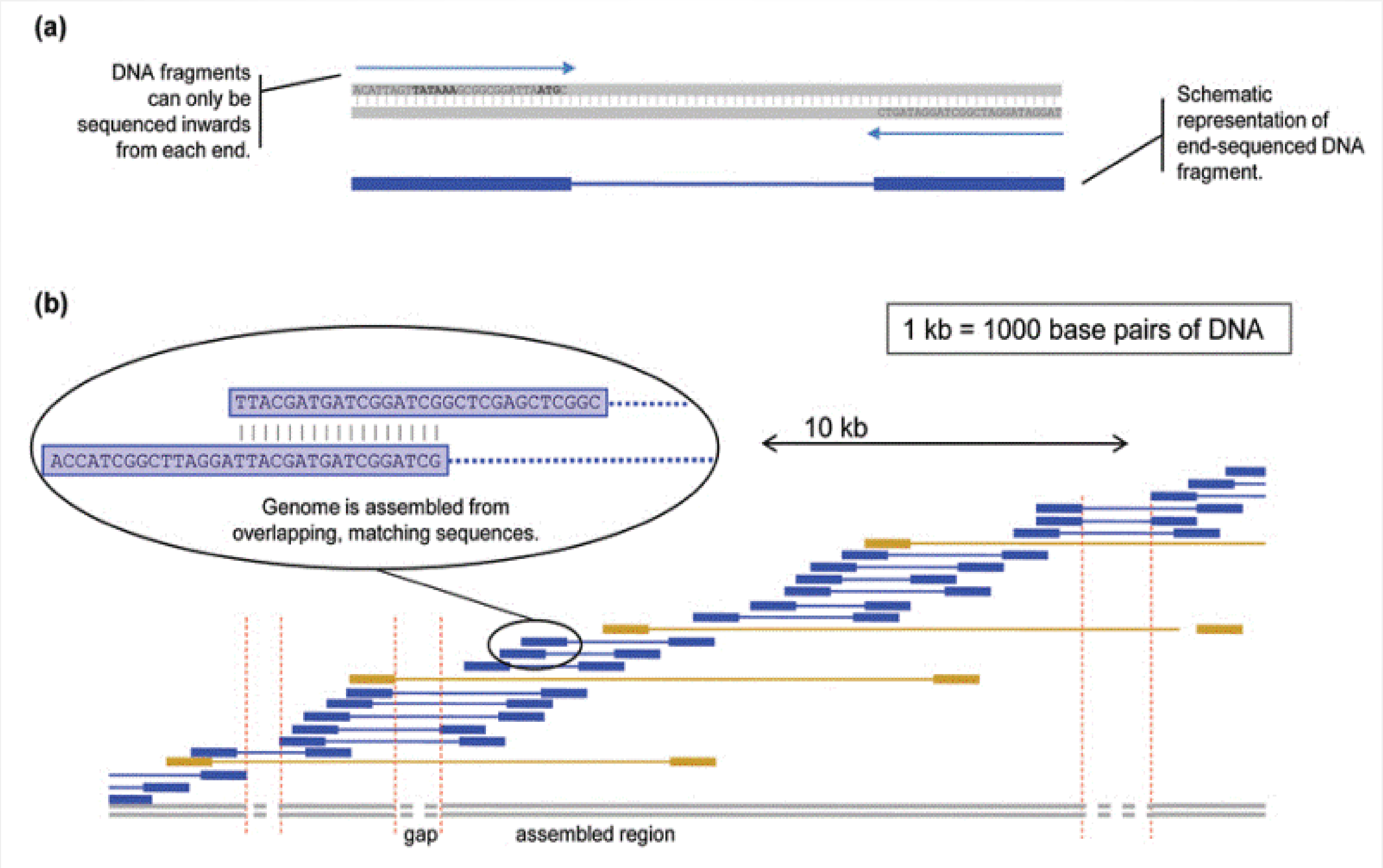
```
ATTGTGCTAGTCGTAGCTAGCT
      ||||||||||||||||
      CTAGTCGTAGCTAGCTGTCAA


TGATGATGCTCTAAGATCTCAT
```

Genome assebly



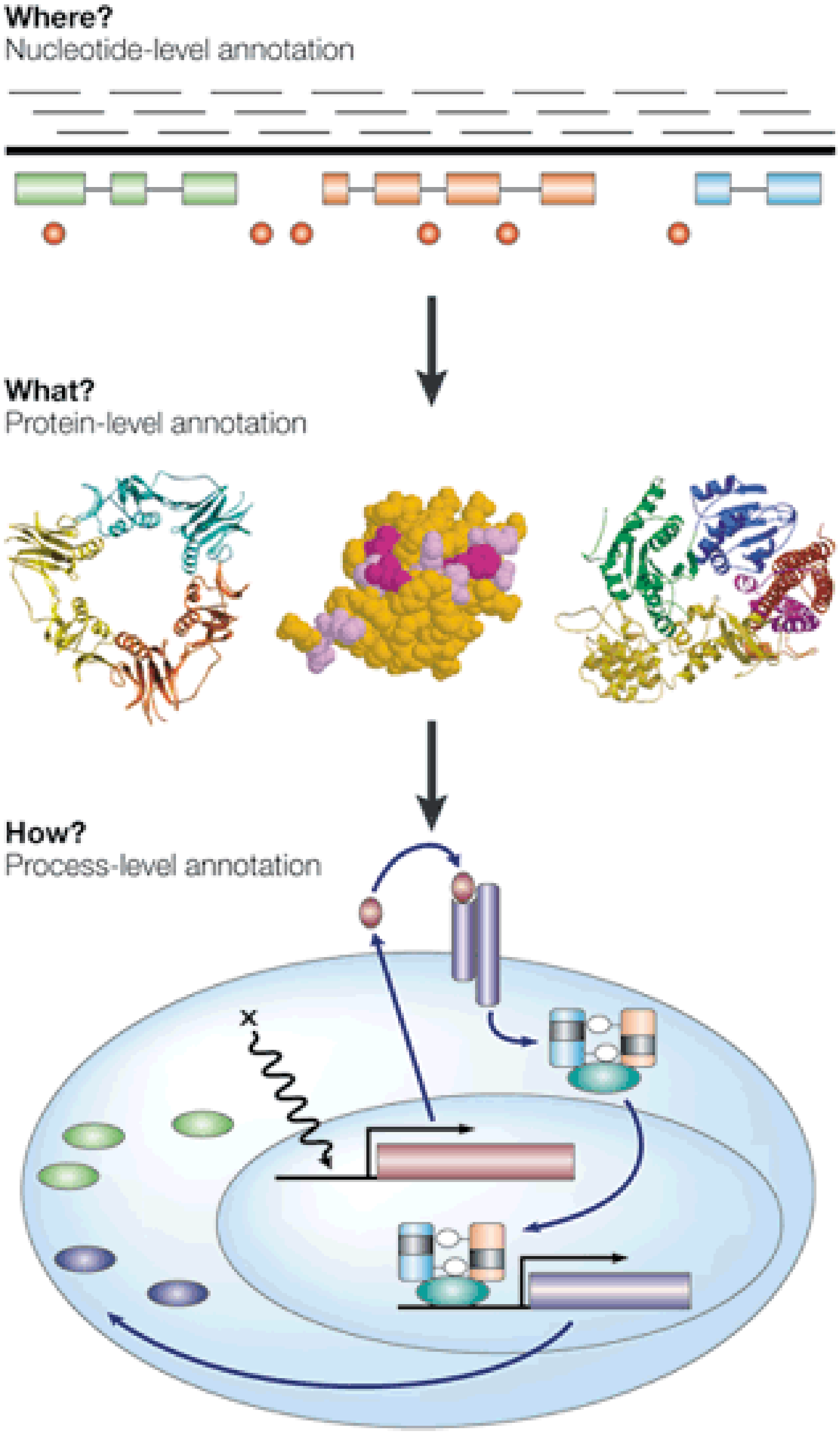
Genome assembly


(a)
DNA fragments can only be sequenced inwards from each end.
ACATTAGTTATAAAGCGGCGGATTAATGC
CTGATAGGATCGGCTAGGATAGGAT
Schematic representation of end-sequenced DNA fragment.
(b)
1 kb = 1000 base pairs of DNA
TTACGATGATCGGATCGGCTCGAGCTCGGC
ACCATCGGCTTAGGATTACGATGATCGGATCG
Genome is assembled from overlapping, matching sequences.
10 kb
gap
assembled region

Assemlby using different fragment sizes: blue - 5kb, yellow - 20kb

Genome annotation

- The process of finding and designating locations of individual genes and other features on raw DNA sequences
- Structural annotation:
  - Searching for ORFs
  - Gene structures (UTF, exon, intron...)
  - Promoter regions: based on motifs
- Functional annotation:
  - Biological functions of the ORFs (genes), ie. BLAST search
  - Gene expression data
  - Regulation networks...
- Annotation projects:
  - ENCyclopedia Of DNA Elements (ENCODE), Entrez Gene, Ensembl, GENCODE, Gene Ontology Consortium, GeneRIF, Uniprot, Vertebrate and Genome Annotation Project (Vega)



Stein L (2001) Genome annotation: from sequence to biology. Nat Rev Genet 2: 493–503

Nature Reviews | Genetics

Steps of genome analysis

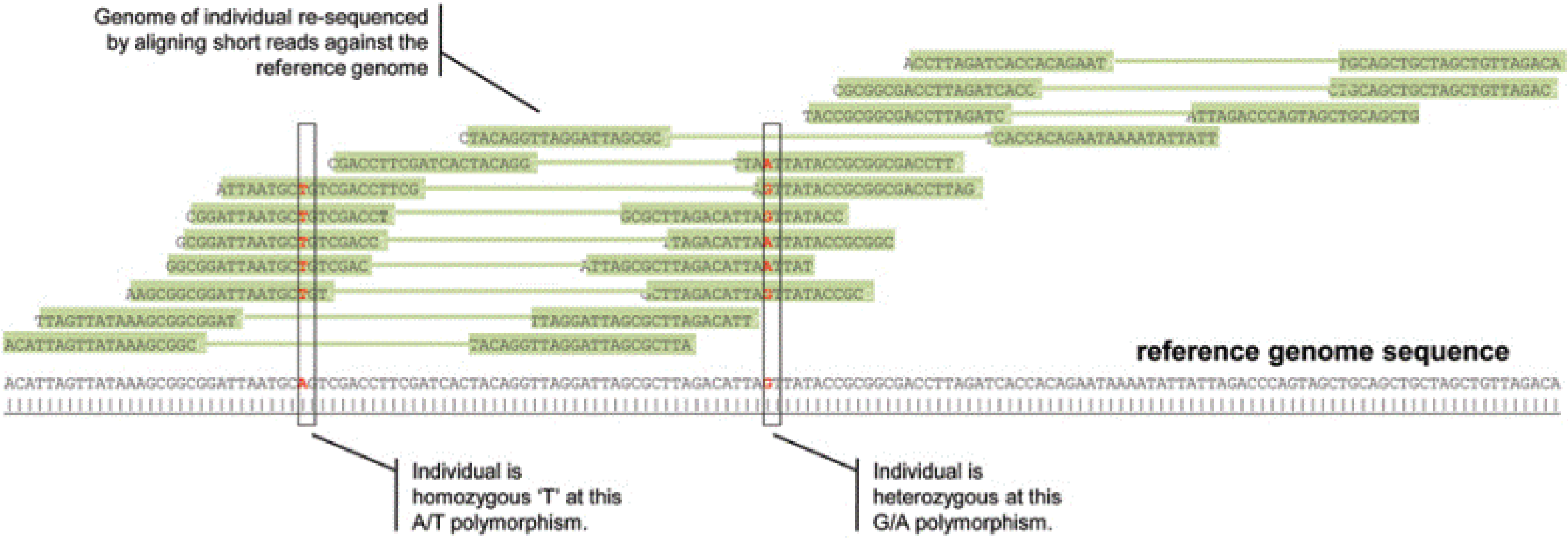
1. Quality checking
2. Trimming: filter out low quality reads (or read parts)

3.a) Newly sequenced genome: *de novo* assembly

3.b) Genome re-sequencing: mapping

4. Unfold genetic diversity: statistical analysis

Re-sequencing



Genome of individual re-sequenced by aligning short reads against the reference genome

Individual is homozygous 'T' at this A/T polymorphism.

Individual is heterozygous at this G/A polymorphism.

Re-sequencing

- Aim: Exploration of genetic diversity
- We map the reads to a known reference geneome
  - Less (but still intense) computation demand
  - genome variability can couse problems
  - Or even remain unobserved – ie. Chromosomal translocations
  - There can be biased or missing regions in the reference genome as well
- Mapping softwares: BWA (Burrow's Wheeler Transform Algorithm), Bowtie, GSNAP, SOAP2, …

Steps of genome analysis

1. Quality checking
2. Trimming: filter out low quality reads (or read parts)

3.a) Newly sequenced genome: *de novo* assembly

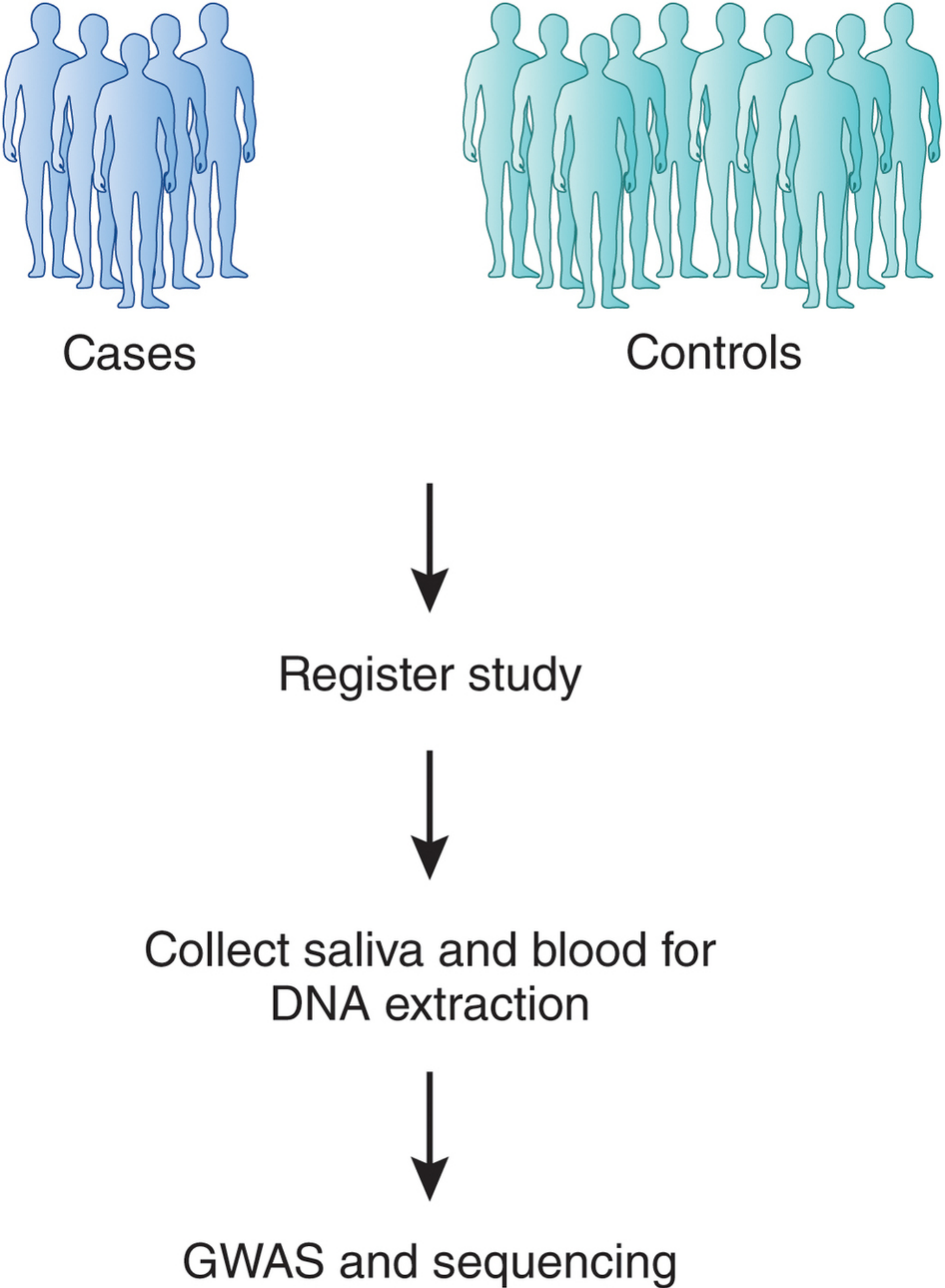
3.b) Genome re-sequencing: mapping

4. Unfold genetic diversity: statistical analysis

Exploring the genetic variability

- Genome differences in between two individuals: ie. SNPs, in/dels, copy number variations, chromosome translocations
  - These can cause different phenotypes or diseases
- SNP analysis / GWAS: genome-wide association study
  - Study a genome-wide set of genetic variants in different individuals to see if any variant is associated with a trait (phenotype)
  - Mostly based on SNPs → allele frequencies
  - Traits: different phenotypes (ie. Size or eye color of individuals) or genetic disorders

- Exploring the genetic variability
- SNP analysis
- GWAS: genome-wide association study



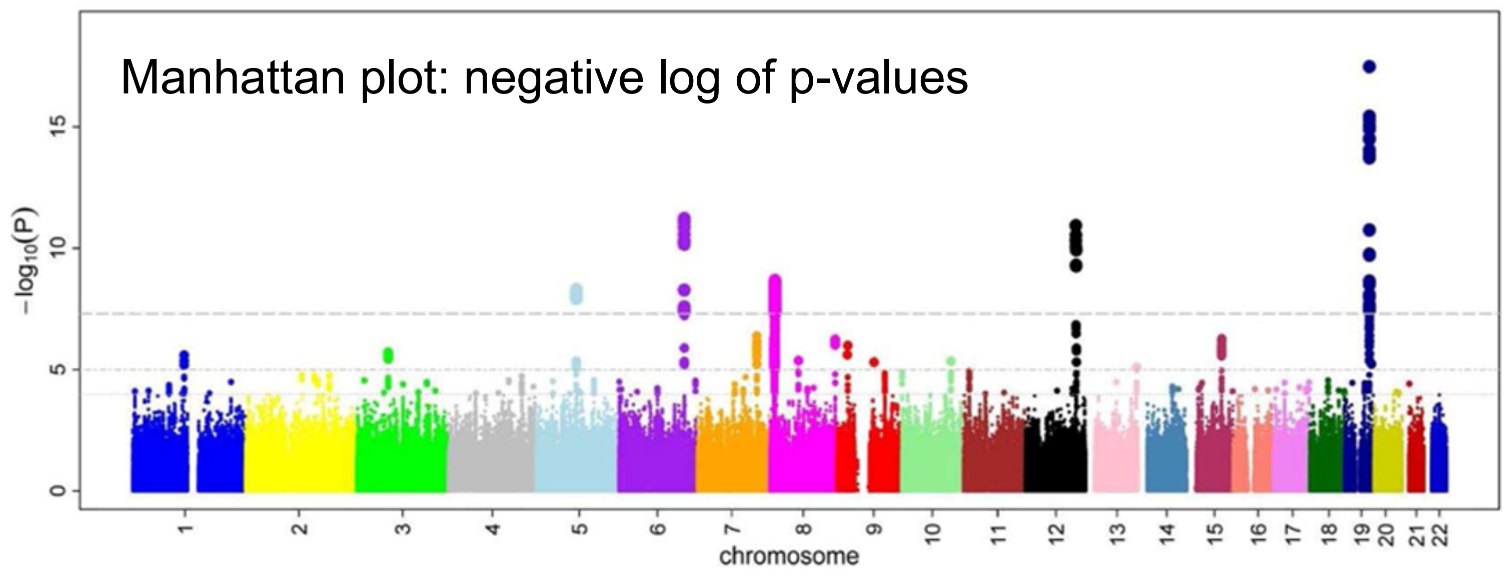
Register study

Collect saliva and blood for DNA extraction

GWAS and sequencing

Exploring the genetic variability

- If the phenotype is caused by a single SNP → it is easy to unfold
- If more than 1 SNP is playing some role to create the phenotype → we should involve many individuals
- We should choose individuals very carefully to exclude possible cofounding factors that would influence our investigation:
  - ie. gender, age, race of individuals, history of populations



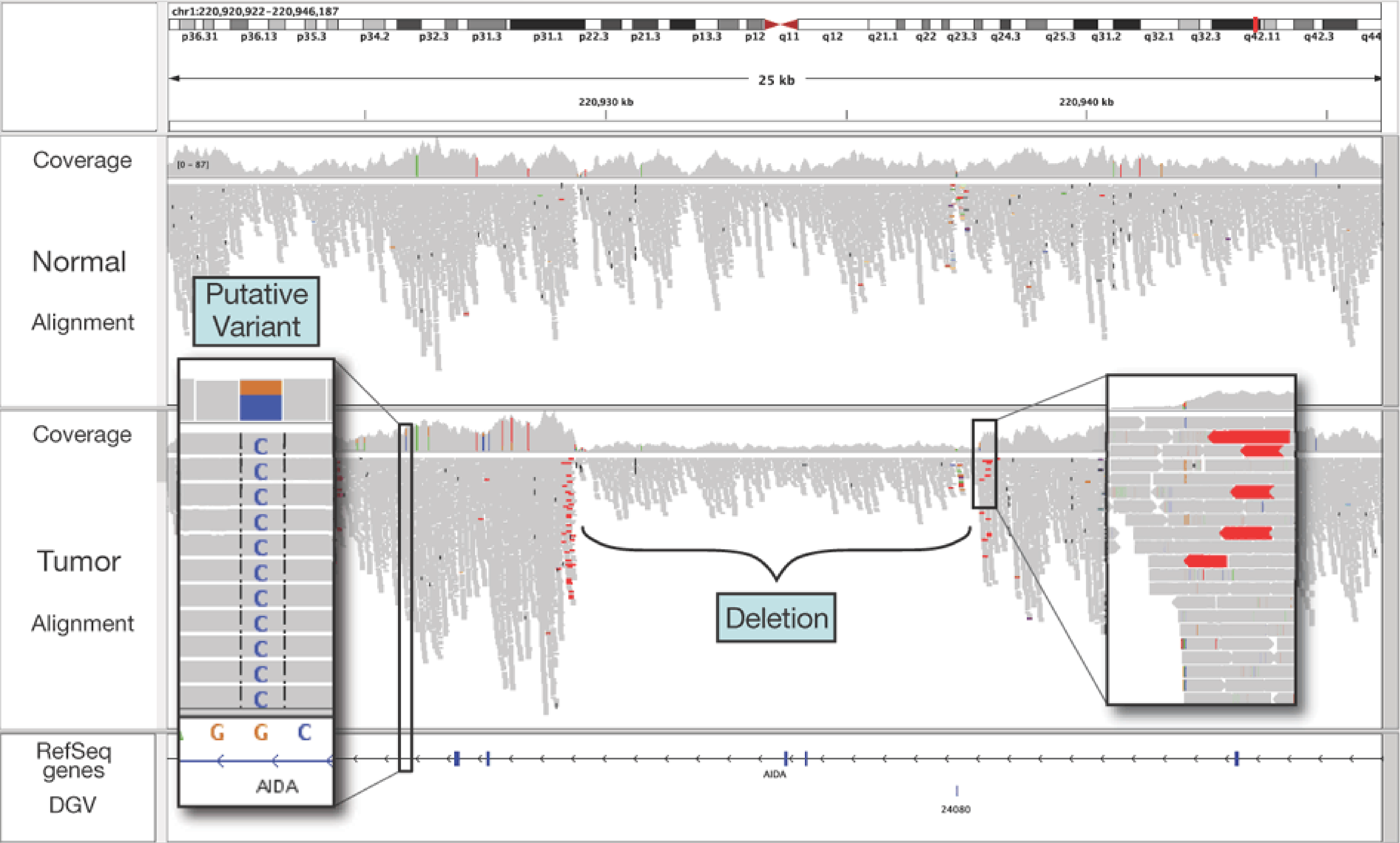
Replicates

- Statistical definition: a fully repeated experiment or set of test conditions
- To calculate statistical tests we need more replicates
  - Replicates: samples got the same “treatment”
  - Depending on the investigation we need 2-3-100 replicates / treatment groups

Genome browsers

- Online, general:
  - http://www.ensembl.org/
  - https://genome.ucsc.edu/
  - http://www.ncbi.nlm.nih.gov/genome/
- Online, species specific:
  - Flybase, WormBase, …
- Offline:
  - Integrative Genomics Viewer (IGV)
  - Golden Helix GenomeBrowse, ...

Offline genome browser


chr1:220,920,922-220,946,187
p36.31
p36.13
p35.3
p34.2
p32.3
p31.3
p31.1
p22.3
p21.3
p13.3
p12
q11
q12
q21.1
q22
q23.3
q24.3
q25.3
q31.2
q32.1
q32.3
q42.11
q42.3
q44
25 kb
220,930 kb
220,940 kb
Coverage
[0 - 87]
Normal
Alignment
Putative Variant
Coverage
Tumor
Alignment
C
G
G
C
AIDA
Deletion
RefSeq genes
DGV
AIDA
24080