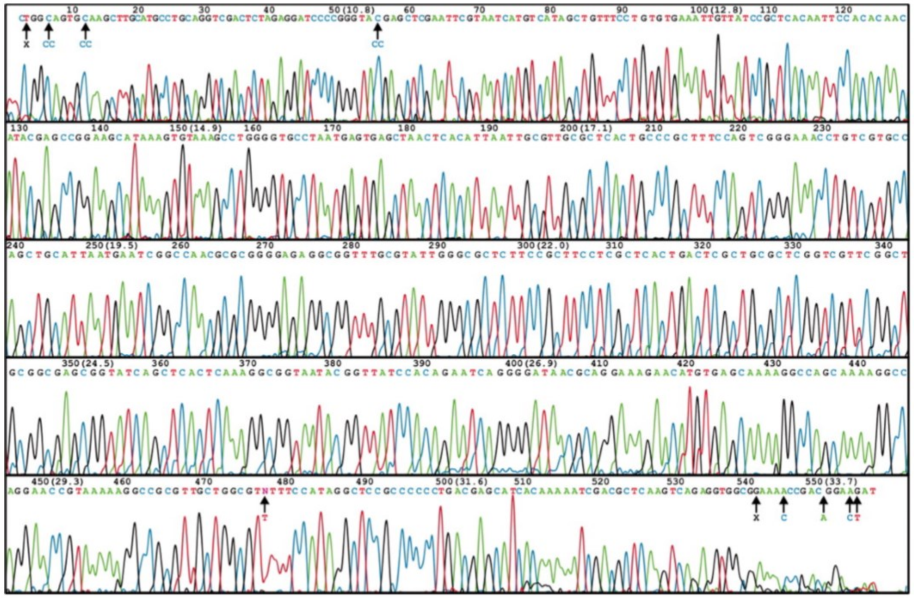
Sequencing II

Radka Reifová

Sequence data

Sanger sequencing

chromatogram



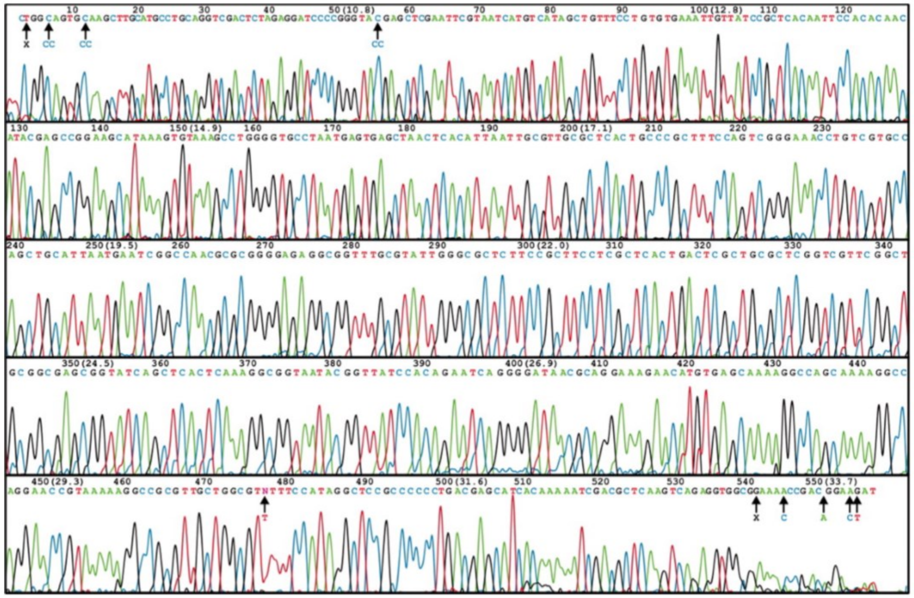
manual editing
(e.g. program Geious)

```
>sekvence1
CGGCAGTGCAAGCTTGCATGCATGCCTGCAGGTCGACTCTAG
AGGATCCCGGGTACGAGCTCGAATTCGTAATCATGTCATAGC
TGTTTCCTGTGTGAAATTGTTATCCGCTCACAATTCCACACA
ACATACGAGCCGGAAGCATAAAGTGTAAAGCCTGGGGTGCCT
GCCTAATGAGTGAGCTAACTCACATTATTGCGTTGCGTTAGT
```

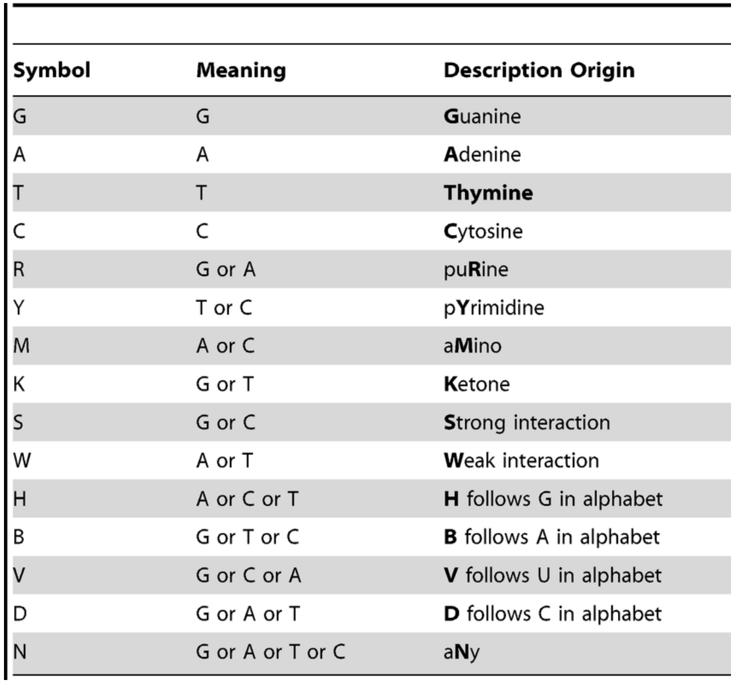
Fasta format

Sanger sequencing

chromatogram



IUPAC nucleotides codes



Symbol	Meaning	Description Origin
G	G	**G**uanine
A	A	**A**denine
T	T	**Thymine**
C	C	**C**ytosine
R	G or A	pu**R**ine
Y	T or C	p**Y**rimidine
M	A or C	a**M**ino
K	G or T	**K**etone
S	G or C	**S**trong interaction
W	A or T	**W**eak interaction
H	A or C or T	**H** follows G in alphabet
B	G or T or C	**B** follows A in alphabet
V	G or C or A	**V** follows U in alphabet
D	G or A or T	**D** follows C in alphabet
N	G or A or T or C	a**N**y

manual editing
(e.g. program Geious)

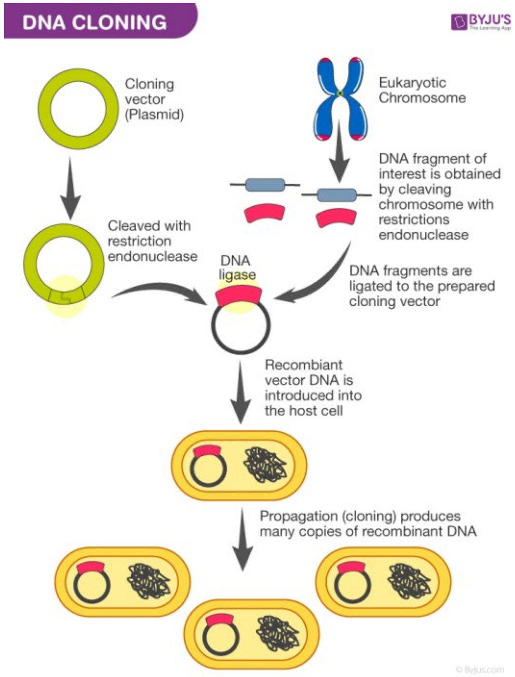
```
>sekvence1
CGGCAGTGCAWGCTTGCATGCATGCSTGCAGGTCGACTCTAG
AGGATCCCGGGTACGAGCTCGAAYTCGTAATRATGTCATAGC
TGTTTCCTGTGTGAAATTGTTATCCGCTCACAATTCCACACA
ACATANNNNCCGGAAGCATAAAGTGTAAAGCCTGGGGTGCCT
GCCTAATGAGTGAGCTAACTCACATTATTGCGTTGCGTTAGT
```

Fasta format

Phasing of diploid sequences

= determination of haplotypes corresponding to sequences of each chromosomes.

- molecular approach: DNA cloning
- statistical approach: inference from population data (program PHASE)

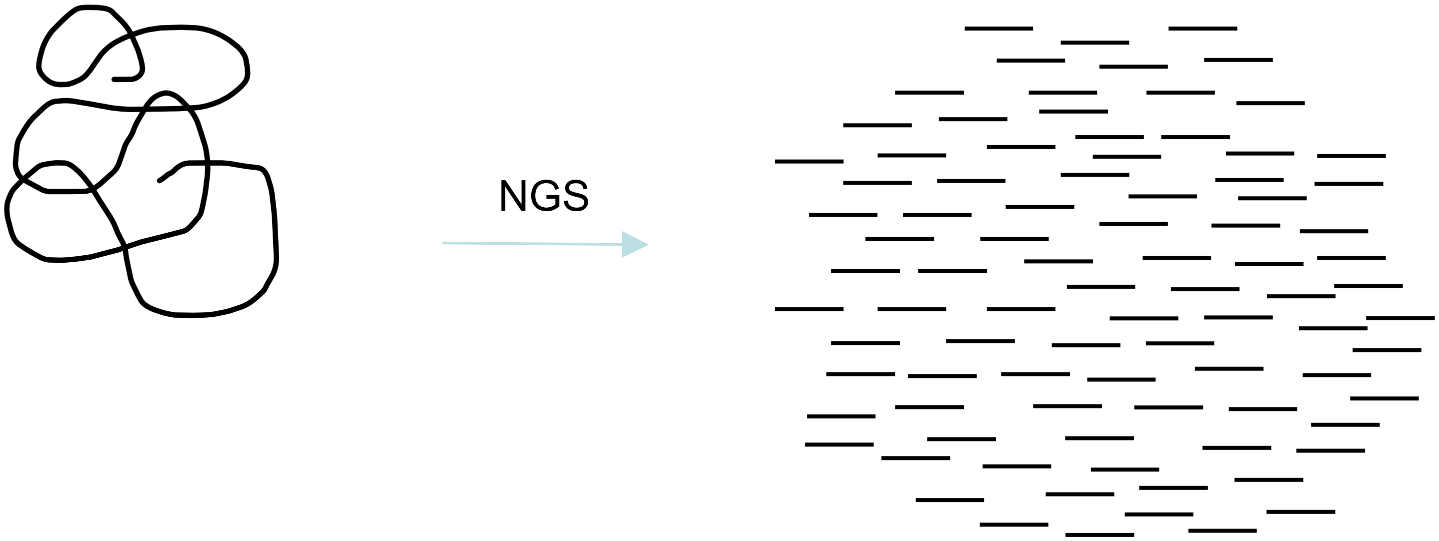

DNA CLONING
BYJU'S
Cloning vector (Plasmid)
Eukaryotic Chromosome
DNA fragment of interest is obtained by cleaving chromosome with restrictions endonuclease
Cleaved with restriction endonuclease
DNA ligase
DNA fragments are ligated to the prepared cloning vector
Recombiant vector DNA is introduced into the host cell
Propagation (cloning) produces many copies of recombinant DNA
© Byjus.com

```
>sekvence1
CGRCAGTGCAWGCTTGCATGCATGCSTGCAGGTCG
ACTCTAGAGGATCCCGGGTACGAGCTCGAAYTCGT
AATRATGTCATAGCTGYTTCCTGTGTGAAATTGTT
```

```
>sekvence1
CGGCAGTGCAAGCTTGCATGCATGCGTGCAGGTCG
ACTCTAGAGGATCCCGGGTACGAGCTCGAATTCGT
AATAATGTCATAGCTGTTTCCTGTGTGAAATTGTT
```

```
>sekvence1
CGACAGTGCATGCTTGCATGCATGCCTGCAGGTCG
ACTCTAGAGGATCCCGGGTACGAGCTCGAACTCGT
AATGATGTCATAGCTGCTTCCTGTGTGAAATTGTT
```

Next generation sequencing



NGS

reads
~ 100bp long

Fastaq format

```
@SEQ_ID
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT
+
!''*((((***+))%%%++)(%%%%).1***-+*''))**55CCF>>>>>>CCCCCCC65
```

line 1: @ sequence name

line 2: sequence

line 3: + additional information about the sequence.

line 4: Phred Quality Scores.

Fastaq format

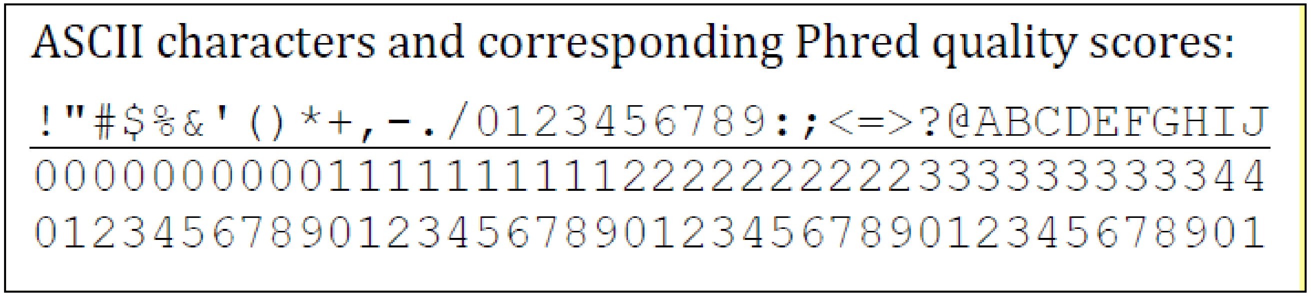
```
@SEQ_ID
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT
+
!''*((((***+))%%%++)(%%%%).1***-+*''))**55CCF>>>>>>CCCCCCC65
```

Phred Quality Scores (*Q*)

$$Q = -10\ \log_{10} P$$

Probability that the base is determined incorrectly.

Q	P	probability that the base is correct
10	0,1	90%
20	0,001	99%
30	0,0001	99.9%
40	0,000001	99.99%

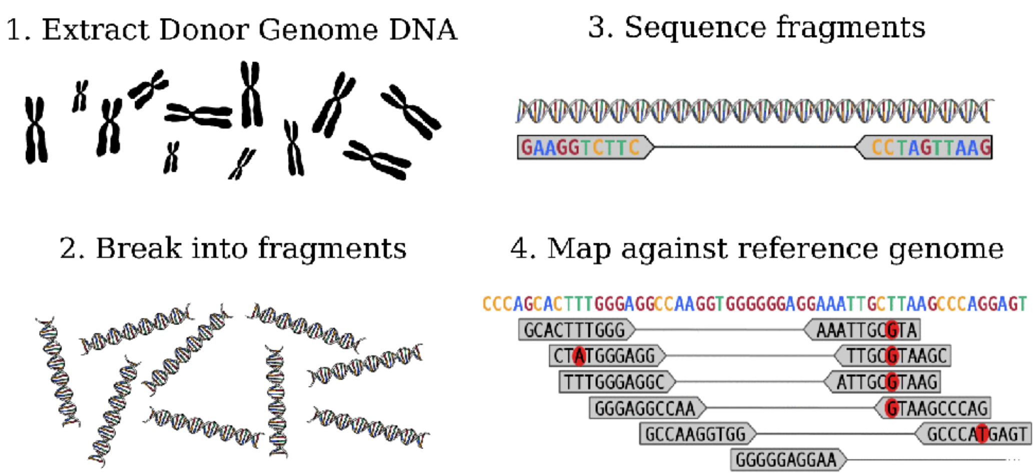


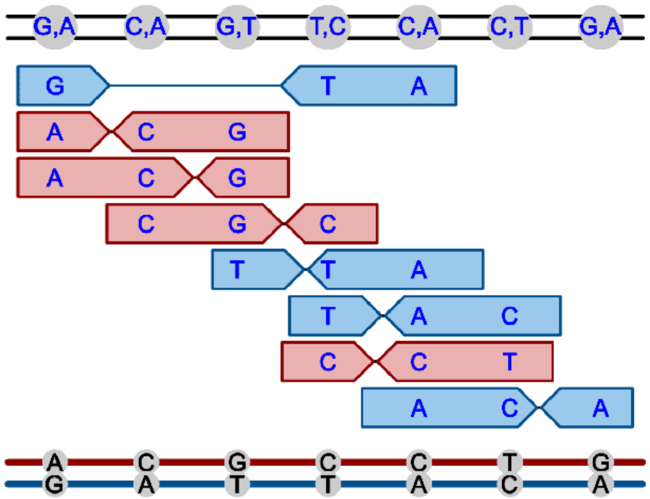
ASCII characters and corresponding Phred quality scores:

```
!"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJ
000000000011111111112222222222333333333344
012345678901234567890123456789012345678901
```

Phased sequencing

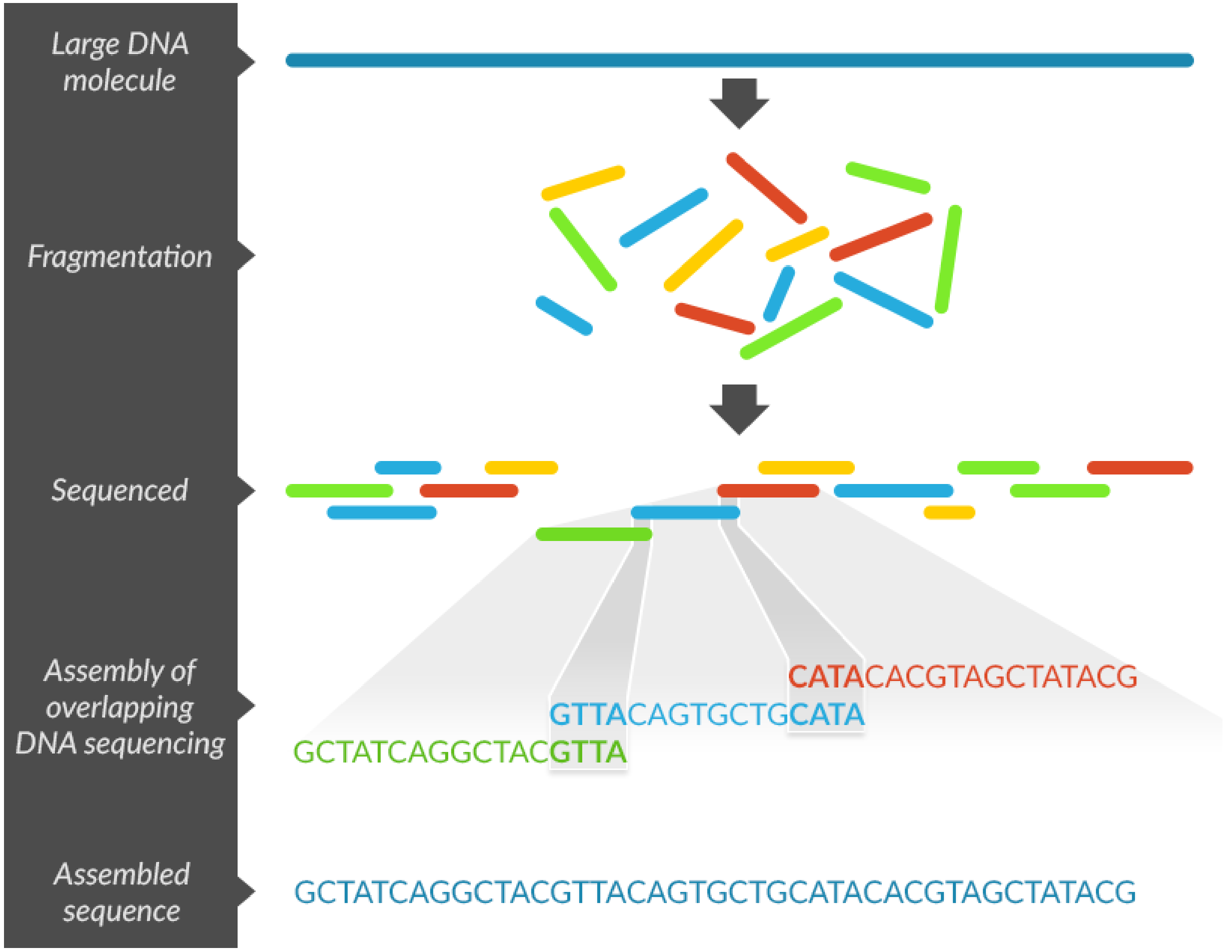
- NGS provides haploid sequences (tj. sequences of individual chromosomes).
- Especially the long reads (Nanopore, Pac Bio) allow to reconstruct the haplotypes.





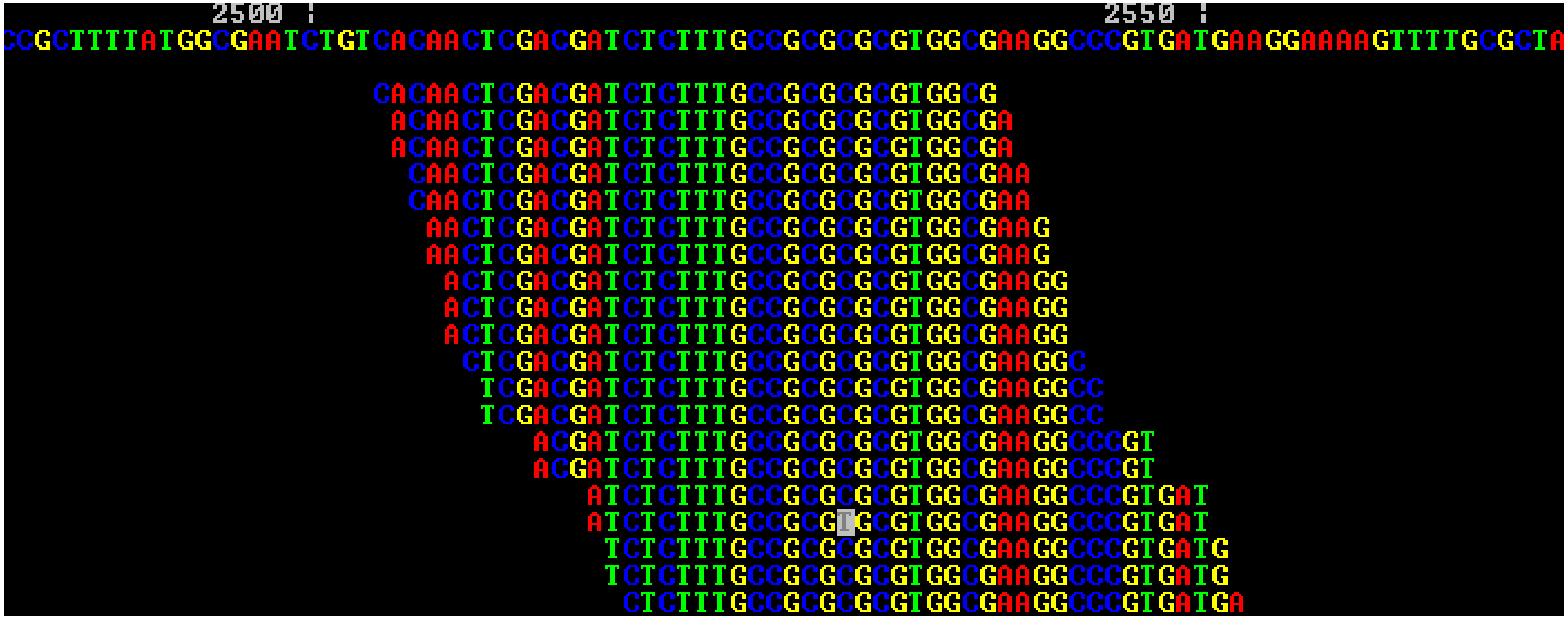
Assembly

- Assembly of the short reads to long sequences corresponding to transcripts (transcriptome sequencing) or chromosomes (genome sequencing)
- Needs sufficient coverage
- Easier with longer reads



Coverage

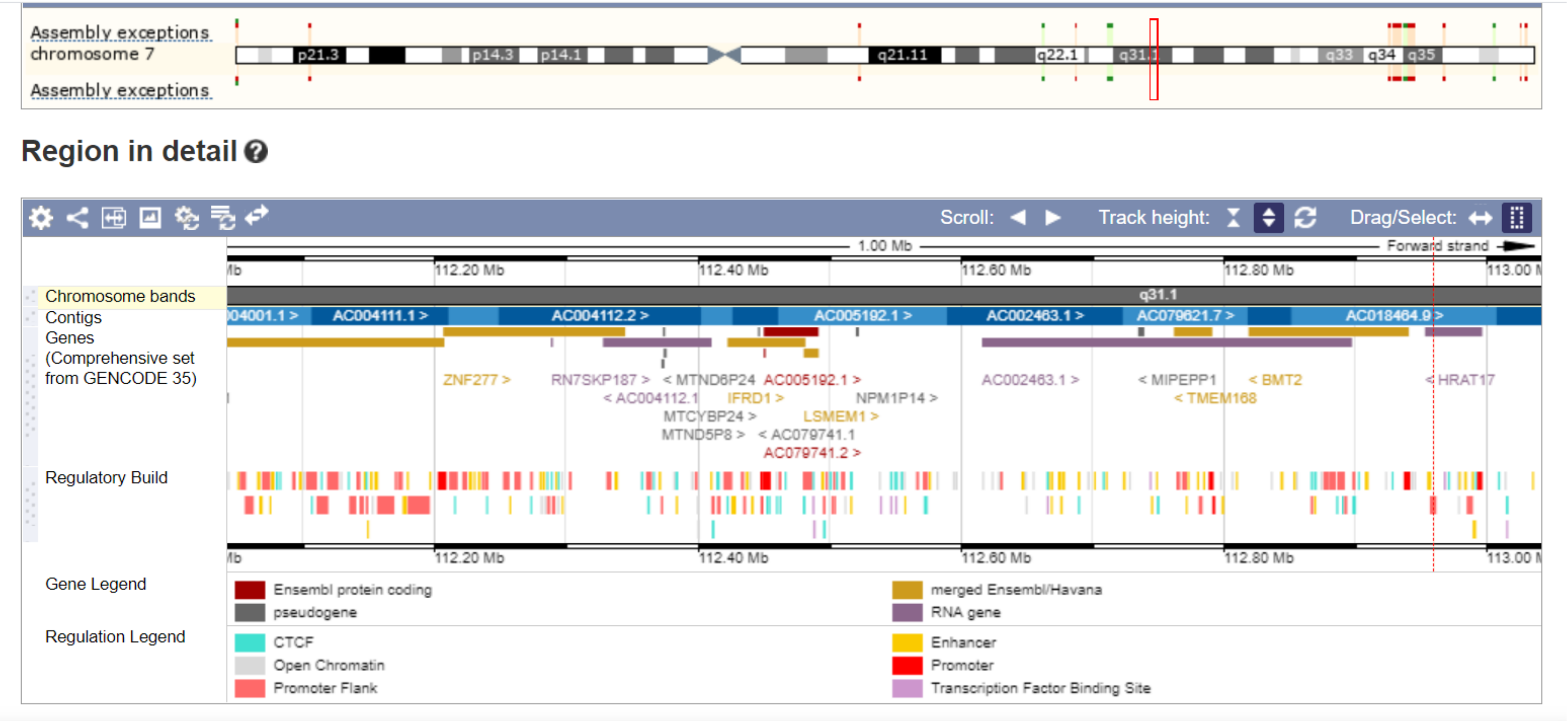
- How many times is the particular base sequenced.
- High coverage (>10x) allows to distinguish sequencing errors from polymorphisms.



```
            2500 !                                            2550 !
CCGCTTTTATGGCGAATCTGTCACAACTCGACGATCTCTTTGCCGCGCGCGTGGCGAAGGCCCGTGATGAAGGAAAAGTTTTGCGCTA

                     CACAACTCGACGATCTCTTTGCCGCGCGCGTGGCG
                      ACAACTCGACGATCTCTTTGCCGCGCGCGTGGCGA
                      ACAACTCGACGATCTCTTTGCCGCGCGCGTGGCGA
                       CAACTCGACGATCTCTTTGCCGCGCGCGTGGCGAA
                       CAACTCGACGATCTCTTTGCCGCGCGCGTGGCGAA
                        AACTCGACGATCTCTTTGCCGCGCGCGTGGCGAAG
                        AACTCGACGATCTCTTTGCCGCGCGCGTGGCGAAG
                         ACTCGACGATCTCTTTGCCGCGCGCGTGGCGAAGG
                         ACTCGACGATCTCTTTGCCGCGCGCGTGGCGAAGG
                         ACTCGACGATCTCTTTGCCGCGCGCGTGGCGAAGG
                          CTCGACGATCTCTTTGCCGCGCGCGTGGCGAAGGC
                           TCGACGATCTCTTTGCCGCGCGCGTGGCGAAGGCC
                           TCGACGATCTCTTTGCCGCGCGCGTGGCGAAGGCC
                              ACGATCTCTTTGCCGCGCGCGTGGCGAAGGCCCGT
                              ACGATCTCTTTGCCGCGCGCGTGGCGAAGGCCCGT
                                 ATCTCTTTGCCGCGCGCGTGGCGAAGGCCCGTGAT
                                 ATCTCTTTGCCGCGTGCGTGGCGAAGGCCCGTGAT
                                  TCTCTTTGCCGCGCGCGTGGCGAAGGCCCGTGATG
                                  TCTCTTTGCCGCGCGCGTGGCGAAGGCCCGTGATG
                                   CTCTTTGCCGCGCGCGTGGCGAAGGCCCGTGATGA
```

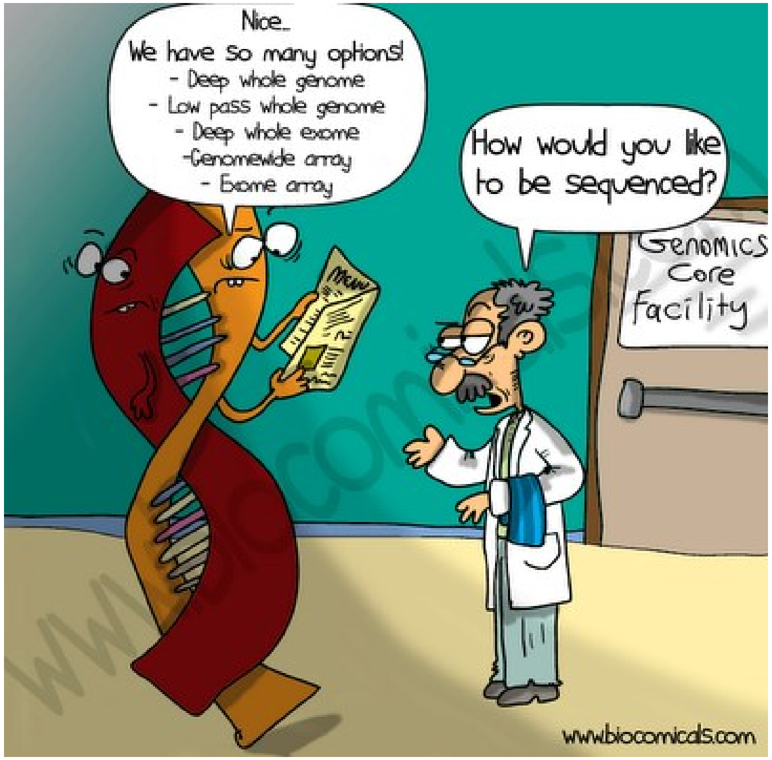
Annotation

- Identification of functional elements in the genome (protein coding and non-coding genes, promotors, repetitive sequences etc.).
- Based on homology to known genes/proteins, RNA data, predictions etc.

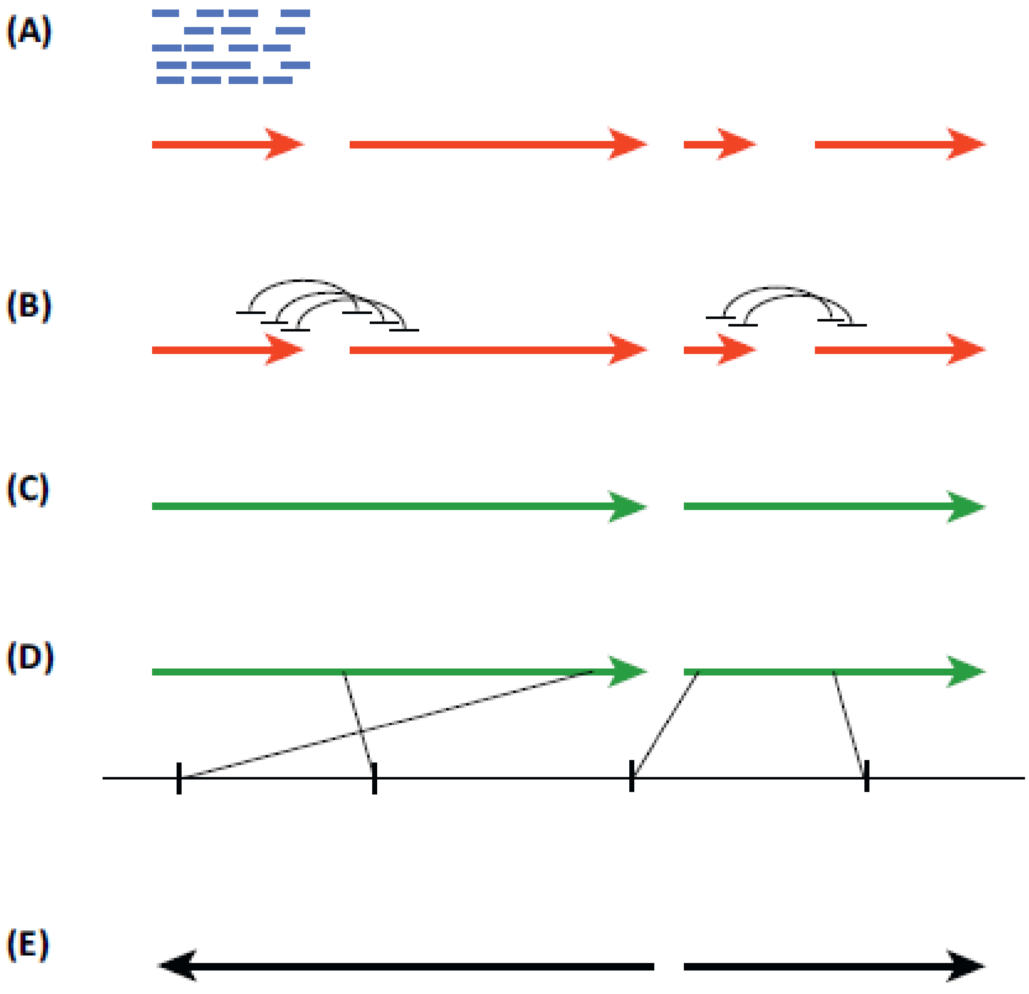


What can we sequence

- Genome sequencing
- RNA sequencing
- Exome sequencing
- Targeted sequencing
- Restriction site associated DNA (RAD) sequencing
- Metagenomics and DNA barcoding



Genome sequencing and assembly



TRENDS in Ecology & Evolution

reads

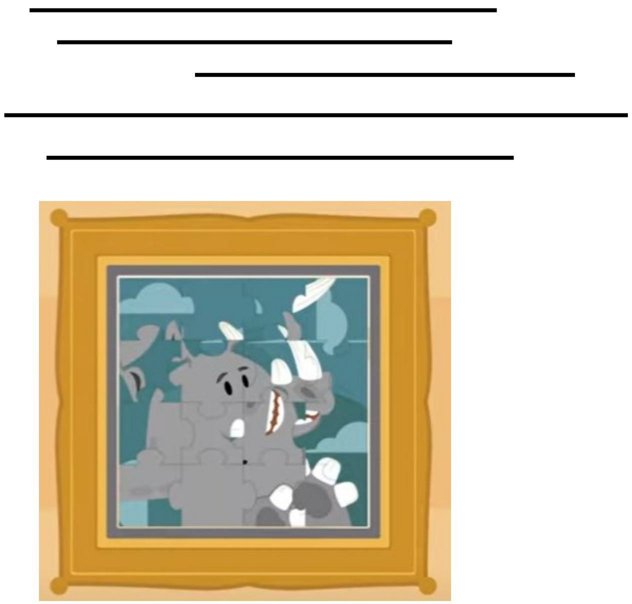
contigs

scaffolds

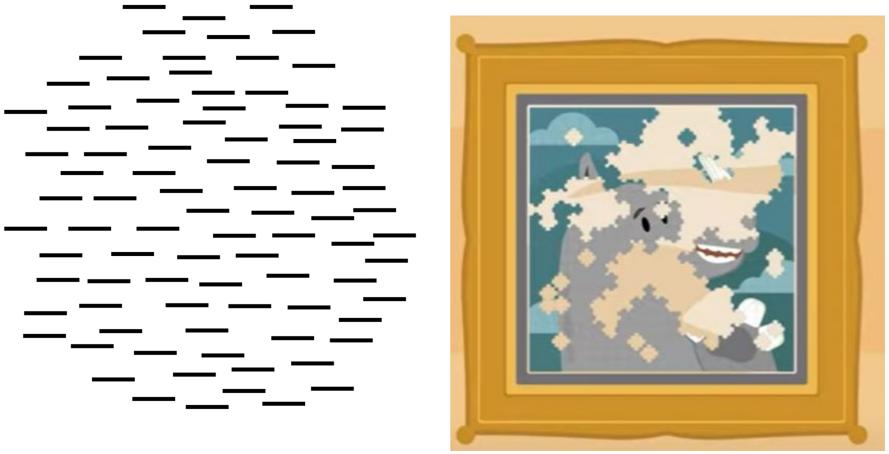
chromosome

Genome assembly is facilitated by long reads (Pac Bio, Nanopore). Relatively large number of errors rate can be “corrected” by short Illumina reads.

Pacific Biosciences/Nanopore

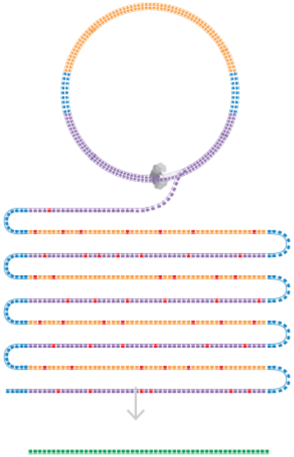


Illumina



OR

Hi-Fi



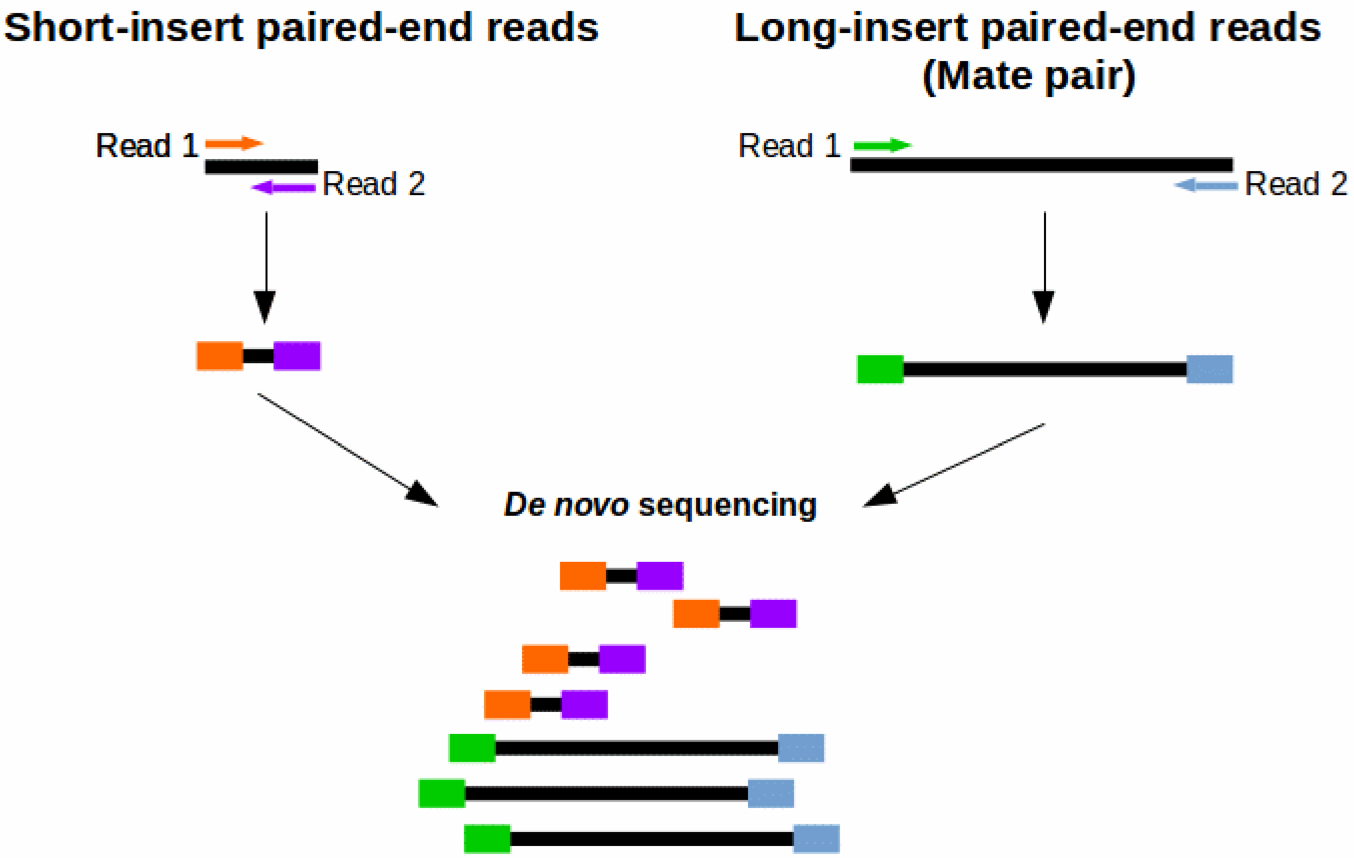
Long reads with low error rate.

single-end
vs.
pair-end and mate-pair sequencing

Sequencing of both ends of the fragments

cca 400 bp

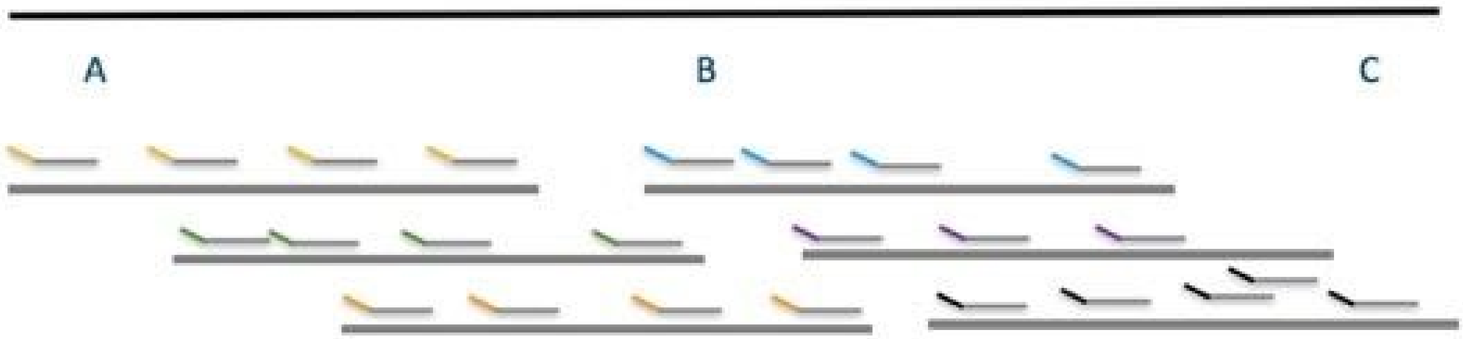
cca 10 000 bp


Short-insert paired-end reads
Long-insert paired-end reads (Mate pair)
Read 1
Read 2
Read 1
Read 2
De novo sequencing

Linked read sequencing

10X Genomics
TELL-Seq

Reads from the same DNA molecule are labelled by the same barcode sequence.



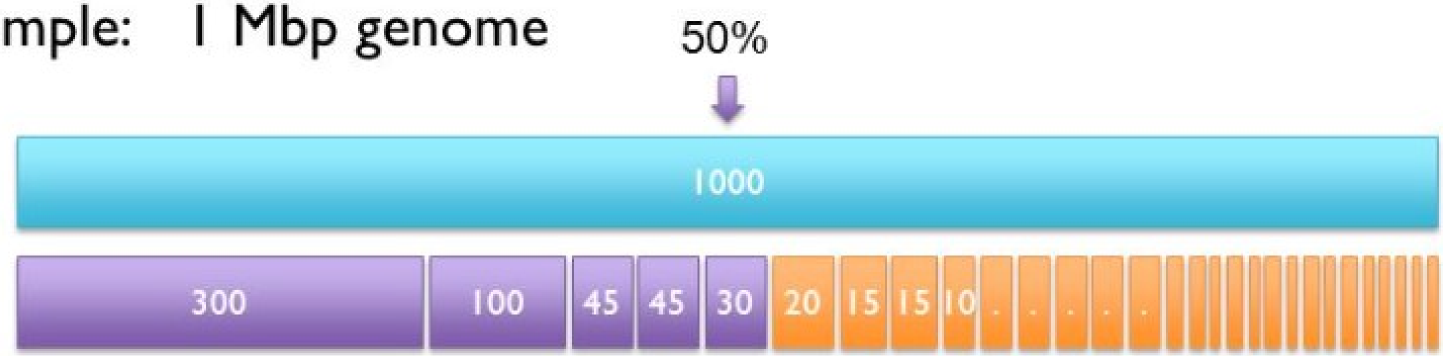
Whole genome assembly
Luscinia megarhynchos

	LM30 assembly ver. 1	LM30 assembly ver. 2	LM30 assembly ver. 3
Number of scaffolds	2 505	3 944	3 727
Total sequence length	1 098 533 284	1 098 533 284	1 098 533 284
Largest scaffold (bp)	77 026 980	76 959 640	95 377 781
Scaffold N50 (bp)	14 623 571	13 437 235	23 710 019
	Nanopore	Nanopore+ Illumina	Nanopore+ Illumina+ 10XGenomics

N50 size

Def: 50% of the genome is in contigs as large as the N50 value

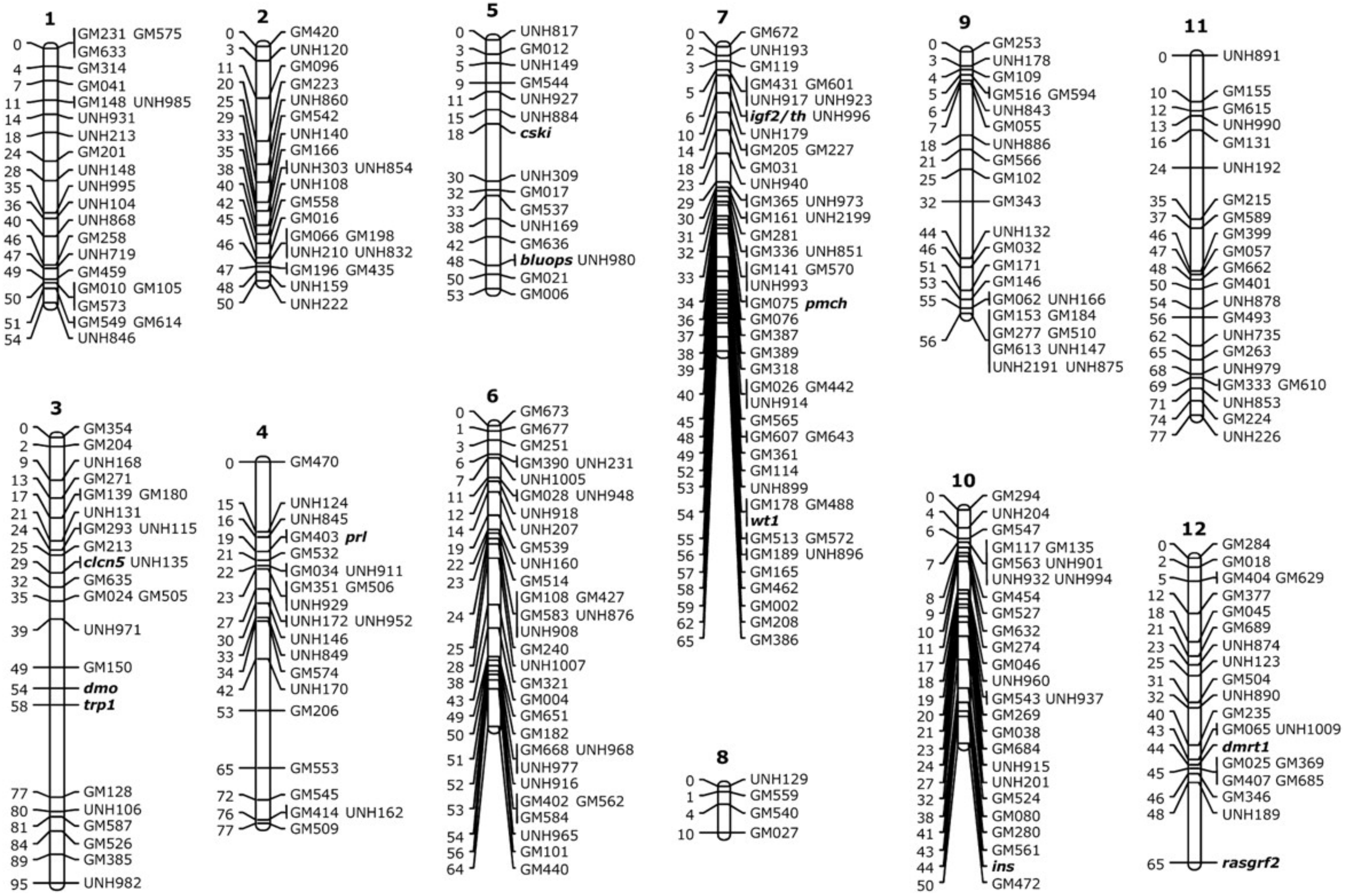
Example: 1 Mbp genome



N50 size = 30 kbp
(300k+100k+45k+45k+30k = 520k >= 500kbp)

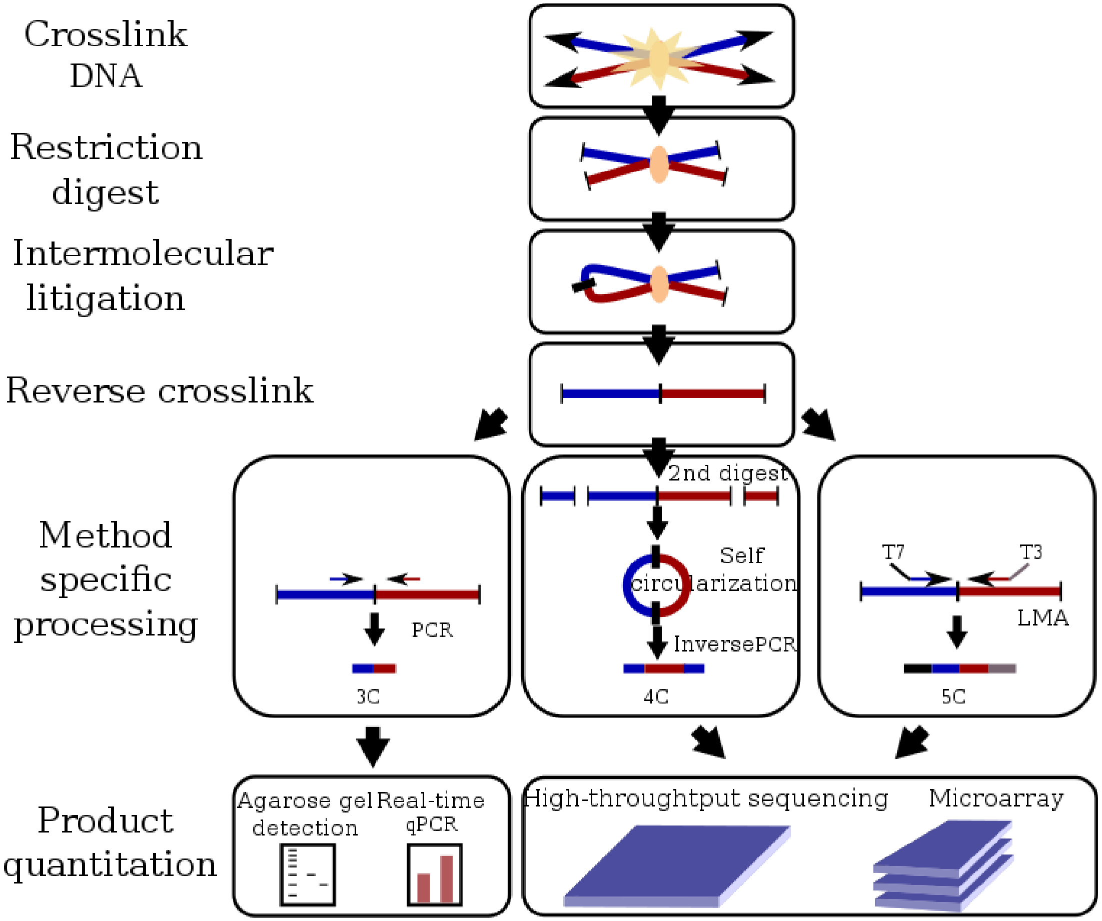
Chromosomal-level assembly can be created using known genetic map

Can be created using laboratory crosses or known pedigrees



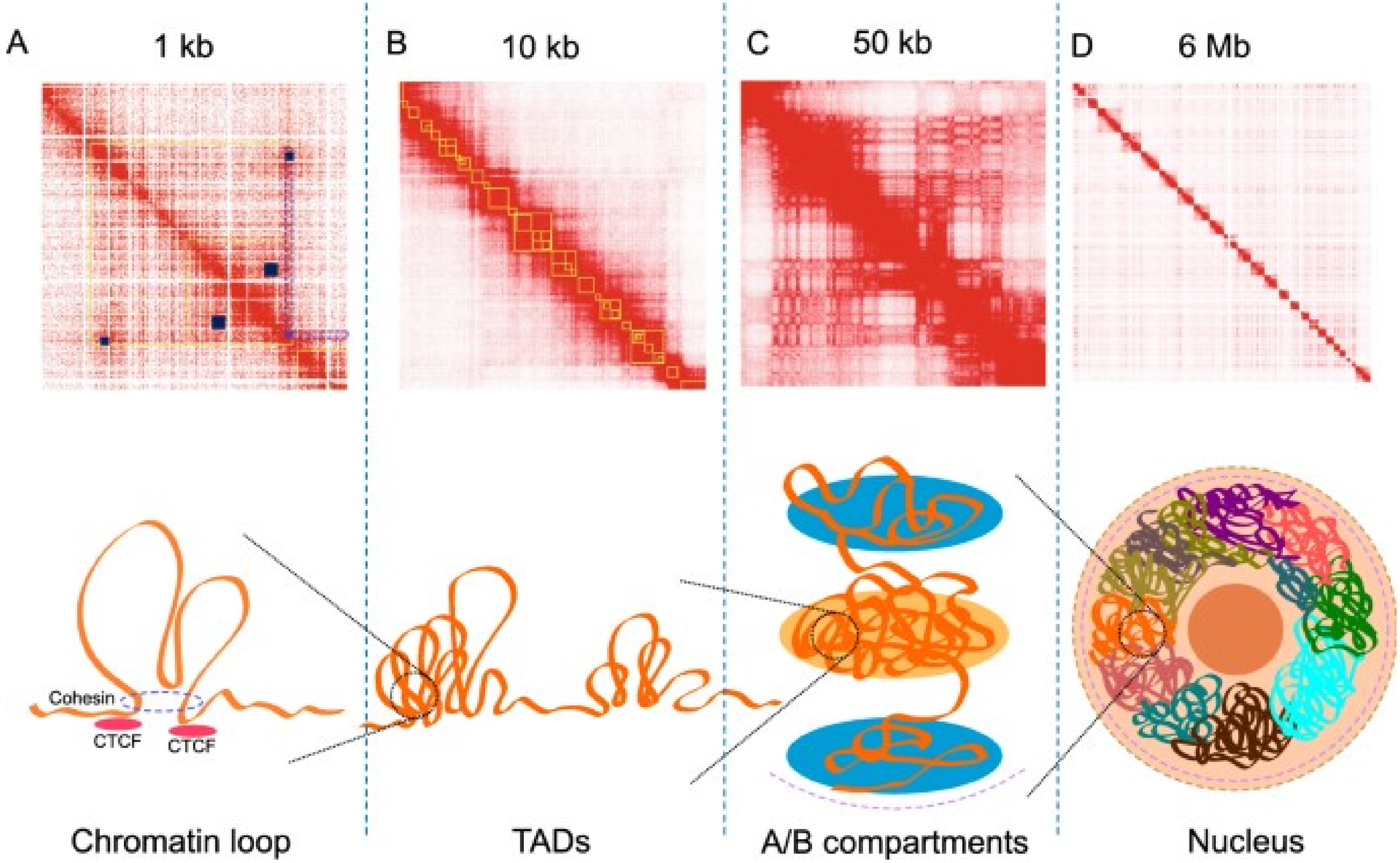
Chromosome conformation capture
Hi-C (Omni-C) sequencing

- Identifies sequences that interacts with each other (are close to each other) in the nucleus. Such sequences are usually from the same chromosome.
- Facilitates creation of the chromosome-level assembly.



3D genomics

Hi-C interaction maps



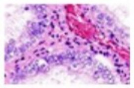
Ensembl Genome Browser

http://www.ensembl.org

Compare genes across species

Find SNPs and other variants for my gene

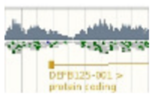
Gene expression in different tissues

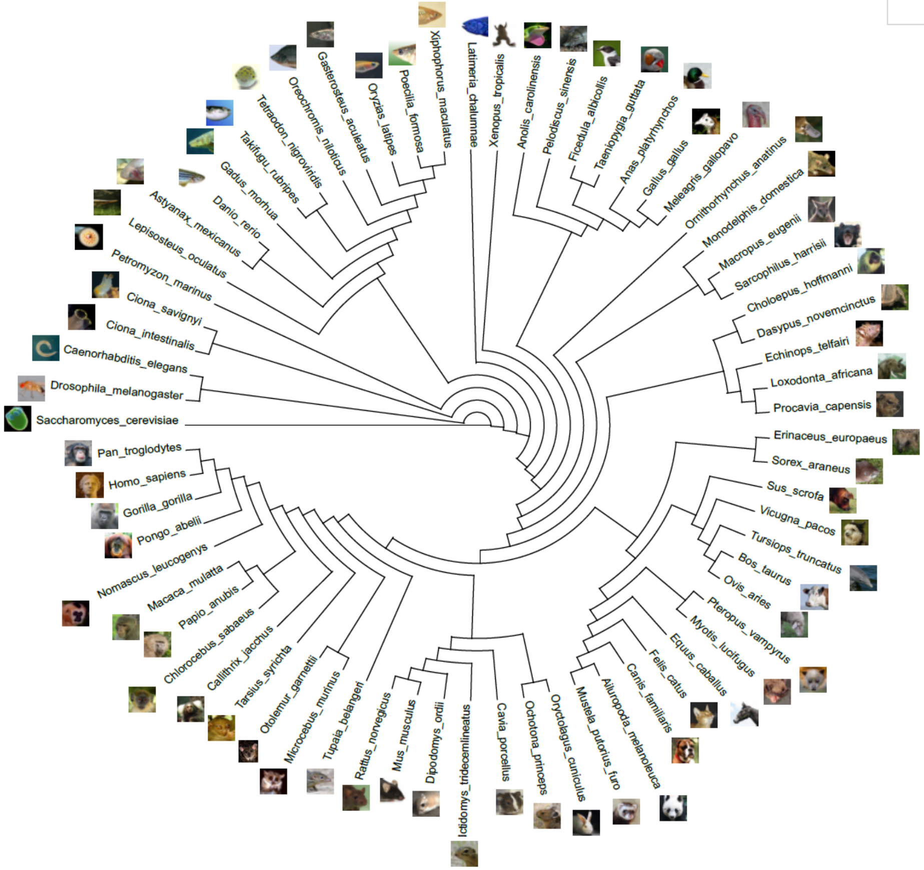


Retrieve gene sequence

Find a Data Display

Use my own data in Ensembl

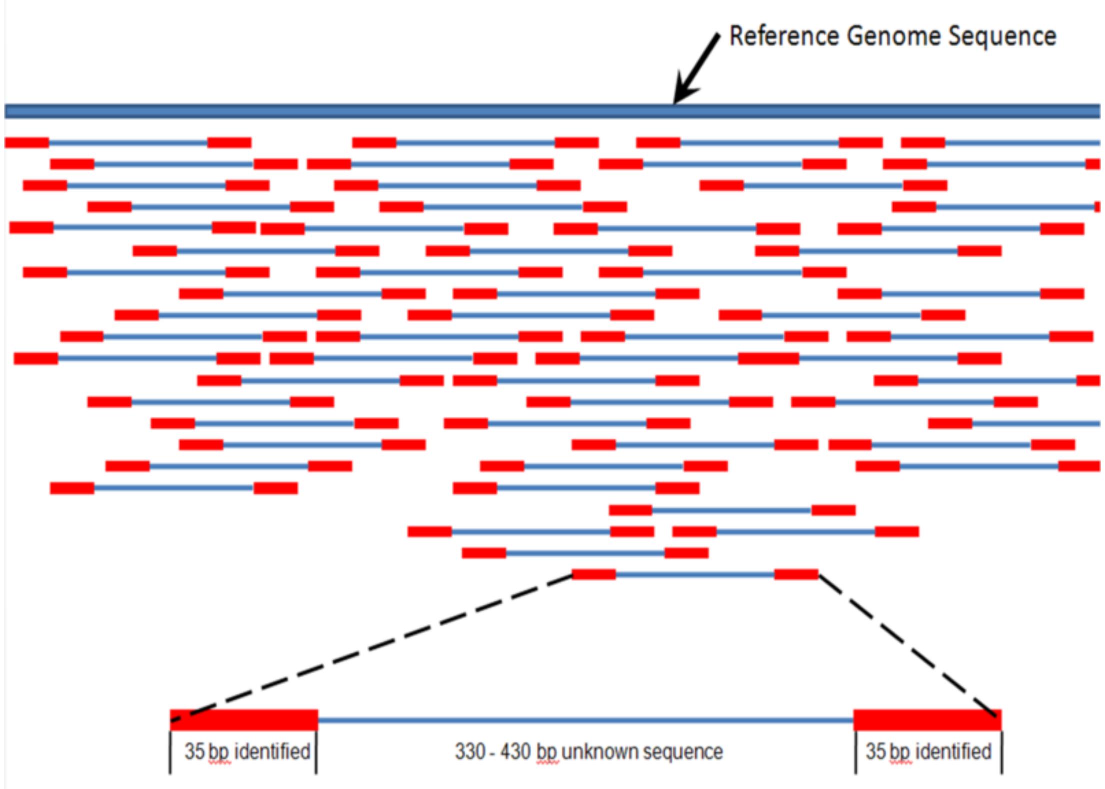


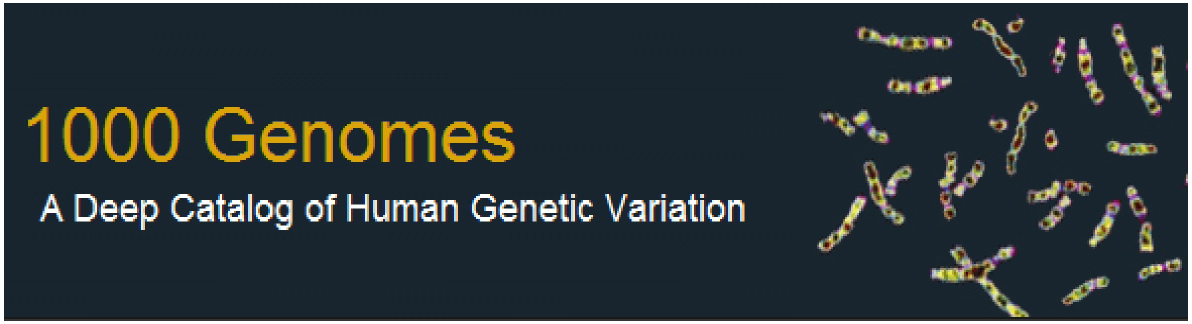


Genome resequencing

Read mapping

Short reads are sufficient. Can be mapped to the reference genome.
Identification of SNP polymorphisms.

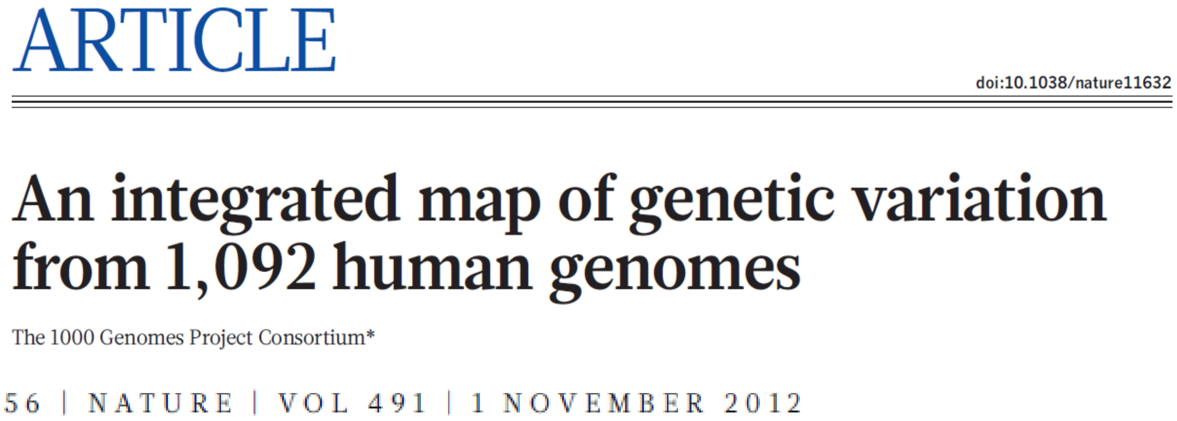




1000 Genomes

A Deep Catalog of Human Genetic Variation

http://www.1000genomes.org



ARTICLE

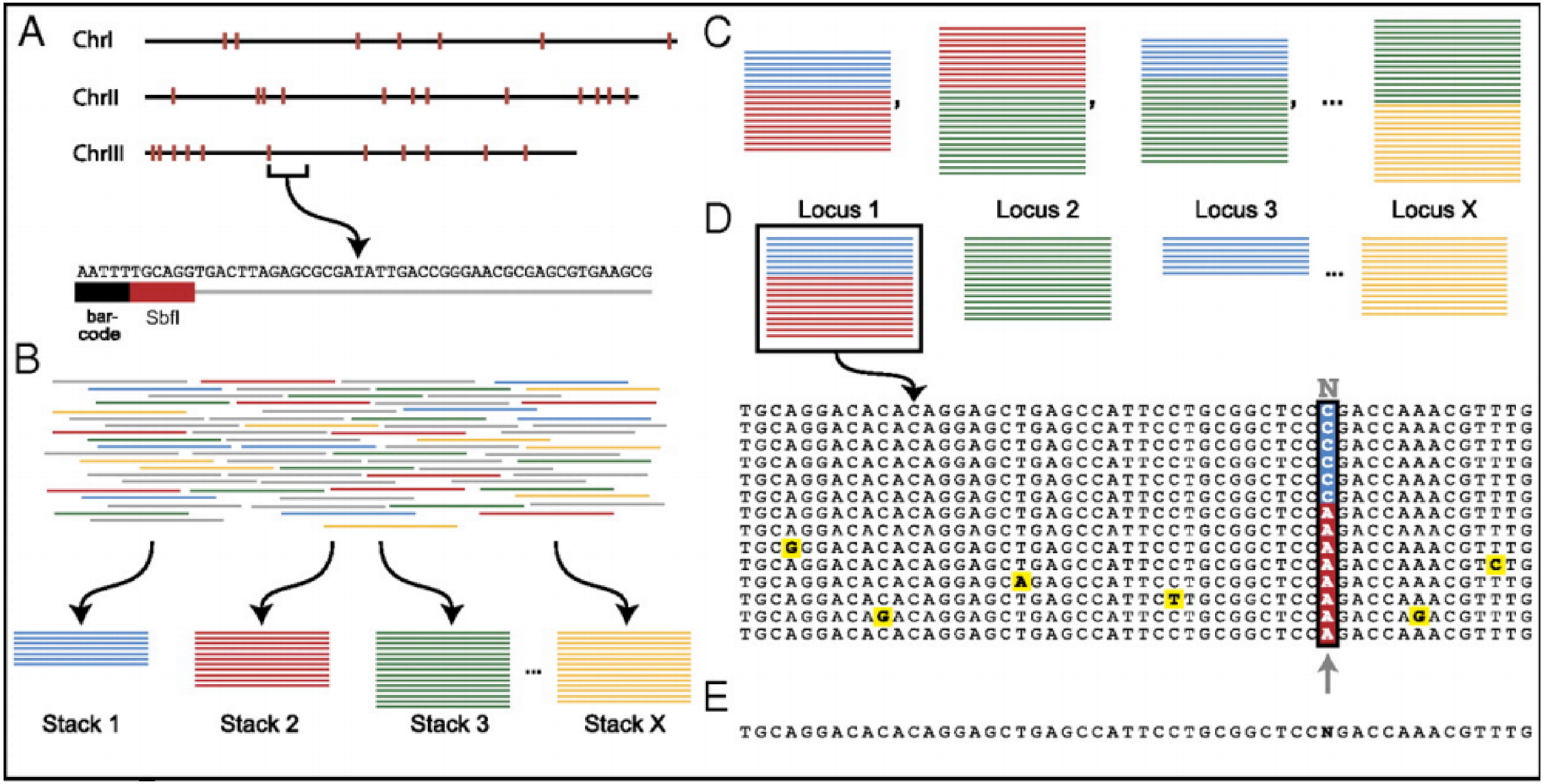
doi:10.1038/nature11632

An integrated map of genetic variation from 1,092 human genomes

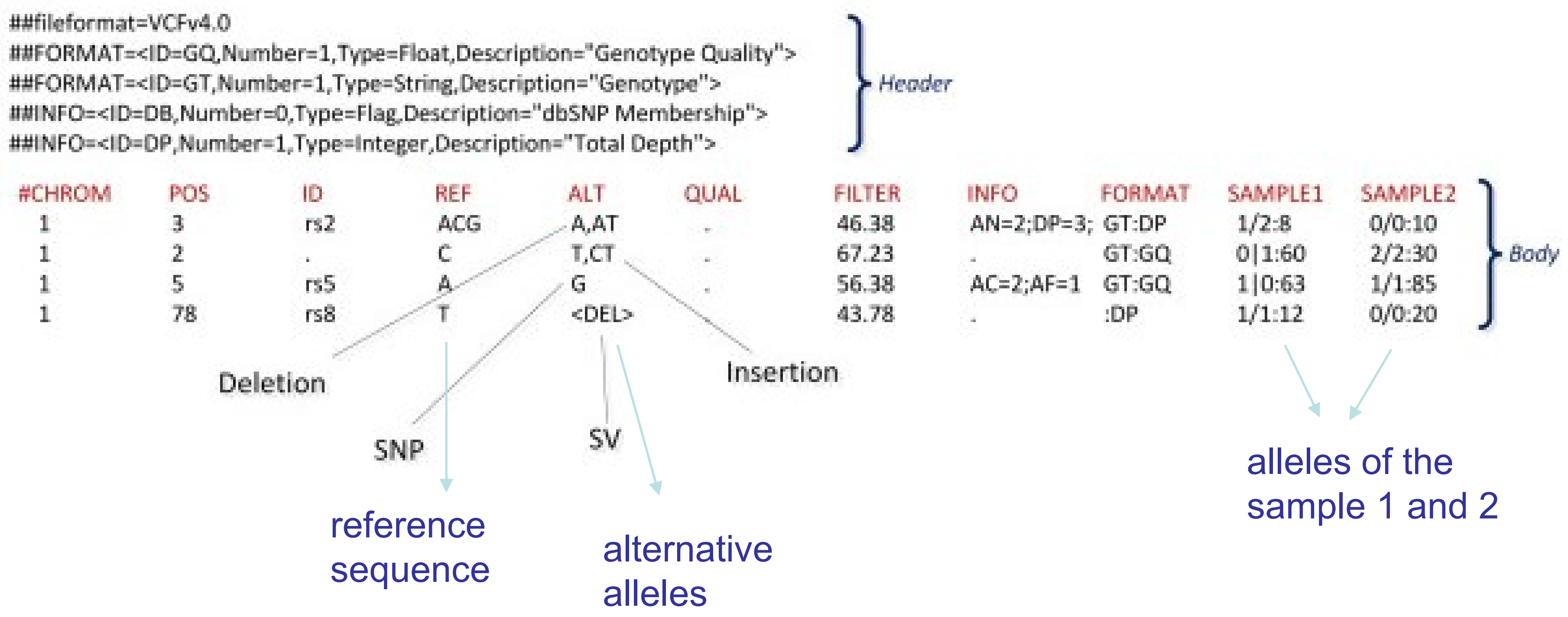
The 1000 Genomes Project Consortium*

56 | NATURE | VOL 491 | 1 NOVEMBER 2012

Identification of SNP polymorphisms (SNP calling)


A
ChrI
ChrII
ChrIII
AATTTTGCAGGTGACTTAGAGCGCGATATTGACCGGGAACGCGAGCGTGAAGCG
bar-code
SbfI
B
Stack 1
Stack 2
Stack 3
...
Stack X
C
...
D
Locus 1
Locus 2
Locus 3
...
Locus X
N
TGCAGGACACACAGGAGCTGAGCCATTCCTGCGGCTCCCGACCAAACGTTTG
TGCAGGACACACAGGAGCTGAGCCATTCCTGCGGCTCCCGACCAAACGTTTG
TGCAGGACACACAGGAGCTGAGCCATTCCTGCGGCTCCCGACCAAACGTTTG
TGCAGGACACACAGGAGCTGAGCCATTCCTGCGGCTCCCGACCAAACGTTTG
TGCAGGACACACAGGAGCTGAGCCATTCCTGCGGCTCCCGACCAAACGTTTG
TGCAGGACACACAGGAGCTGAGCCATTCCTGCGGCTCCCGACCAAACGTTTG
TGCAGGACACACAGGAGCTGAGCCATTCCTGCGGCTCCAGACCAAACGTTTG
TGCAGGACACACAGGAGCTGAGCCATTCCTGCGGCTCCAGACCAAACGTTTG
TGCGGGACACACAGGAGCTGAGCCATTCCTGCGGCTCCAGACCAAACGTTTG
TGCAGGACACACAGGAGCTGAGCCATTCCTGCGGCTCCAGACCAAACGTCTG
TGCAGGACACACAGGAGCAGAGCCATTCCTGCGGCTCCAGACCAAACGTTTG
TGCAGGACACACAGGAGCTGAGCCATTCTTGCGGCTCCAGACCAAACGTTTG
TGCAGGACAGACAGGAGCTGAGCCATTCCTGCGGCTCCAGACCAGACGTTTG
TGCAGGACACACAGGAGCTGAGCCATTCCTGCGGCTCCAGACCAAACGTTTG
E
TGCAGGACACACAGGAGCTGAGCCATTCCTGCGGCTCCNGACCAAACGTTTG

Vcf file



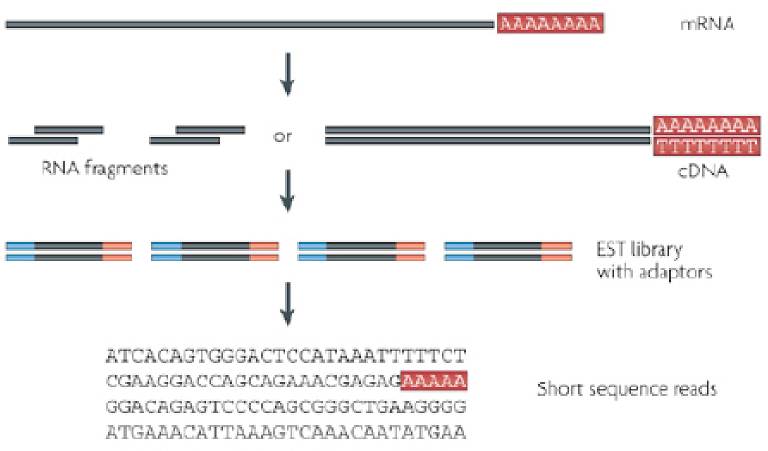
```
##fileformat=VCFv4.0
##FORMAT=<ID=GQ,Number=1,Type=Float,Description="Genotype Quality">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##INFO=<ID=DB,Number=0,Type=Flag,Description="dbSNP Membership">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth">

#CHROM   POS   ID    REF   ALT     QUAL   FILTER   INFO          FORMAT   SAMPLE1   SAMPLE2
  1      3     rs2   ACG   A,AT    .      46.38    AN=2;DP=3;    GT:DP    1/2:8     0/0:10
  1      2     .     C     T,CT    .      67.23    .             GT:GQ    0|1:60    2/2:30
  1      5     rs5   A     G       .      56.38    AC=2;AF=1     GT:GQ    1|0:63    1/1:85
  1      78    rs8   T     <DEL>          43.78    .             :DP      1/1:12    0/0:20
```

reference sequence

alternative alleles

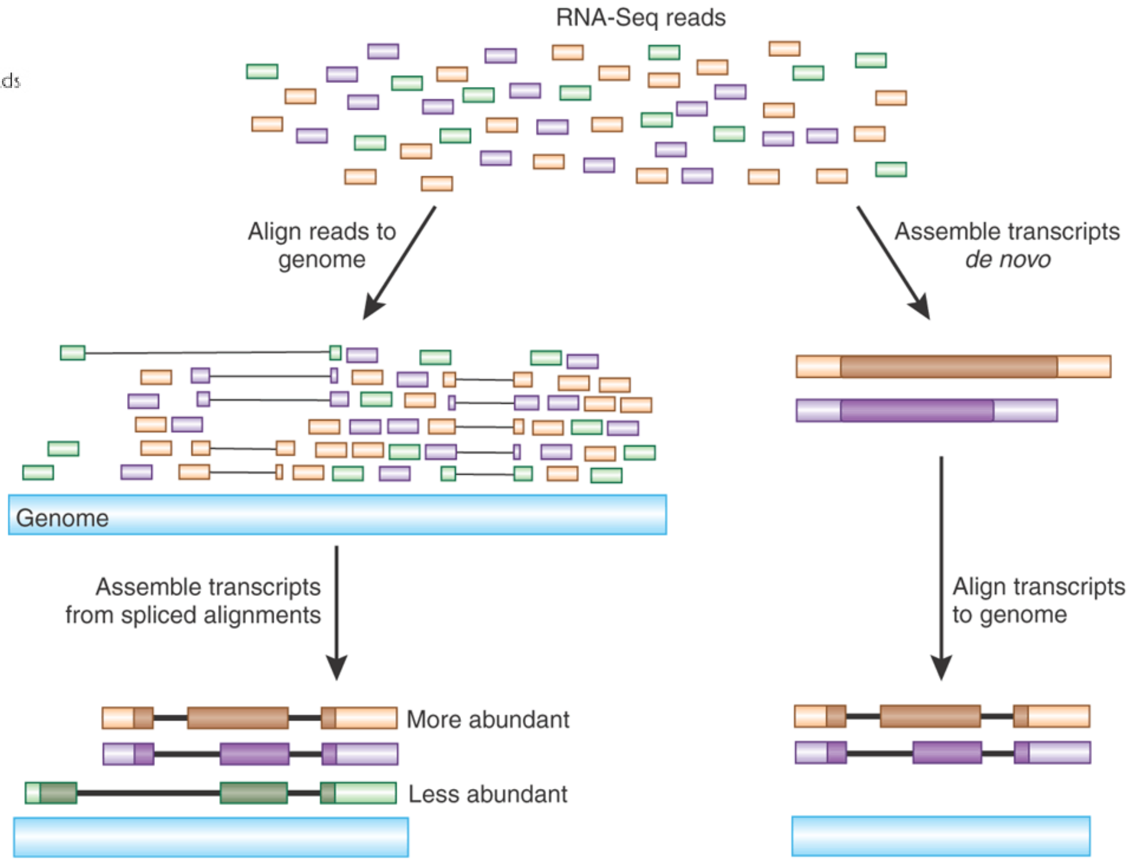
alleles of the sample 1 and 2

RNA sequencing


AAAAAAAA
mRNA
RNA fragments
or
AAAAAAAA
TTTTTTTT
cDNA
EST library
with adaptors
ATCACAGTGGGACTCCATAAATTTTTCT
CGAAGGACCAGCAGAAACGAGAGAAAAA
GGACAGAGTCCCCAGCGGGCTGAAGGGG
ATGAAACATTAAAGTCAAACAATATGAA
Short sequence reads

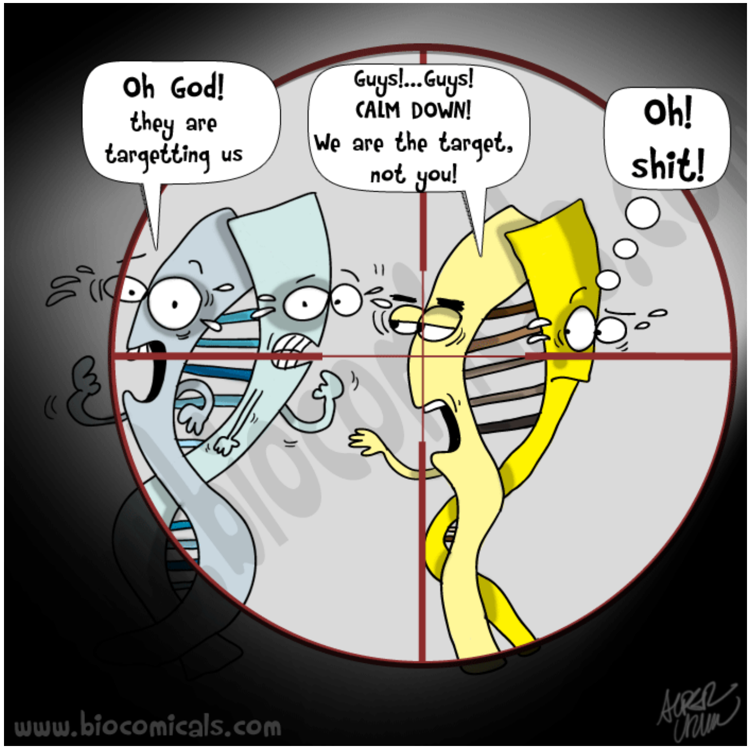
- Allows to obtain sequence of only transcribed parts of the genome.
- Coding sequencing, non-coding RNAs
- We need to isolate RNA from the tissue.

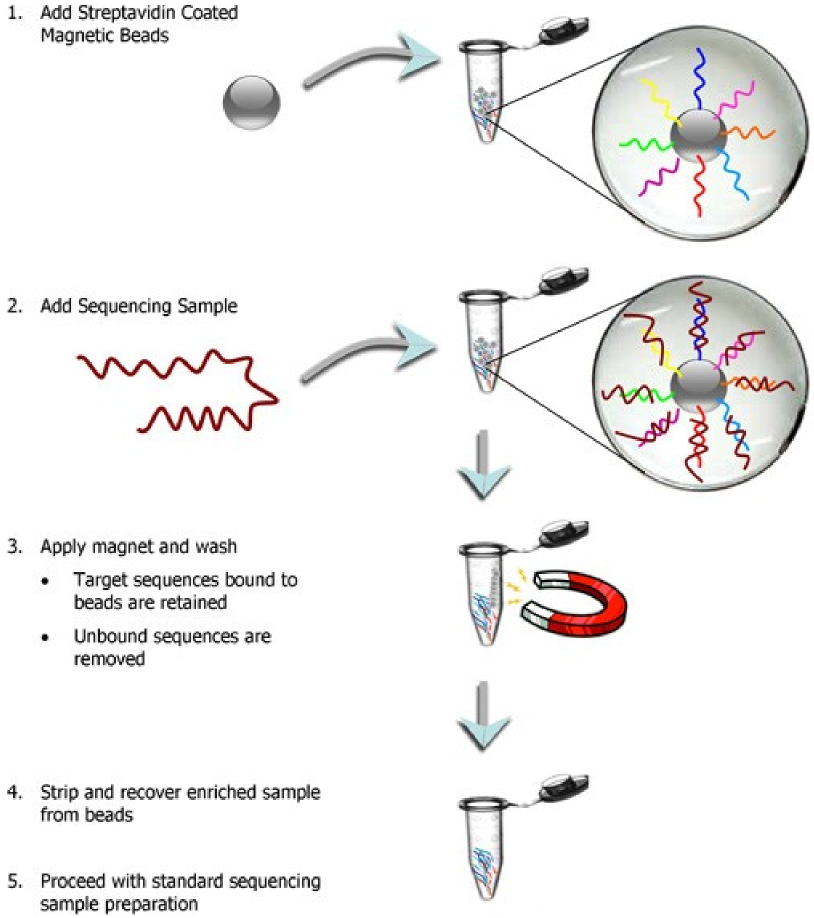
- Multiple samples can be pooled in the same run.
- Can be differentiated by tags.


RNA-Seq reads
Align reads to genome
Assemble transcripts *de novo*
Genome
Assemble transcripts from spliced alignments
Align transcripts to genome
More abundant
Less abundant

Targeted sequencing
Hybridization-based capture

- Based on hybridization to designed probes, we first select sequences that will be later sequenced.
- We can design probes to individual genes, whole chromosome, or exome (exome sequencing).

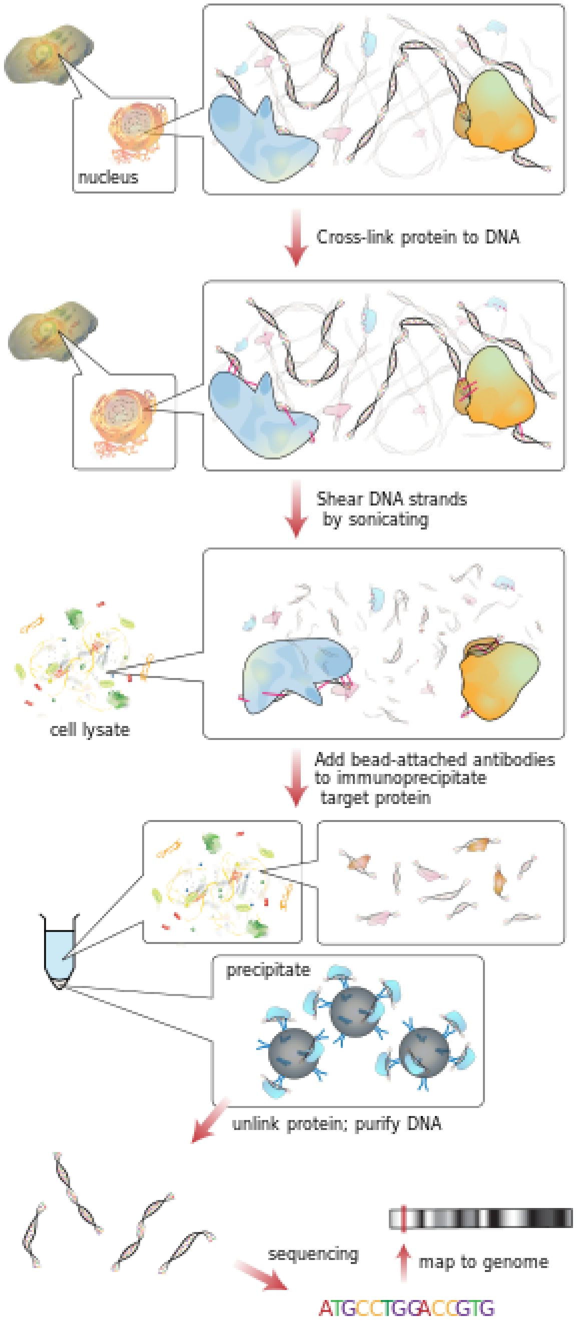

Oh God! they are targetting us
Guys!...Guys! CALM DOWN! We are the target, not you!
Oh! shit!
www.biocomicals.com



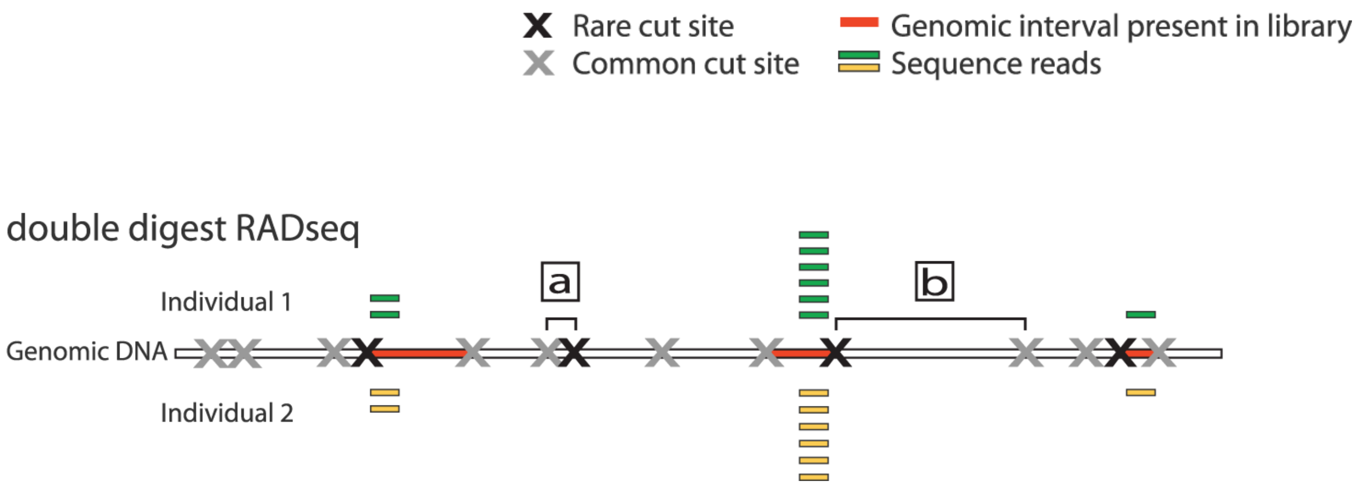
1. Add Streptavidin Coated Magnetic Beads
2. Add Sequencing Sample
3. Apply magnet and wash
   - Target sequences bound to beads are retained
   - Unbound sequences are removed
4. Strip and recover enriched sample from beads
5. Proceed with standard sequencing sample preparation

ChIP (Chromatin Imunoprecipitation) sequencing

- Identification of sequences recognized by particular DNA biding proteins (transcription factors etc.).

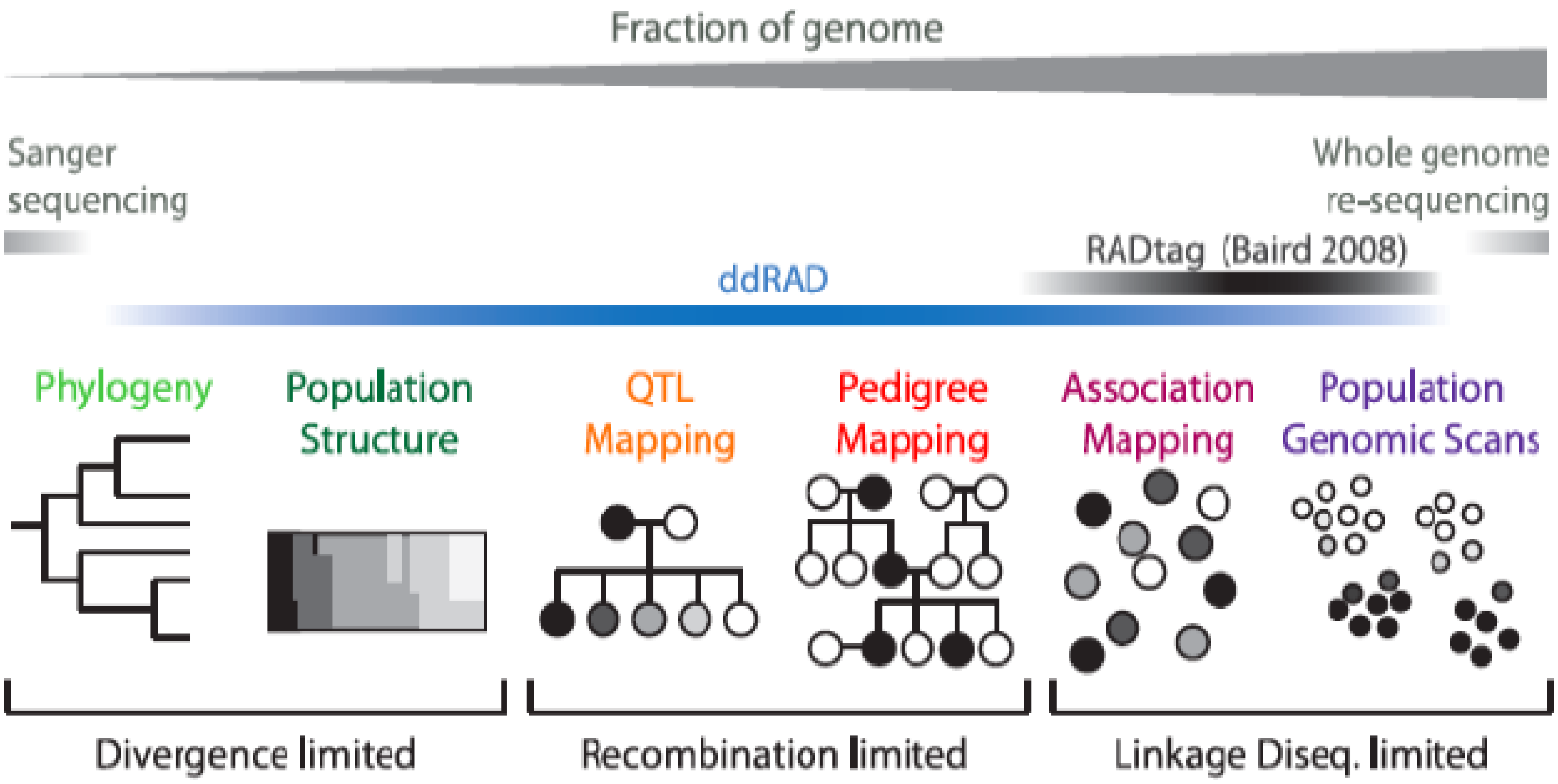

nucleus
Cross-link protein to DNA
Shear DNA strands by sonicating
cell lysate
Add bead-attached antibodies to immunoprecipitate target protein
precipitate
unlink protein; purify DNA
sequencing
map to genome
ATGCCTGGACCGTG

Restriction site associated DNA sekvenování (RAD seqeuencing)



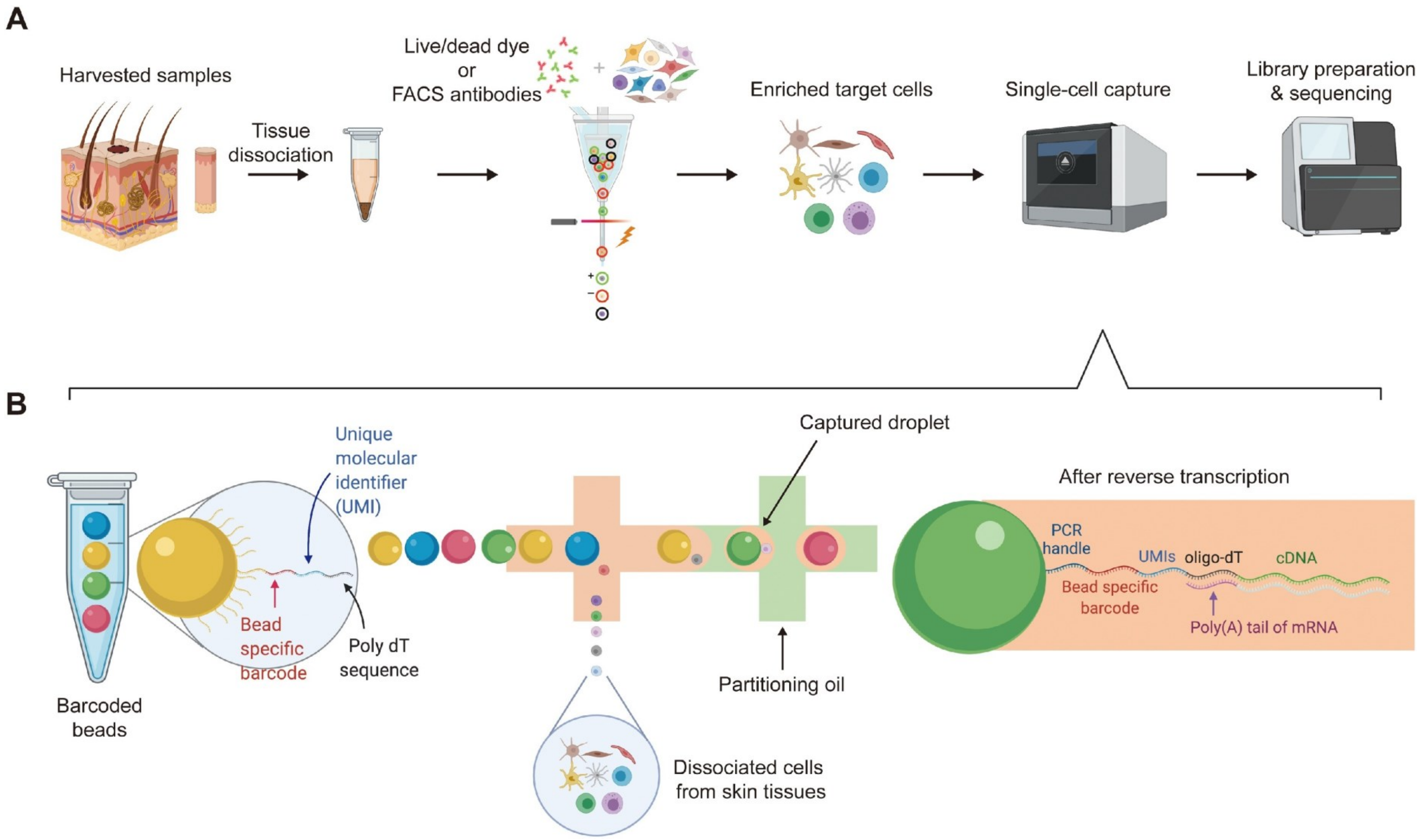
- Štěpení genomové DNA pomocí jednoho či dvou (double-digest) restrikčních enzymů.
- Výběr restrikčních fragmentů jen určité velikosti
- Sekvenování krátkých úseků vybraných fragmentů.
- Umožňuje získat stejné sekvence z mnoha jedinců.
- Do jednoho runu lze poolovat stovky až tisíce jedinců.

Využití ddRAD sekvenování

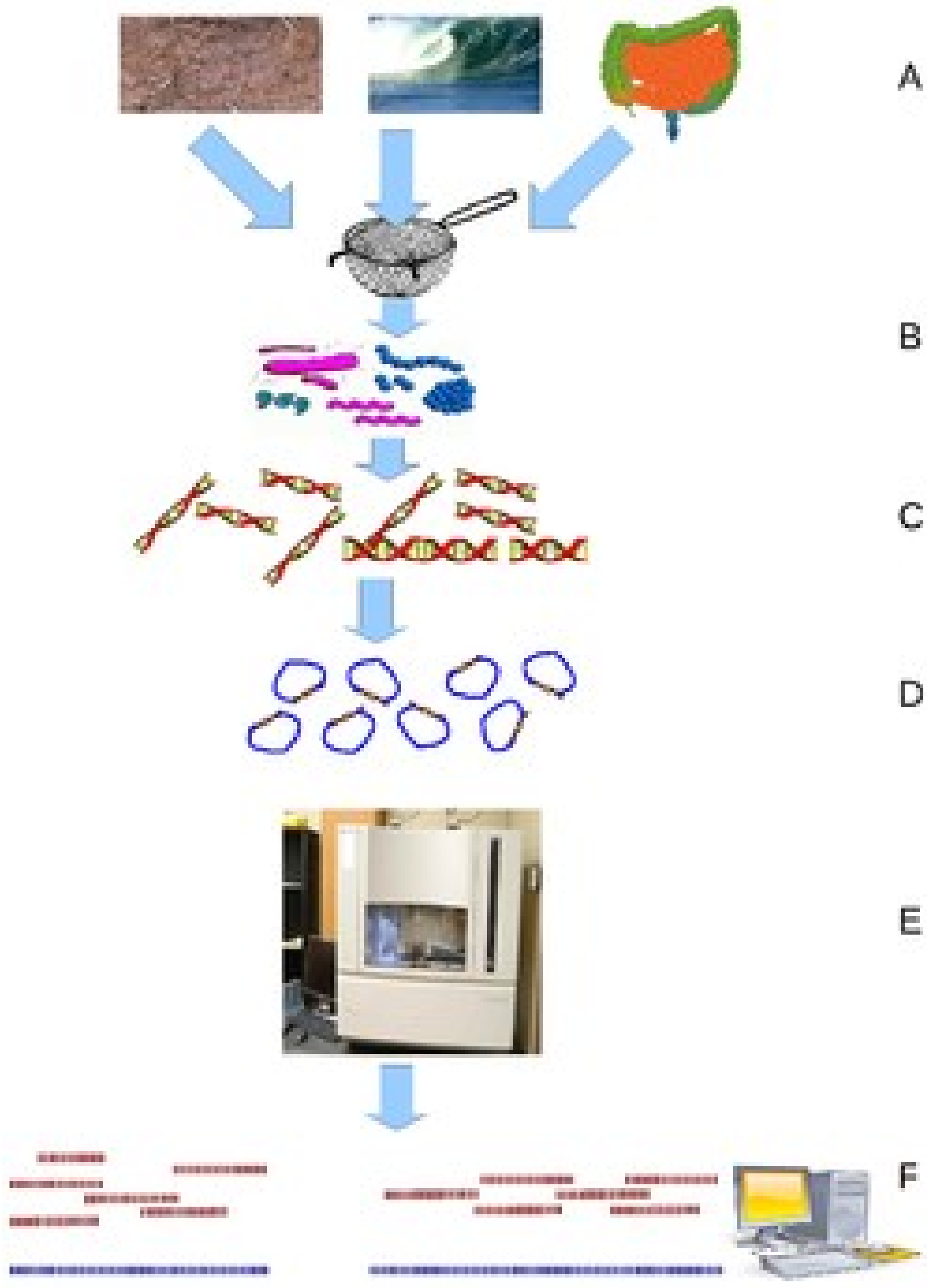


10X Genomics

Single cell sequencing



- Zjištění genové exprese v jednotlivých buňkách.
- Získání sekvencí DNA z jednotlivých buněk.
- Identifikace mutací v nádorových buňkách.

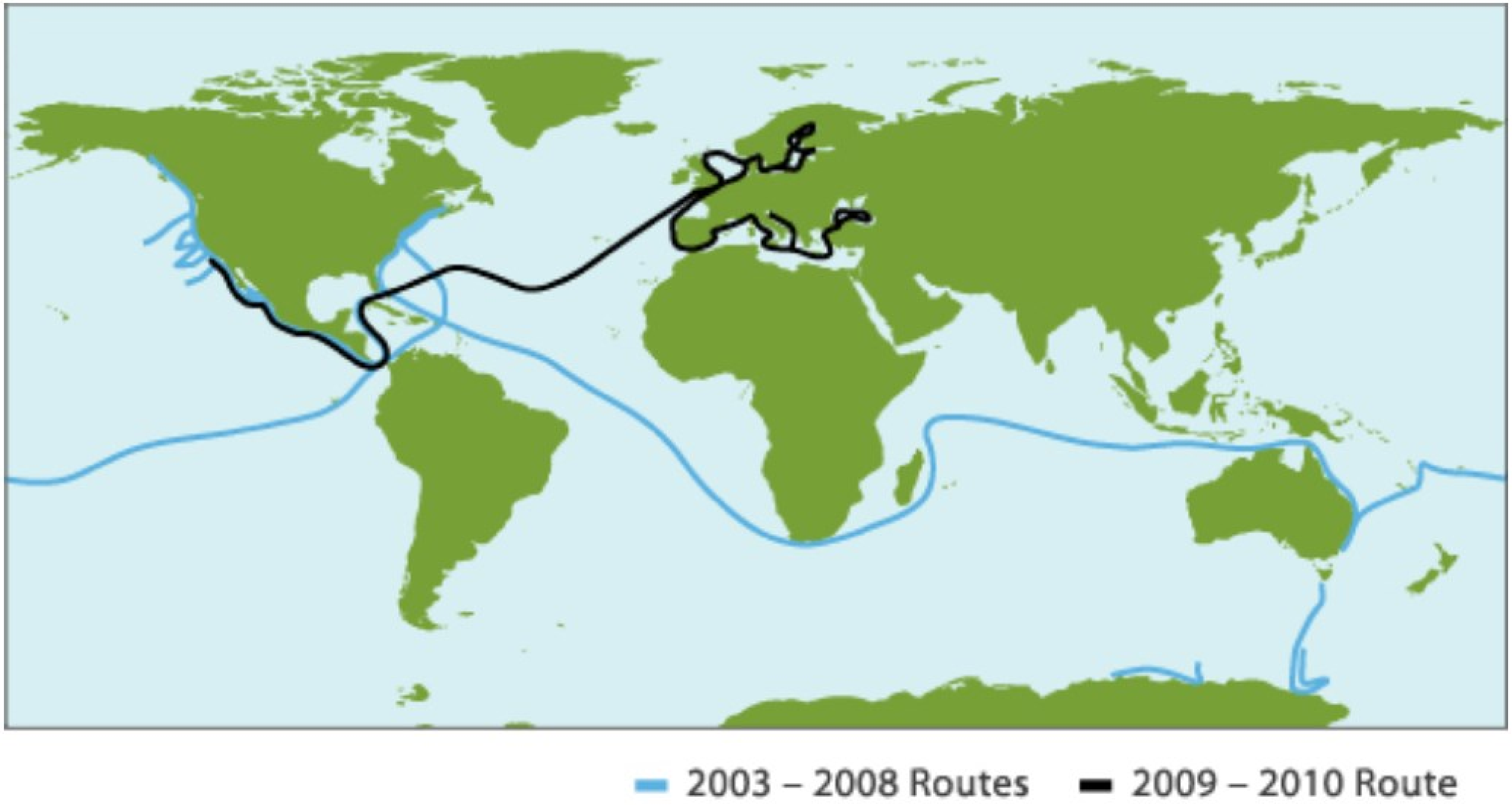


Metagenomics

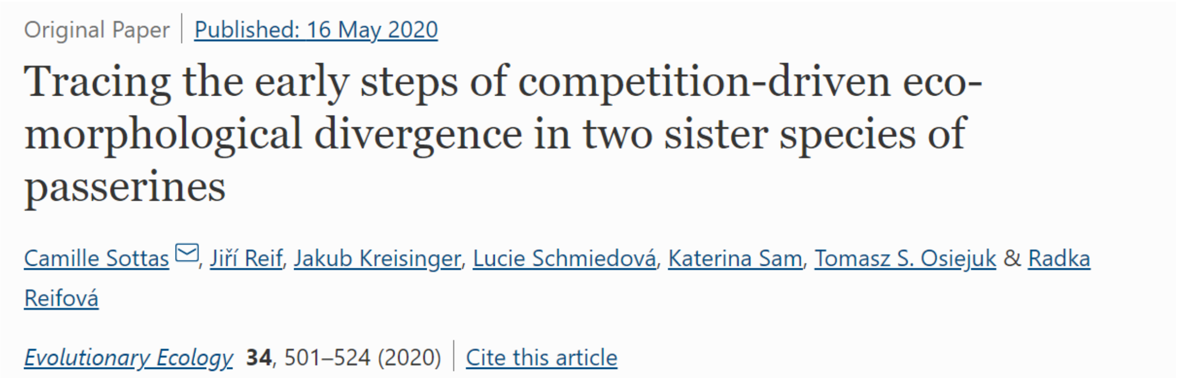
- Identification of organisms in various samples (soil, water, gut samples etc.)
- Enables identification of species which cannot be cultivated.
- Barcoding. PCR amplification of specific genes: 16S rRNA, cytochrome c oxidase I (COI).
- Comparison of obtained sequences with available databases.

Metagenomics

- Craig Venter (2003 - 2010) - Global Ocean Sampling Expedition



Identification of prey species



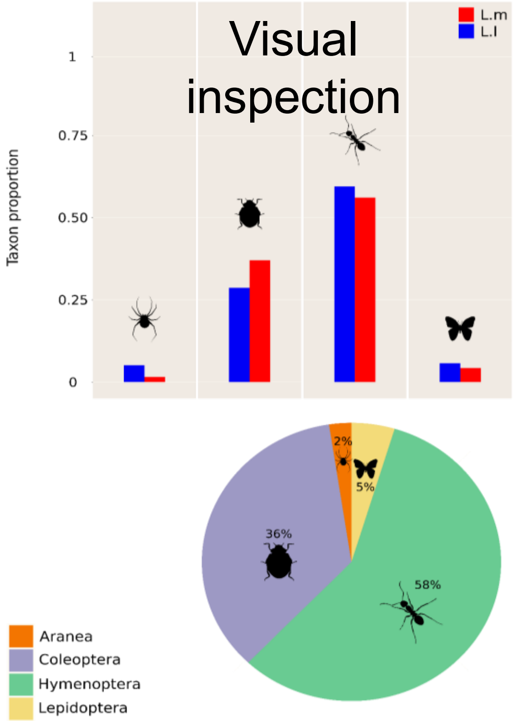
Original Paper | Published: 16 May 2020

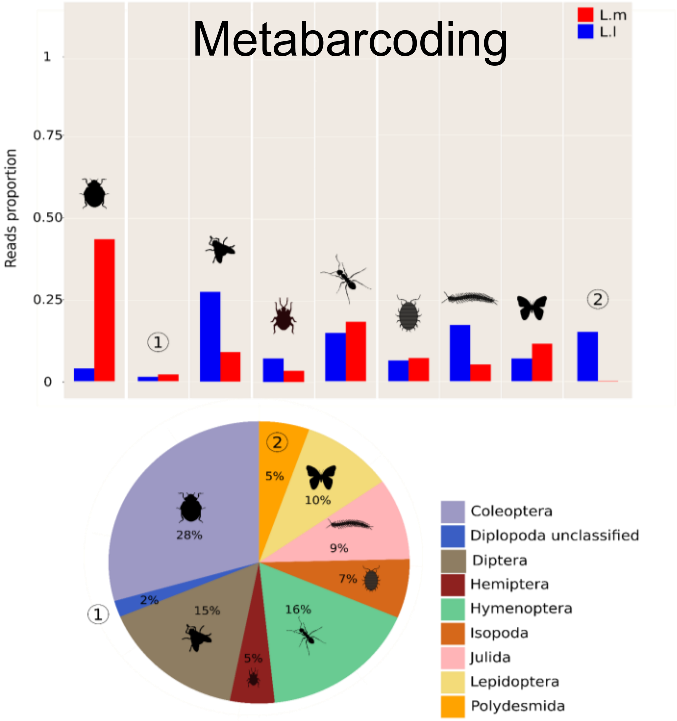
Tracing the early steps of competition-driven eco-morphological divergence in two sister species of passerines

Camille Sottas, Jiří Reif, Jakub Kreisinger, Lucie Schmiedová, Katerina Sam, Tomasz S. Osiejuk & Radka Reifová

Evolutionary Ecology **34**, 501–524 (2020) | Cite this article

PCR amplification of cytochrome c oxidase I (COI)
Using primers targeting a broad range of invertebrate taxa

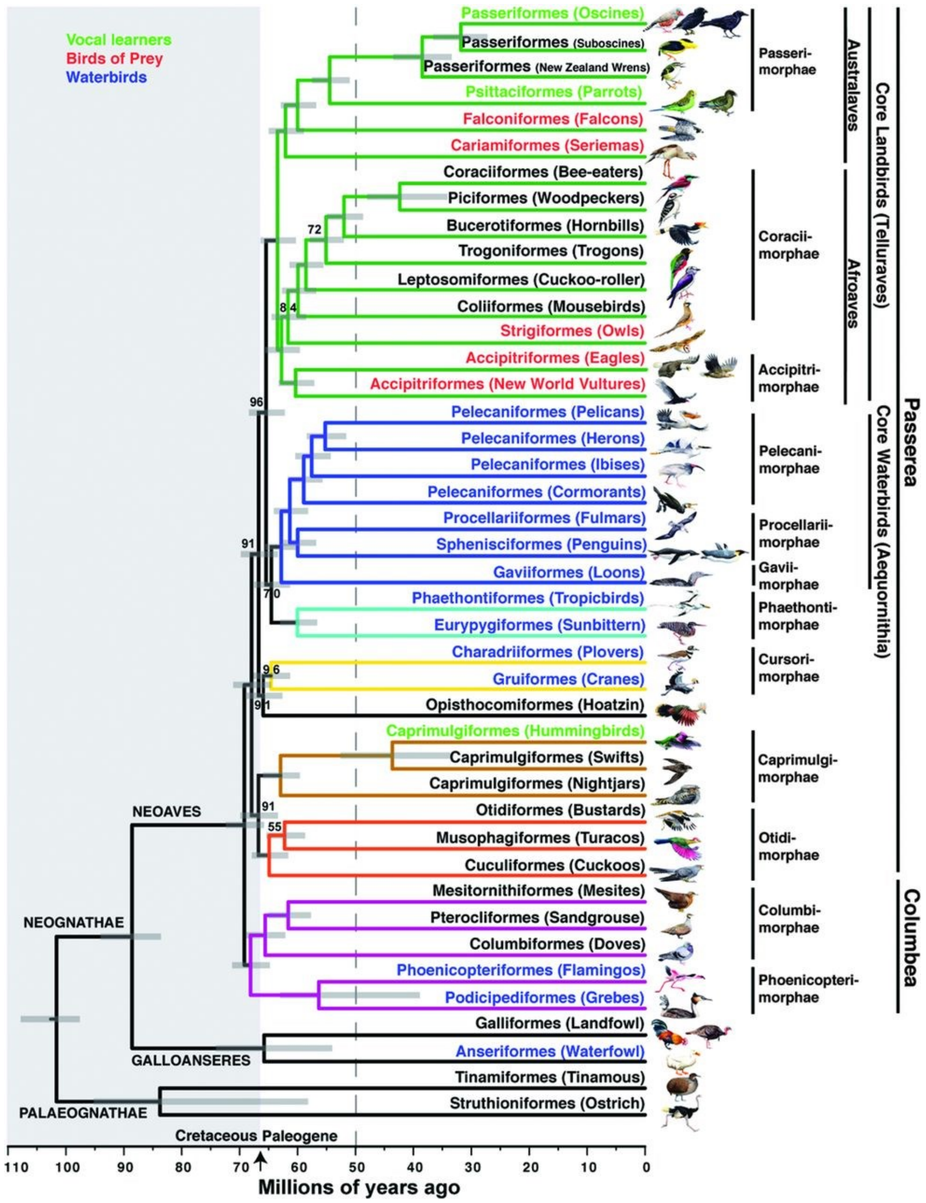




How can be sequence data used in zoology?

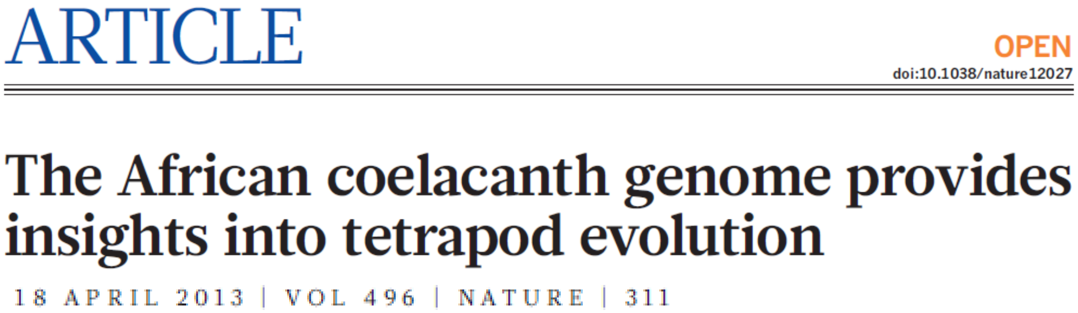
Fylogenomika

- Fylogeneze ptáků založená na celogenomových sekvencích 48 zástupců všech ptačích řádů.



Jarvis et al. Science 2014

Komparativní genomika



ARTICLE

OPEN

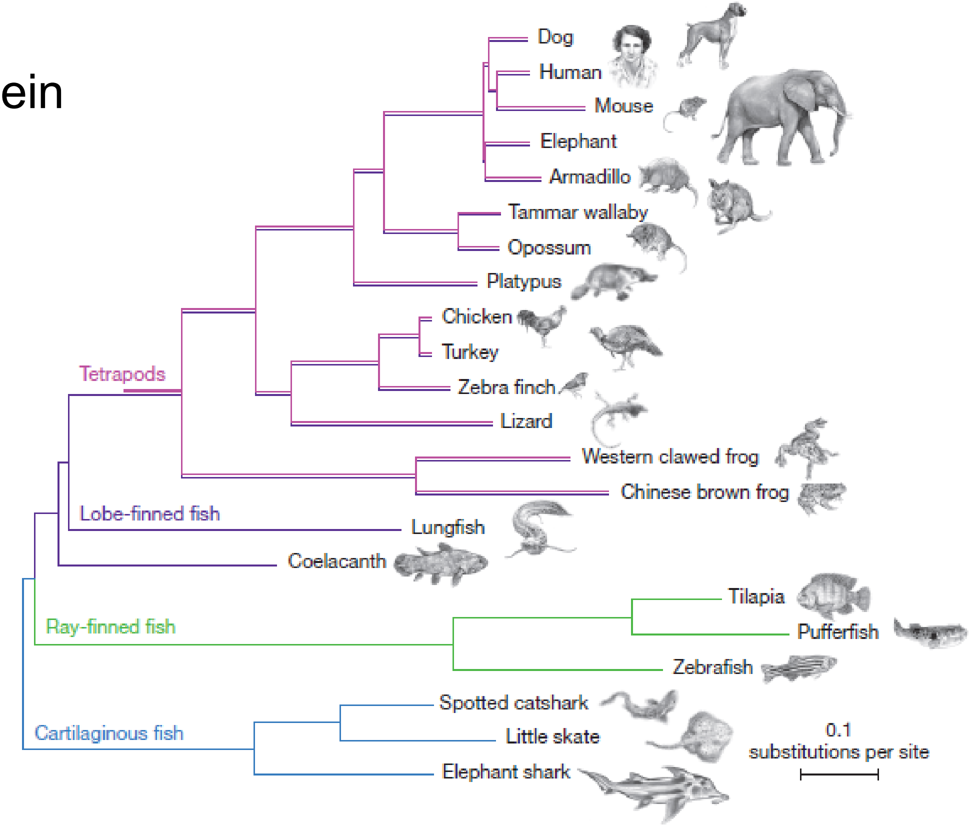
doi:10.1038/nature12027

The African coelacanth genome provides insights into tetrapod evolution

18 APRIL 2013 | VOL 496 | NATURE | 311

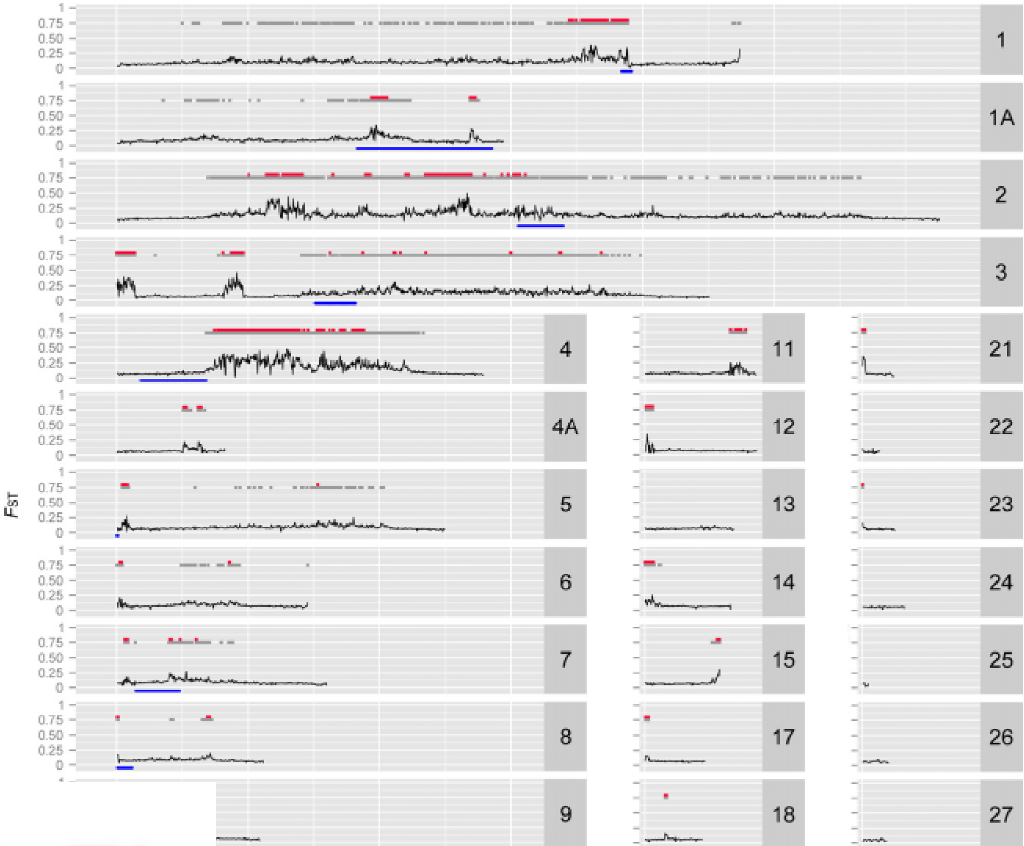
2x pomalejší substituční rychlost protein kódujících sekvencí ve srovnání s ostatními tetrapody.

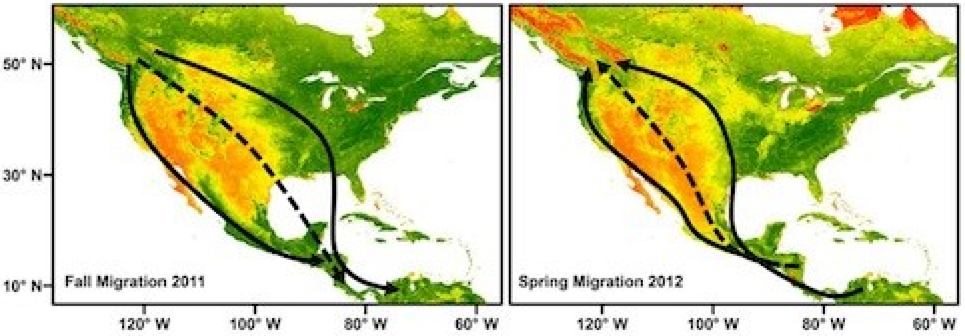
Latimérie podivná



Populační a evoluční genomika

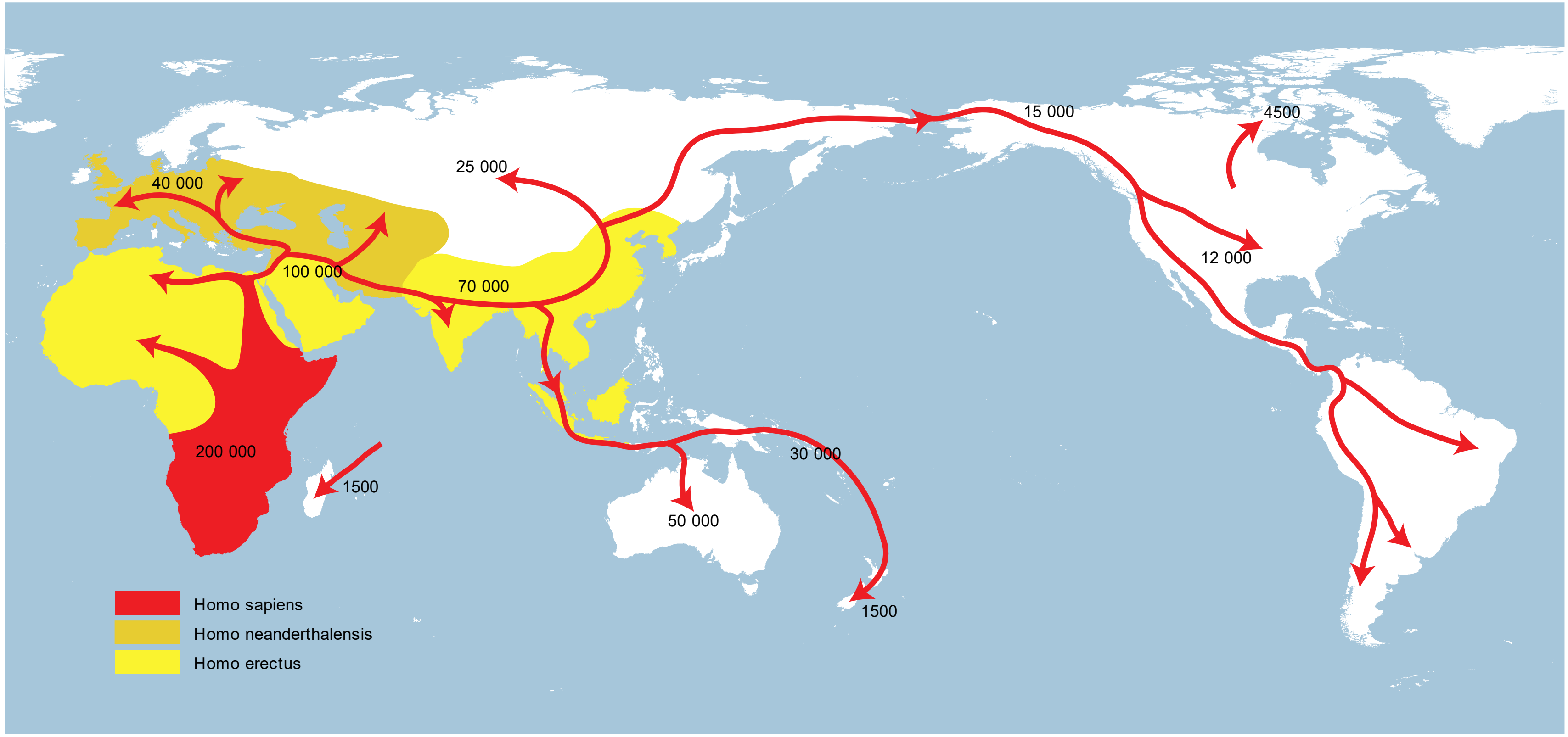
Migrace a speciace u drozda malého





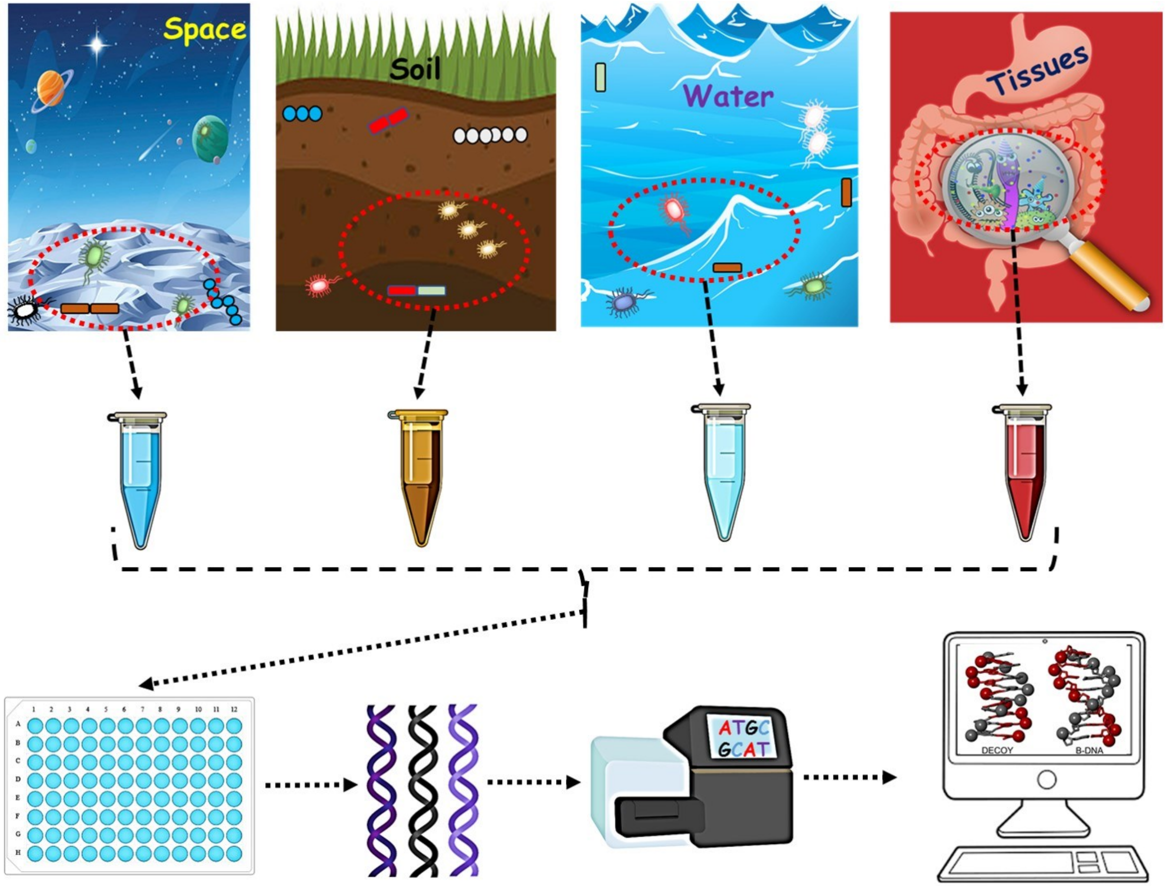
Delmore et al. 2015

Fylogeografie

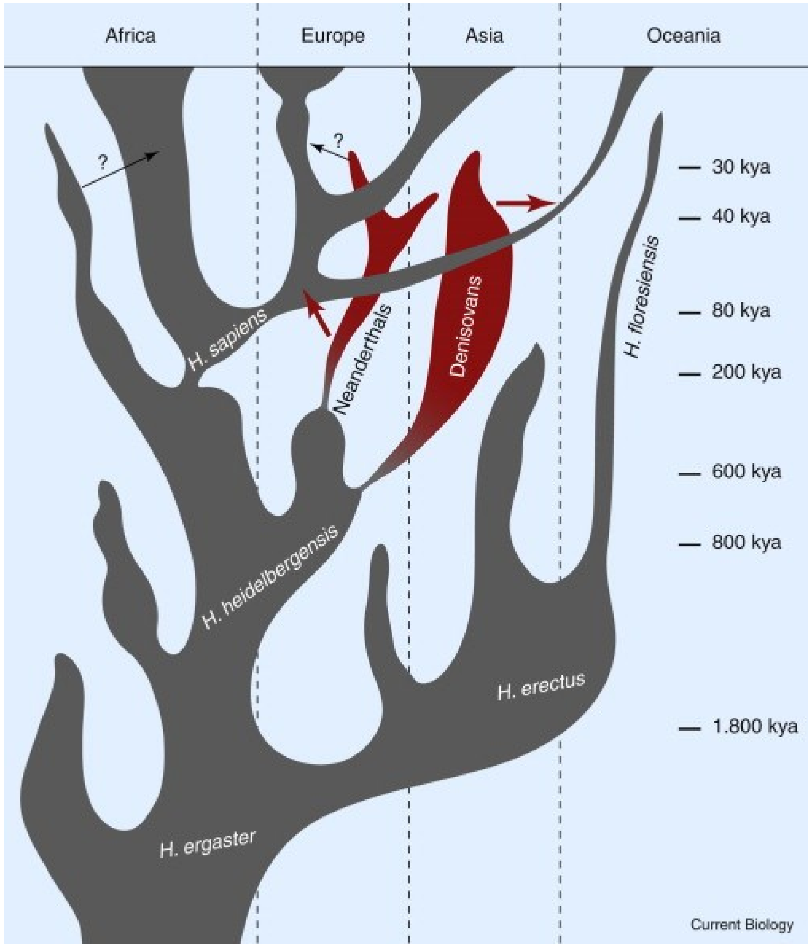


Metagenomika

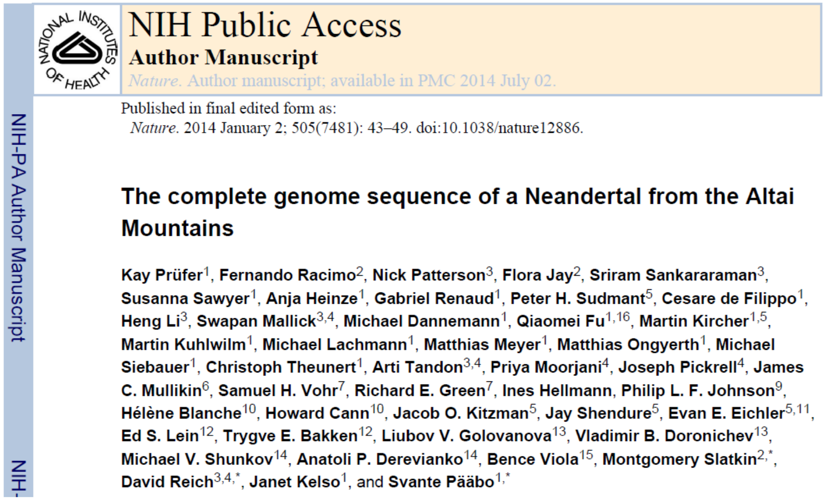
- Identifikace mikroorganismů žijicích v určitých prostředích.
- Lze identifikovat i nekultivovatelné bakterie a jiné mikroorganismy.
- Identifikace potravy.



Paleogenomika



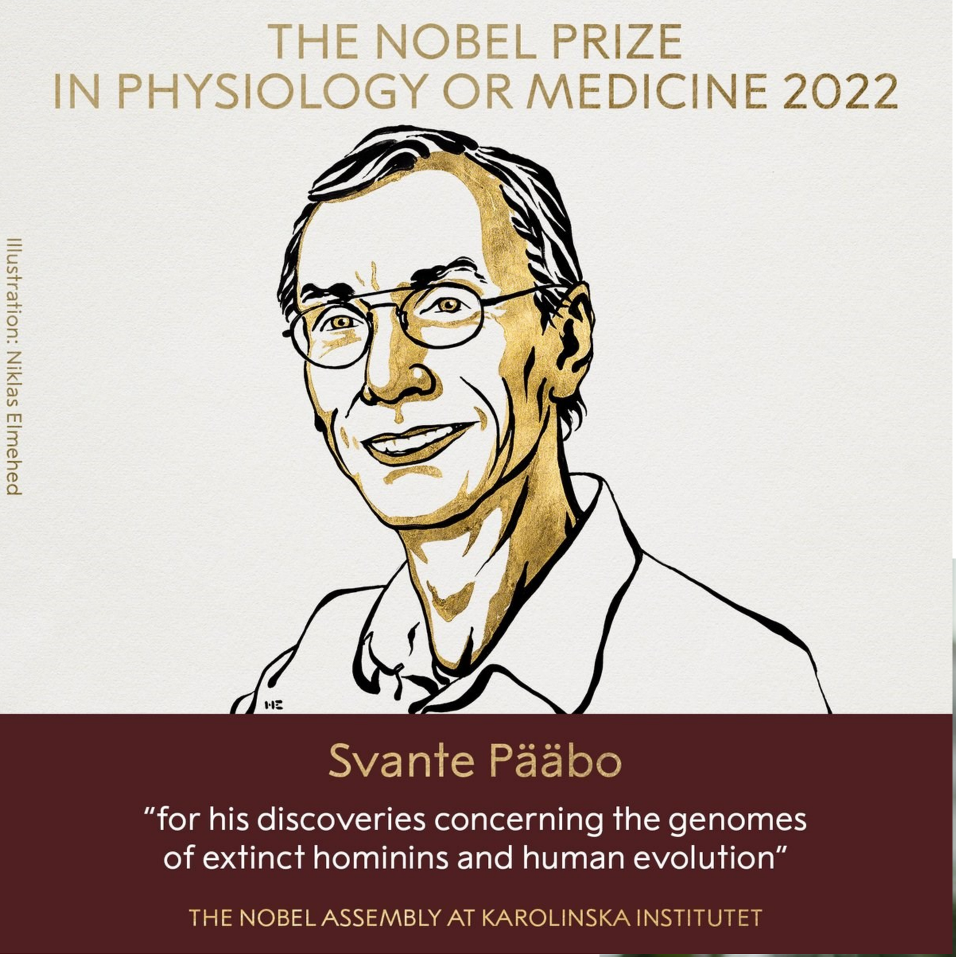
Current Biology



Published in final edited form as:
Nature. 2014 January 2; 505(7481): 43–49. doi:10.1038/nature12886.

The complete genome sequence of a Neandertal from the Altai Mountains

Kay Prüfer[1], Fernando Racimo[2], Nick Patterson[3], Flora Jay[2], Sriram Sankararaman[3], Susanna Sawyer[1], Anja Heinze[1], Gabriel Renaud[1], Peter H. Sudmant[5], Cesare de Filippo[1], Heng Li[3], Swapan Mallick[3,4], Michael Dannemann[1], Qiaomei Fu[1,16], Martin Kircher[1,5], Martin Kuhlwilm[1], Michael Lachmann[1], Matthias Meyer[1], Matthias Ongyerth[1], Michael Siebauer[1], Christoph Theunert[1], Arti Tandon[3,4], Priya Moorjani[4], Joseph Pickrell[4], James C. Mullikin[6], Samuel H. Vohr[7], Richard E. Green[7], Ines Hellmann, Philip L. F. Johnson[9], Hélène Blanche[10], Howard Cann[10], Jacob O. Kitzman[5], Jay Shendure[5], Evan E. Eichler[5,11], Ed S. Lein[12], Trygve E. Bakken[12], Liubov V. Golovanova[13], Vladimir B. Doronichev[13], Michael V. Shunkov[14], Anatoli P. Derevianko[14], Bence Viola[15], Montgomery Slatkin[2,*], David Reich[3,4,*], Janet Kelso[1], and Svante Pääbo[1,*]



THE NOBEL PRIZE
IN PHYSIOLOGY OR MEDICINE 2022

Illustration: Niklas Elmehed

Svante Pääbo

“for his discoveries concerning the genomes of extinct hominins and human evolution”

THE NOBEL ASSEMBLY AT KAROLINSKA INSTITUTET