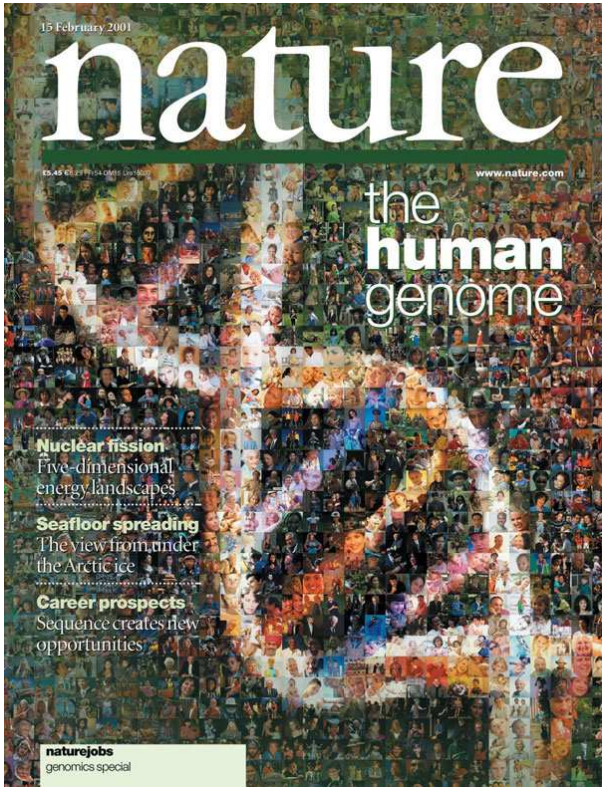
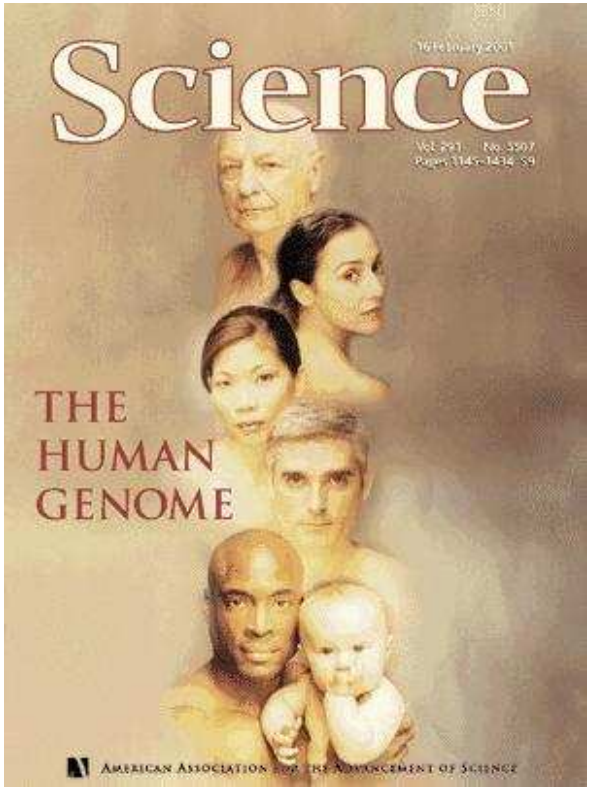
Sequencing and Genome Assembly
(part 2 - long reads)

Tomáš Vinař

30.09.2021





Overview of Sequencing Technologies

Technology	Read length	Errors	Output per day	Cost per MB
1st generation				
Sanger	up to 1000bp	$< 1\%$	3 MB	\$4000
2nd (next) generation (cca 2004)				
Illumina	250bp	$< 0.1\%$	150 GB	\$0.03
3rd generation (emerging)				
PacBio	cca 14kbp	10%	700 GB	\$0.02
PacBio HiFi	cca 15kbp	$< 1\%$	70 GB	\$0.20
Oxford Nanopore	really long	up to 10%	50 GB	\$0.02

From the last lecture

- Genome is assembled from sequencing reads
- Genome assembly using de Bruijn graphs
- de Bruijn graphs not suitable for long reads with high error rate
  - "Disassembly" to k-mers throws away too much information (read length 10000+, k is usually between 30 and 70)
  - Error rate around 10% makes de Bruijn graph unwieldy (for $k = 31$, k-mer 3 errors on average)

Overlap–Layout–Consensus approach

- **Overlap:** Find overlaps between reads
  and create an **overlap graph**
- **Layout:** Simplify the overlap graph and find paths which will correspond to **contigs**
- **Consensus:** For each contig locate overlapping reads and construct a sequence as a consensus at each position
  (corrects local errors)

Overlap: Finding read overalps

```
CATCTCTAGGCCAGC
      ||||||  ||
      TAGGCCTGCTTCTTG
```

- special case of the sequence alignment (next lecture)
- overlaps **will contain errors**
  (in our case approx. 1 error per 10bp of the overlap)
- **there are many reads:** $30\times$ human genome coverage
  $\Rightarrow$ cca 9 mil. of reads of length 10000
  we cannot afford to compare all pairs of reads
- practical approac:
  - fast pre-filtering of **suitable candidate pairs of reads**
    (for example those containing a common k-mer)
  - followed by a slower alignment for candidate pairs

Layout: Creating the overlap graph

- Example result from the previous phase:
  CATCTCTAGGCCAGC / TAGGCCTGCTTCTTG, overlap 9 bp
  ...
- Create **overlap graph**:
  vertices: reads weighted edges: overlaps and lengths

Example:

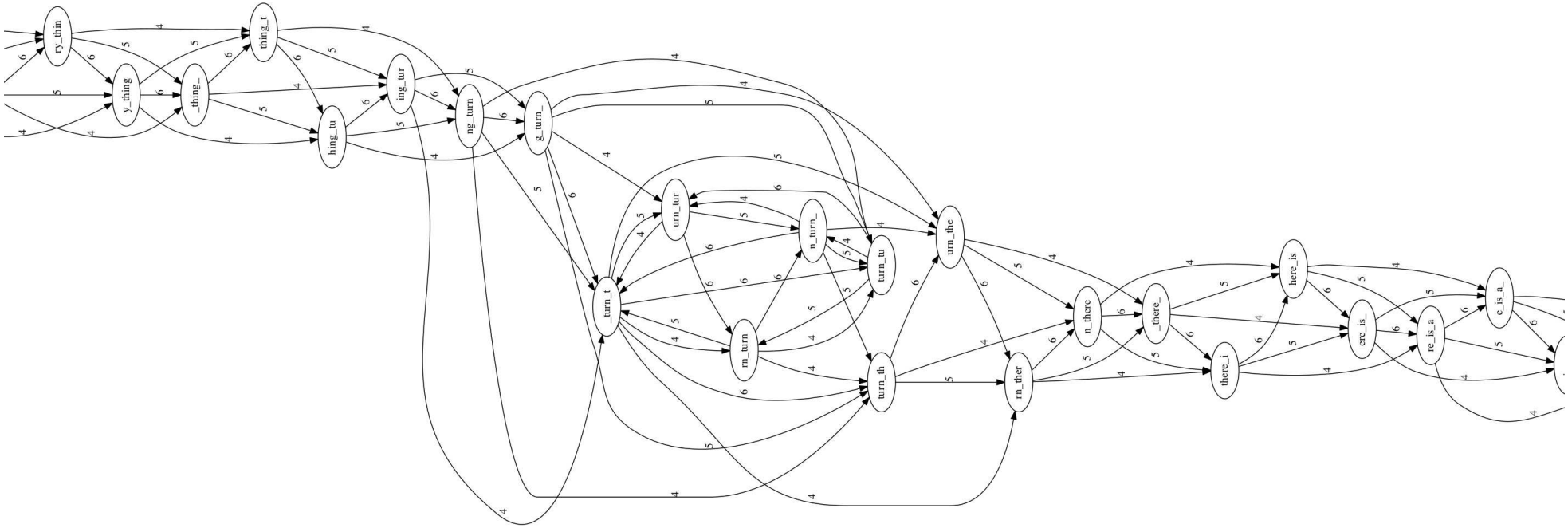
`to_every_thing_turn_turn_turn_there_is_a_season`

read length 7, minimum required overlap 4

Example:

`to_every_thing_turn_turn_turn_there_is_a_season`

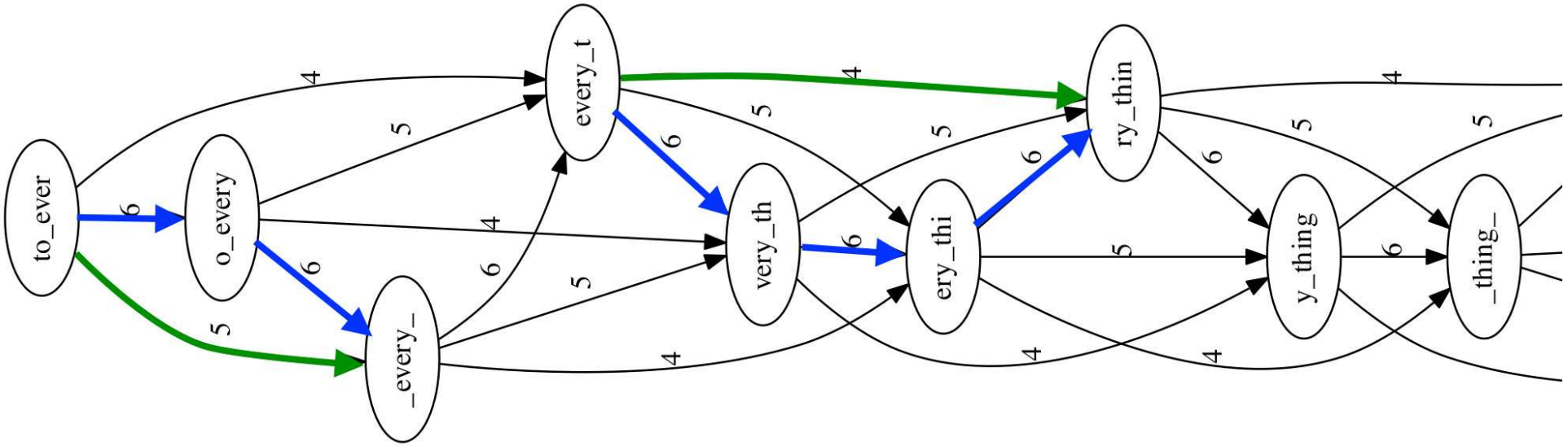
read length 7, minimum required overlap 4



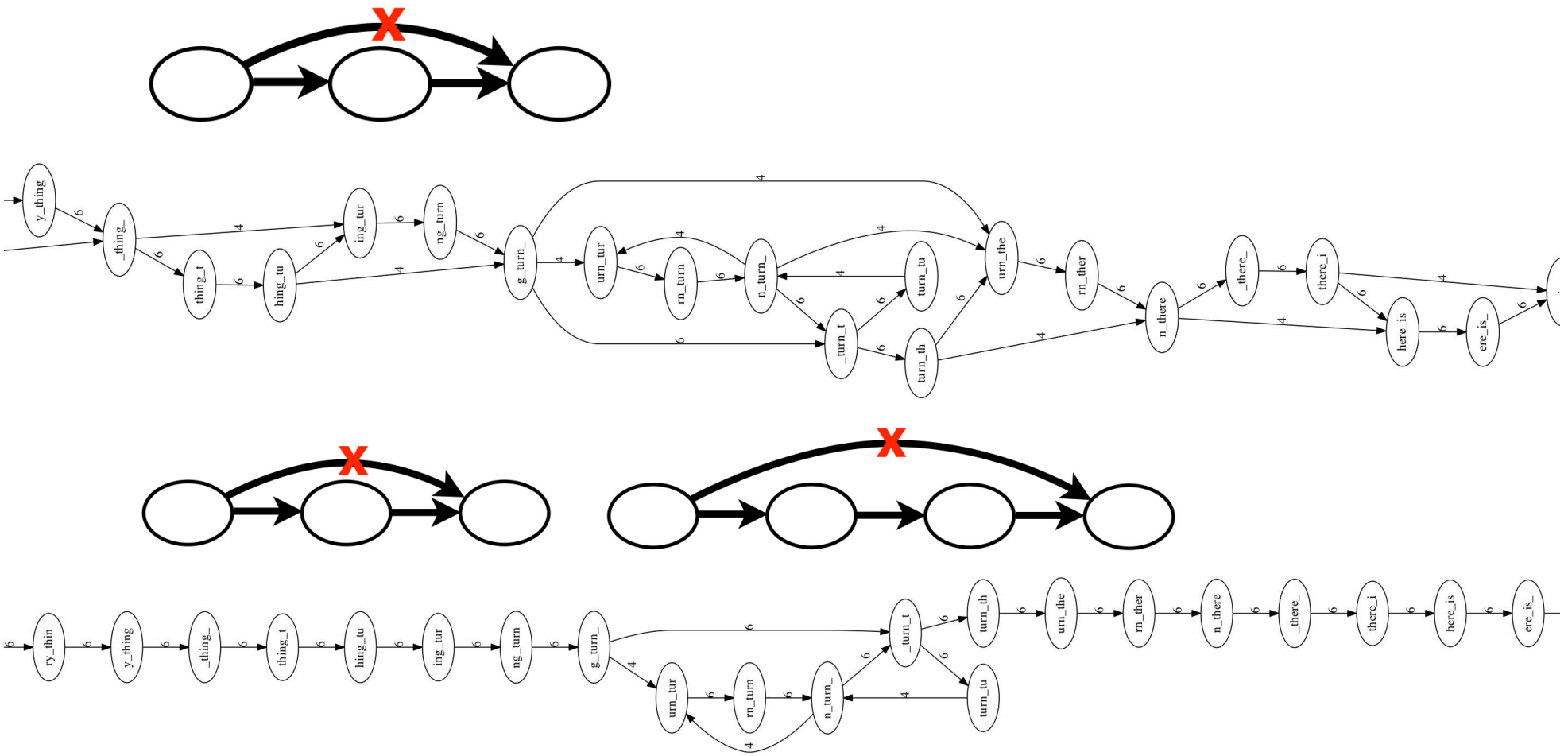
Example and figures by Ben Langmead

Layout: Transitive edges

- Some edges are superflous because they say the same thing as other edges



Layout: Removal of transitive edges

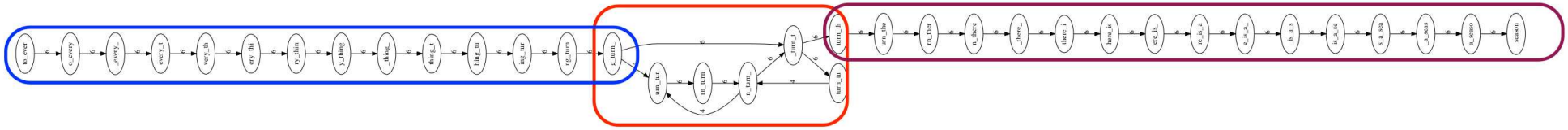


Layout: Identifying contigs

Original sequence:

`to_every_thing_turn_turn_turn_there_is_a_season`

Non-branching paths represent contigs



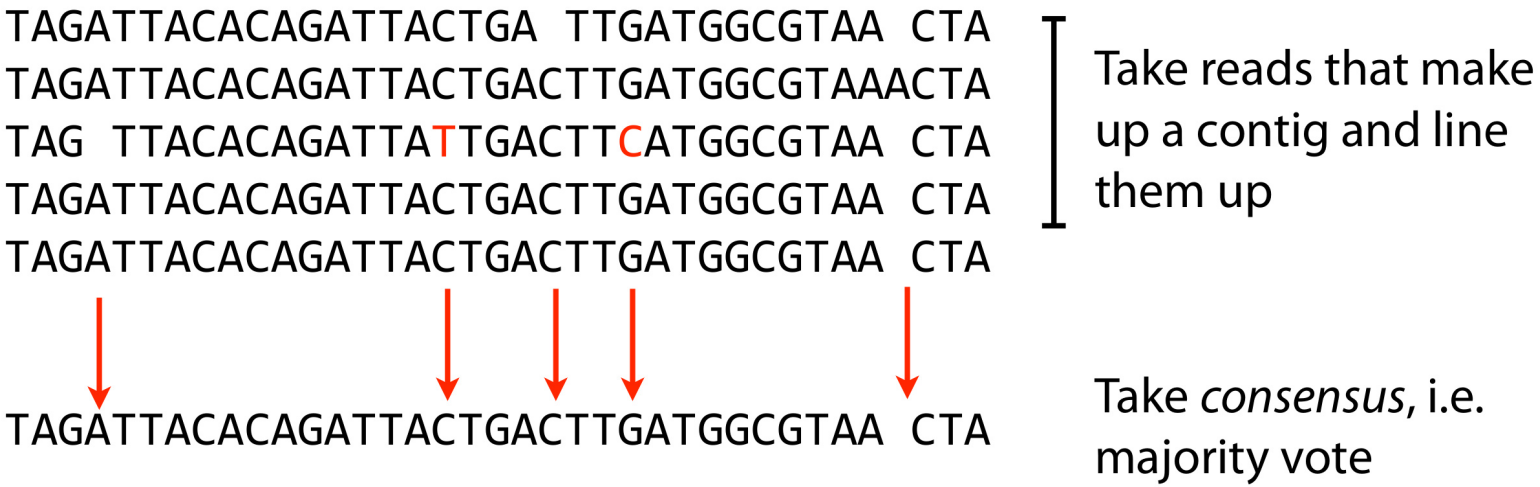
Result:

Contig 1
`to_every_thing_turn_`

Contig 2
`turn_there_is_a_season`

Unresolvable repeat

Consensus: Obtaining the final sequence



Differences between de Bruijn graphs and the overlap graph

de Bruijn graphs

- fixed length of overlaps
- throw away information about contiguity spanning more than k bp
- genome represented by paths
- errors: bubbles and tips
- errors resolved in pre-processing
- contigs cover almost all edges

Overlap graphs

- variable length of overlaps
- use most of the information derived from overlaps
- genome represented by paths
- errors are "hidden"
- errors resolved in post-processing (consensus)
- transitive edges need to be removed

Example: Assembling genome of *Magnusiomyces capitatus*

(genome length 19.6 Mbp, 4 chromosomes + mtDNA)

Technology	Coverage	# contigs	largest	avg	N50
Illumina / Spades	250x	1102	172.6 Kbp	17.6 Kbp	62.0 Kbp
PacBio / Canu	37x	17	4.7 Mbp	1.2 Mbp	1.7 Mbp
PacBio + MinION	65x	11	4.4 Mbp	1.8 Mbp	2.0 Mbp

Summary

- Long reads allow us to assemble much more contiguous genome sequences compared to short reads
- Fast algorithms required to locate read overlaps
  (more in the next lecture)
- Overlap graphs and de Bruijn graphs are similar concepts
  attempts at unifying the two

Genome Sequencing Milestones

1976	MS2 (RNA virus) 40 kB
1988	Human genome sequencing project (15 years)
1995	bacterium H. influenzae 2 MB, shotgun (TIGR)
1996	S. cerevisiae 10 MB, BAC-by-BAC (Belgium, UK)
1998	C. elegans 100 MB, BAC-by-BAC (Wellcome Trust)
1998	Celera: human genome in three years!
2000	D. melanogaster 180 MB, shotgun (Celera, Berkeley)
2001	2x human genome 3 GB (NIH, Celera)
after 2001	mouse, rat, chicken, chimpanzee, dog,. . .
2007	Genomes of Watson and Venter (454)
2012	1000 human genomes
soon	10k vertebrate genomes, sequencing as a diagnostic tool
2021	3.5 million SARS-CoV-2 genomes

Use of NGS: Population genetics

- Obtain sequence reads from one individual
- What are the differences of the individual from the "reference" genome?
- How do genetic change influence phenotype?
- Personalized medicine
- Population structure and history
- Ethical questions

Bioinformatics problems:

- Mapping short reads to reference sequence
- Identification of differences (both local and large-scale)

Use of NGS: Environmental sequencing – metagenomics

- What microorganisms live in our bodies?
  gut flora, mouth, skin, ...
- Microbial diversity in different ecosystems
- It is difficult to isolate individual species
- We can sequence a mixture of different genomes
- Then we try to assemble at least short contigs

Bioinformatics problems:

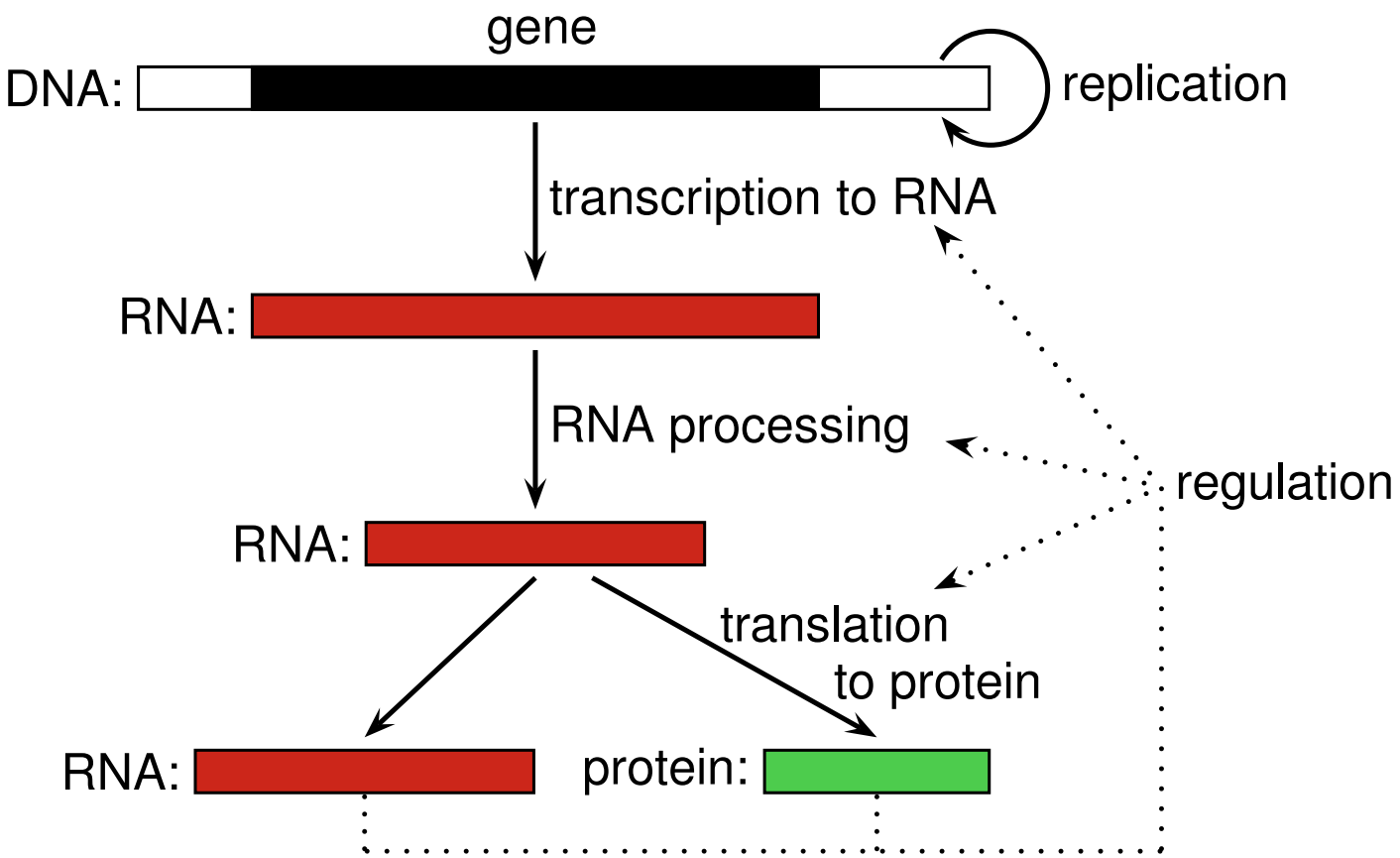
- Binning: Separation of reads from different genomes

Use of NGS: identification of genes, binding sites,...

- RNA-seq: sequencing mRNAs, obtaining positions of genes and their expression levels
- Chip-Seq: filtering DNA bound by a certain protein, sequening them and mapping to the genome

Bioinformatics problems:

- Identification of splice sites
- Identification of binding sites using read coverage


gene
DNA:
replication
transcription to RNA
RNA:
RNA processing
regulation
RNA:
translation
to protein
RNA:
protein: