Next Generation Sequencing: Adjusting to Big Data

Daniel Nicorici, Dr.Tech.

Statistikot Suomen Lääketeollisuudessa

29.10.2013

ORION

Building well-being

Outline

- Human Genome Project
- Next-Generation Sequencing
- Personalized Medicine
- RNA-seq
- Statistical analysis of genomic data
- Conclusions

One Human Genome

- 6 billion letters written in DNA
- four letters in the DNA alphabet – A, C, G, T - carry the instructions to make all living organisms
- 23 chromosomes

- written out, the one Human Genome would:
  - fill one million pages, or
  - fill 5000 books stacked 60 meters high, or
  - take a century to read out (24 hours a day)

Human Genome

- Angelina Jolie's BRCA1 gene and her decision to have mastectomy
- Ten years ago, James Watson (co-discoverer of the structure of DNA) said "*Once we had the genome sequenced, we would have the language of life*" but it turns out it is a language we do not understand
- Assymetry where the medical doctor has all the information and the patient has just a little bit => this is starting to change

Sequencing human genome

- Human Genome Project
  - first completed draft of human genome announced in year 2003 => cost: 2.7 billion $

- Today, year 2013:
  - sequencing using next-generation sequencing the genome of one person costs ~4000$ (that is 675 000 times cheaper than ten years ago!!!)

Sequencing human genome – cont'd

- Nice house (omakotitalo) in Turku

350 000 € in year 2003

ORION
Building well-being

Sequencing human genome – cont'd

- Nice house (omakotitalo) in Turku

350 000 € in year 2003 → **0.52 € in year 2013**

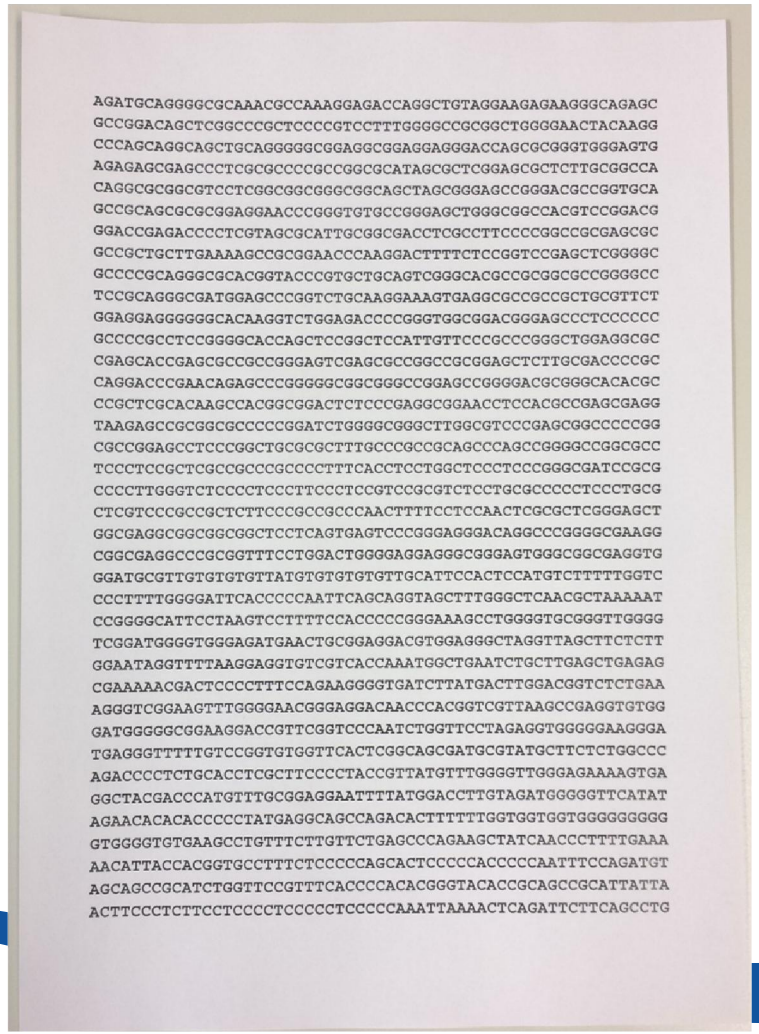
The U.K. plans to sequence the genome of 100 000 patients by year 2017!!!

ORION
Building well-being

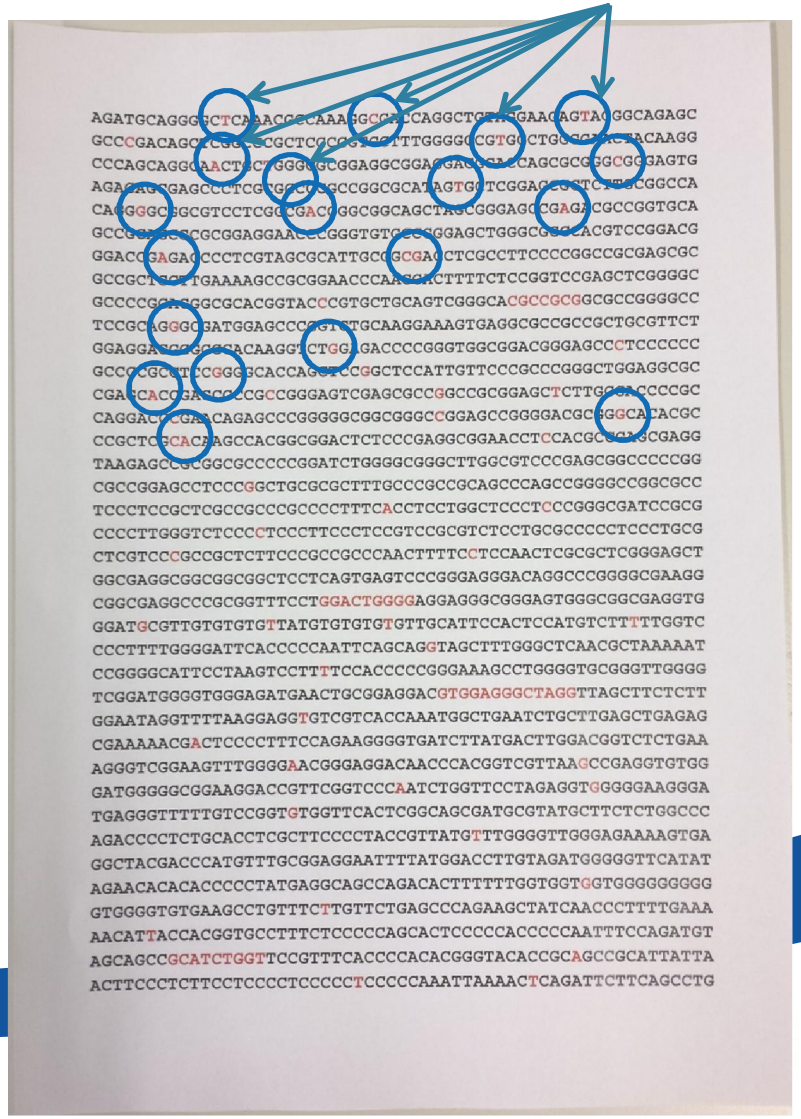
Next Generation Sequencing – cont'd

differences versus the reference genome

Reference genome

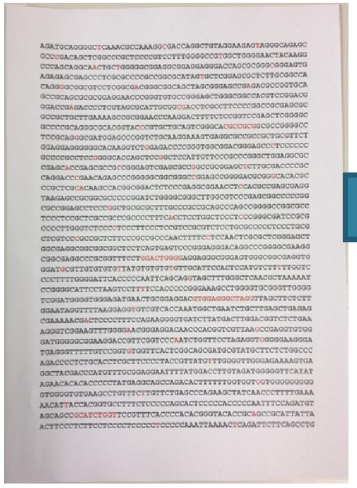


Genome of person A

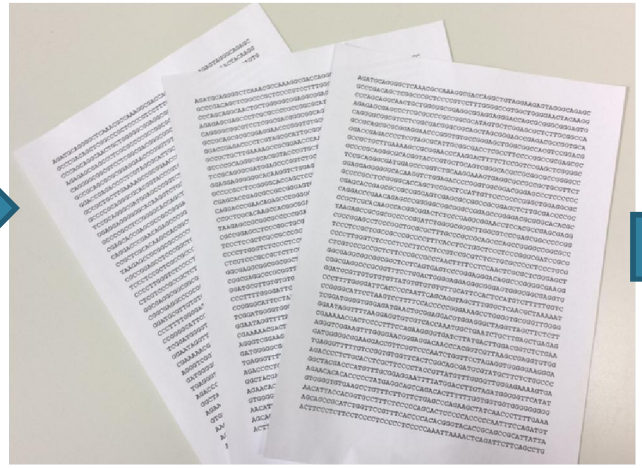


Next Generation Sequencing – cont'd

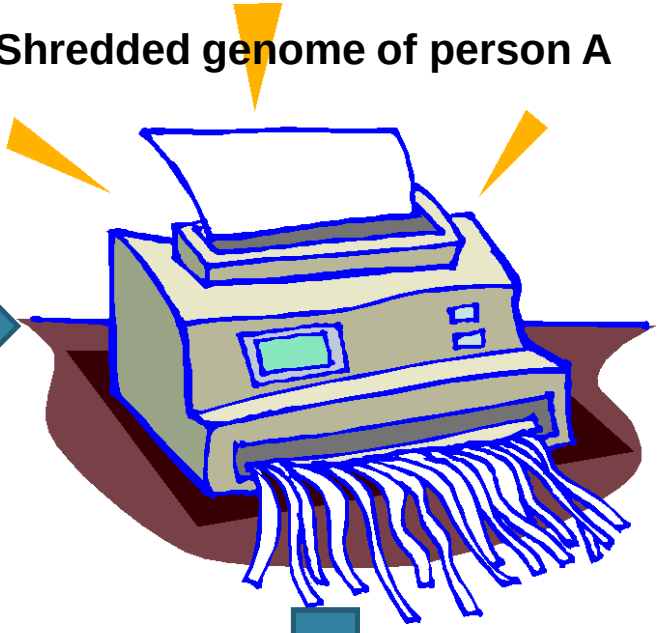
Genome of person A



3 copies x Genome of person A



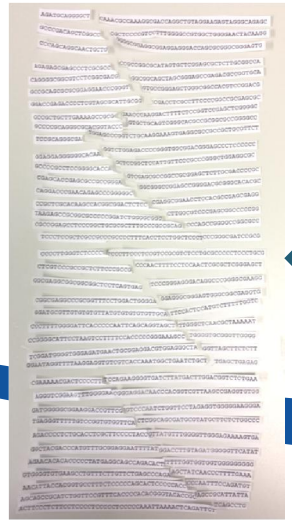
Shredded genome of person A



Next-Generation Sequencer

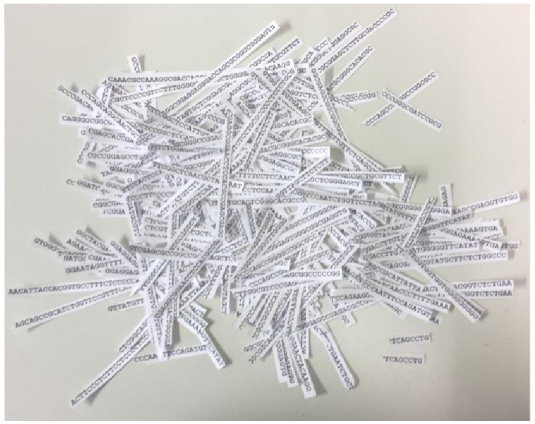
Output data
(short reads DNA sequences)

We want this!



Challenge: make sense of the data!

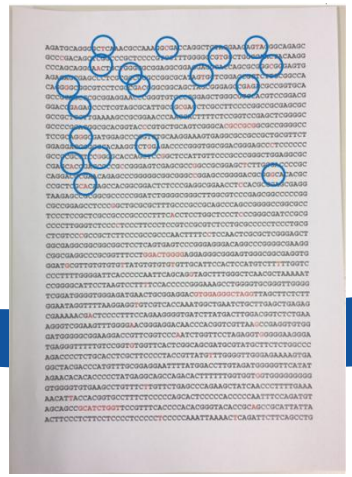
Shredded genome of person A



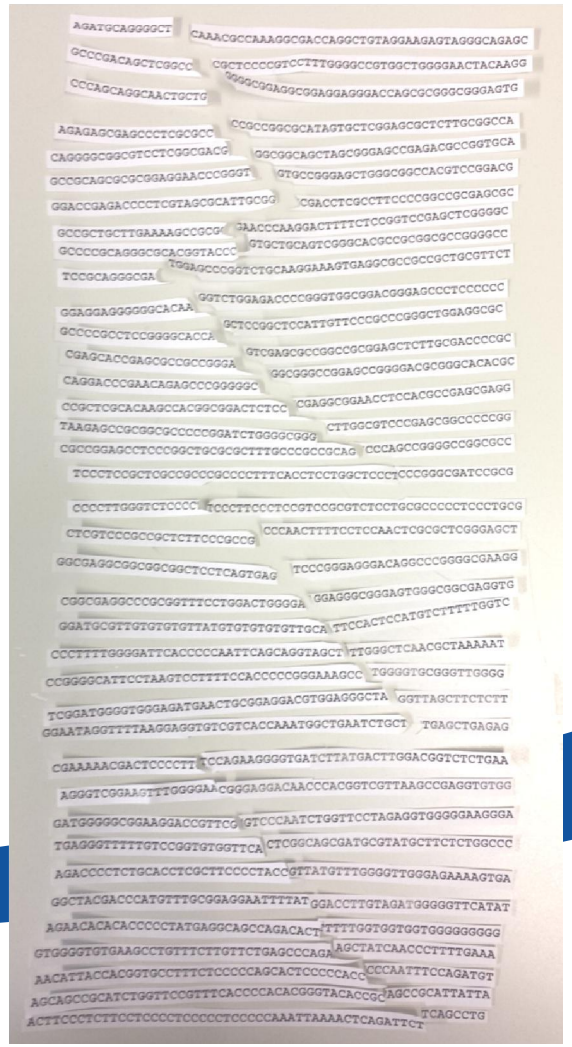
+

Reference genome

Genome of person A



<u>Reconstructed</u> genome of person A



Central Dogma of Biology

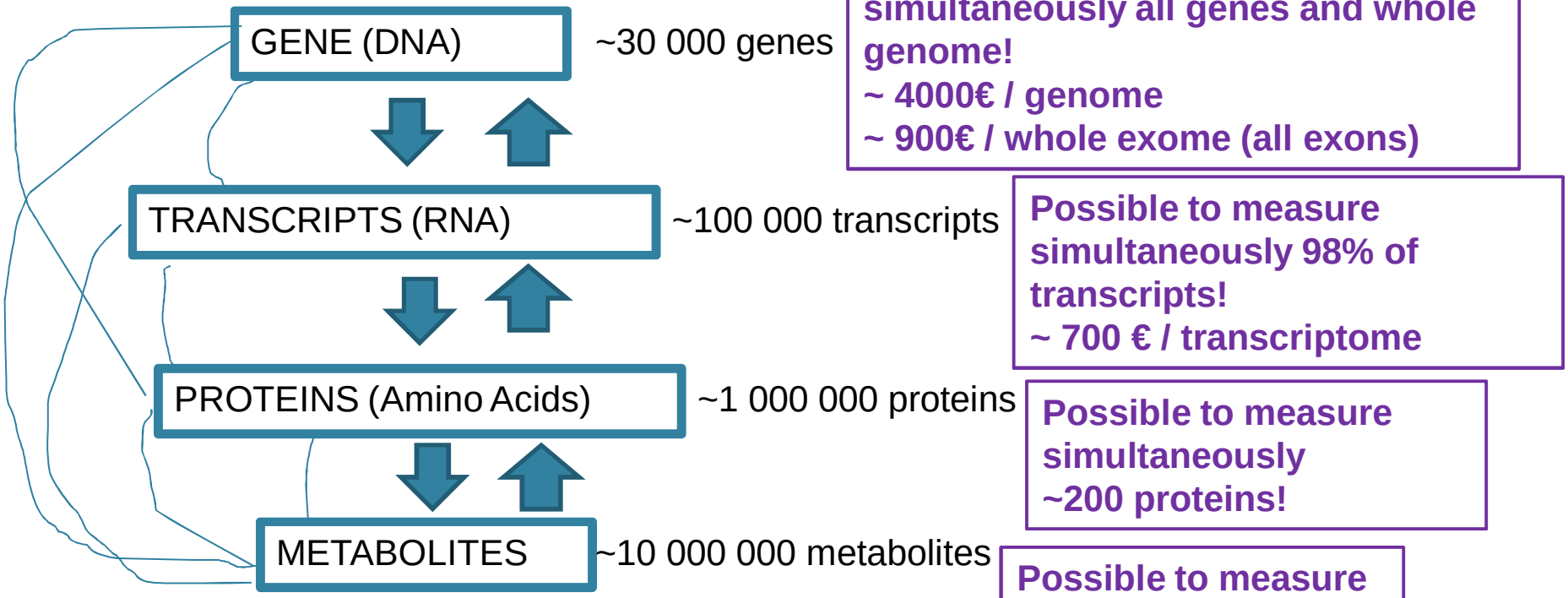


Fig 1 – Central Dogma of biology

ORION
Building well-being

Next Generation Sequencing

- DNA-seq - sequence the whole human genome
- Exome-seq - sequence only the parts of human genome that code for proteins
- RNA-seq - sequence the whole transcriptome
- Chip-seq - chromatin immunoprecipitation-sequencing

Statistical analysis of genomic data

- Data (i.e. hyperspace):
  - approx. 100 samples (two groups: placebo vs treated)
  - approx. 1 000 000 SNPs (one million categorical variables!!!)

Max. 10 possible values

Statistical analysis of genomic data – cont'd

~1 000 000 variables (SNPs)

Condition	Biological Sample	*SNP 1*	*SNP 2*	*SNP 3*	*…*	*SNP 1 000 000*
Healthy	Person 1	0	2	1	…	0
Healthy	Person 2	1	1	1	…	1
…	…	…	…	…	…	…
Disease	Person 50	2	1	2	…	2
Disease	Person 51	2	0	1	…	0
…	…	…	…	…	…	…

~ 50 samples

~ 50 samples

0 = AA; 1 = GA; 2 = GG

0 = CC; 1 = CG; 2 = GG

0 = CC; 1 = CT; 2 = TT

0 = AA; 1 = AT; 2 = TT

ORION Building well-being

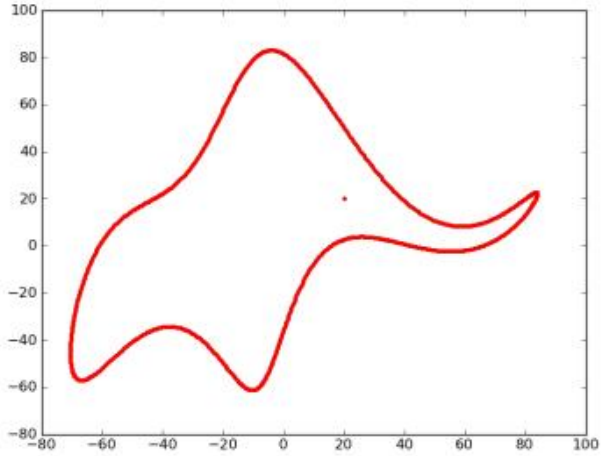
Statistical analysis of genomic data

- Challenge:
  - The SNPs (or features) found to be statistically significant in a dataset of 100 samples and 1 000 000 SNPs tend to not get validated when the number of samples is increased further to 1000 samples or more
- Data is not distributed uniformly in the hyperspace and it tends to concentrate in few parts of the hyperspace!
- *New statistical tools/methods are needed!*

Statistical analysis of genomic data – cont'd

- John von Neumann famously said:

With four parameters I can fit an elephant, and with five I can make him wiggle his trunk!



ORION
Building well-being

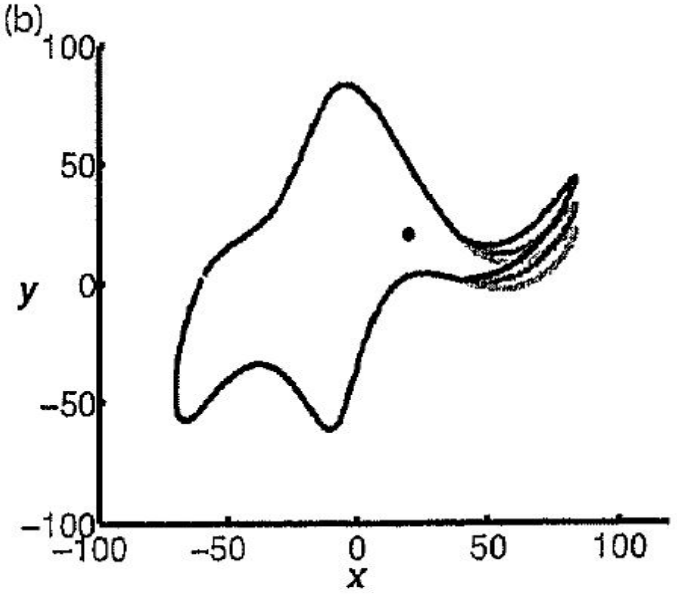
Statistical analysis of genomic data – cont'd

From article: "Drawing an elephant with four complex parameters" by Jurgen Mayer, Khaled Khairy, and Jonathon Howard, Am. J. Phys. 78, 648 (year 2010)

$$x(t) = \sum_{k=0}^{\infty} \left(A_k^x \cos(kt) + B_k^x \sin(kt)\right),$$

$$y(t) = \sum_{k=0}^{\infty} \left(A_k^y \cos(kt) + B_k^y \sin(kt)\right),$$

Parameter	Real part	Imaginary part
$p_1=50-30i$	$B_1^x=50$	$B_1^y=-30$
$p_2=18+8i$	$B_2^x=18$	$B_2^y=8$
$p_3=12-10i$	$A_3^x=12$	$B_3^y=-10$
$p_4=-14-60i$	$A_5^x=-14$	$A_1^y=-60$
$p_5=40+20i$	Wiggle coeff.=40	$x_{eye}=y_{eye}=20$



ORION
Building well-being

Personalized medicine

- Conceptually:
  - Read the DNA of a given patient
  - Identify the mutations in the DNA of the given patient
  - Compare the mutations against a catalogue of disease-causing mutations
  - Decide wheter a diagnosis can be confidently made
  - Compare the mutated genes against a catalogue of drugs
  - Decide wheter a treatment can be given

- For example, there are 2000 known mutations in ”cystic fibrosis transmembrane conductance regulator gene (CFTR)” but only few tens have been shown to cause cystic fibrosis.
- What data should be returned to patients? (gray area; guidelines for returning secondary/incidental findings)

Personalized medicine – cont'd

- Disease
  - Smart => most of the times is difficult to treat (several genes, several mutations)
  - Dumb => most of the times is easy to treat (one gene, one mutation)

- Medicine and biology's big problem => there is too much data to handle

Personalized medicine – examples

- Pharmacogenomics has indentified approx. 70 genes for which specific mutations cause humans to metabolize drugs ineffectively
- there are hundreds of actionable gene mutations – mutations that cause disease but whose consenquences can be avoided by medical strategies with knowledge of their presence
- In some cases, cancer-driving mutations in tumors, once identified, can be counteracted by treatments with currently available drugs

Next Generation Sequencing

- DNA-seq - sequence the whole human genome
- Exome-seq - sequence only the parts of human genome that code for proteins
- RNA-seq - sequence the whole transcriptome
- Chip-seq - chromatin immunoprecipitation-sequencing

Why RNA-seq?

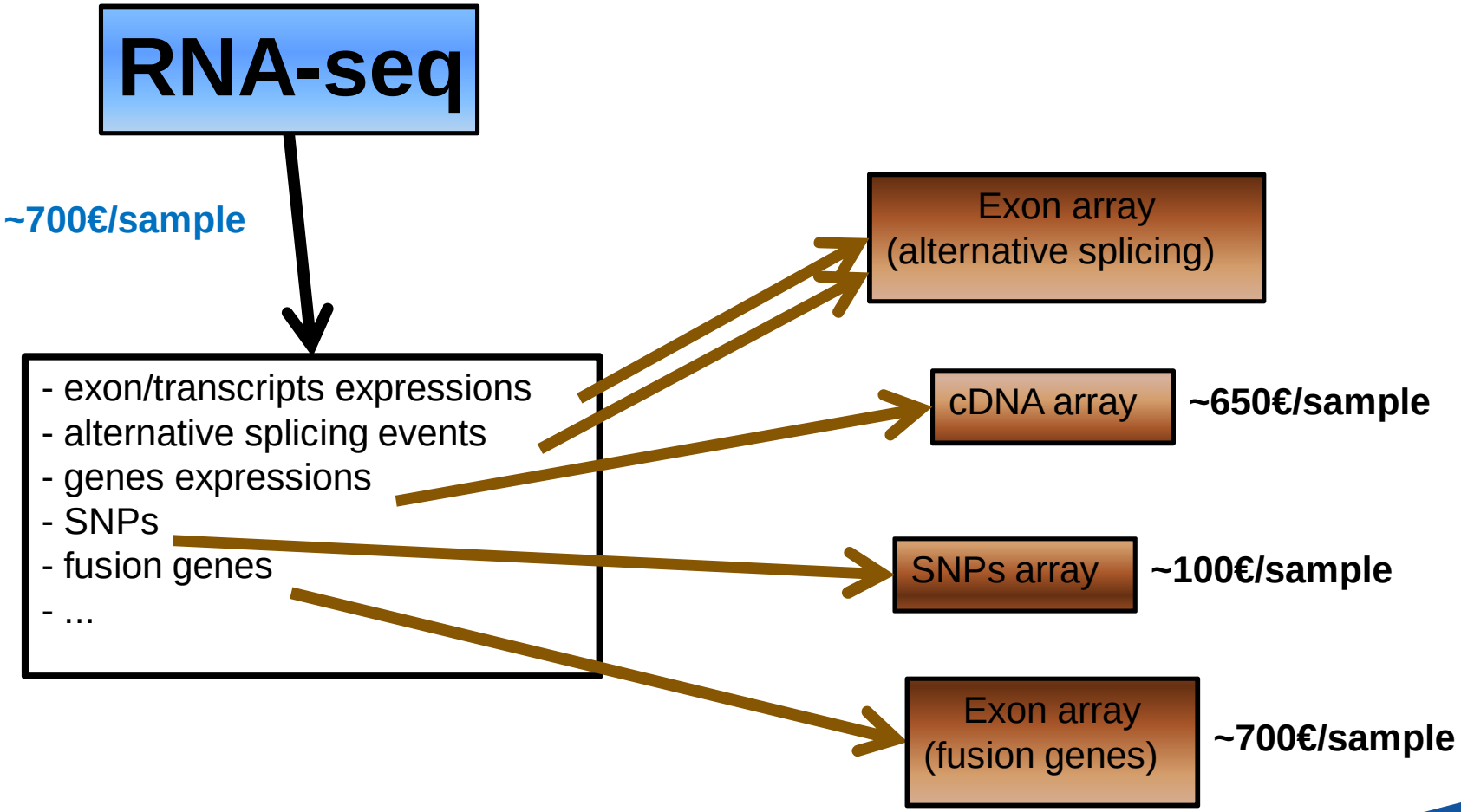


Fig. 2 – RNA-seq vs other technologies

ORION
Building well-being

RNA-seq

- RNA-seq dataset example:
  - 4 samples (control group with 2 samples; treated group has 2 samples) => compressed raw data ~29 GB (~7 DVDs)
  - Once sample compressed raw data => ~7,2 GB (~2 DVDs)
  - Once sample raw data => ~18,5 GB (~4 DVDs)

- TODO:
  - Find Differentially Expressed Genes (DEG) between the two groups (control VS treated)
  - Find fusion genes in all samples

Example input / raw NGS data (FASTQ file)

```
@F1000000000001/1
CCAAAGGTAACCNCCCAGGTGTTCAGANNNAGNNANATTTAAGNTGNNNN
+
ab`_`a`\aaa\B^aaa^QQ][aa_TXBBBVTBBZB[aBBBBBBBBBBB
@F1000000000005/1
GCCCGCCTTGGCCTCCCTAAGTGCTGGNNNCANNGNCGTGAGCNACNNNG
+
aaaa`aa_``\aa`aaaa`W`Zaaa]BBBBBBBBBBBBBBBBBBBBBB
@F1000000000009/1
GTGATATTTAATCAAGACATTAACATGNNNAGNNGNTTGTTGANTTNNNA
+
a^_^aXa`a[_``_``Xa]``V]`Ra[BBBUWBB[BBBBBBBBBBBBBB
@F1000000000013/1
TGACCTCCGGCCTAGCCATGTGATTTCNNNTCNNCNCCATAACNCTNNNC
+
abbbbabbaabaa`[a`Q_`^_Z_aa^BBB\^BB^B^a^`VO\B^[BBBB
……
```

Example input / raw NGS data (FASTQ file) – cont'd

Usually **tens/hundreds of millions of reads** are present in one FASTQ file

ORION
Building well-being

Align reads on transcriptome

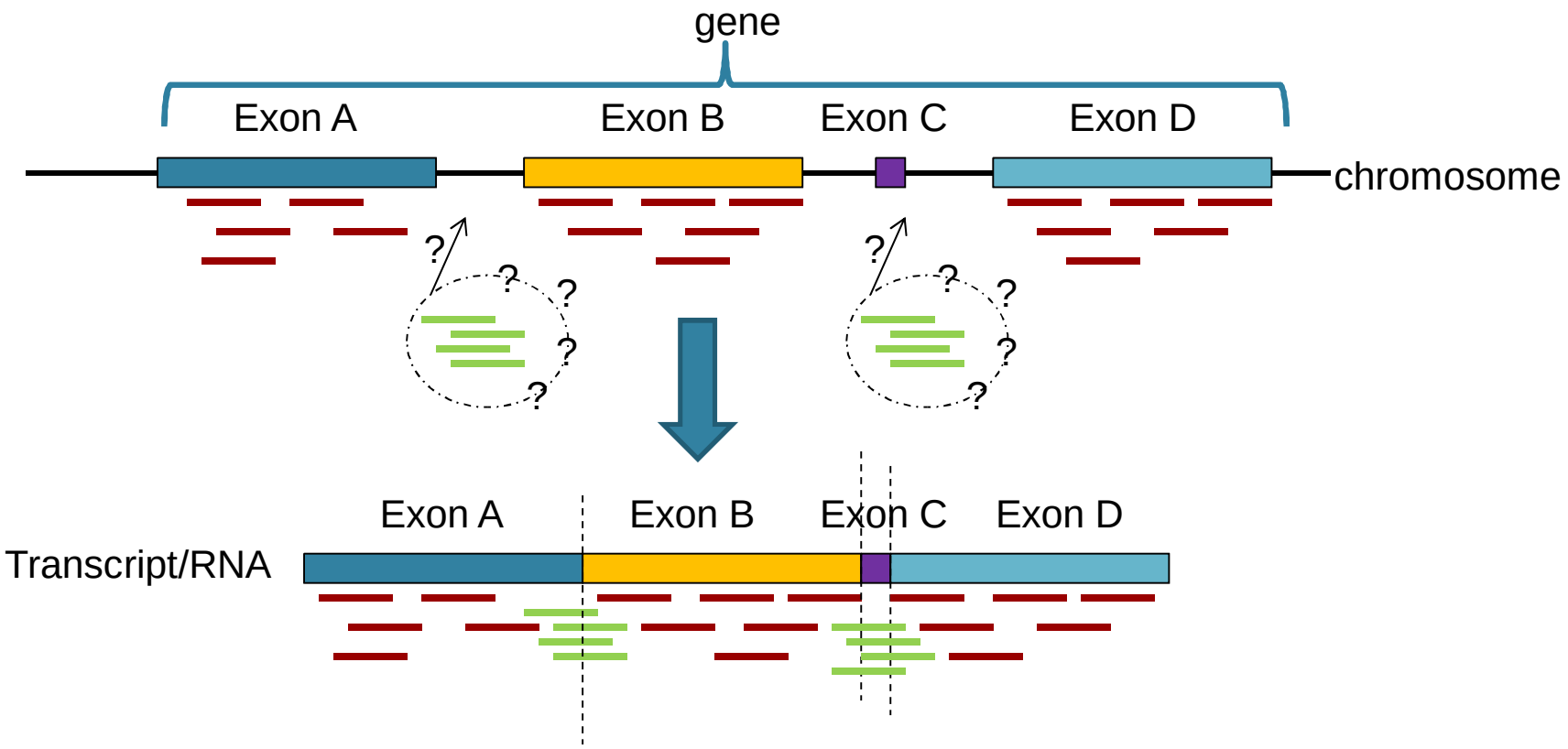


Fig. 3 – Alignment of reads on transcriptome

Align reads on genome

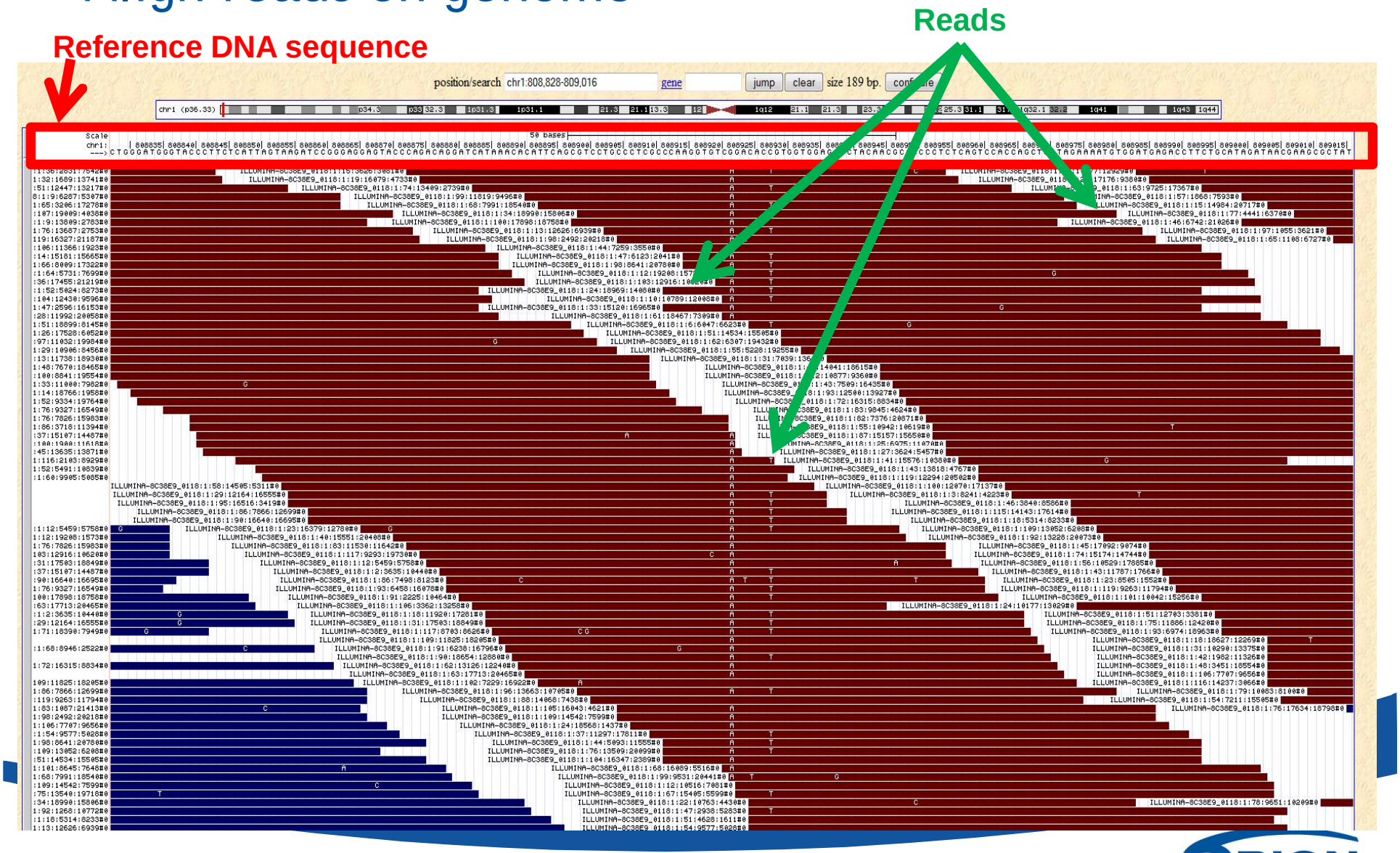


Fig. 4 – Reads aligned on genome

ORION
Building well-being

Reads coverage visualization

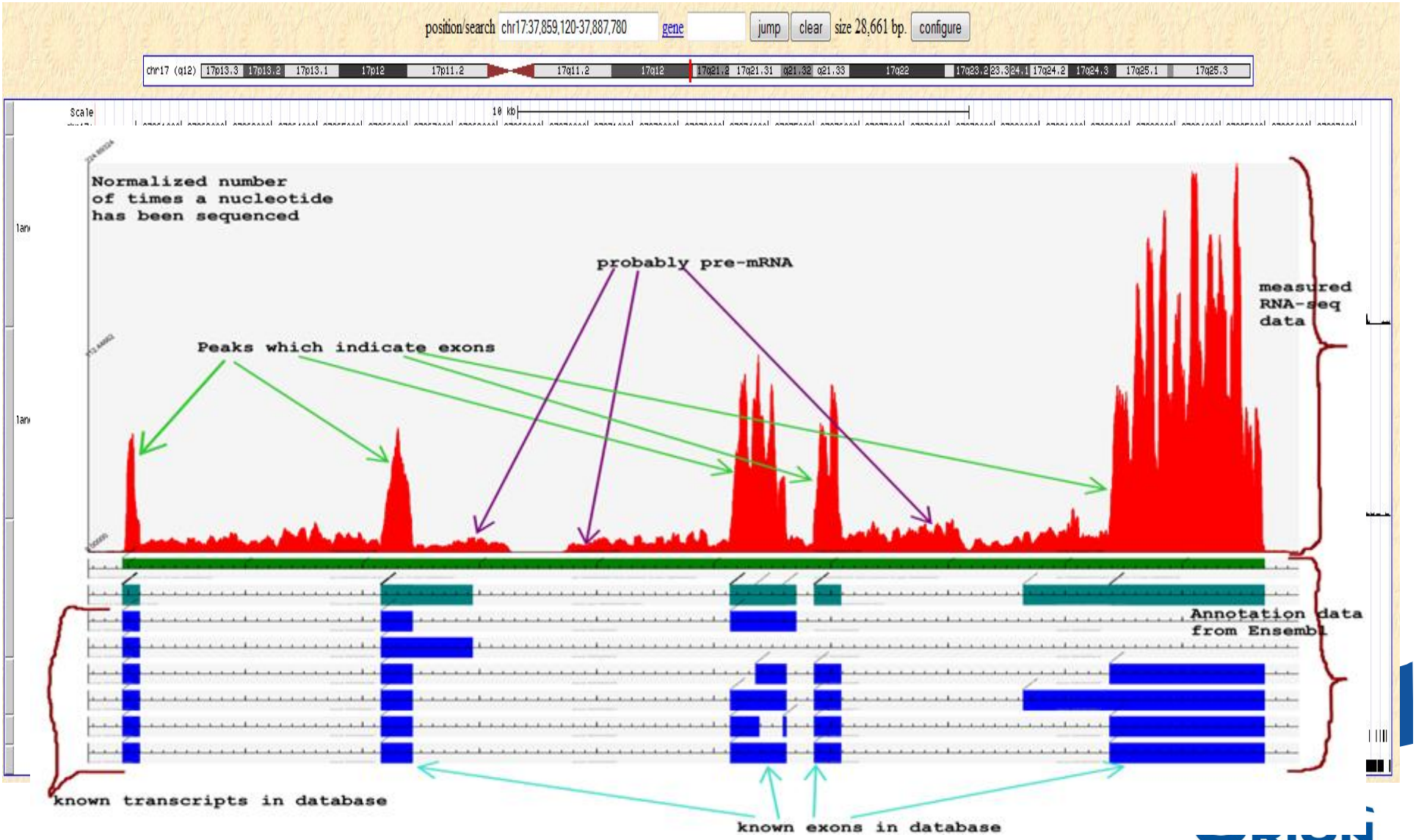


Fig. 5 – Genome coverage

Finding Differentially Expressed Genes

1. Quality filtering of reads (adaptor trimming, rRNA removal, etc.)
2. Aligning reads on genome (computational intensive)
3. Count number of reads per each known gene in genome
4. Normalize the counts
5. Find the Differentially Expressed Genes
6. Pathway enrichment using the found Differentially Expressed Genes

Counts data in RNA-seq

~30 000 variables

Condition	Biological Sample	*Gene 1*	*Gene 2*	*Gene 3*	*…*	*Gene 30 000*	Sum counts
Control	Tumor cells 1	50 000	10	8134	…	0	40 000 000
Control	Tumor cells 2	1 000	0	7533	…	1	70 000 000
Treated	Tumor cells 3	224	48	34	…	193	45 000 000
Treated	Tumor cells 4	0	3511	341	…	0	55 000 000

Count of reads mapping on each gene in each sample

Challenge: *Find the Differentially expressed genes between the Control and Treated!*

Challenges with count data from RNA-seq

- Discrete positive (skewed) data => no (log-) normal model
- Coverage (i.e. count all reads in a sample) between samples varies a lot => normalization between the samples should be used
- Very large dynamic range (0 -> 10 000 000) between genes => heteroskedasticity should be taken into account!
- Very small number of replicates => no rank based methods, no permutation methods

Variance depends on mean

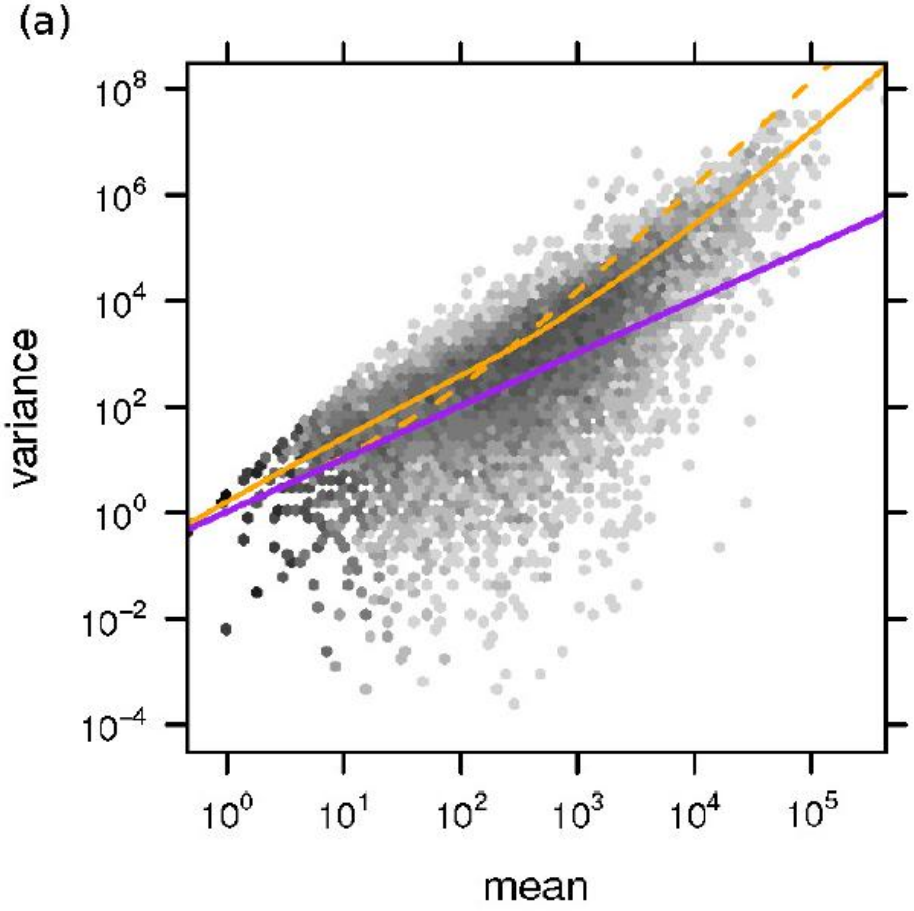


Fig. 6 - Variance and mean computed for two biological replicates using count data from RNA-seq

Anders et al, ”Differential expression analysis for sequence count data”, Genome Biology, 2010.

Poisson	$v = \mu$
Poisson + constant CV	$v = \mu + \alpha \mu^2$
Poisson + local regression	$v = \mu + f(\mu^2)$

Model fitting:

- estimate variance from replicates
- fit a line to obtain the dependence between variance and mean (local regression for gamma-family generalized linear model, more math needed due to handle the differences between sample coverages)

The negative-binomial distribution known to work the best!

ORION Building well-being

Magnitude of effect VS statistical significance

- Preferable to use *modified t-statistics* than the *ordinary t-statistics* when looking for DEGs, which means that instead of comparing means using

$$T_i = \frac{\bar{x_i} - \bar{y_i}}{s_i}$$

where s_i is the standard deviation of the replicates x_i (respectively y_i) of gene i in two different conditions it is better to use

$$T'_i = \frac{\bar{x_i} - \bar{y_i}}{s_i + s_0}$$

where s_0 minimizes the coefficient of variation of T'_i

Magnitude of effect VS statistical significance – cont'd

- In DEGs analysis is custom to select the DE based on
  - Fold-change (used because it makes the results more reproducible), and
  - P-values

Alternative splicing

- Is an important process which affecs the result of gene expression

Fig. 7 – A biological hypothesis

ORION
Building well-being

Alternative splicing – cont'd

- Is an important process which affecs the result of gene expression

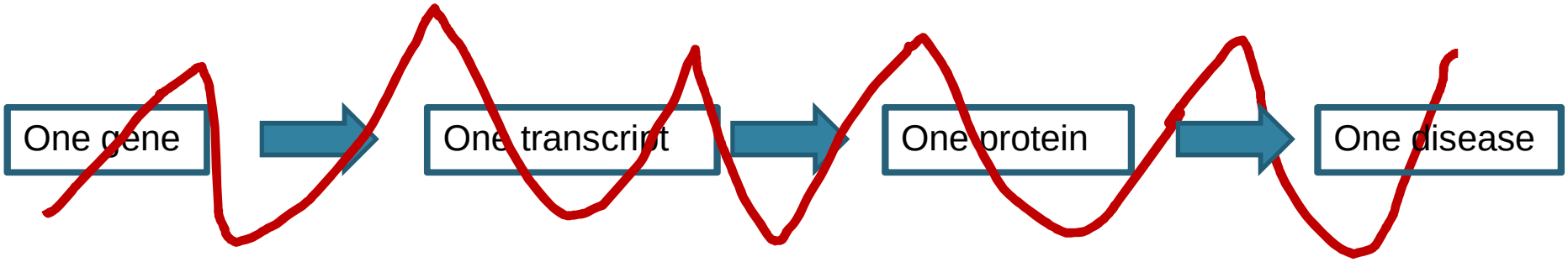


Fig. 7 – A biological hypothesis

ORION
Building well-being

Alternative splicing – cont'd

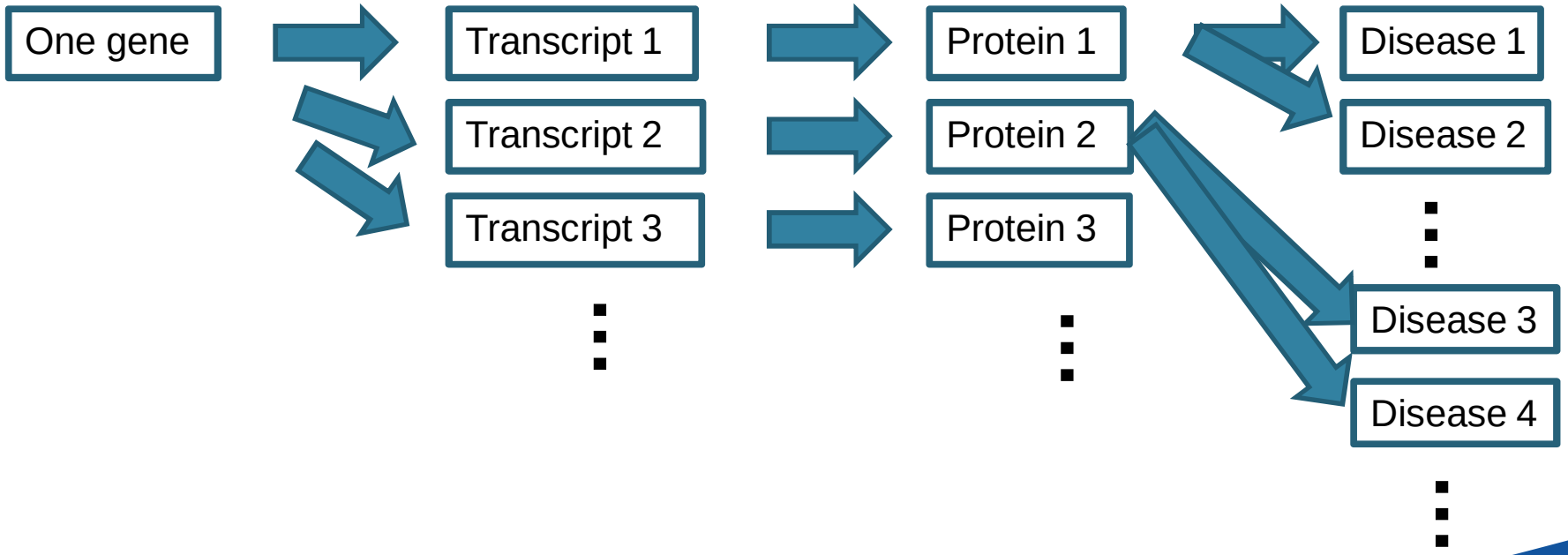


Fig. 8 – Another biological hypothesis

ORION
Building well-being

Finding Fusion Genes in RNA-seq

- Fusion genes:
  - results from chromosomal rearrangements and have been found in leukaemias, prostate cancer, lung cancer, brain cancer, etc.
  - Play important role in onset and development of cancer
  - until very recently was belived that they are very rare
  - ~60% of prostate cancer are characterized by presence of fusion genes (e.g. TMPRSS2-ERG) fusion

Finding Fusion Genes in RNA-seq – cont'd

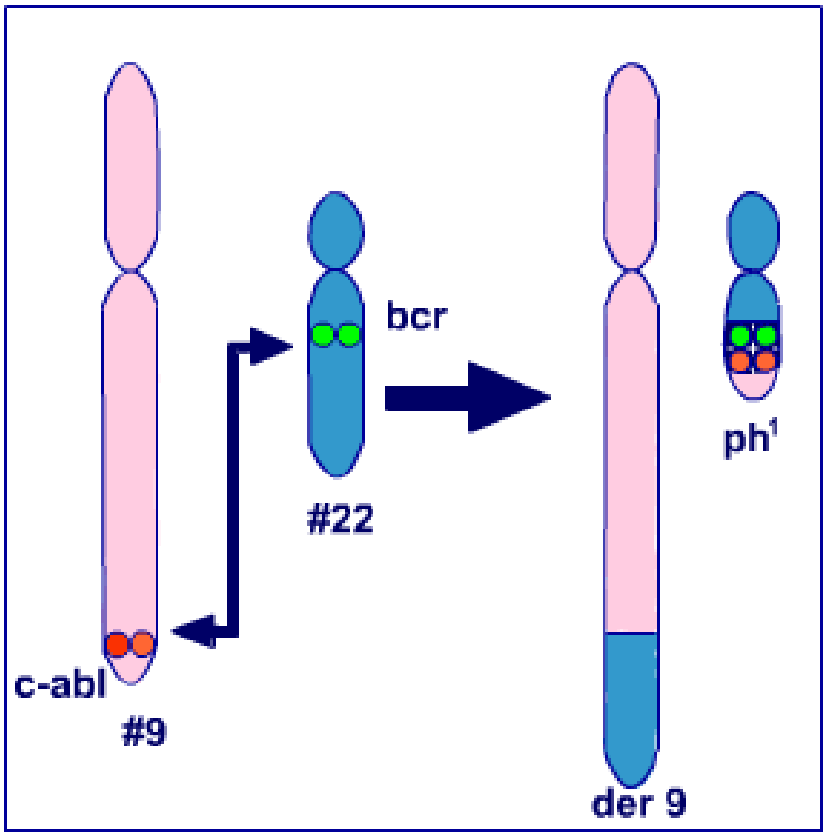


Fig. 9 – Example of BCR-ABL translocation/fusion-gene

Finding Fusion Genes in RNA-seq – cont'd

1. Reads filtering (quality filtering, adaptor removal, etc.)
2. Align all reads on genome
3. Aligning against the transcriptome all the reads which
   - map uniquely on genome, or
   - do not map on genome
4. Find candidates fusion-genes by:
   1. looking for pair-end reads which map simultaneusly on two different transcripts of two different genes
   2. using biological knowledge (e.g. type of RNA, paralog genes, pseudogenes, database of known fusions and conjoined genes, etc.)
5. Using the preliminary list of candidate fusion genes, find the fusion junction:
   1. generating exon-exon combinations, and/or
   2. using local alignments (e.g. BLAT)

Finding Fusion Genes in RNA-seq – cont'd

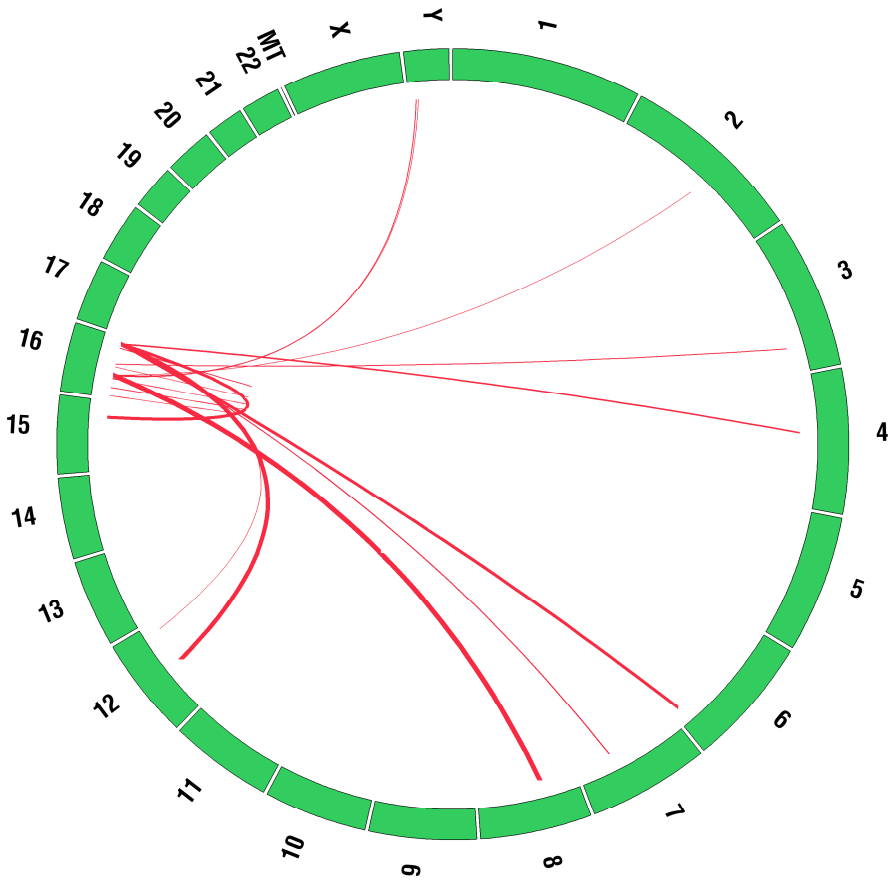


Fig. 10 – Visualization of candidate fusion genes

ORION
Building well-being

Personalized medicine

99 USD per sample => 1 million SNPs !!!

ORION
Building well-being

Personalized medicine

99 USD per sample => 1 million SNPs !!!

$99

ORION
Building well-being

Personalized medicine

99 USD per sample => 1 million SNPs !!!

ORION
Building well-being

Personalized medicine

99 USD per sample => 1 million SNPs !!!

n^{th} kit

$80

ORION
Building well-being

Personalized medicine – cont’d

+ raw data

ORION
Building well-being

NGS Challenges

- Price of storing/transferring the data can be expensive
- Price of analyzing the data can be expensive
- Storing the data, moving the data, analyzing it
- Expertise of engineers, computer scientists, statisticians, biologists, medical doctors is need!
- NGS data is not organized in any coherent way which allows one to compute across it
- Need for more computing power, need to move more efficient the data (hard drive sent via postal mail, store the biological sample than the sequence data
- the cost of computing is threatening to become the limiting factor in genomics (computing costs 10 times more than the research)

NGS challenges

- A gene regulatory network, which attempts to map how different genes control the expression of other genes, is smaller than a social network, with thousands rather than millions of nodes but the data is harder to define => when we look to biological/genomic data, we many times do not know exactly what we are looking at yet!

Conclusions

- NGS is here today!
- NGS is used today used in Finnish hospitals (HY, Tampere, Oulu, etc.)
- NGS is computationally intensive
- New statistical methods are needed for analyzing NGS data