

Next Generation
Sequencing:
Adjusting to Big
Data

Daniel Nicorici, Dr. Tech.
Statistikot Suomen Lääketeollisuudessa
29.10.2013

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



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!!!**

Next Generation Sequencing – cont'd

Reference genome

```

AGATGCAGGGGCGCAAACGCCAAAGGAGACACAGGCTGTAGGAAGAGAAGGGCGAGAGC
GCCGGACAGCTCGCCCGCTCCCGCTCCTTTGGGGCCCGGGCTGGGGAACTACAAGG
CCCAGCAGGCAGCTGCAGGGGGCGGAGGCGGAGGAGGGACACGCGGGTGGGAGTG
AGAGAGCGAGCCCTCGCCCGCCCGGGCCATAGCGCTCGGAGCGCTCTTGGCCCA
CAGGCGCGGCTCCTCGGCGCGGGCGGACGCTAGCGGAGCCGGGACGCGGTTGCA
GCCGACGCGCGGAGGAACCCGGGTGTCCGGGAGCTGGCGGCCACGTCGGGACG
GGACCGAGACCCCTCGTAGCGCATTGCGGCGACCTCGCCTTCCCGGCCCGAGGCGC
GCCGCTGCTTAAAAAGCCGCGGAACCCAGGACTTTTCCCGTCCGAGCTCGGGGC
GCCCCGCAAGGGCGCACGGTACCCGTGCTGCAGTCGGGCACGCCGCGGCCCGGGGCC
TCCCGAGGGCGATGGAGCCCGTCTGCAAGGAAAGTGAGGCGCCCGCTCGCTTCT
GGAGGAGGGGGGCAAGGCTGGAGACCCCGGGTGGCGGACGGGAGCCCTCCCGCC
GCCCGCTCCGGGGCACCGTCCCGCTCCATTGTTCCCGCCGGGCTGGAGGCGC
CGAGCACCGAGCGCCCGGGAGTCGAGCGCCGGCCGGGAGCTCTTGGCACCCCGC
CAGGACCCGAAACAGAGCCCGGGGGCGGCGGGCGGAGCCGGGACCGCGGACACCC
CCGCTCGCACAAAGCCACGGCGGACTCTCCCGAGCGGAACTCCACGCGAGCGAGG
TAAGAGCCGCGGCCCCGGATCTGGGGCGGGCTTGGCGTCCCGAGCGGCCCCGG
CGCCGAGCCTCCGGGCTCGCGCTTTGCCCGCCGAGCCAGCCGGGGCCGGCGCC
TCCCTCCGCTCGCCCGCCGCCCTTCACTCCTGGCTCCCTCCCGGGCGATCCGCG
CCCTTGGGCTCCTCCCTCCCTCCCTCCGTCGCGTCTCCTGCGCCCTCCCTGCG
CTCGTCCCGCGCTCTTCCCGCCGCCAACTTTTCTCCAACTCGCGCTCGGGAGCT
GGCGAGGGCGCGGCTCCTCAGTGAATCCCGGAGGACAGGCCCCGGCGAAGG
CGCGAGGCCCGCGGTTTCTGACTGGGGAGGAGGGCGGAGTGGGCGCGAGGTG
GGATGCGTTGTGTGTATGTGTGTGTGTGATCCACTCCATGTCTTTTGGTC
CCCTTTGGGGATTACCCCAATTCAGCAGGTAGCTTTGGGCTCAACGCTAAAAAT
CCGGGCATTCCTAAGTCTTTTCCACCCCGGAAAGCCTGGGGTCCGGGTGGGG
TCGGATGGGTGGAGATGAATGCGGAGGACGTGGAGGGCTAGGTTAGCTTCTTT
GGAATAGGTTTTAAGGAGGTGCTGACCAATGGCTGAATCTGCTTAGAGTGAAG
CGAAAAACGACTCCCTTTCCAGAAGGGGTGATCTTATGACTTGGACGCTCTGAA
AGGTCGGAAGTTTGGGAAACGGGAGGACRACCCACGTCGTTAAGCCGAGGTGTGG
GATGGGGCGGAAAGACCGTTCGGTCCCAATCTGGTTCCTAGAGGTGGGGAAAGGA
TGAGGGTTTTTGTCCGCTGTGGTTCACTCGGACGATGCGTATGCTTCTCTGGCCC
AGACCCCTGACACCTCGCTTCCCTACCGTTATGTTTGGGGTGGGAGAAAAGTGA
GGCTACGACCCATGTTTGGGAGGAATTTATGGACCTTGTAGATGGGGTTTATAT
AGAACACACACCCCTATGAGGACGCCAGACACTTTTTTGGTGGTGGGGGGGGGG
GTGGGTGTGAAGCCTGTTTCTTGTCTGAGCCAGAAGCTATCAACCCCTTTGAAA
AACATTACCACGGTGCCTTTCTCCCGAGCACTCCCCACCCCAATTTCCAGATGT
AGCAGCCGATCTGGTTCCGTTTCAACCCACAGGGTACCCGAGCCGATATTA
ACTTCCCTTCTCCCTCCCTCCCTCCCAAAATAAACTCAGATTCTTCAGCCTG
    
```

Genome of person A *differences versus the reference genome*

```

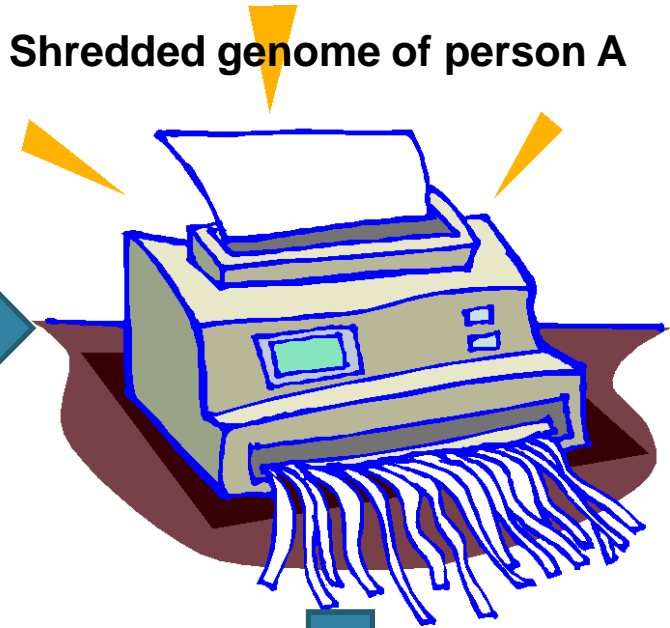
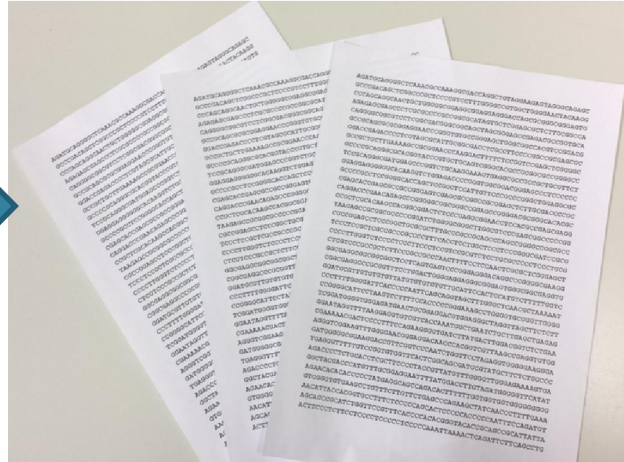
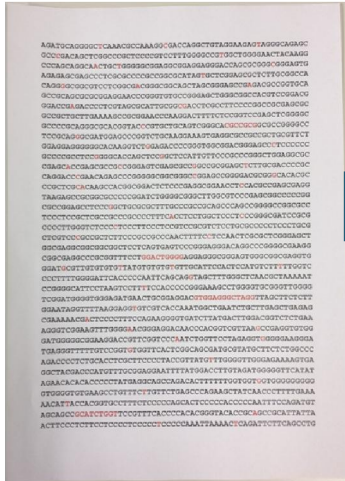
AGATGCAGGGGCTCAACCCCAAAGGGCTCAGGCTGTGGAAGAGTAAGGCAGAGC
GCCCGACAGCTCGCCCGCTCCCGCTCCTTTGGGGCCCGGGCTGGGGAACTACAAGG
CCCAGCAGGCAGCTGCAGGGGGCGGAGGCGGAGGAGGGACACGCGGGTGGGAGTG
AGAGAGCGAGCCCTCGCCCGCCCGGGCCATAGCGCTCGGAGCGCTCTTGGCCCA
CAGGCGCGGCTCCTCGGCGCGGGCGGACGCTAGCGGAGCCGGGACGCGGTTGCA
GCCGACGCGCGGAGGAACCCGGGTGTCCGGGAGCTGGCGGCCACGTCGGGACG
GGACCGAGACCCCTCGTAGCGCATTGCGGCGACCTCGCCTTCCCGGCCCGAGGCGC
GCCGCTGCTTAAAAAGCCGCGGAACCCAGGACTTTTCCCGTCCGAGCTCGGGGC
GCCCGCTCCGGGGCACCGTCCCGCTCCATTGTTCCCGCCGGGCTGGAGGCGC
TCCCGAGGGCGATGGAGCCCGTCTGCAAGGAAAGTGAGGCGCCCGCTCGCTTCT
GGAGGAGGGGGGCAAGGCTGGAGACCCCGGGTGGCGGACGGGAGCCCTCCCGCC
GCCCGCTCCGGGGCACCGTCCCGCTCCATTGTTCCCGCCGGGCTGGAGGCGC
CGAGCACCGAGCCCGCGGGAGTCGAGCGCCGGCCGGGAGCTCTTGGCACCCCGC
CAGGACCCGAAACAGAGCCCGGGGGCGGCGGGCGGAGCCGGGACCGCGGACACCC
CCGCTCGCACAAAGCCACGGCGGACTCTCCCGAGCGGAACTCCACGCGAGCGAGG
TAAGAGCCCGCGGCCCCGGATCTGGGGCGGGCTTGGCGTCCCGAGCGGCCCCGG
CGCCGAGCCTCCGGGCTCGCGCTTTGCCCGCCGAGCCAGCCGGGGCCGGCGCC
TCCCTCCGCTCGCCCGCCGCCCTTCACTCCTGGCTCCCTCCCGGGCGATCCGCG
CCCTTGGGCTCCTCCCTCCCTCCCTCCGTCGCGTCTCCTGCGCCCTCCCTGCG
CTCGTCCCGCGCTCTTCCCGCCGCCAACTTTTCTCCAACTCGCGCTCGGGAGCT
GGCGAGGGCGCGGCTCCTCAGTGAATCCCGGAGGACAGGCCCCGGCGAAGG
CGCGAGGCCCGCGGTTTCTGACTGGGGAGGAGGGCGGAGTGGGCGCGAGGTG
GGATGCGTTGTGTGTATGTGTGTGTGTGATCCACTCCATGTCTTTTGGTC
CCCTTTGGGGATTACCCCAATTCAGCAGGTAGCTTTGGGCTCAACGCTAAAAAT
CCGGGCATTCCTAAGTCTTTTCCACCCCGGAAAGCCTGGGGTCCGGGTGGGG
TCGGATGGGTGGAGATGAATGCGGAGGACGTGGAGGGCTAGGTTAGCTTCTTT
GGAATAGGTTTTAAGGAGGTGCTGCTCACCAATGGCTGAATCTGCTTAGAGTGAAG
CGAAAAACGACTCCCTTTCCAGAAGGGGTGATCTTATGACTTGGACGCTCTGAA
AGGTCGGAAGTTTGGGAAACGGGAGGACRACCCACGTCGTTAAGCCGAGGTGTGG
GATGGGGCGGAAAGACCGTTCGGTCCCAATCTGGTTCCTAGAGGTGGGGAAAGGA
TGAGGGTTTTTGTCCGCTGTGGTTCACTCGGACGATGCGTATGCTTCTCTGGCCC
AGACCCCTGACACCTCGCTTCCCTACCGTTATGTTTGGGGTGGGAGAAAAGTGA
GGCTACGACCCATGTTTGGGAGGAATTTATGGACCTTGTAGATGGGGTTTATAT
AGAACACACACCCCTATGAGGACGCCAGACACTTTTTTGGTGGTGGGGGGGGGG
GTGGGTGTGAAGCCTGTTTCTTGTCTGAGCCAGAAGCTATCAACCCCTTTGAAA
AACATTACCACGGTGCCTTTCTCCCGAGCACTCCCCACCCCAATTTCCAGATGT
AGCAGCCGATCTGGTTCCGTTTCAACCCACAGGGTACCCGAGCCGATATTA
ACTTCCCTTCTCCCTCCCTCCCTCCCAAAATAAACTCAGATTCTTCAGCCTG
    
```

Next Generation Sequencing - cont'd

Genome of person A

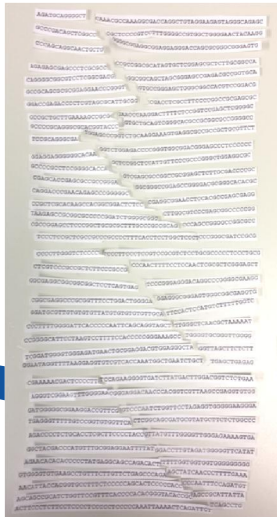
3 copies x Genome of person A

Shredded genome of person A



We want this!

Output data
(short reads DNA sequences)



Next-Generation Sequencer

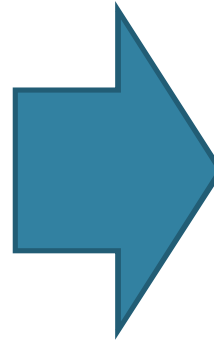
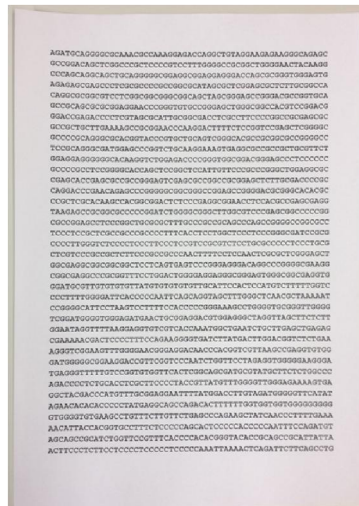


Challenge: make sense of the data!

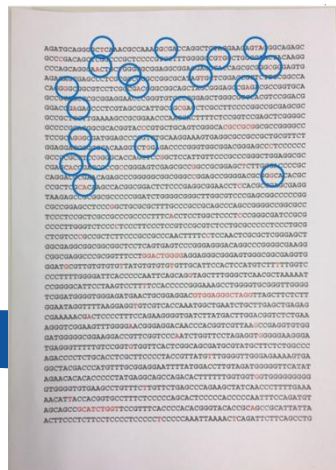
Shredded genome of person A



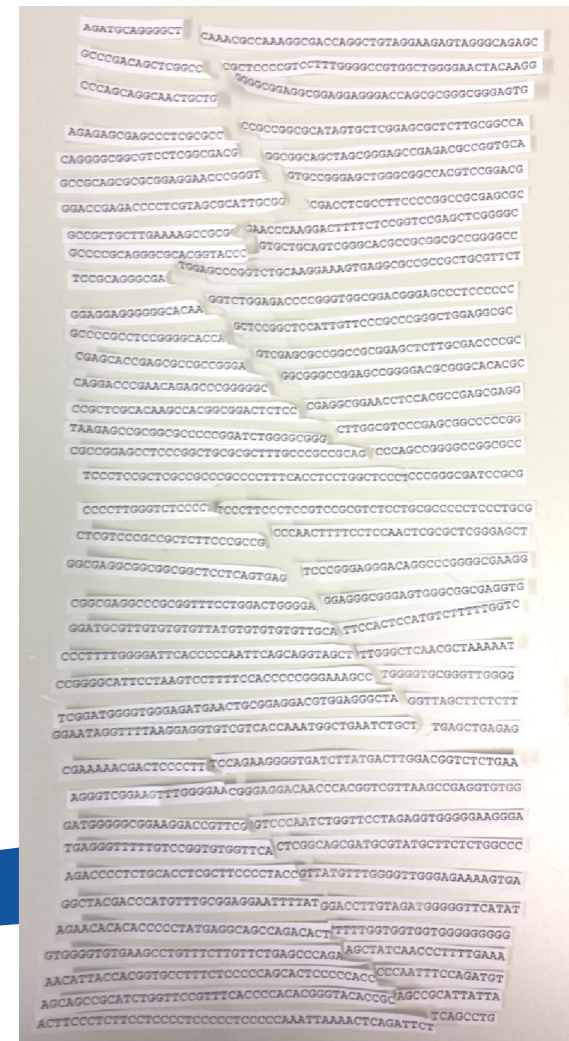
Reference genome



Genome of person A



Reconstructed genome of person A



Central Dogma of Biology

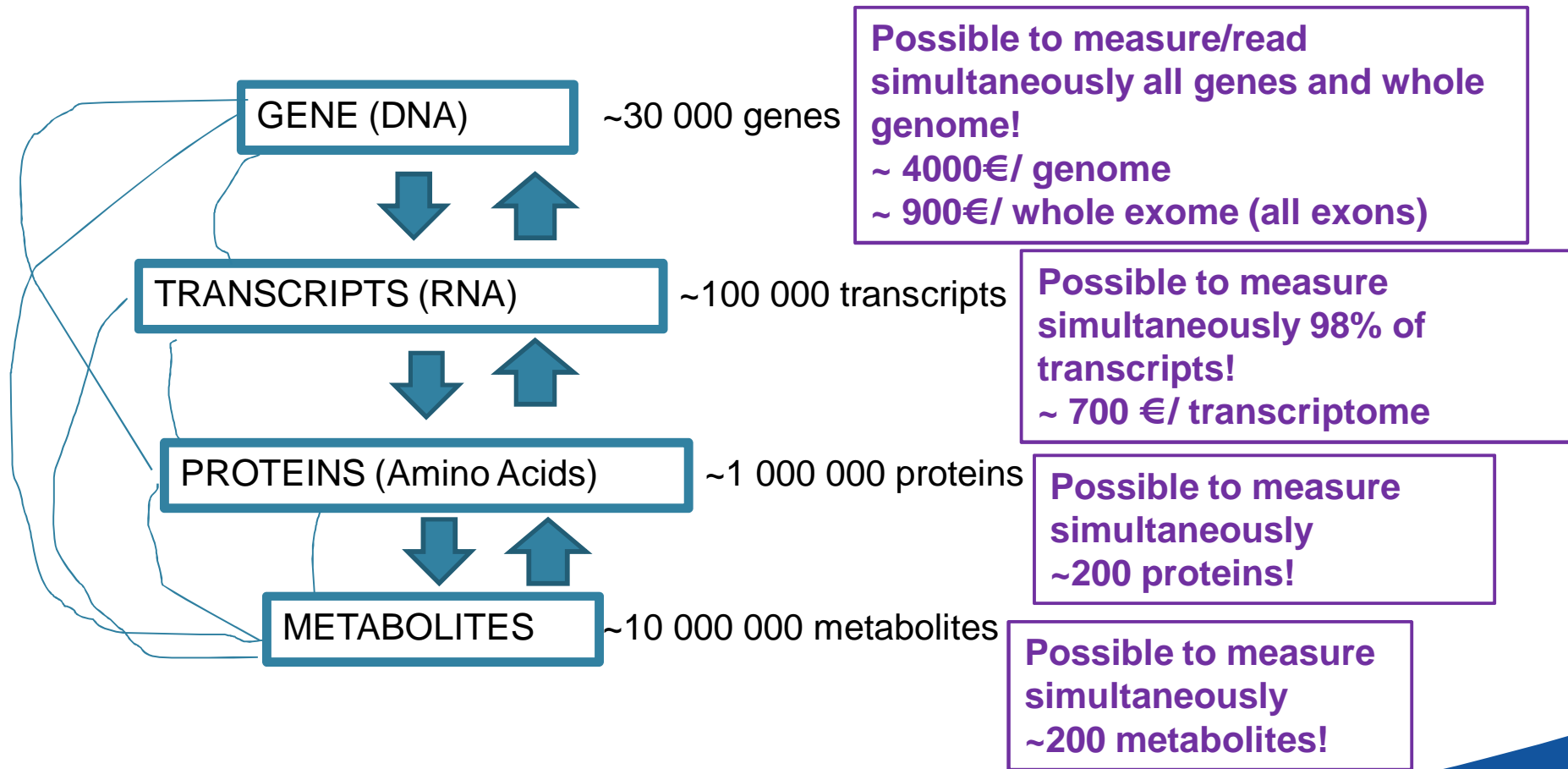


Fig 1 – Central Dogma of biology

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

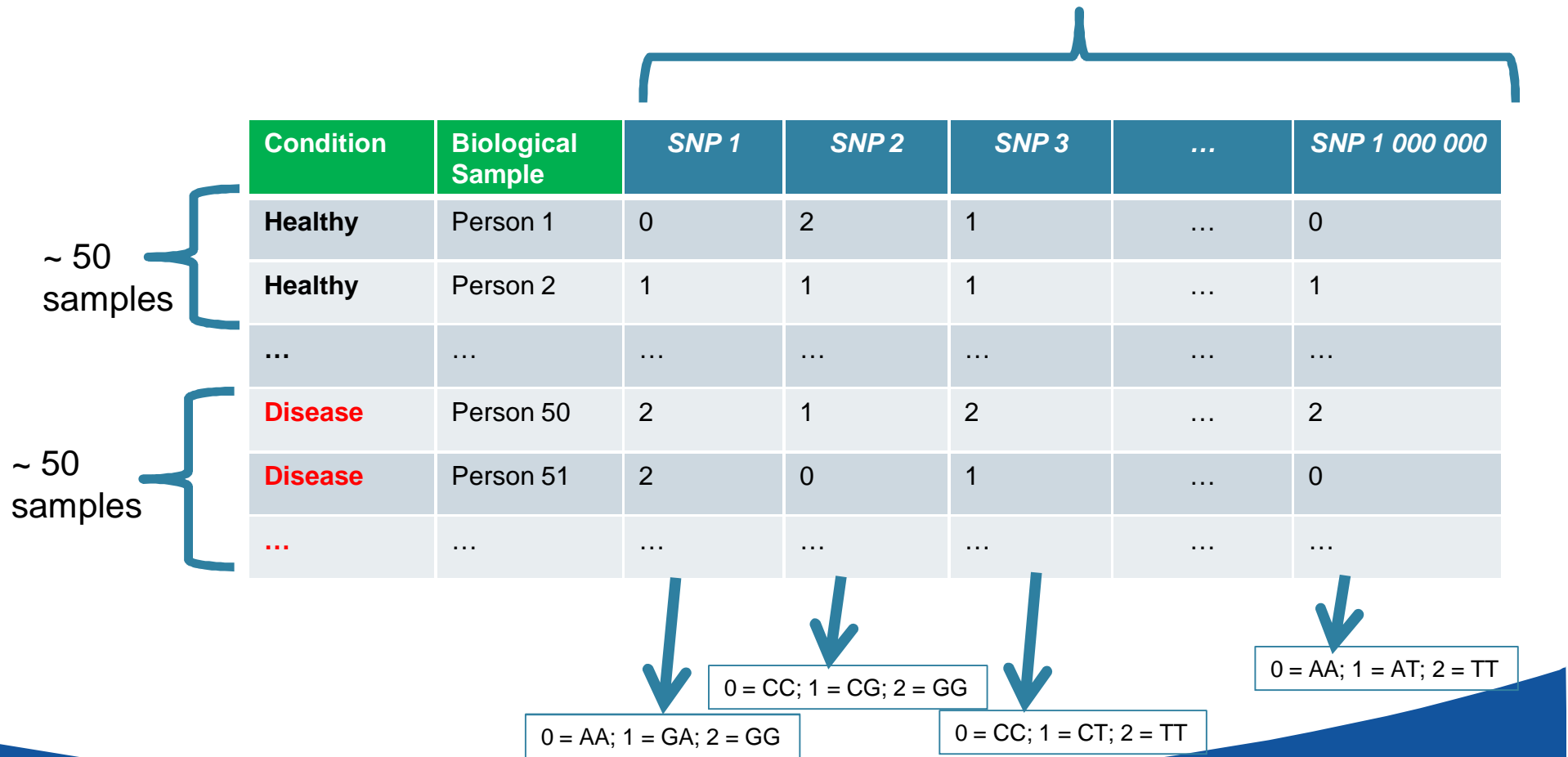
Statistical analysis of genomic data

- 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)



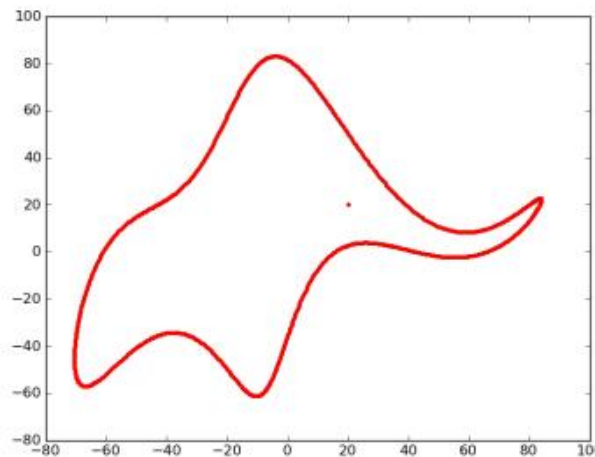
Statistical analysis of genomic data

- 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!



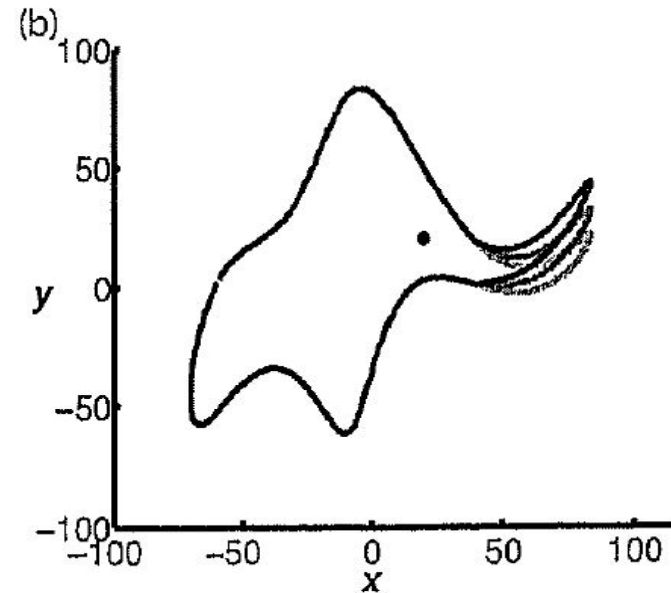
Statistical analysis of genomic data - cont'd

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

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

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

Parameter	Real part	Imaginary part
$p_1 = 50 - 30i$	$B_1^x = 50$	$B_1^y = -30$
$p_2 = 18 + 8i$	$B_2^x = 18$	$B_2^y = 8$
$p_3 = 12 - 10i$	$A_3^x = 12$	$B_3^y = -10$
$p_4 = -14 - 60i$	$A_4^x = -14$	$A_4^y = -60$
$p_5 = 40 + 20i$	Wiggle coeff. = 40	$x_{eye} = y_{eye} = 20$



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 whether a diagnosis can be confidently made
 - Compare the mutated genes against a catalogue of drugs
 - Decide whether 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?

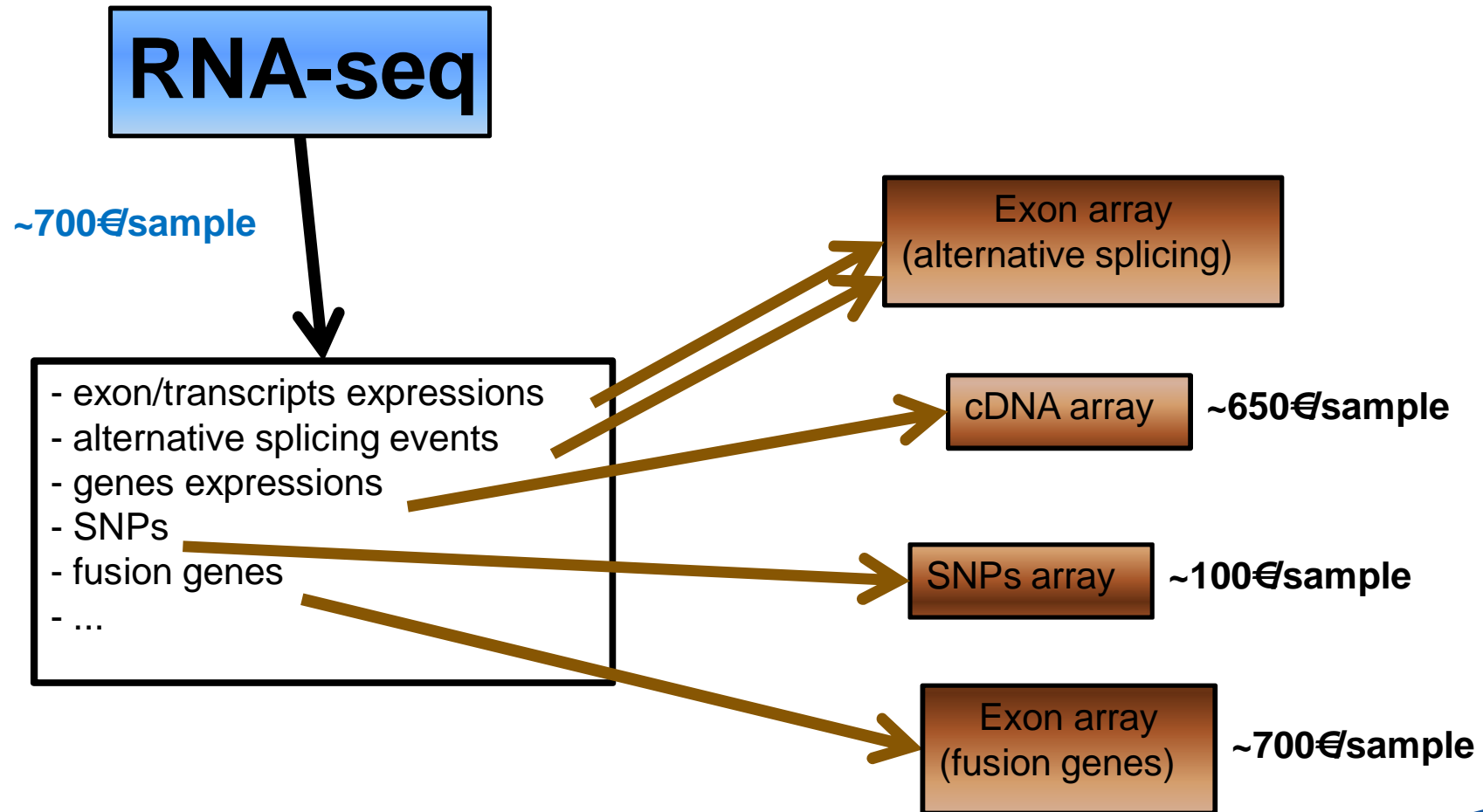


Fig. 2 – RNA-seq vs other technologies

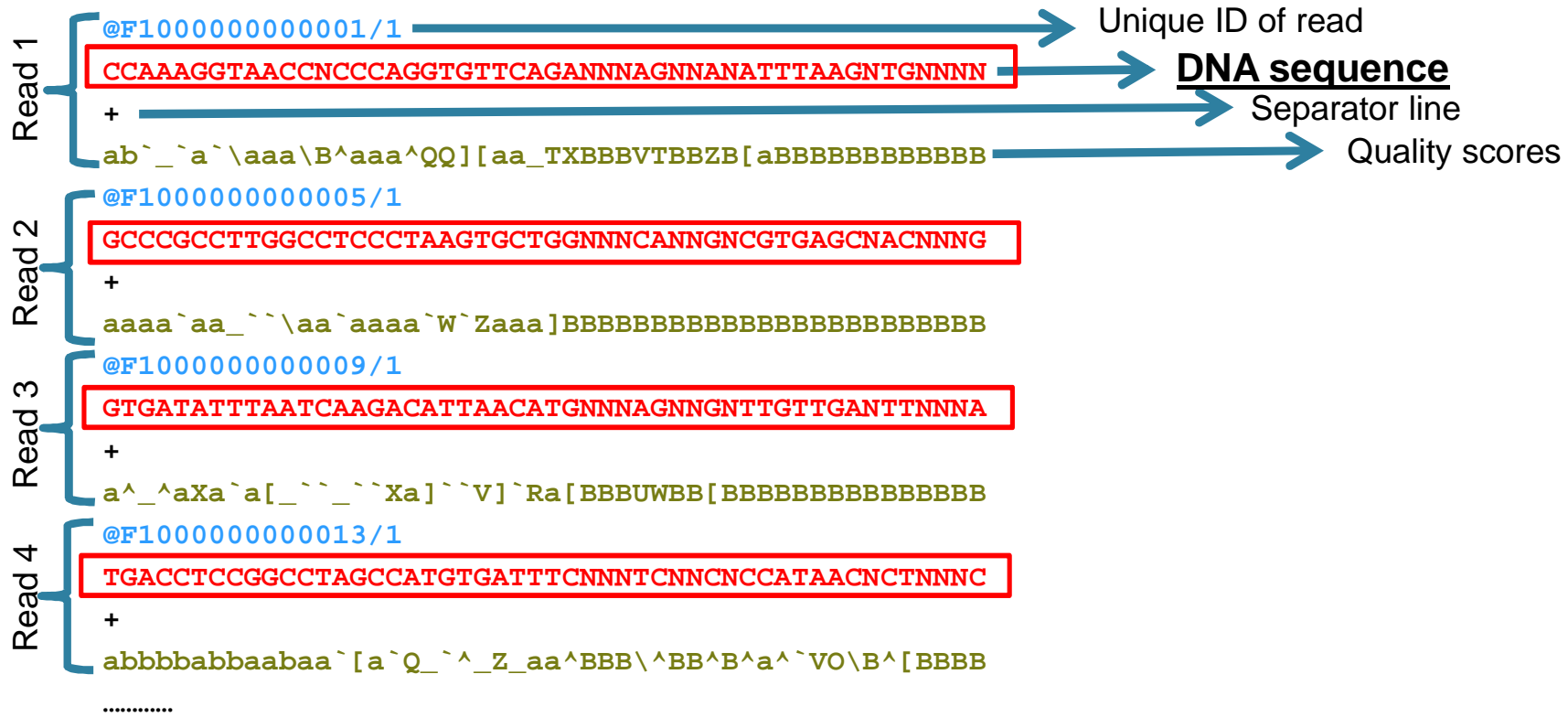
RNA-seq

- 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
CCAAAGGTAACCNCCCAGGTGTTTCAGANNAGNNANATTTAAGNTGNNNN
+
ab`_`a`aaa`B^aaa^QQ][aa_TXBBBVTBBZB[aBBBBBBBBBBBB
@F1000000000005/1
GCCCCCCTTGGCCTCCCTAAGTGCTGGNNNCANNGNCGTGAGCNACNNNG
+
aaaa`aa_``\aa`aaaa`W`Zaaa]BBBBBBBBBBBBBBBBBBBBBB
@F1000000000009/1
GTGATATTTAATCAAGACATTAACATGNNNAGNNGNTTGTTGANTTNNA
+
a^_`aXa`a[_``_``Xa]``V]`Ra[BBBUWBB[BBBBBBBBBBBBBB
@F1000000000013/1
TGACCTCCGGCCTAGCCATGTGATTCNNNTCNNCNCATAACNCTNNNC
+
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

Align reads on transcriptome

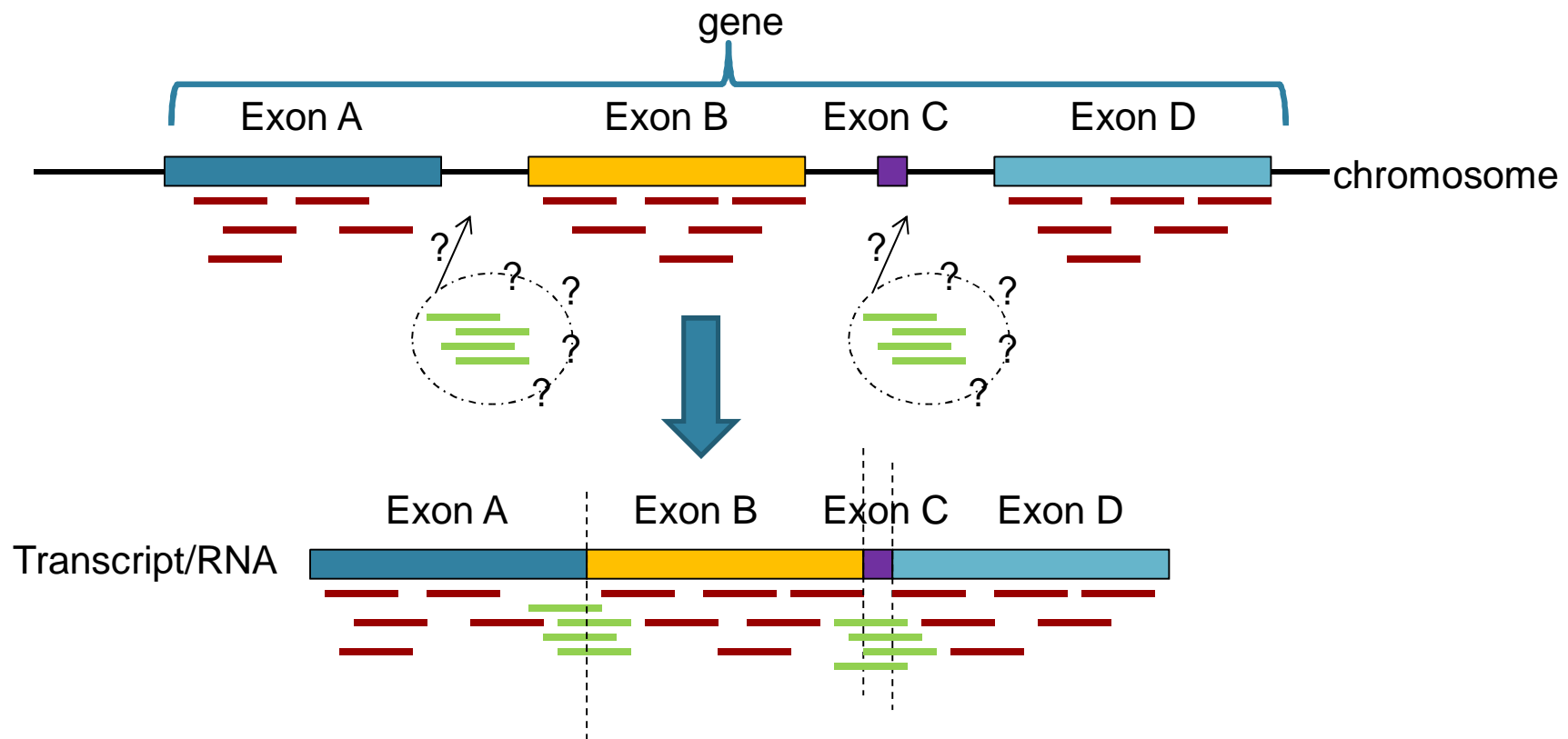


Fig. 3 – Alignment of reads on transcriptome

Align reads on genome

Reference DNA sequence

Reads

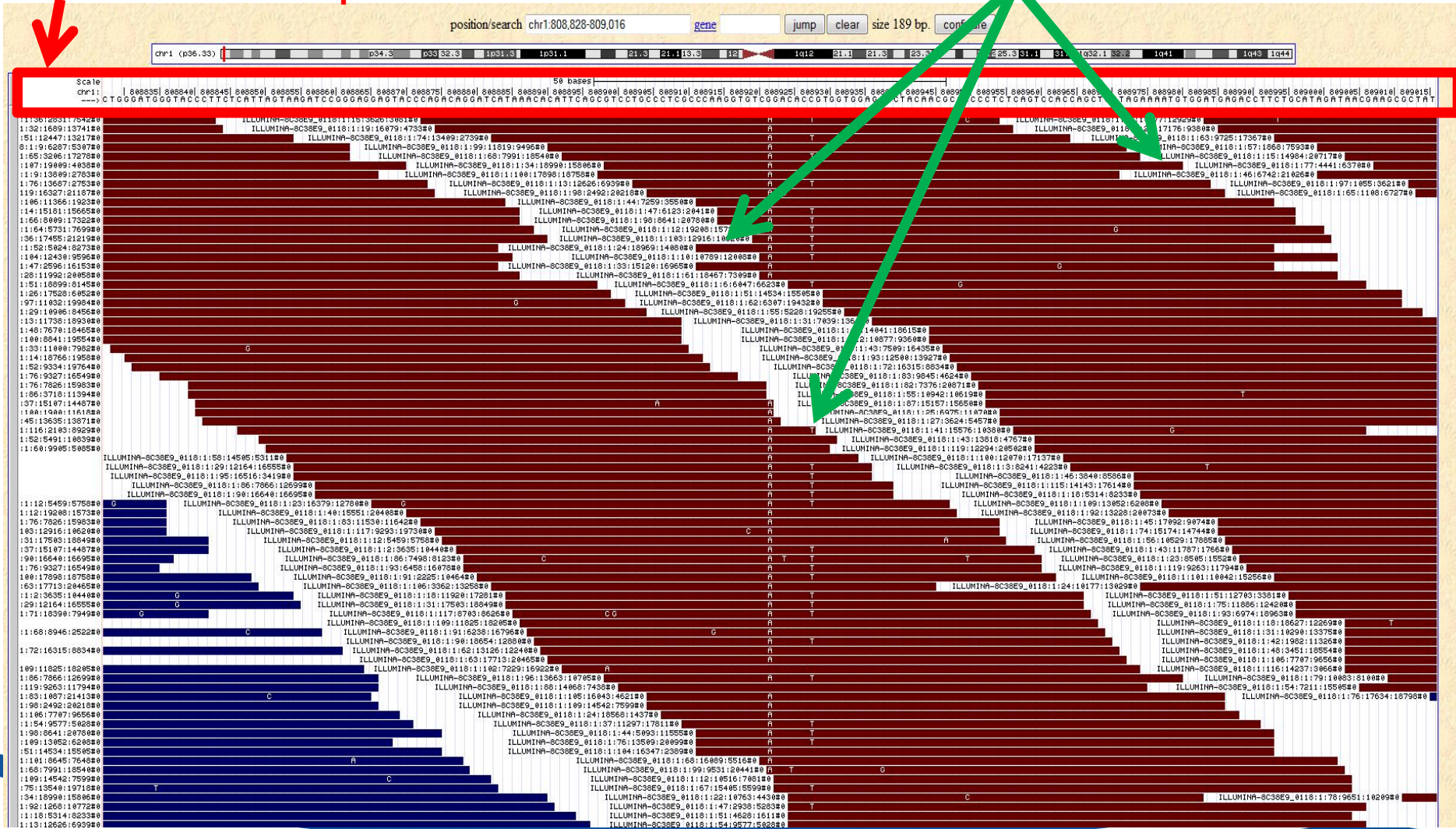


Fig. 4 – Reads aligned on genome

Reads coverage visualization

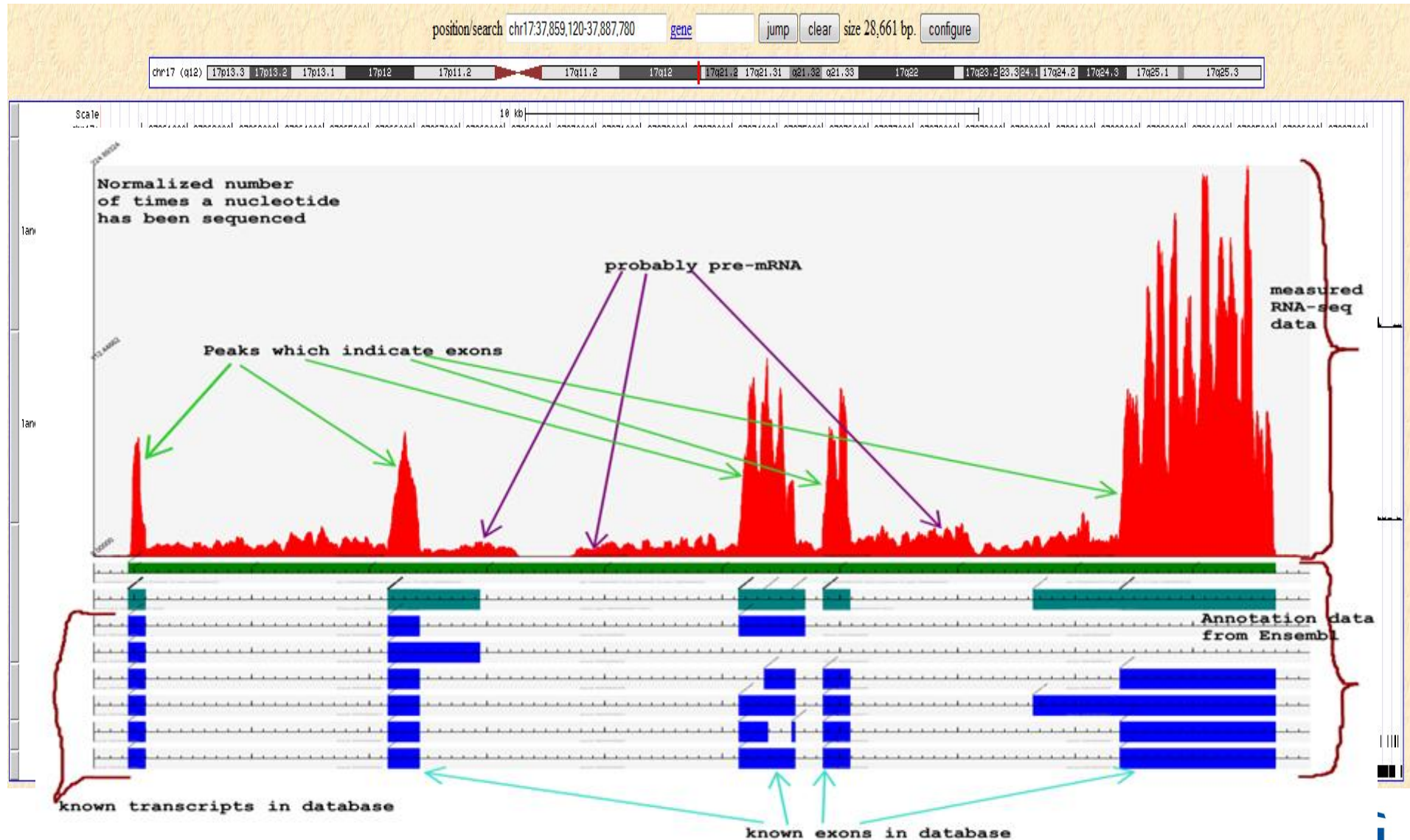


Fig. 5 – Genome coverage

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

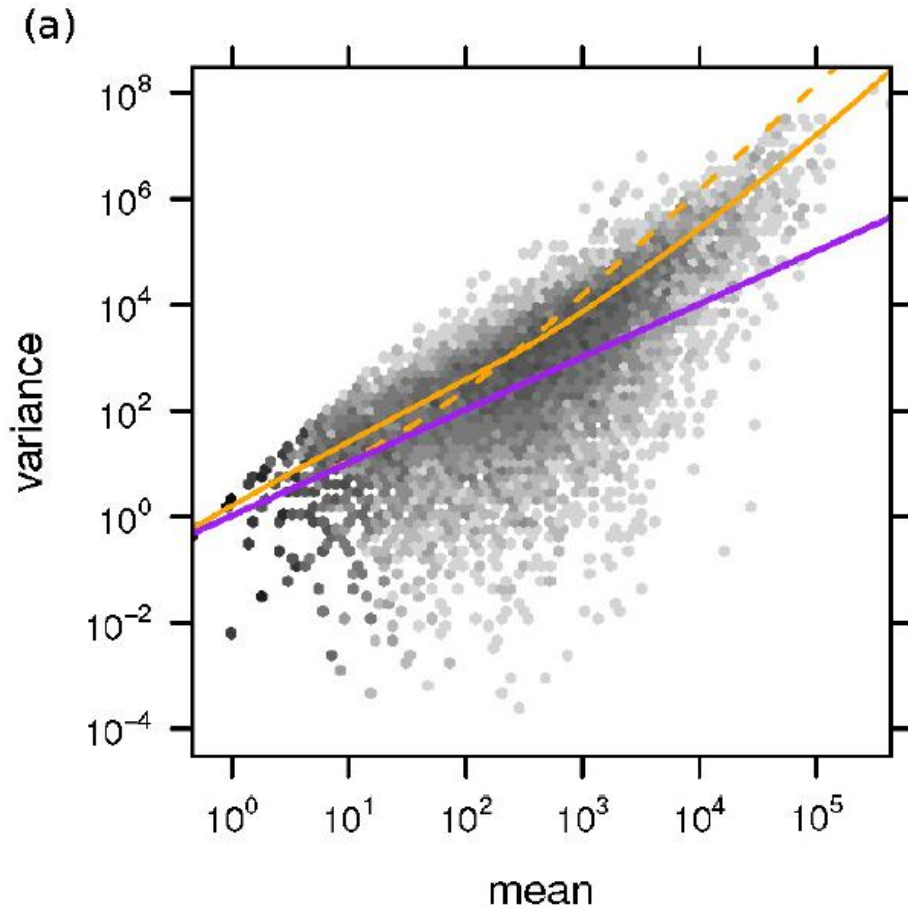
Count of reads mapping on each gene in each sample

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



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!

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

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 affects the result of gene expression



Fig. 7 – A biological hypothesis

Alternative splicing - cont'd

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

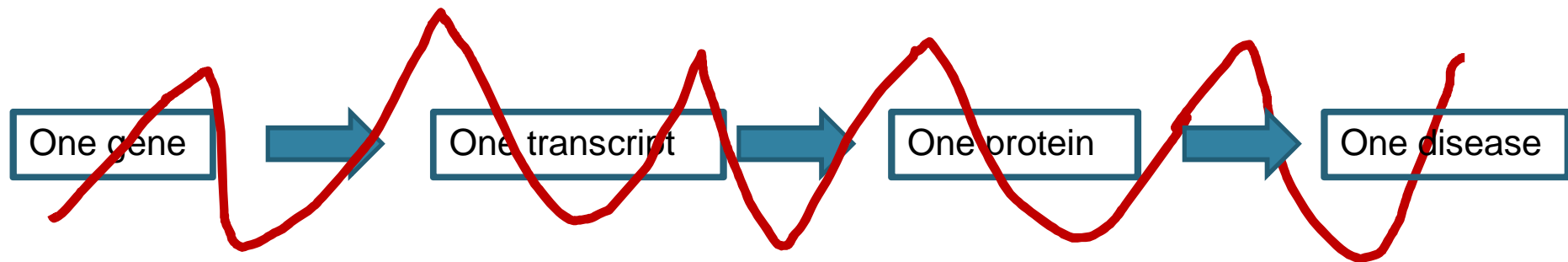


Fig. 7 – A biological hypothesis

Alternative splicing - cont'd

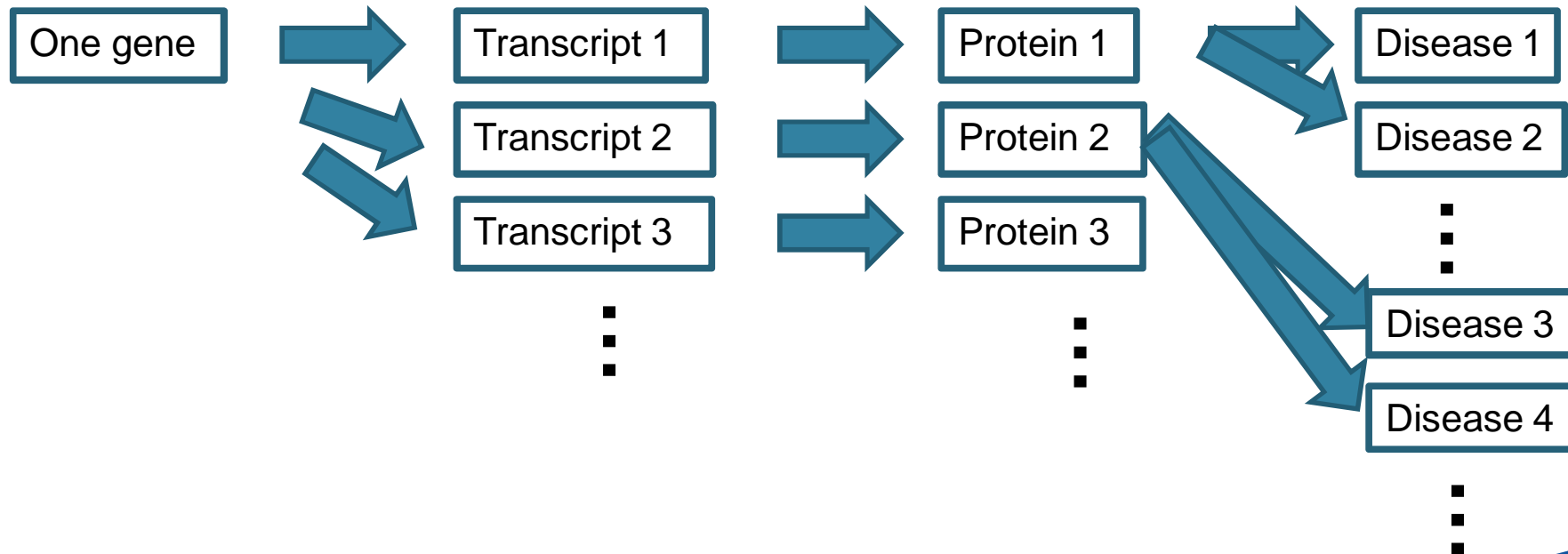


Fig. 8 – Another biological hypothesis

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 believed 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

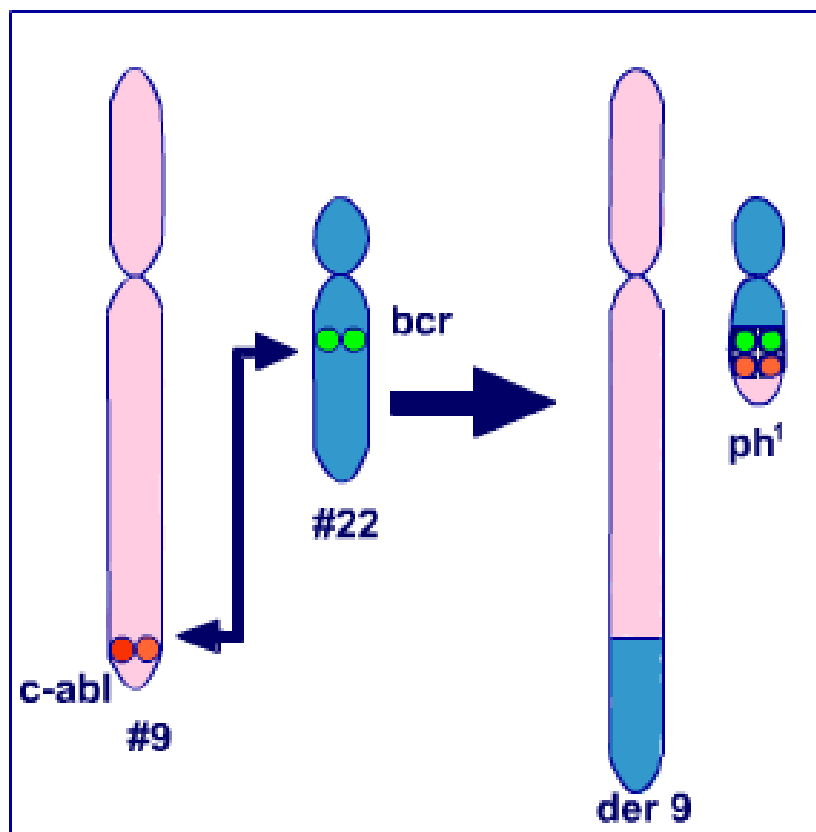


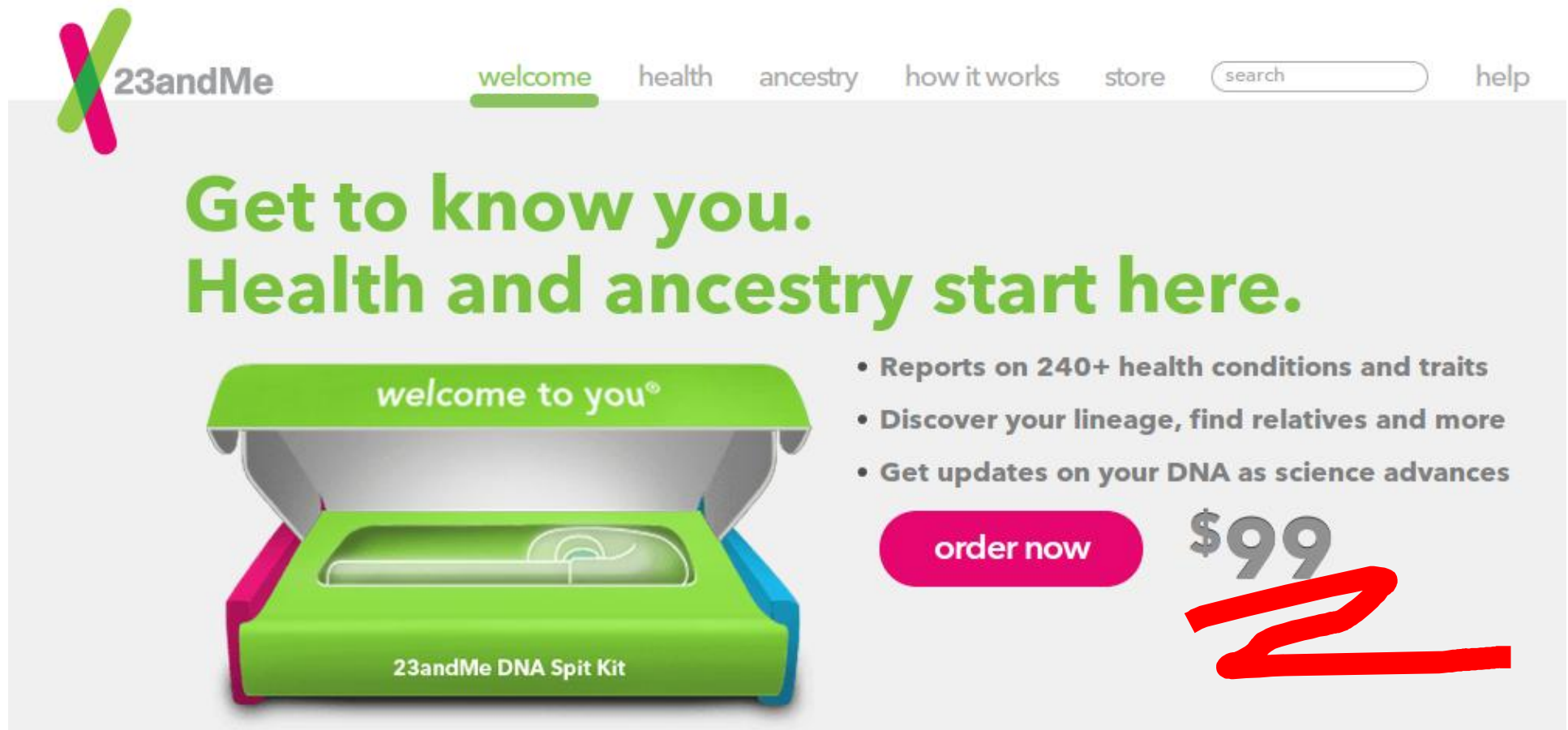
Fig. 9 – Example of BCR-ABL translocation/fusion-gene

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 simultaneously 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)

Personalized medicine

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



The image shows a screenshot of the 23andMe website. At the top left is the 23andMe logo, consisting of a stylized 'X' made of two overlapping lines (one pink, one green) followed by the text '23andMe'. To the right of the logo is a navigation menu with links for 'welcome', 'health', 'ancestry', 'how it works', 'store', a search bar, and 'help'. The 'welcome' link is underlined in green. Below the navigation is a large green banner with the text 'Get to know you. Health and ancestry start here.' in a bold, green, sans-serif font. In the center of the banner is an image of an open green DNA Spit Kit box. The lid of the box is open and has the text 'welcome to you®' on it. The box itself is green with a blue and pink accent on the sides. Below the box is the text '23andMe DNA Spit Kit'. To the right of the box is a list of three bullet points: '• Reports on 240+ health conditions and traits', '• Discover your lineage, find relatives and more', and '• Get updates on your DNA as science advances'. Below the list is a pink button with the text 'order now'. To the right of the button is the price '\$99' in a large, bold, black font. A thick red scribble is drawn over the price '\$99'.

23andMe

welcome health ancestry how it works store search help

Get to know you. Health and ancestry start here.

welcome to you®

23andMe DNA Spit Kit

- Reports on 240+ health conditions and traits
- Discover your lineage, find relatives and more
- Get updates on your DNA as science advances

order now

\$99

Personalized medicine

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



The image shows a screenshot of the 23andMe website. At the top left is the 23andMe logo, followed by navigation links for 'welcome', 'health', 'how it works', and 'help'. A search bar is also visible. The main headline reads 'Get to know you. Health and ancestry starts here.' Below this, there is a list of features: 'Reports on 240+ health conditions and traits', 'Discover your heritage, find relatives and more', and 'Get updates on your DNA as science advances'. A large blue '\$99' is overlaid on the page, and a pink 'order now' button is visible. In the bottom left of the screenshot, a green DNA Spit Kit is shown with the text '23andMe DNA Spit Kit' on it.

Personalized medicine

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

23andMe

Second kit

Get to know you
Health and ancestry insights here.

- Results on 24 health conditions and traits
- Discover your heritage, find relatives and more
- Get updates on your DNA as science advances

23andMe DNA Spit Kit

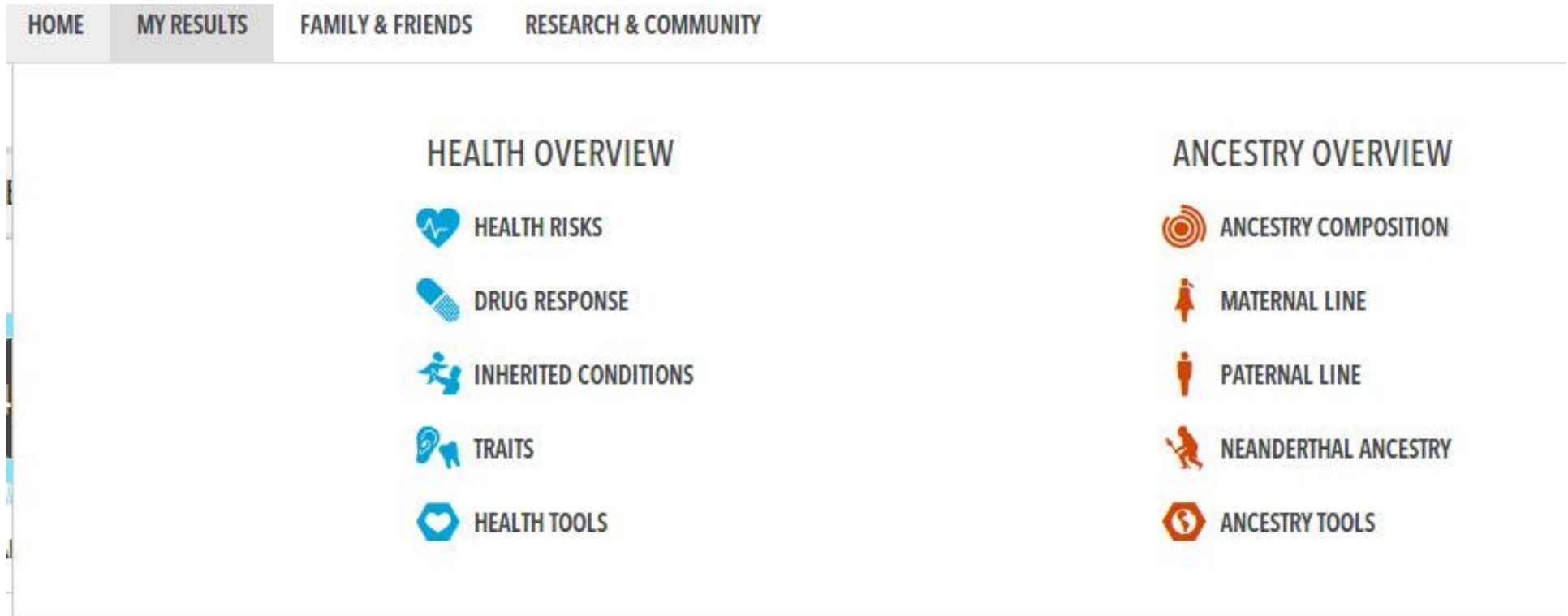
\$99

Personalized medicine

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

The image shows a screenshot of the 23andMe website. At the top left is the 23andMe logo, followed by navigation links: 'welcome', 'ancestry', 'how it works', 'store', a search bar, and 'help'. The main heading reads 'Get to know you. Health and ancestry. It all starts here.' Below this is a list of features: '• Results on 24 health conditions and traits', '• Discover your heritage, find relatives and more', and '• Gets updates on your DNA as science advances'. A green DNA Spit Kit is shown in the foreground with the text '23andMe DNA Spit Kit' on its base. A pink button with the text 'Get it now' is visible next to the price '\$99'. Large blue text '\$80' is overlaid on the image, with the 'nth kit' text above it.

Personalized medicine - cont'd



+ raw data

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