Statistical Genomics: Master of Science in Bioinformatics and Master of Science in Statistical Data Analysis

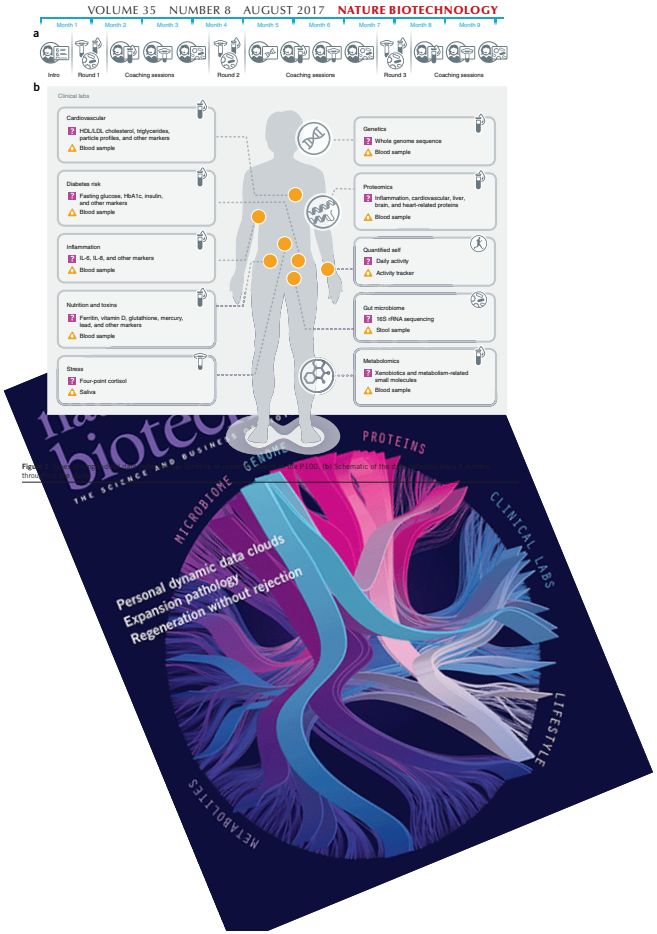
Lieven Clement
Ghent University, Belgium

Applications

- Forensics: matching crime scene DNA with suspects
- Pharmacogenomics: influence of genetic variation on drug response
- Agriculture: crop yield, tolerance to drought
- Precision medicin
- Microbiome
- History
- Bio-economy
- ...

Scientific Integrity and Reproducible Research

Bio-informatics research is based on empirical data



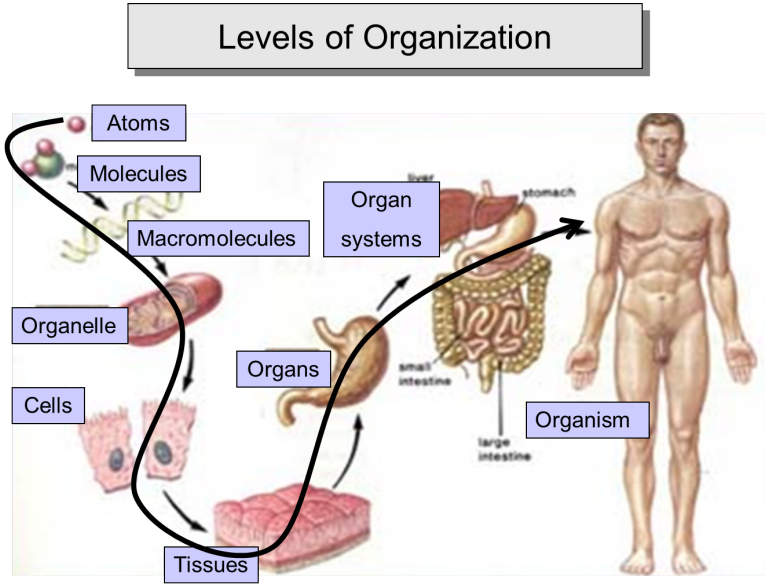
Scientific Integrity and Reproducible Research

Bio-informatics research is based on empirical data

→ Need for statistics

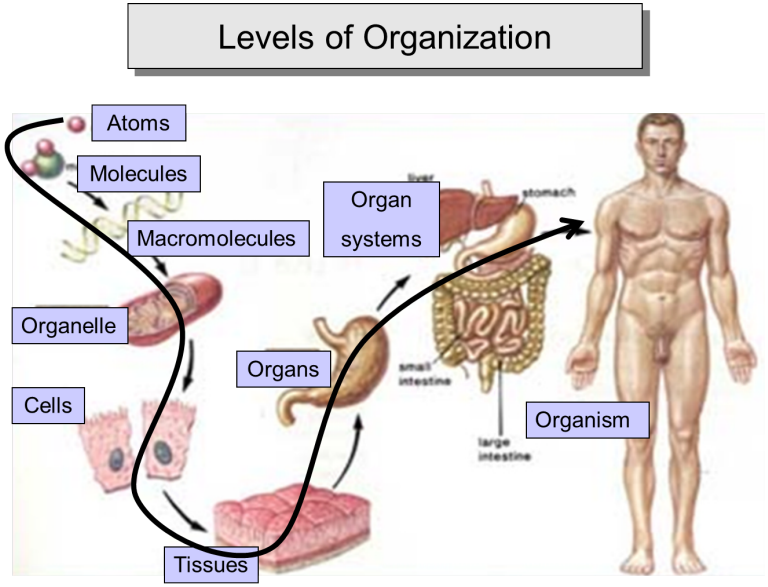
Genomics

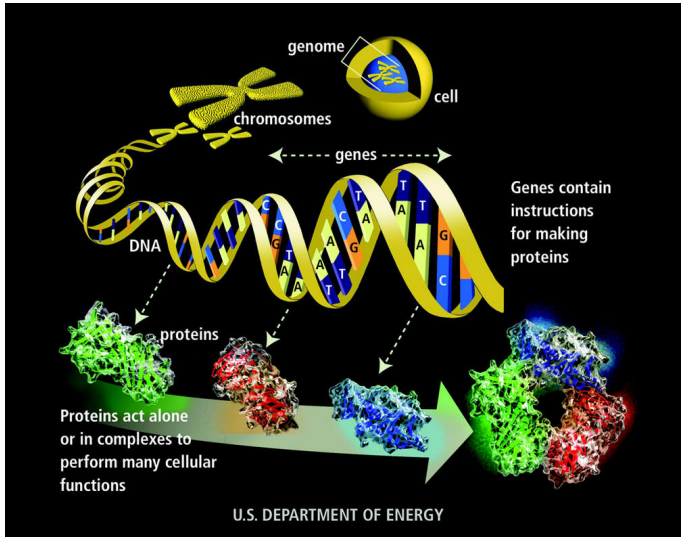
- The genome is entire hereditary information of an organism
- Contains all info needed for each function of an organism
- Most of the functions are carried out by proteins
- **Gene** is genomic region that directs synthesis of a **protein**
- **Genomics** studies all genetic information of an organism together: specific code, effects, functions and interactions



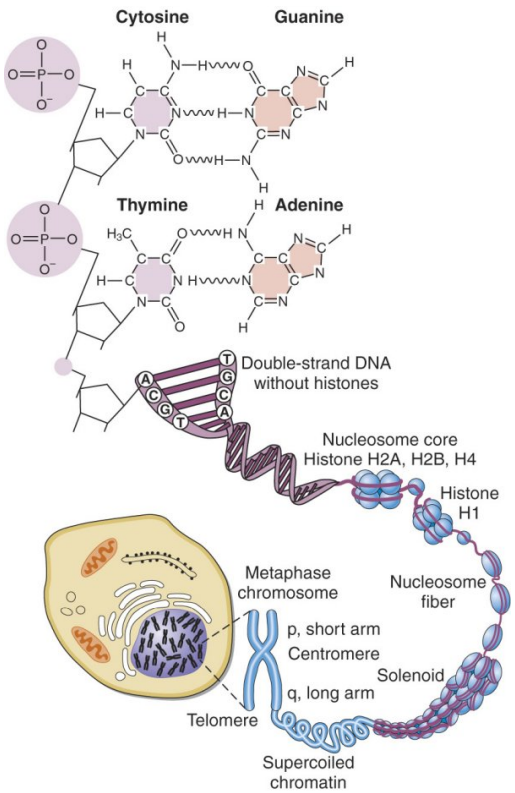
Genomics

- The genome is entire hereditary information of an organism
- Contains all info needed for each function of an organism
- Most of the functions are carried out by proteins
- **Gene** is genomic region that directs synthesis of a **protein**
- **Genomics** studies all genetic information of an organism together: specific code, effects, functions and interactions



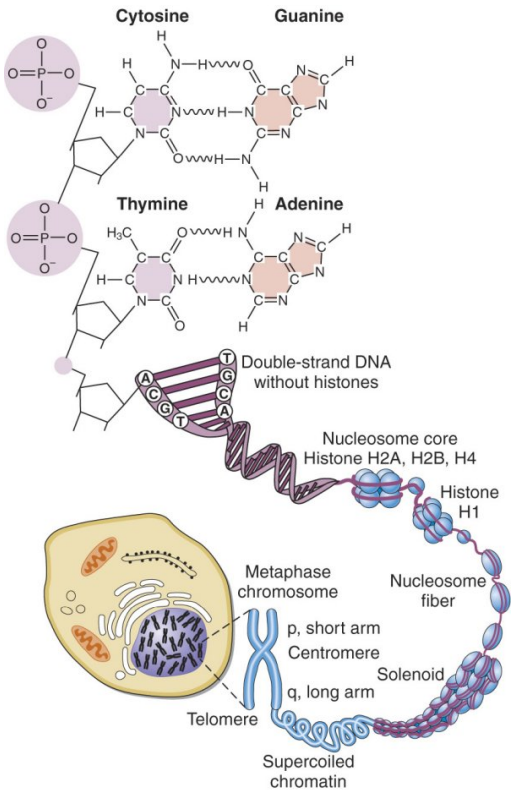


Genome - DNA



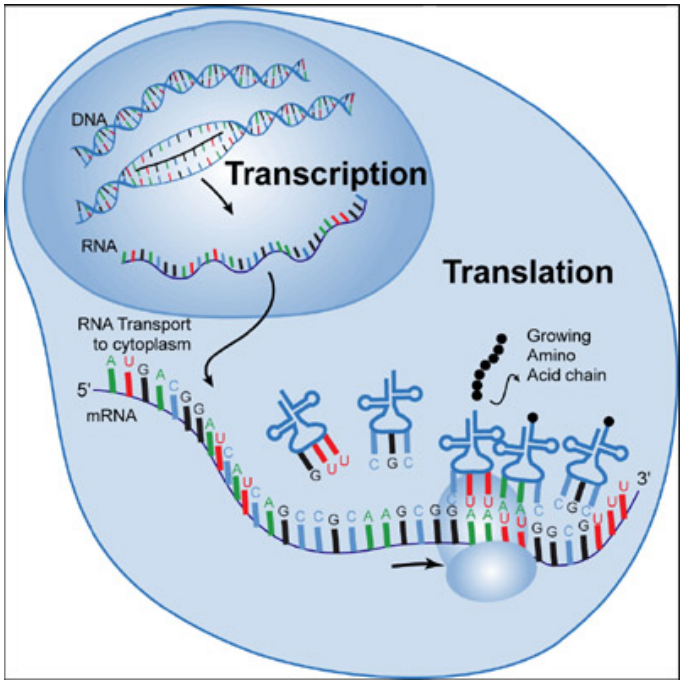
- Is stored in DNA (DeoxyriboNucleic Acid)(for many types of viruses in RNA)
- A code of 4 nucleotides
  - purines: adenine (A) and guanine (G)
  - pyrimidines: thymine (T) and cytosine (C)
  - a phosphate group;
  - a deoxyribose sugar;
- Double helix structure (2-3 hydrogen bounds)
- Organized in chromosomes
- Most of it is in the nucleus, also a part in mitochondrion (energy organelle of the cell)

DNA structure



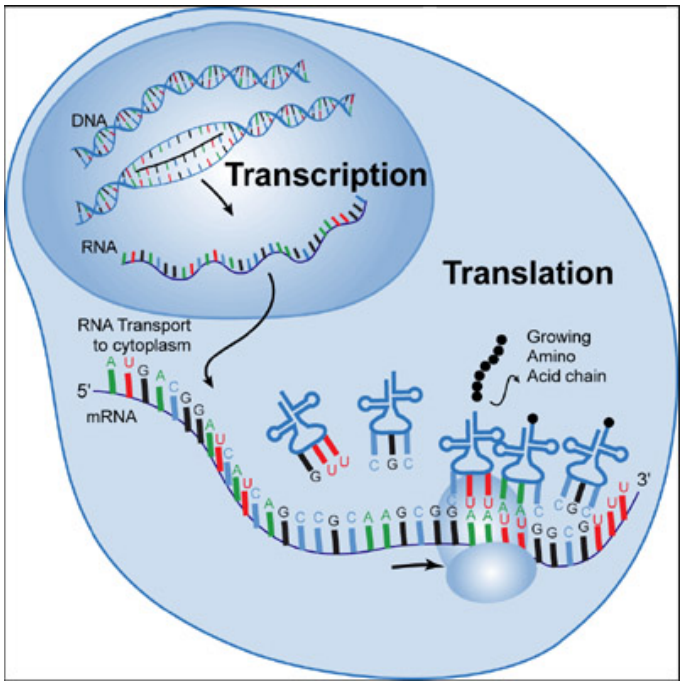
- Polynucleotide chains are directional molecules with slightly different ends: 3' end and 5' end.
- 3' and 5' refers to carbon atom numbering in the sugar ring. (3' hydroxyl group, 5' phosphate group)
- Complementary DNA strands are antiparallel (i.e, 5' to 3' ends for each strand are opposite)
- Most of it is coiled and condensed: very stable

Transcription-Translation



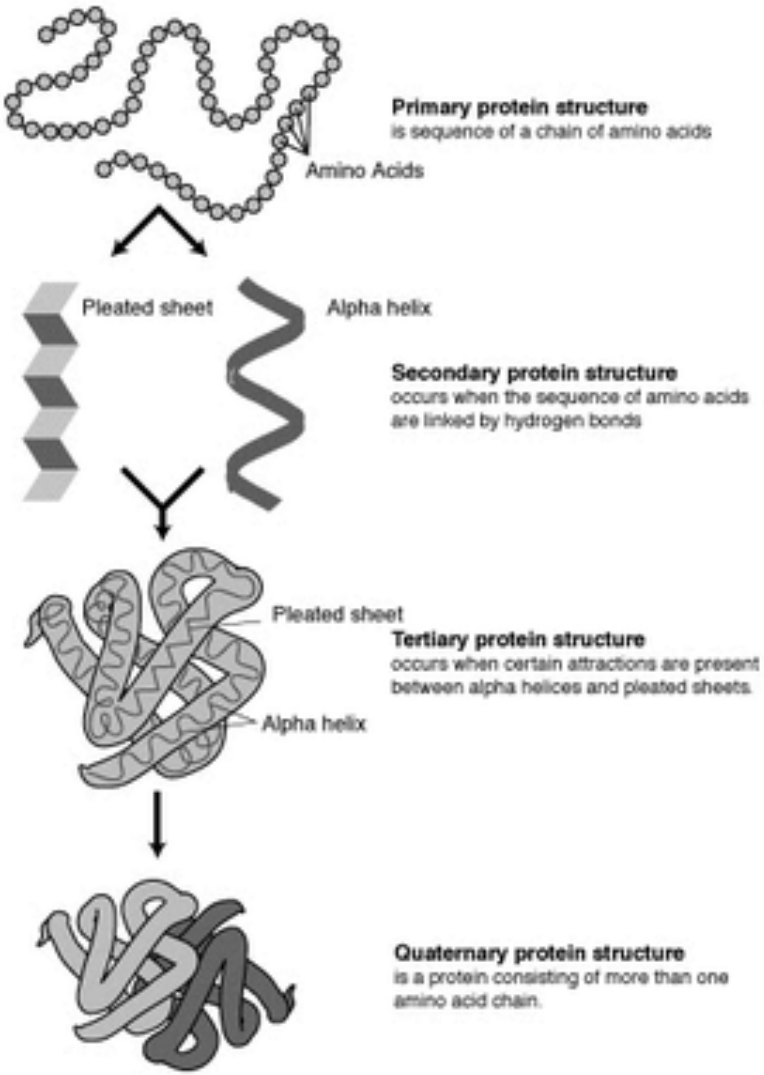
- Genome/DNA all genetic info in each cell: "Hard Drive". Four letter code: A, C, T, G
- Transcriptome/RNA: genetic info actively used by cell: "RAM"
- Transcription
  - Unwinding of DNA
  - RNA polymerase
  - DNA template: antisense strand
  - Single complementary RNA strand
  - Splicing

Transcription-Translation



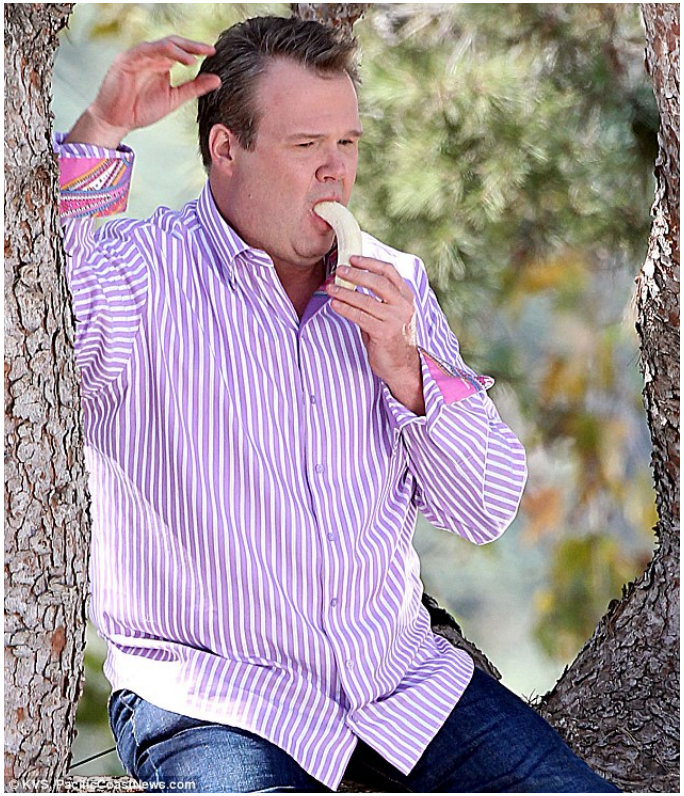
- Genome/DNA: "Hard Drive"
- Transcriptome/RNA: active genetic info in cell: "RAM"
- Proteome
- Translation RNA→Protein
  - At ribosomes: factories of the cell
  - 24 amino acids (aa)
  - 3 consecutive bases codon
  - tRNA: with antisense codon, caries one type of aa
  - several codons exist for same aa
  - start codon AUG (methionine, often removed)
  - stop codon UAG, UAA, UGA
- Post-translational modification+protein folding

Proteins



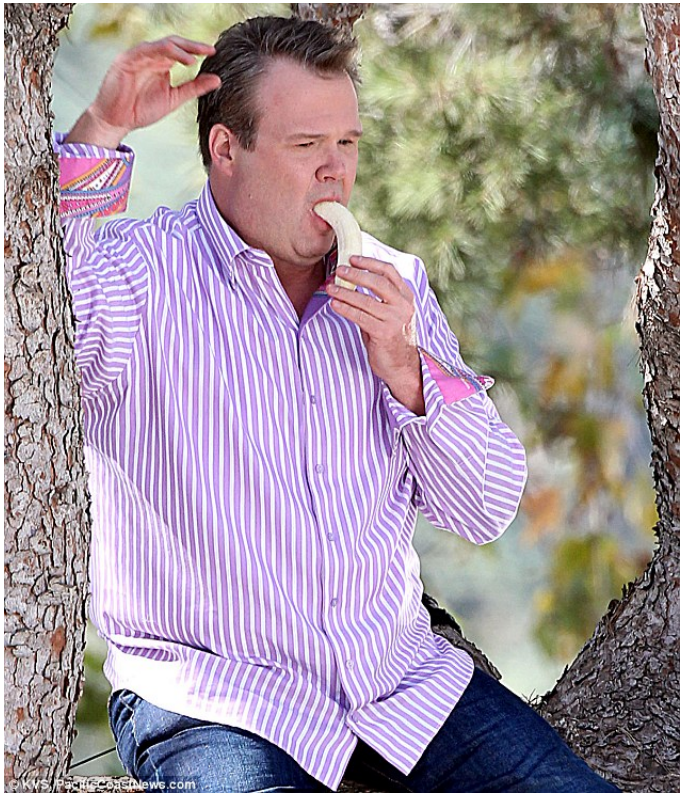
science daily

The human genome



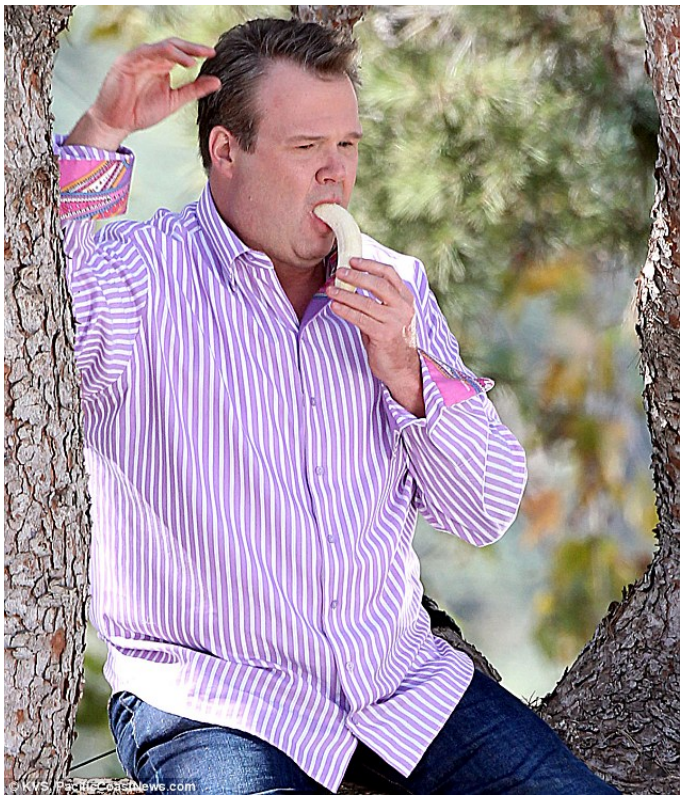
- Humans: 2×3 billion base pairs
- 2 meters of DNA
- $\pm$ 20.000 protein coding genes (500-4000/ chromosome)
- 99.9% in common with each-other
- Only 2% is protein coding

The human genome



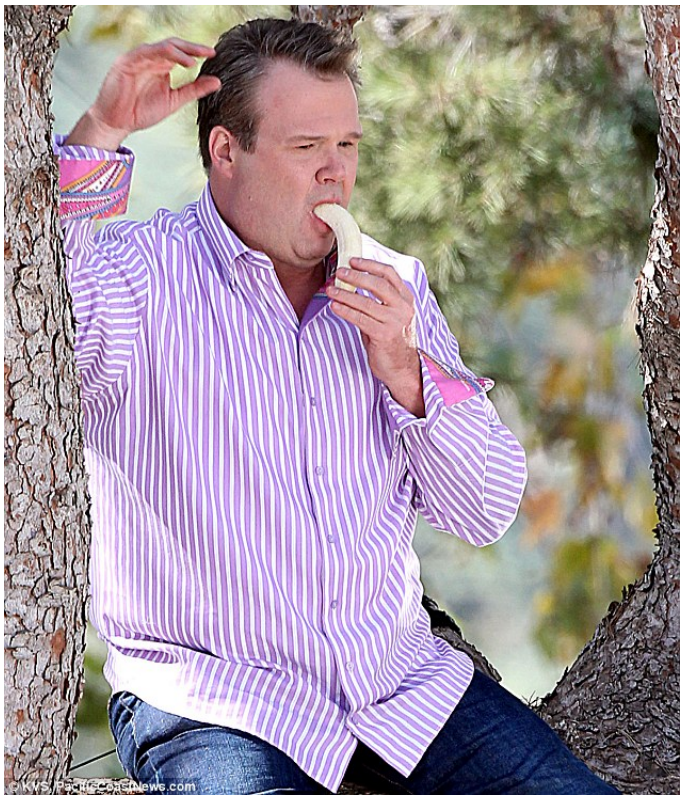
- Humans: 2×3 billion base pairs
- 2 meters of DNA
- $\pm$ 20.000 protein coding genes (500-4000/ chromosome)
- 99.9% in common with each-other
- Only 2% is protein coding
- 96% in common with chimp

The human genome



- Humans: 2×3 billion base pairs
- 2 meters of DNA
- $\pm$ 20.000 protein coding genes (500-4000/ chromosome)
- 99.9% in common with each-other
- Only 2% is protein coding
- 96% in common with chimp
- 50% in common with banana

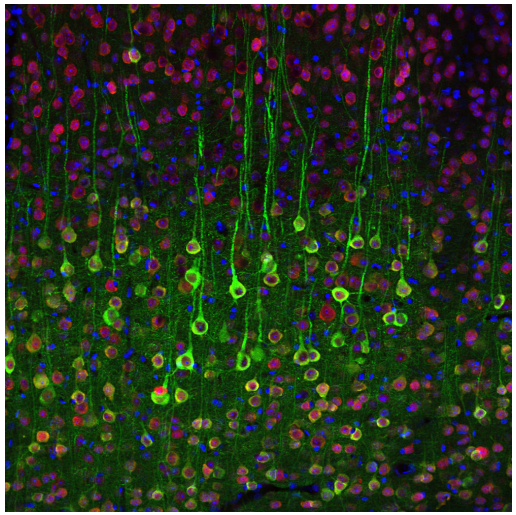
The human genome

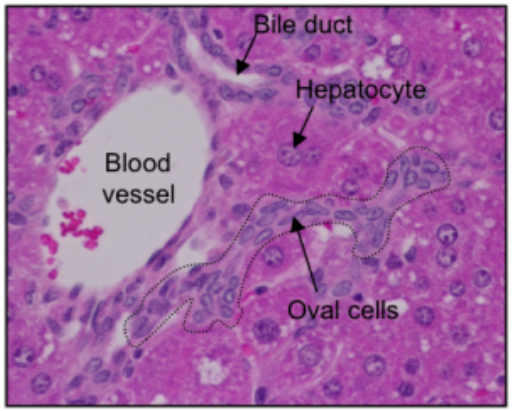


- Humans: 2×3 billion base pairs
- 2 meters of DNA
- $\pm$ 20.000 protein coding genes (500-4000/ chromosome)
- 99.9% in common with each-other
- Only 2% is protein coding
- 96% in common with chimp
- 50% in common with banana
- Organized in 23 pairs of chromosomes
  - 22 autosomal pairs
  - One sex chromosome pair: XX for females and XY for males
  - In each pair, one paternally other maternally inherited (cf. meiosis)

All cells of organism have same genome: still huge differences between different cells and over time?

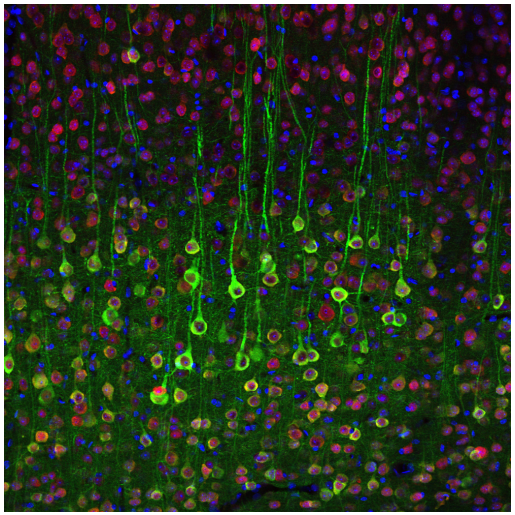
Brain vs liver cell

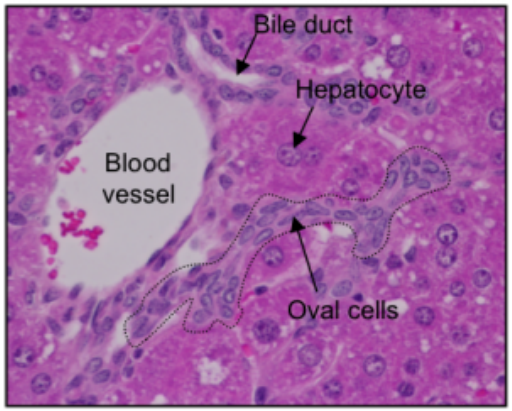




All cells of organism have same genome: still huge differences between different cells and over time?

Brain vs liver cell





Development of butterfly

Differential Gene Expression

- Different genes are expressed in different cells and at different times
- Genes are expressed at different levels in different cells and over time

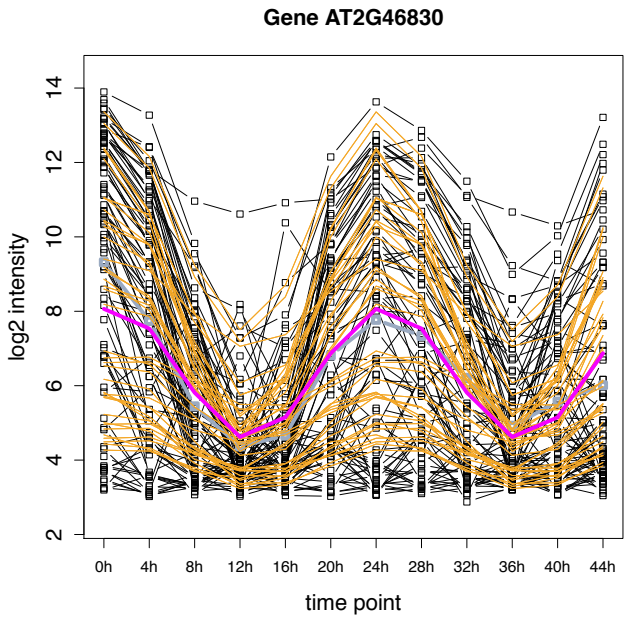
Human

Tissue/Cell	Number of genes*	Fraction of genes*	Ensembl genes†
Skeletal muscle[1]	11,276	0.61	11,953
Liver[1,3]	11,392	0.61	12,191
BT474[4]	11,844	0.64	12,808
MB435[4]	11,847	0.64	12,726
HME[5]	12,084	0.65	12,920
T47D[4]	12,205	0.66	12,983
Heart	12,209	0.66	13,159
MCF7[4]	12,281	0.66	13,216
Adipose tissue	12,553	0.68	13,503
Colon	13,016	0.70	14,052
Cerebellum[2,3]	13,132	0.70	14,043
Kidney	13,235	0.71	14,177
Brain[1]	13,298	0.71	14,107
Breast	13,406	0.72	14,537
Lymph node	13,534	0.73	14,686
Testes	15,518	0.84	16,869

*annotations from RefSeq, protein-coding genes.
†number of protein-coding genes, annotations from Ensembl.
[1]number of genes detected in mouse: skeletal muscle 11,799; liver 11,201; brain 13,626.
[2]standard deviation for samples from different individuals: 106.
[3]mean number for different individuals.
[4]breast cancer cell line.
[5]human mammary epithelial cell line.
doi:10.1371/journal.pcbi.1000598.t002

Ramsköld D et al. (2009) PLoS Comput Biol

Arabidopsis Clock Gene



De Beuf et al. (2012) BMC bioinformatics

Differential gene expression

Pomeroy et al. (2002) Nature 415

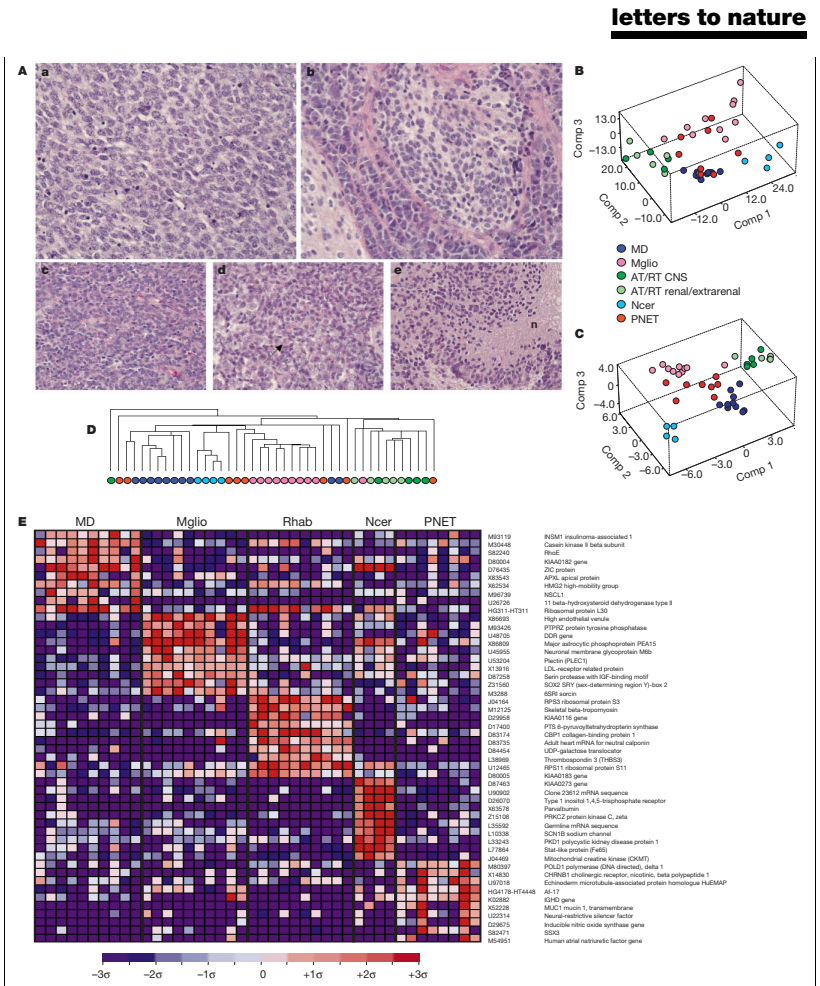
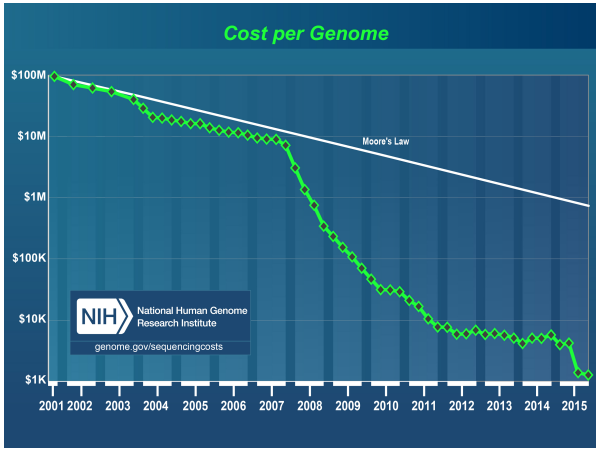


Figure 1 Classification of embryonal brain tumours by gene expression. **A**, Representative photomicrographs of embryonal and non-embryonal tumours. **a**, Classic medulloblastoma; **b**, desmoplastic medulloblastoma; **c**, supratentorial primitive neuroectodermal tumour (PNET); **d**, atypical teratoid/rhabdoid tumour (AT/RT; arrow indicates rhabdoid cell morphology); and **e**, glioblastoma with pseudopalisading necrosis (n). Magnification at 400×. **B**, Principal component analysis (PCA) of tumour samples using all genes exhibiting variation across the data set. The axes represent the three linear combinations of genes that account for most of the variance in the original data set (see Supplementary Information I and III). MD, medulloblastoma; Mglio, malignant glioma; Ncer, normal cerebella. **C**, PCA using 50 genes selected by signal-to-noise metric to be most highly associated with each tumour type (the top 10 for each tumour are listed in **E**). **D**, Clustering of tumour samples by hierarchical clustering using all genes exhibiting variation across the data set. **E**, Signal-to-noise rankings of genes comparing each tumour type to all other types combined (see Supplementary Information I). For each gene, red indicates a high level of expression relative to the mean; blue indicates a low level of expression relative to the mean. Rhab, rhabdoid. The standard deviation (σ) from the mean is indicated.

'omics profiling



- Study all of the genome simultaneously by high throughput 'omics profiling
- Huge number of variables/features for every sample (p features)
- Number of observations $n <<< p$
- Statistics is key to distinguish real patterns from random patterns that are observed because of we look in high dimensional data

Topics

Module I: Quantitative Proteomics

1. Data exploration and quality control using plots
2. Preprocessing: log-transformation, Filtering, Normalization, Summarization
3. Dealing with batch effects and other confounders
4. Statistical Concepts
    1. Linear models/Linear mixed models
    2. Trade-off between biological relevance/effect size vs statistical significance
    3. Empirical Bayes Methods
    4. Multiple testing

Module II: Next generation sequencing (NGS, Transcriptomics)

1. NGS Data exploration
2. Preprocessing/normalization
3. Additional Statistical Concepts
    1. Generalized linear models (GLM) for binary data
    2. GLM for count data
    3. Overdispersion

Organisation

1. Theory and Tutorials are blended
   - Module I: week 1-5
   - Module II: week 6-10
   - Project: week 1-10 via small assignments + week 11-12

2. Communication and submission of projects via minerva
3. All tutorials are based on R/Bioconductor
   - via R-studio
   - Scripts are made in R/markdown notebooks: a file format to combine text, R code and R output.

   $\rightarrow$ This makes it very easy to document your analysis and to distribute them in a way which is reproducible.

Organisation

4. Project
   - Projects: 10/20
   - Written Exam: 10/20.
     - Open book
     - Deep insight expected
     - Critical assessment of R-output,

Projects + Master thesis

- Project 201415, Master thesis 201516:

zingeR: unlocking RNA-seq tools for zero-inflation and single cell applications

Koen Van den Berge, Charlotte Soneson, Michael I. Love, Mark D. Robinson, Lieven Clement
doi: https://doi.org/10.1101/157982

- Project 201516: Neurogenomic profiling reveals distinct gene expression profiles between brain parts that are consistent across cichlid species of the genus Ophthalmotilapia. Derycke et al. 2018.
- Project 201516: Manuscript in preparation. A leap of the hurdle in mass spectrometry based proteomics. (Presentation at HUPO conference 2017).

Mass spectrometrists should search for all peptides, but assess only the ones they care about

NATURE METHODS | VOL.14 NO.7 | JULY 2017 | 643

Adriaan Sticker[1–4], Lennart Martens[2–5] & Lieven Clement[1,4,5]

- Master thesis 201516:
- Design Project 201718: paper in preparation.
- Continuing on statistical genomics project for thesis is possible.