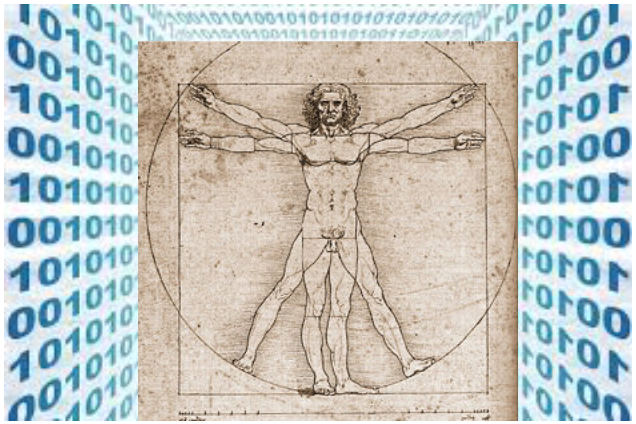


Introduction to biomedical data

MERJA HEINÄNIEMI

ASSOCIATE PROFESSOR IN BIOINFORMATICS

INSTITUTE OF BIOMEDICINE, SCHOOL OF MEDICINE

UNIVERSITY OF
EASTERN FINLAND

Big data in biomedicine

- Genomics data matrices: ***long & thin***

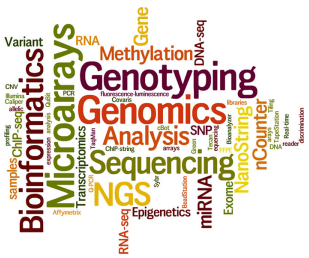
  10^4 to 10^6 variables ; 10 to 10^2 patients

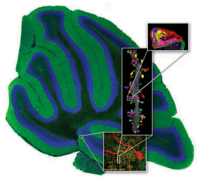
- Health record data matrices: ***short & broad***

  10 to 10^2 variables ; 10^4 to 10^6 patients

- Biomedical images:

  high-resolution digital images and movies





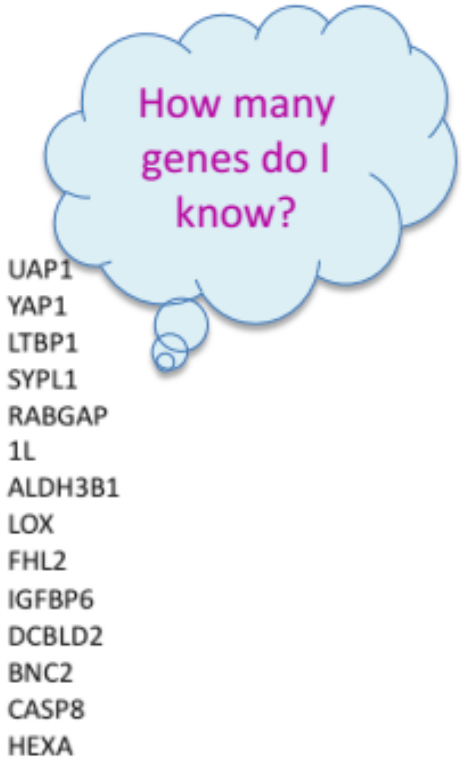
Future challenges

Modeling these together

Biomedical researcher needs new skills

*A biomedical researcher (a humble human being with limitations) would typically distill genome-wide data down to **the hundred or so genes** we think we know something about*

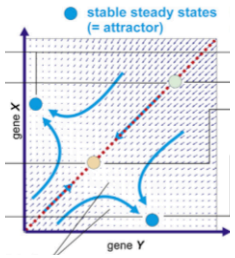
Motivation By modelling the data in clever ways, it is also possible to discover key genes(/cell types/…) in a data-driven manner

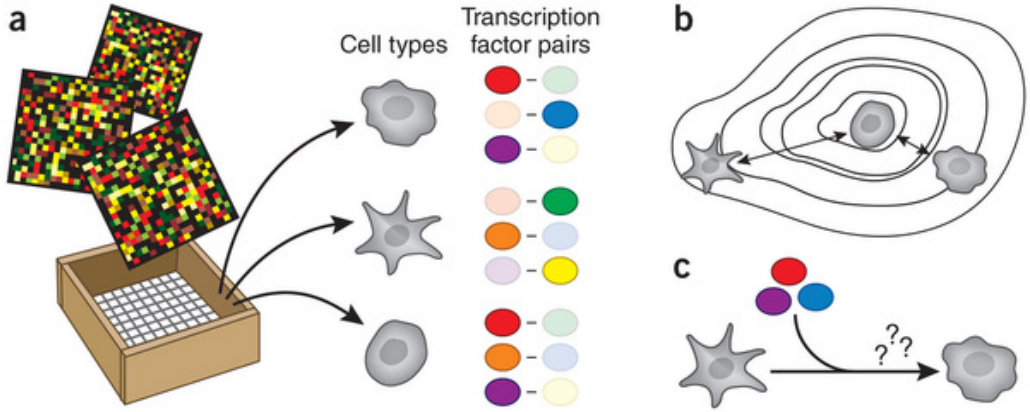


Path to success

Interdisciplinarity

Physics, computer science + molecular biology



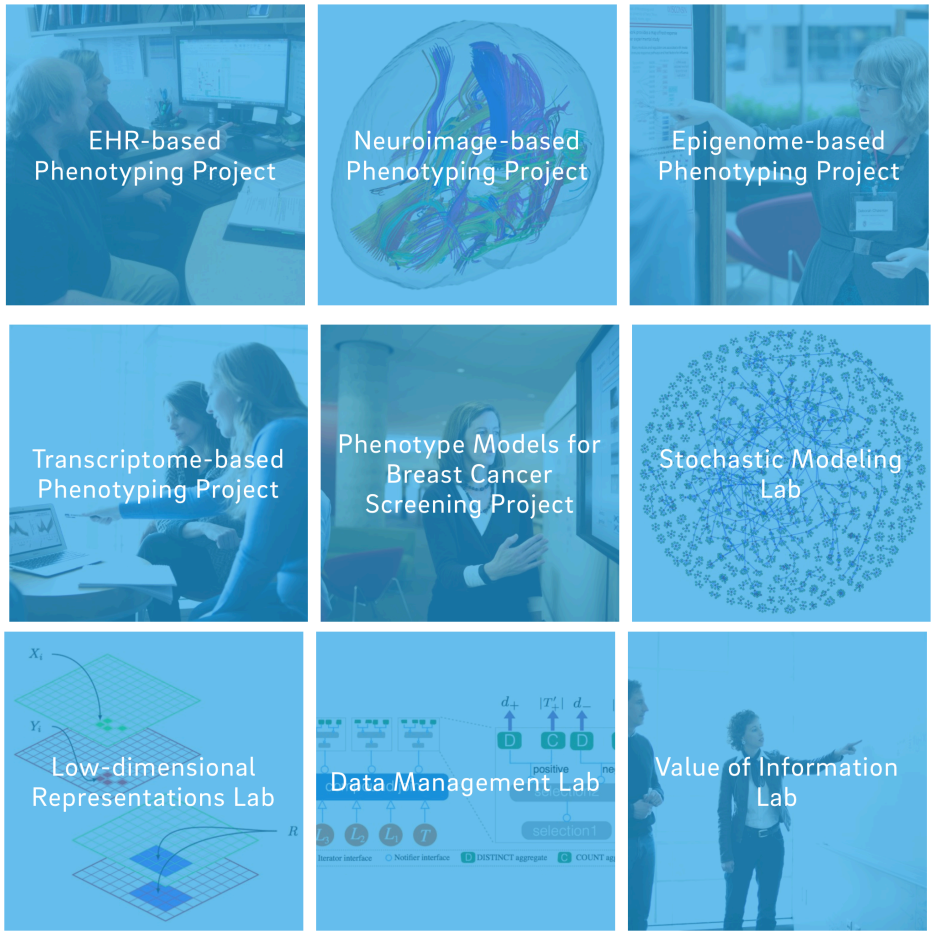


Discovered key regulators of normal cell types from data !

Heinäniemi et al Nature Methods 2013

Interdisciplinarity

NIH funded centre of excellence in Big Data



Center for Predictive Computational Phenotyping

http://cpcp.wisc.edu

Input from the field of machine learning

Unsupervised methods

Example application: Discovery of molecular disease subtypes

Juha and Robert will present examples this afternoon

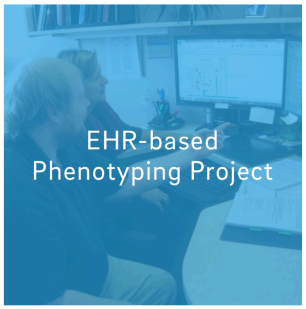
2D

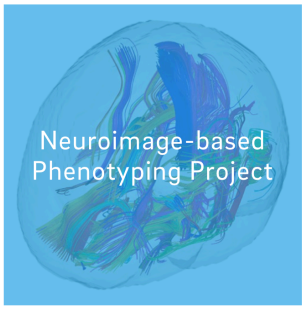
Dimensionality reduction, clustering

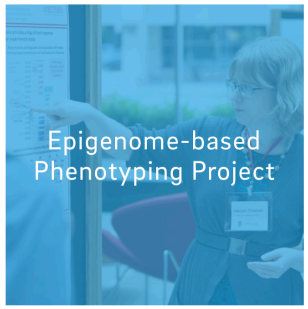
Unsupervised methods

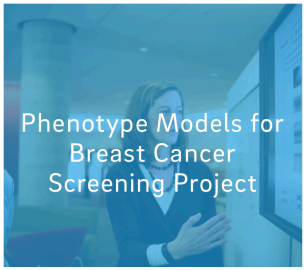
A multi-task graph-clustering approach for chromosome conformation capture data sets identifies conserved modules of chromosomal interactions

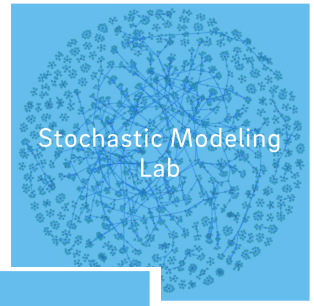
A MAD-Bayes Algorithm for State-Space Inference and Clustering with Application to Querying Large Collections of ChIP-Seq Data Sets

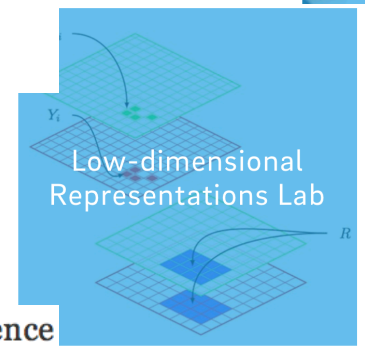


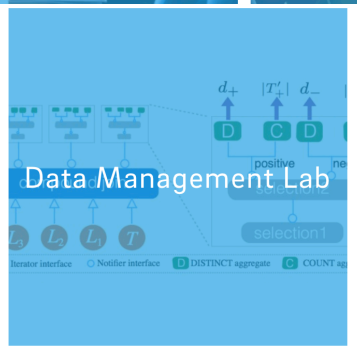


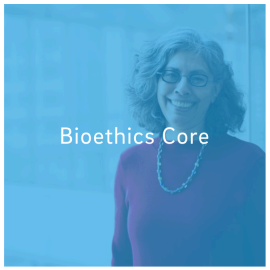






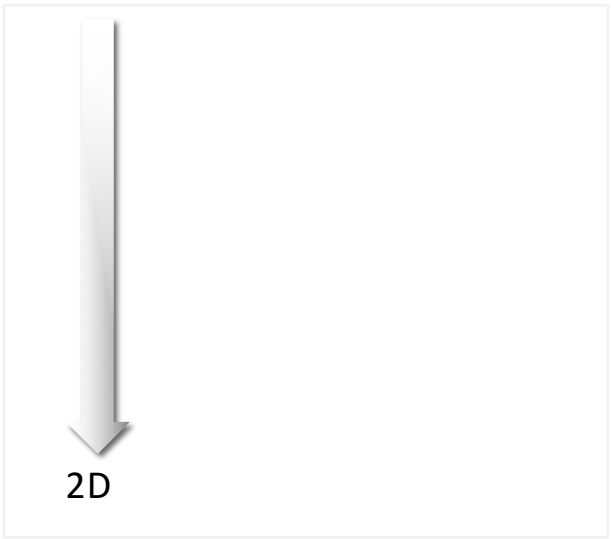






Input from the field of machine learning

Unsupervised methods



Dimensionality reduction, clustering

Supervised methods

Can you think of possible issues related to large datasets?

Classifiers, regression models

Supervised methods – training data

ELSEVIER

EBioMedicine

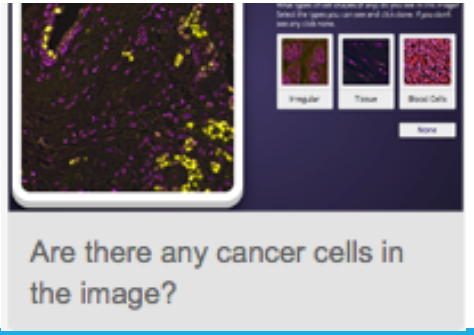
Volume 2, Issue 7, July 2015, Pages 681–689

Open Access

Original Article

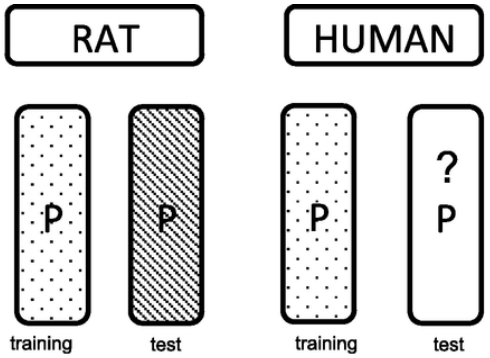
Crowdsourcing the General Public for Large Scale Molecular Pathology Studies in Cancer

The human eye is especially good at pattern recognition



Modeling in an interdisciplinary setting

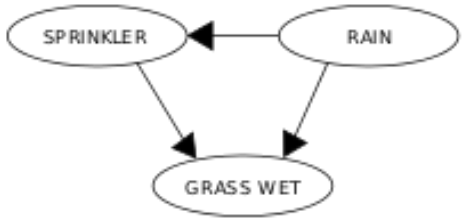
All models are wrong, but some of them are useful – George Box



Trans-species learning of cellular signaling systems with bimodal deep belief networks **Bioinformatics, 31, 2015**

Modeling in an interdisciplinary setting

SPRINKLER		
RAIN	T	F
F	0.4	0.6
T	0.01	0.99

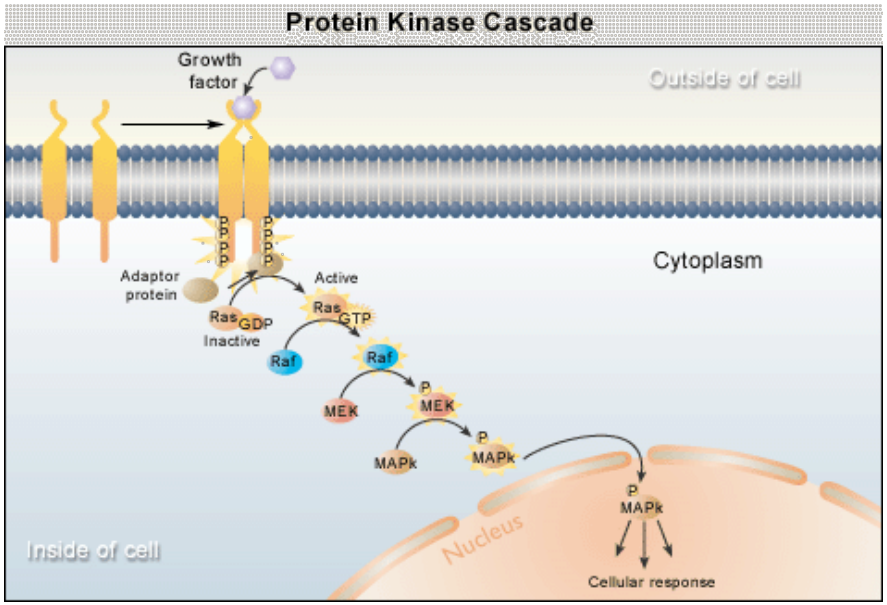


RAIN		
T	F	
0.2	0.8	

		GRASS WET	
SPRINKLER	RAIN	T	F
F	F	0.0	1.0
F	T	0.8	0.2
T	F	0.9	0.1
T	T	0.99	0.01

Arrow: conditional probability

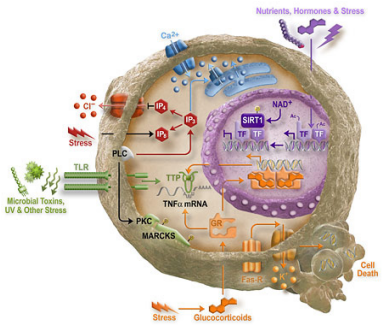
The model can answer questions like "What is the probability that it is raining, given the grass is wet?"



Arrow: molecular interaction

The model represents a molecular mechanism in transmitting a signal

Molecular biology in a nutshell



Learning goals

These are the questions that we will be addressing:

1. What is the function of key biomolecules in cells?
2. What is meant by gene expression?
3. What type of data matrix do we get from gene expression measurement?

Examples of changes in biomolecules that underlie disease

DNA – RNA - protein

To produce a functional part of the molecular machinery inside a cell, typically a protein molecule:

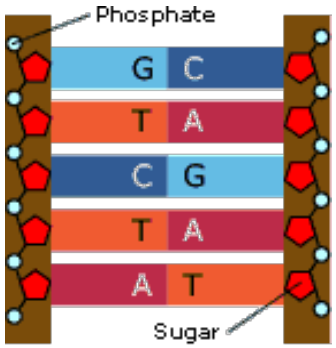
Access the segment of DNA encoding the ”recipe” (gene region)

Copy the DNA ”recipe” into a messenger RNA molecule (transcription)

Translate the message using the genetic code into a sequence of amino acids that are joined together -> protein then folds into its functional shape

Additional ”tags” (modifications, signal peptides) serve to place the protein inside the cell and may regulate its activity

DNA – 4 letter alphabet (4 bases)



- The information exists as sequence of bases that are paired
- Two strands of the double helix run in opposite directions: each can be used to generate the other using the base-pairing rule: G pairs with C and A pairs with T

(an adenine-thymine pair, or a cytosine-guanine pair)

-> The base pairing is restricted!

From nobelprize.org

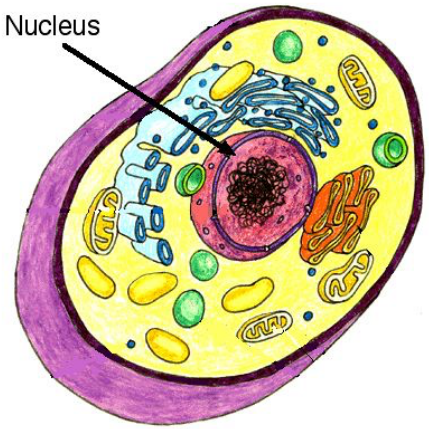
The human genome

The complete DNA content found in cells of an organism is called a genome (referring to the collection of genes that species can use)

The total length of human DNA is > 3 billion base pairs (”letters”)

Human DNA is not one long molecule, we have 22+X/Y chromosomes in two copies (maternal, paternal)

In eukaryotic cells (like human cells), DNA is in the nucleus



RNA

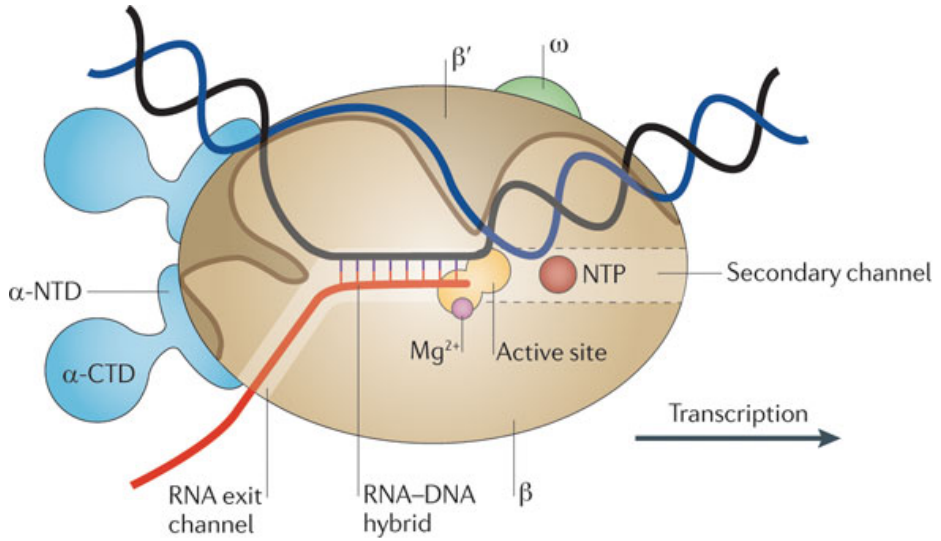
In all eucaryotic cells DNA never leaves the nucleus

The genetic code (the genes) is copied into RNA

RNA, then in turn is decoded (translated) into proteins in the cytoplasm

The process of copying the messages from gene regions is called **gene expression** and the message is also referred to as **transcript**

(mRNA is only one type of RNA, also structural RNAs exist!)



Nature Reviews | Microbiology
doi:10.1038/nrmicro2560

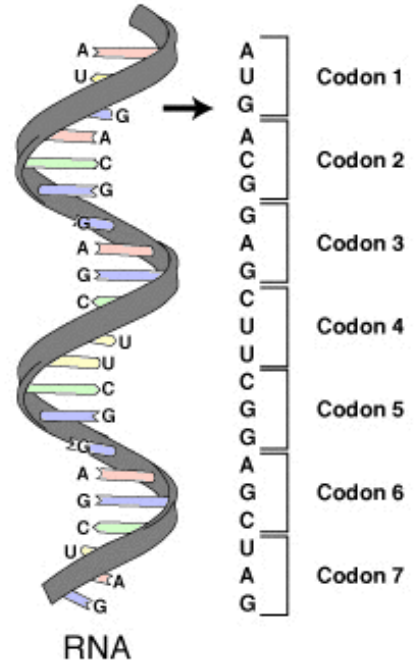
Genetic code to make a protein

mRNA is composed of four different nucleotides whereas a protein is built up from 20 amino acids

The nucleotide sequence is interpreted in codons, groups of three nucleotides

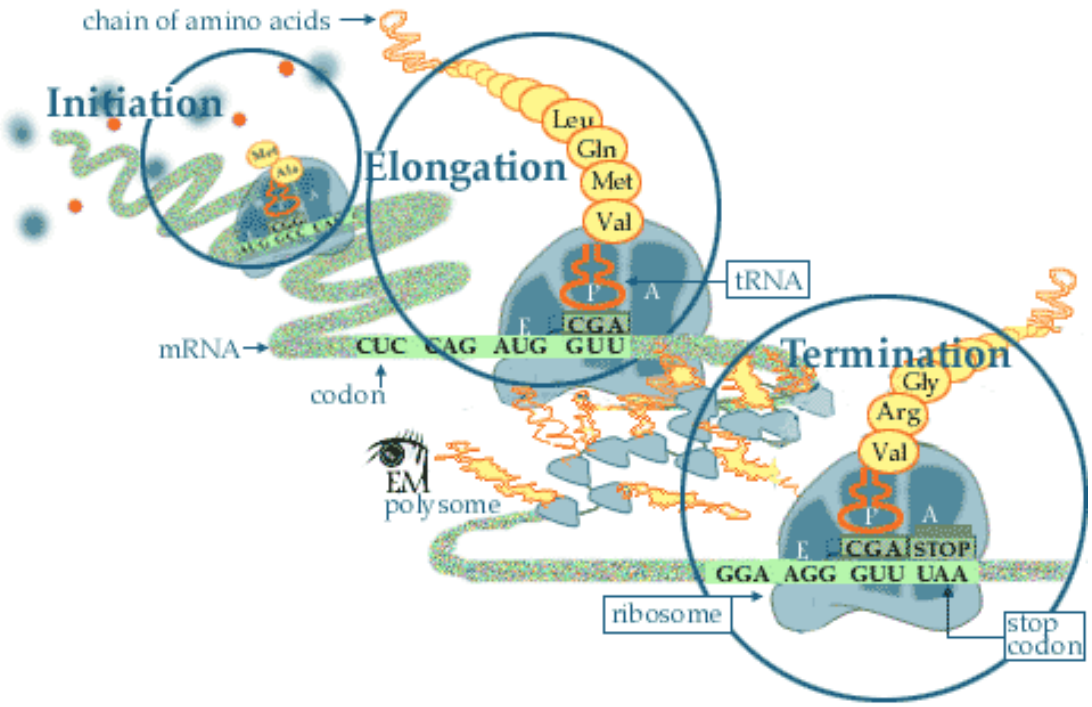
These codons have their corresponding anticodon in the tRNA, linked to one particular amino acid

This is referred to as the genetic code

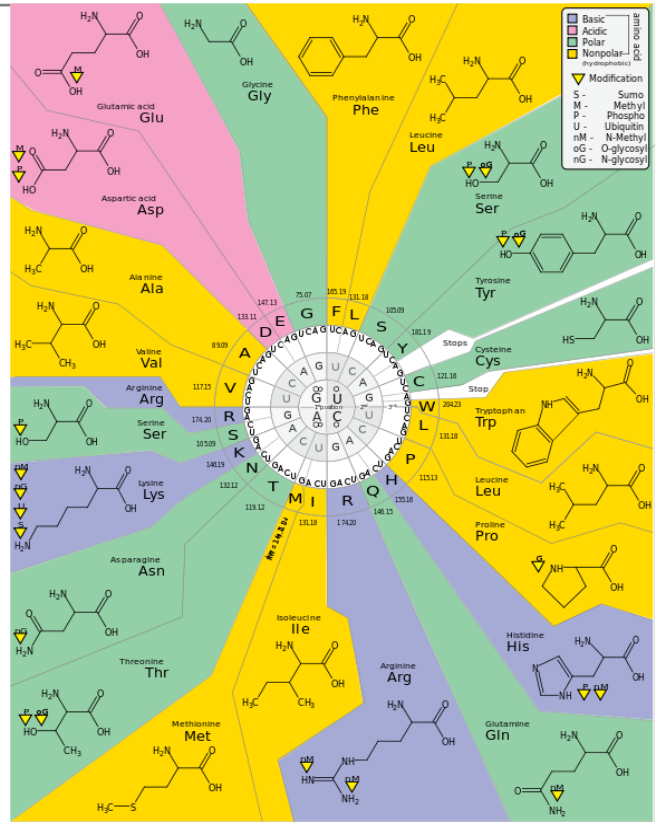


From Wikipedia

From letters into structure

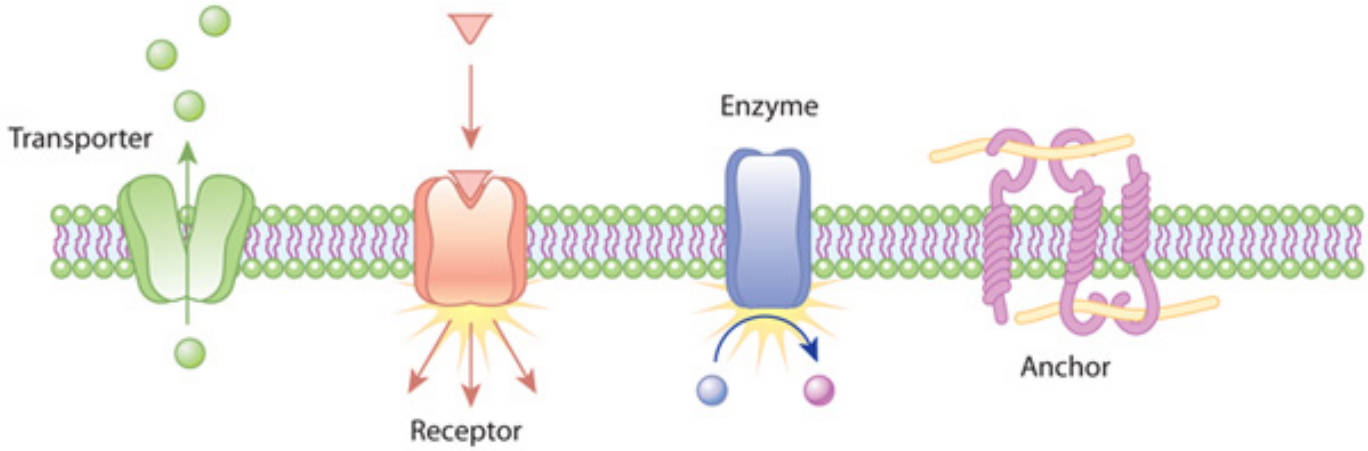


nobelprize.org

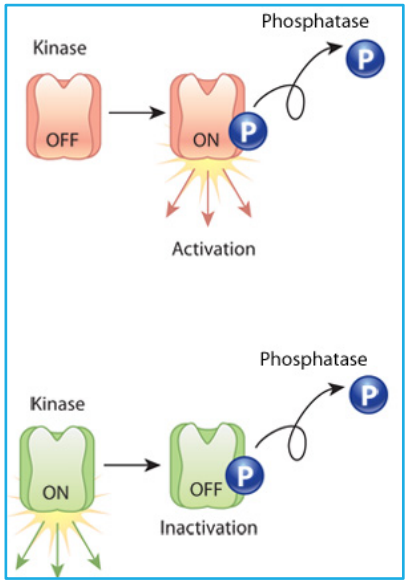


Protein

- Proteins are responsible for nearly every task of cellular life
- Different structures have evolved for different tasks

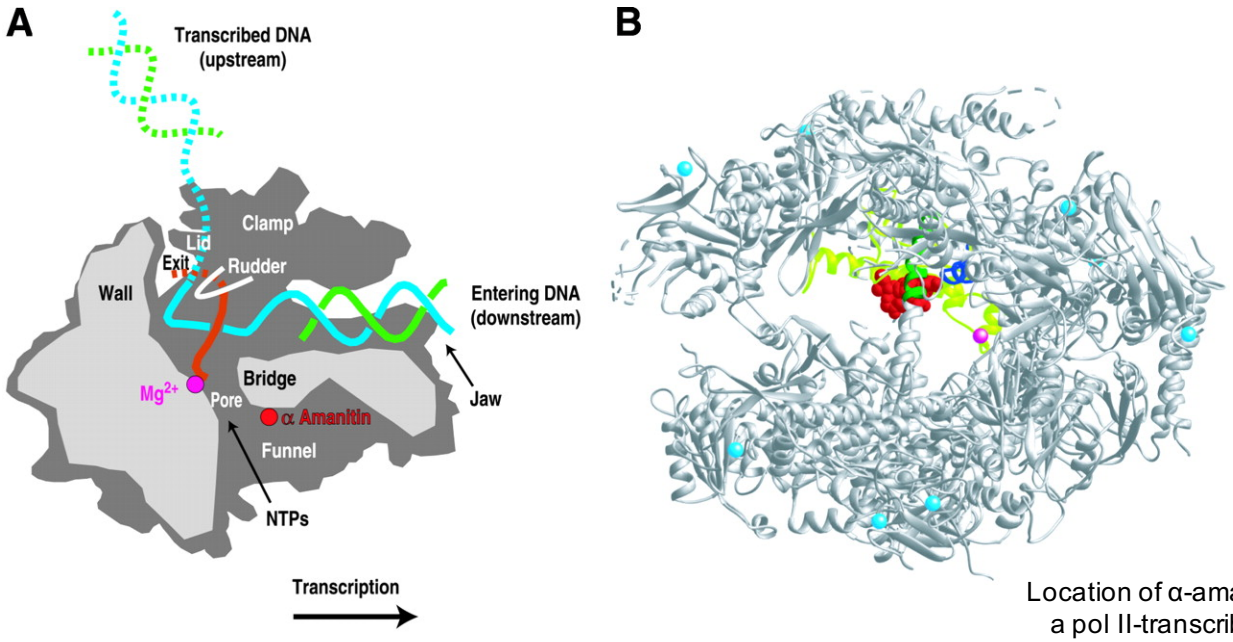


http://www.nature.com/scitable/topicpage/protein-function-14123348



Protein activity can be regulated

Protein complexes – molecular machines



Location of α-amanitin bound to pol II. (A) Cutaway view of a pol II-transcribing complex showing the location of α-amanitin binding (red dot) in relation to the nucleic acids and functional elements of the enzyme.

David A. Bushnell et al. PNAS 2002;99:1218-1222

PNAS

Summary 1

By now you know the roles of DNA, RNA and proteins in cells

Important for course task: Different cell types (e.g. neurons, muscle or liver cells) carry out different functions – they need different proteins to do this -> **gene expression is cell-specific**

What is read into RNA

Gene expression is tightly regulated in cells

- proteins known as transcription factors play a key role: they recognize a DNA sequence motif

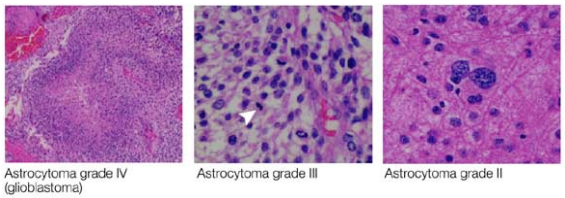
- activity of transcription factors in turn is regulated by the cell signaling cascades

- activity of cell signaling is regulated by their environment (other cells, hormones, nutrients, stress signals)

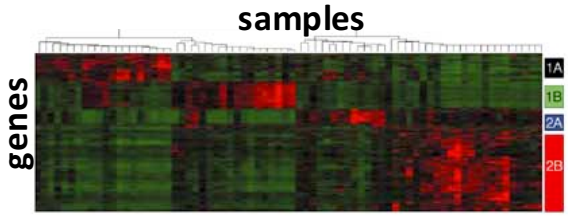
-> by measuring RNAs we get a snapshot of the ”recipies” a given cell is using, and thereby its state and identity

Gene expression profiles in diagnostics?

Image data – used in pathology



vs



Molecular profiles

Future challenges

Modeling these together

Hierarchical clustering of samples
Color indicates RNA level

N.B. In both – a population of cells measured

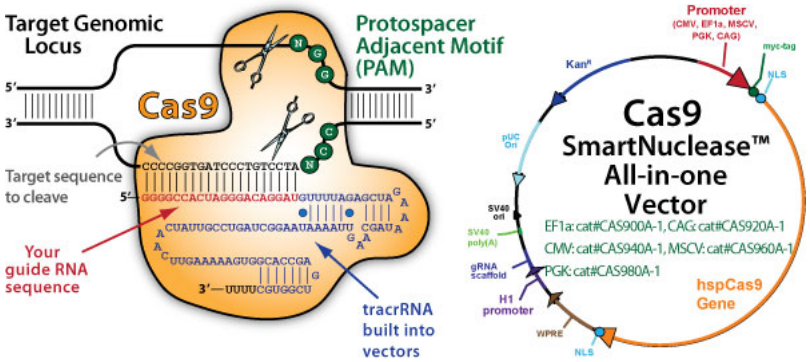
Molecular biology toolbox

Molecular biology – working with DNA, RNA and proteins

DNA represents a biomolecule that is relatively easy to work with

The molecular biology tool box is filled with various methods to extract and manipulate DNA (cut, paste, mutate etc)

State-of-the-art genome editing



https://www.systembio.com/crispr-cas9-plasmids

Creating disease models and studying mechanisms of a disease - possible

… even fixing disease !

SHARON LEES/GREAT ORMOND STREET HOSPITAL

Layla Richards is in remission from leukaemia after getting cells treated with DNA-cutting enzymes.

GENETIC MODIFICATION

Gene-editing wave hits clinic

Genome editing with Cas9 in adult mice corrects a disease mutation and phenotype

Hao Yin, Wen Xue, Sidi Chen, Roman L Bogorad, Eric Benedetti, Markus Grompe, Victor Koteliansky, Phillip A Sharp, Tyler Jacks & Daniel G Anderson

Affiliations | Contributions | Corresponding author

Nature Biotechnology **32**, 551–553 (2014) | doi:10.1038/nbt.2884
Received 12 December 2013 | Accepted 20 March 2014 | Published online 30 March 2014
| Corrected online **31 March 2014**
Corrigendum (September, 2014)

Abstract

Abstract · Accession codes · Change history · References · Author information · Supplementary information

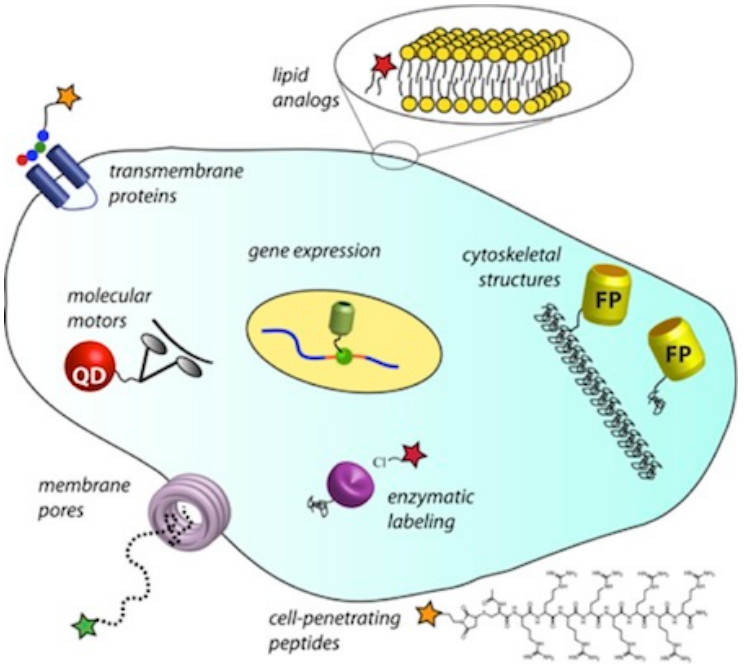
We demonstrate CRISPR-Cas9–mediated correction of a *Fah* mutation in hepatocytes in a mouse model of the human disease hereditary tyrosinemia. Delivery of components of the CRISPR-Cas9 system by hydrodynamic injection resulted in initial expression of the wild-type Fah protein in ~1/250 liver cells. Expansion of Fah-positive hepatocytes rescued the body weight loss phenotype. Our study indicates that CRISPR-Cas9–mediated genome editing is possible in adult animals and has potential for correction of human genetic diseases.

Molecular data available on disease

What kind of large-scale **(omics)** data is available?

Availability and ease of use of the technologies matters: many **DNA ja RNA level measurements** (gene expresssion levels, genetic variants)

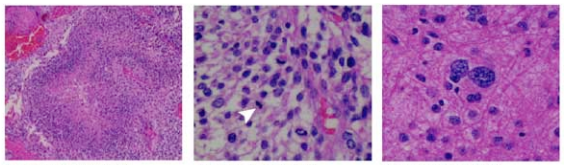
Less data on proteins and metabolites (so far)



We call global profiles of a certain molecule type ***omics*** *profiles*

RNA profiles – gene expression levels - transcriptomics

Most common data type in biomedicine



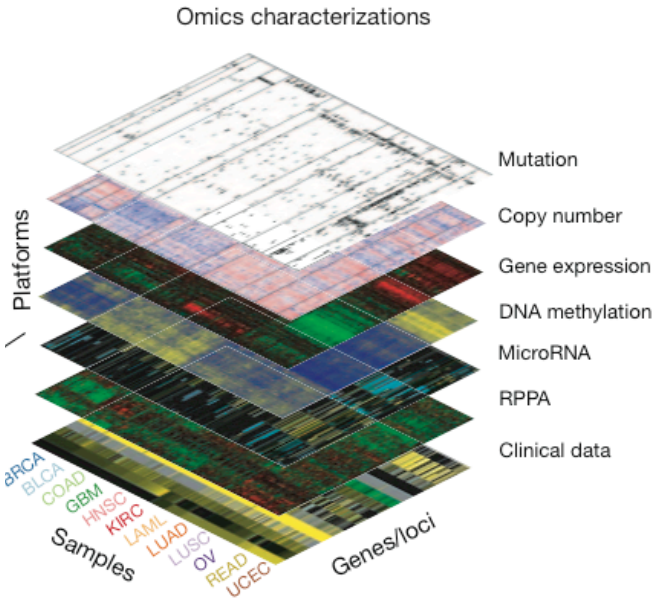
Tissue sample from patient

a snapshot of the ”recipies” present in tissue, reflecting the state and identity of cell types within

Isolate RNA

PROJECTS	PRIMARY SITE	CASES	FILES
39	29	14,531	250,498

Multi-omics: The Cancer Genome Atlas

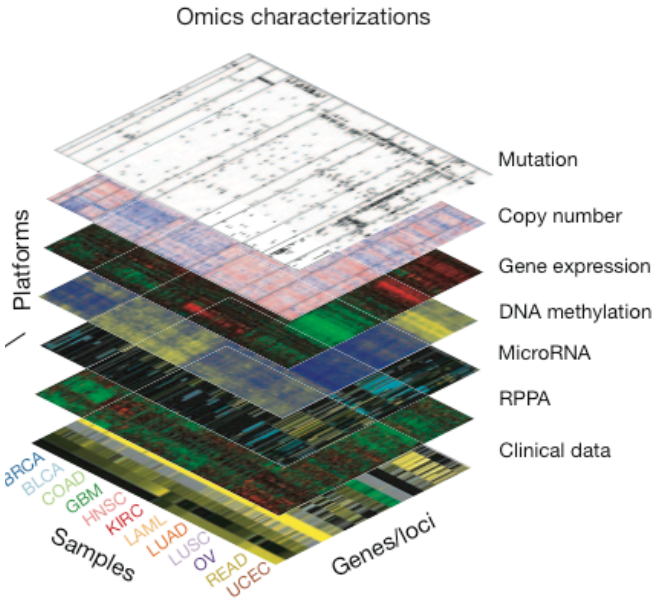


http://cancergenome.nih.gov/

Could we use multiple omics profiles to distinguish even better between different types of cancers?

PROJECTS	PRIMARY SITE	CASES	FILES
39	29	14,531	250,498

Multi-omics: The Cancer Genome Atlas



http://cancergenome.nih.gov/

Unsupervised methods

Multimodal Dimensionality reduction presented by Robert

Defining important parts of DNA

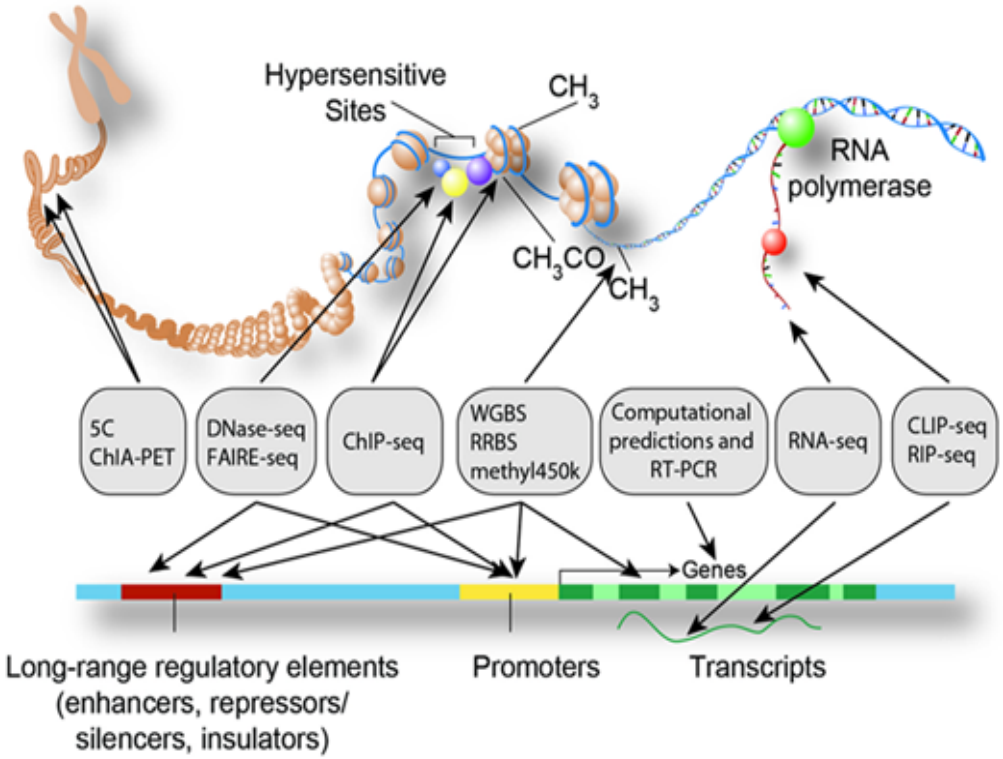
Gene regions

- Transcription start sites (TSS)

Regulatory regions

- Promoters (near TSS)
- Enhancers (distal)

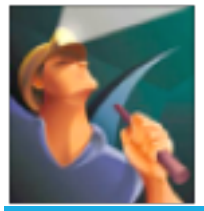
Annotating these across the 3 billion base pairs of human DNA is a major task!



Solution: genome-wide data, Deep sequencing datasets:
signal at each position along DNA measured for a large set of specific activity markers (e.g. Gene start marker H3K4me3)

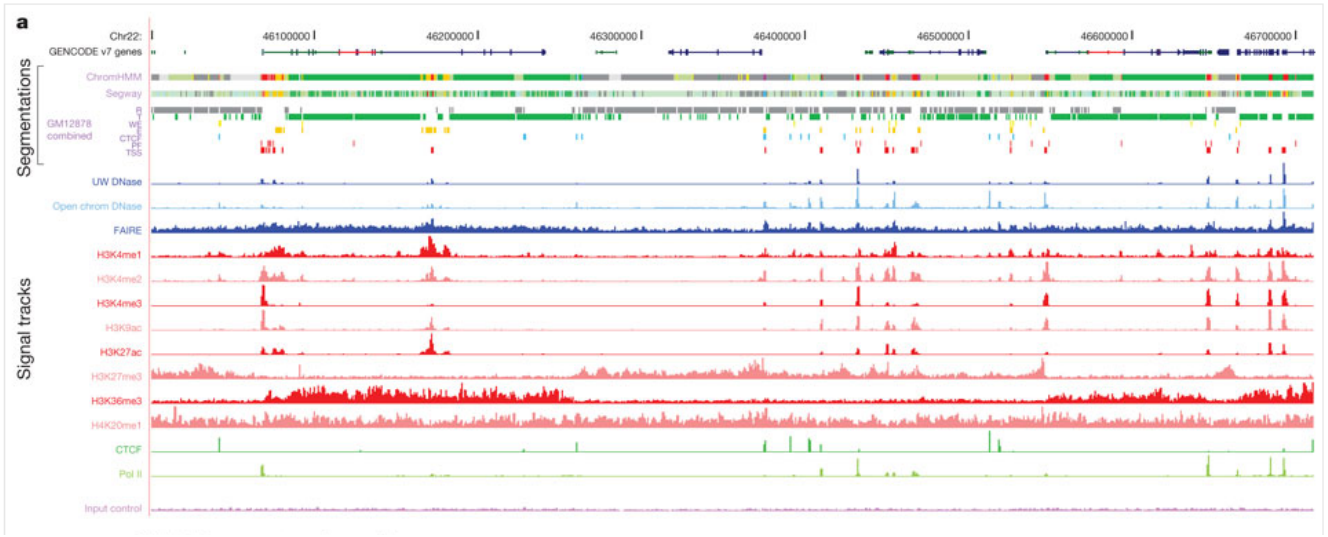
Assay		Assay category		Target of assay		Date released		Available data	
ChIP-seq	6267	DNA binding	6267	histone	2954	July, 2013	3113	fastq	7854
DNase-seq	832	Transcription	2935	histone modification	2954	March, 2014	887	bam	6604
polyA mRNA RNA-seq	705	DNA accessibility	900	transcription factor	2016	July, 2016	611	bigWig	5792
RNA-seq	526	DNA methylation	681	control	1677	May, 2016	577	bed narrowPeak	2881
shRNA RNA-seq	477	RNA binding	512	broad histone mark	1619	October, 2011	456	bigBed narrowPeak	2809
+ See more...		+ See more...		+ See more...		+ See more...		+ See more...	

Data
– Knowledge?

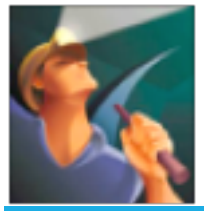


Solution: genome-wide data, Deep sequencing datasets:
signal at each position along DNA measured for a large set of specific activity markers (e.g. Gene start marker H3K4me3)

From
An integrated encyclopedia of DNA elements in the human genome
The ENCODE Project Consortium
Nature **489**, 57–74 (06 September 2012) | doi:10.1038/nature11247



Data
– Knowledge?



ML application: Automated segmentation of chromosomes into functional regions

e.g. HMM has been used – can you propose deep neural network alternatives based on Sequence modeling lecture?

Impact – analyzing changes in DNA

DNA can undergo changes, e.g. UV light can lead to base pair changes (mutation)

-> Inside gene regions, such changes can alter the coding sequence

-> Outside gene regions, changes can affect regulatory regions thereby altering RNA levels from nearby genes

Summary 2

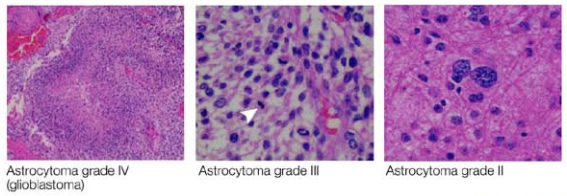
By now you learned about large-scale data available from disease models and patients

- Omics data refers to profiling the full content of DNA, RNA (and proteins) from the sample
- From one sample, it is possible to generate many measurement types

Important for course task: RNA-level measurements available across Alzheimer's disease patient cohort – this data gives you a snap-shot of the "recipies" used by cells in the tissue in disease vs control tissue samples, and thereby means to identify cell types/states that change in disease

RNA profiles – gene expression levels - transcriptomics

Most common data type in biomedicine



Tissue sample from patient

Isolate RNA

RNA levels:

Data matrix

10^4-10^5 rows

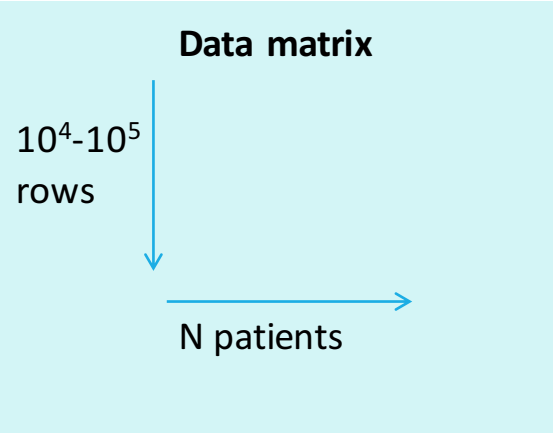
N patients

20 300 genes code for proteins

24 885 code functional RNAs

RNA profiles – gene expression levels

RNA levels



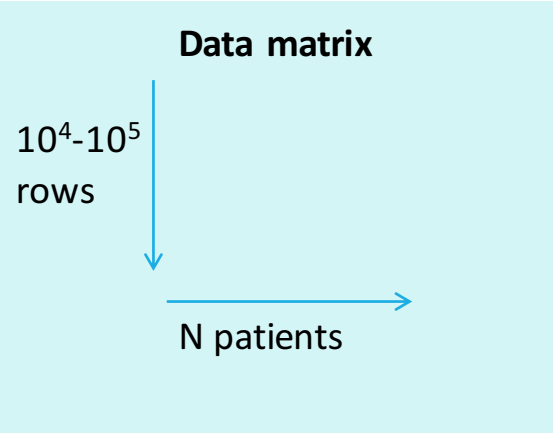
Two types of measurement methods

Microarray

RNA-sequencing (RNA-seq)

RNA profiles – gene expression levels

RNA levels



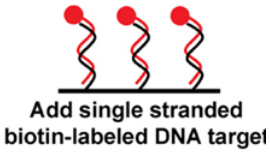
Two types of measurement methods

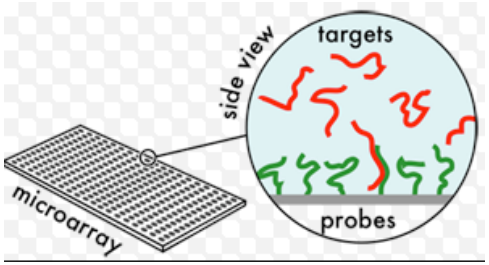
Microarray – continuous data

RNA-sequencing (RNA-seq) – discrete data

Microarray: How is the continuous signal generated?

Sample RNA molecules are converted to labeled cDNAs
–> microarray: matching DNA probes on a glass slide





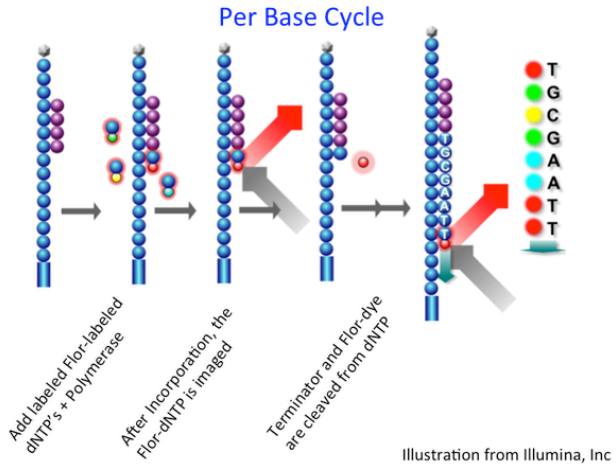
when two molecules ”match” (hybridize) light is emitted

Challenges
- Detecting specific hybridization, i.e. specific base pair pairing vs unspecific binding to probe
- Need to know in advance what molecules to detect (design of probes)

Data preprocessing: Most genes are expressed at very low levels; few genes high -> the distribution is skewed to the right. Statisticians often deal with highly skewed data on a logarithmic scale

RNA-seq: How is the discrete signal generated?

Sample RNA molecules are converted to cDNAs –> they attach to a flow cell where one letter by letter the sequence is read (50-100 base pairs from one or both ends)



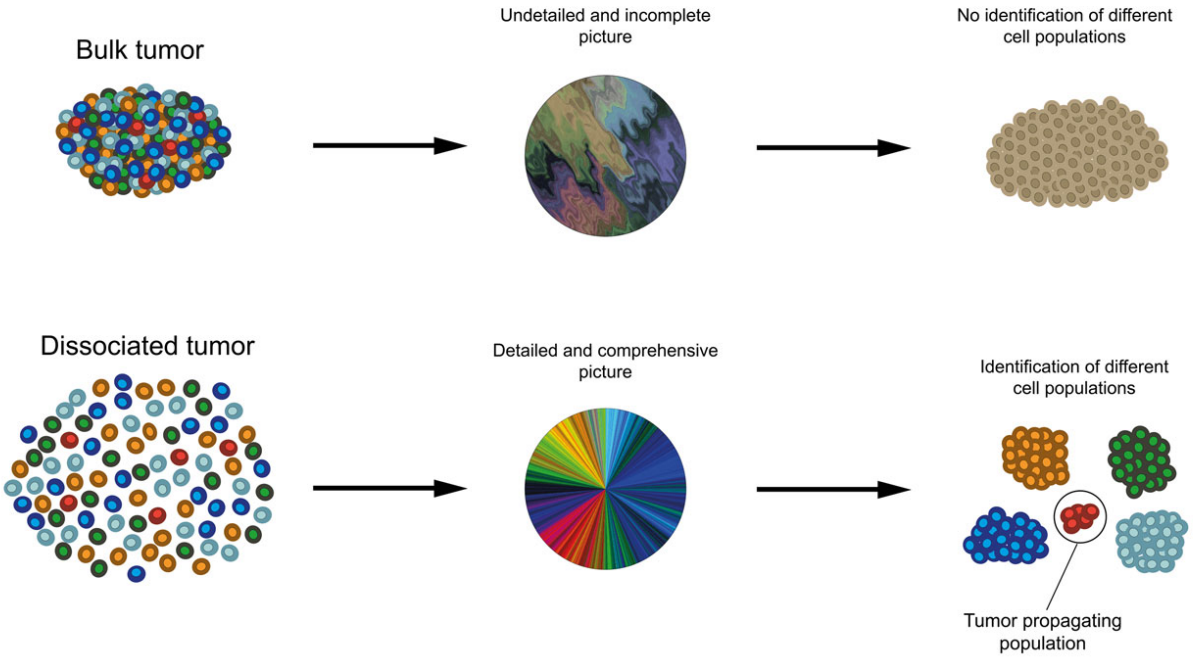
The obtained sequence is then matched against known RNAs (can also be used to find new) in a step called alignment, and for each RNA we count the sequence reads obtained for it

Challenges

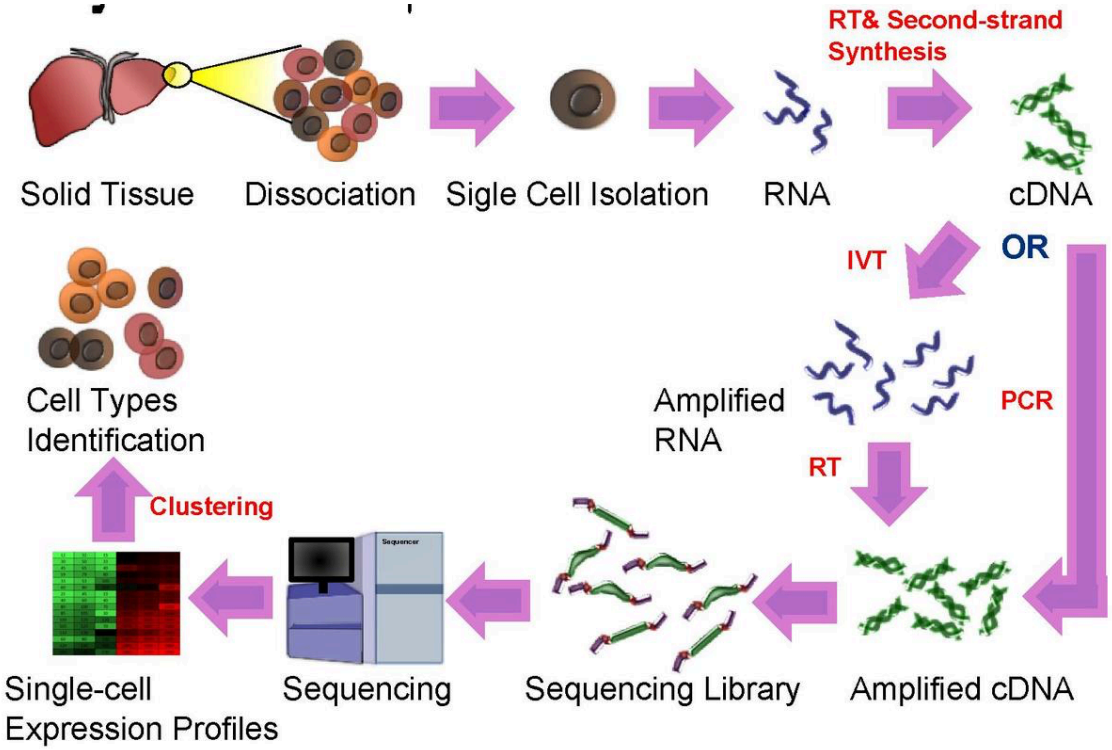
- The "recipe" RNAs, i.e mRNAs represent only a fraction of total RNA in cells – structural RNA needs to be depleted: differences between samples and experiment in success

Discrete count signal is typically modelled using negative binomial distribution (or Poisson distribution)

RNA-seq: bulk vs single cell



RNA-seq: bulk vs single cell



Analysis

https://www.nature.com/news/single-cell-sequencing-made-simple-1.22233

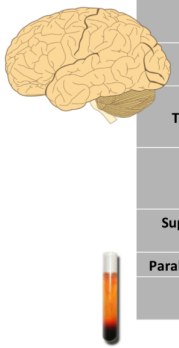
Bioinfo track independent assignment

You will analyze and model data generated by AMPAD-AD, *the Accelerating Medicines Partnership-Alzheimer's Disease*

Scientific hypothesis: neuro-degenerative diseases such as Alzheimer's disease affect the different brain cell types, resulting e.g. in neuronal loss in certain brain regions

Question: Which cell types are up-, down-regulated or unchanged in AD?

Human Tissue



HUMAN TISSUE	Diagnosis	Assay
Prefrontal Cortex	• Alzheimer's Disease • Mild Cognitive Impairment • Parkinson's Disease • Amyothrophic Lateral Sclerosis • Corticobasal Degeneration • Frontotemporal Dementia • Dementia with Lewy Bodies	• RNAseq • Gene Expression array • miRNA array • ChIPseq • DNA Methylation array • Proteomics • Confocal Imaging • SNP genotypes • Proteomics • Whole Exome Seq
Visual Cortex	• Alzheimer's Disease	• Gene Expression Array • SNP genotypes
Temporal Cortex	• Alzheimer's Disease • Progressive Supranuclear Palsy • Parkinson's Disease	• RNAseq • SNP genotypes
Cerebellum	• Alzheimer's Disease • Progressive Supranuclear Palsy • Parkinson's Disease	• RNAseq
Superiour Temporal Gyrus	• Alzheimer's Disease	• RNAseq • Whole Exome Seq
Parahippocampal Gyrus	• Alzheimer's Disease	• RNAseq
Serum	• Alzheimer's Disease • Mild Cognitive Impairment	• Metabolomics

Data: RNA-sequencing from postmortem tissue samples collected across 7 different brain regions
+ Associated metadata (clinical and technical)

Bioinfo track independent assignment

We can distinguish different cell types from measuring RNA levels

The problem: tissue samples represent the average RNA level profile across all cells (and cell types) – these are called bulk measurements

-> cell type deconvolution problem: discussed next

Bioinfo track independent assignment

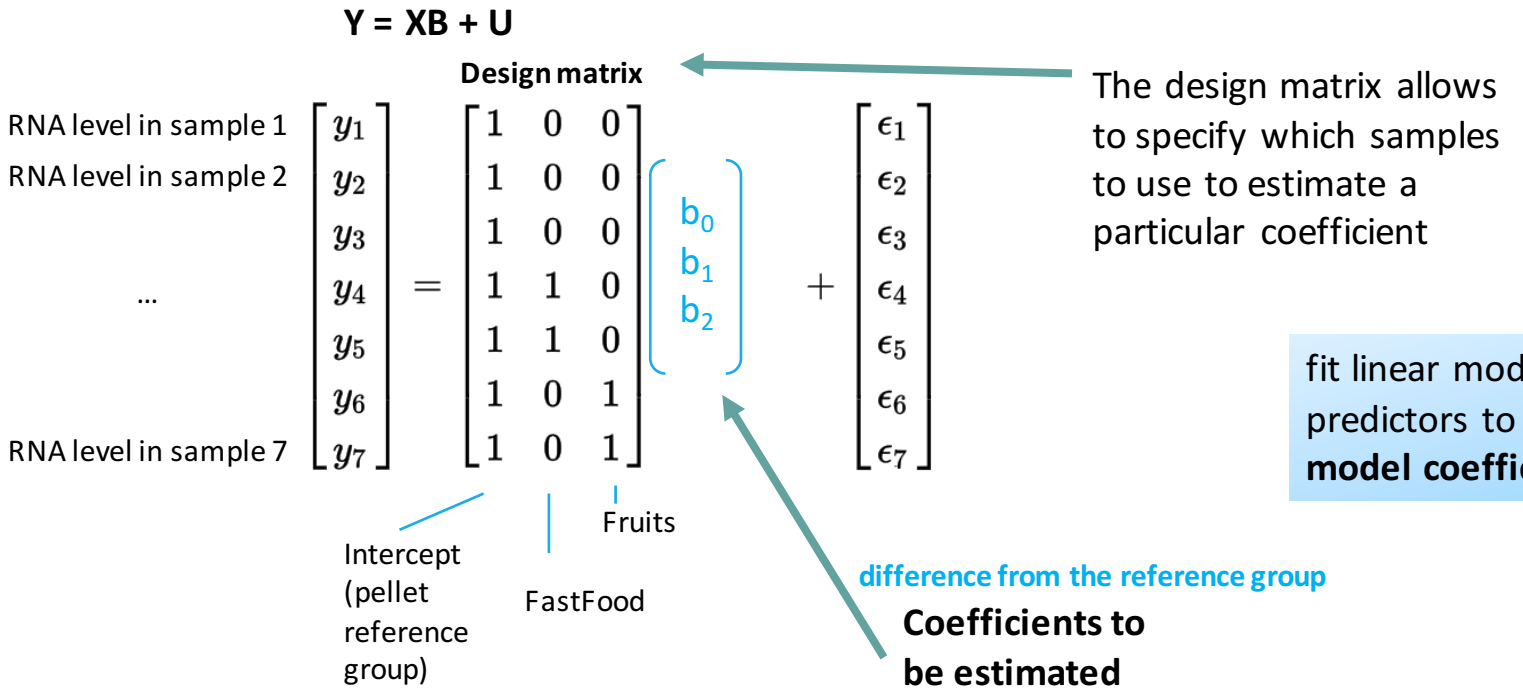
New type of data is now available: single cell RNA-sequencing (scRNAseq)

The aim is to predict the proportion of different cell types from bulk RNASeq using clever approach utilizing data from single cell RNASeq

Challenge: ***Can you come up with a clever way to utilize scRNAseq to deconvolve the signal in the Alzheimer brain tissue samples?***

Available also: Differential expression analysis comparing disease to control

Application of statistics: linear model fit $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$



Summary 4

By now you should be familiar with the data types you will work with in the independent assignment:

Bulk RNA-sequencing

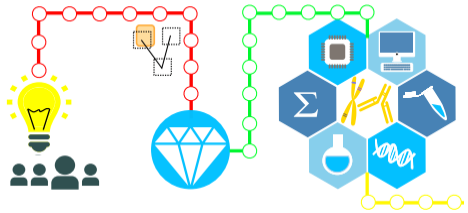
Single cell RNA-sequencing

(Differentially expressed genes table)

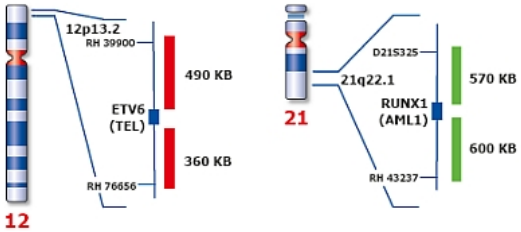
-> Next – the deconvolution problem

Heinäniemi lab

Systems Genomics



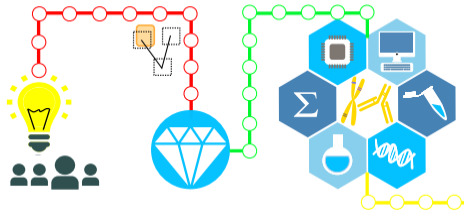
Let's test your molecular biology know-how & briefly discuss one of our current research projects - Studying childhood leukemia



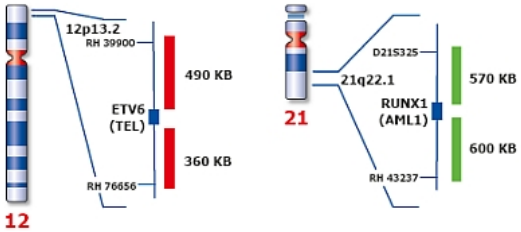
In many leukemias we find a change in DNA that joins two RNAs from different chromosomes -> **what do you think could happen ?**

Heinäniemi lab

Systems Genomics



Where could machine learning methods help?



A fusion protein is generated that **binds to DNA** and thereby regulates RNA levels of important genes

Measure where it bound and extract DNA sequence as input data

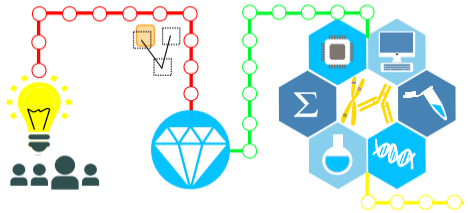
Supervised methods

Deep neural networks to discover DNA sequence motif bound by this fusion protein

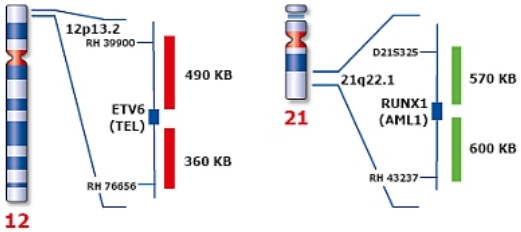
Demo by Juha in the afternoon

Heinäniemi lab

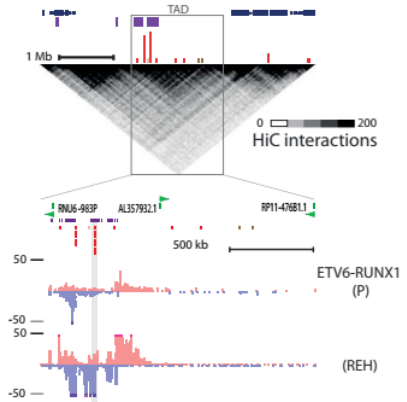
Systems Genomics



Where could machine learning methods help?



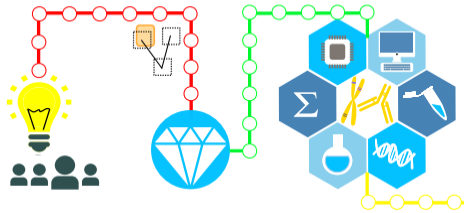
A fusion protein is generated that binds to DNA and thereby **regulates RNA levels** of important genes



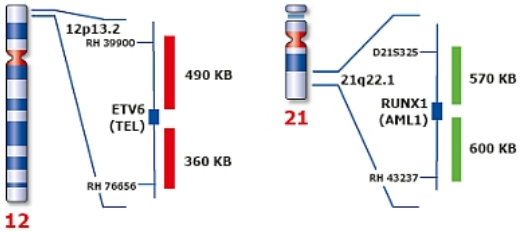
Could we use RNA profiles to distinguish between different types of leukemia and other blood cancers?

Heinäniemi lab

Systems Genomics

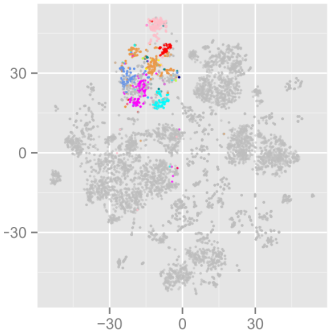


Where could machine learning methods help?



A fusion protein is generated that binds to DNA and thereby **regulates RNA levels** of important genes

10 000 patient RNA profiles



Unsupervised methods

Dimensionality reduction