Computational Analysis of Biomedical High-Throughput Data Sets

Yvonne Lichtblau, Saskia Trescher

Who should be here

- Master Informatik, Biophysik, Diplominformatik

- Ability to read English papers

- Interest in biological questions

- Knowledge in algorithms
  - Trees, graphs, dynamic programming, complexity, …

- Basic statistics

How it will work

- Today: Presentation and choice of topics
- Meet advisor by 22.11.16 to discuss topic and papers
- Send flash-presentation to your advisor by 06.12.16
- Present topic in 5min flash-presentation 13.12.16
- Meet your advisor by 24.1.17 to discuss slides
- Present your topic (30min) at the Blockseminar (07.02.2017)
- Write seminar thesis (15 pages, english) by 31.3.2017

Agenda

- **Introduction**
- Topics and assignment
- Hints on presenting your topic and writing your thesis

Cell – Chromosomes - Genes

Chromosomes:

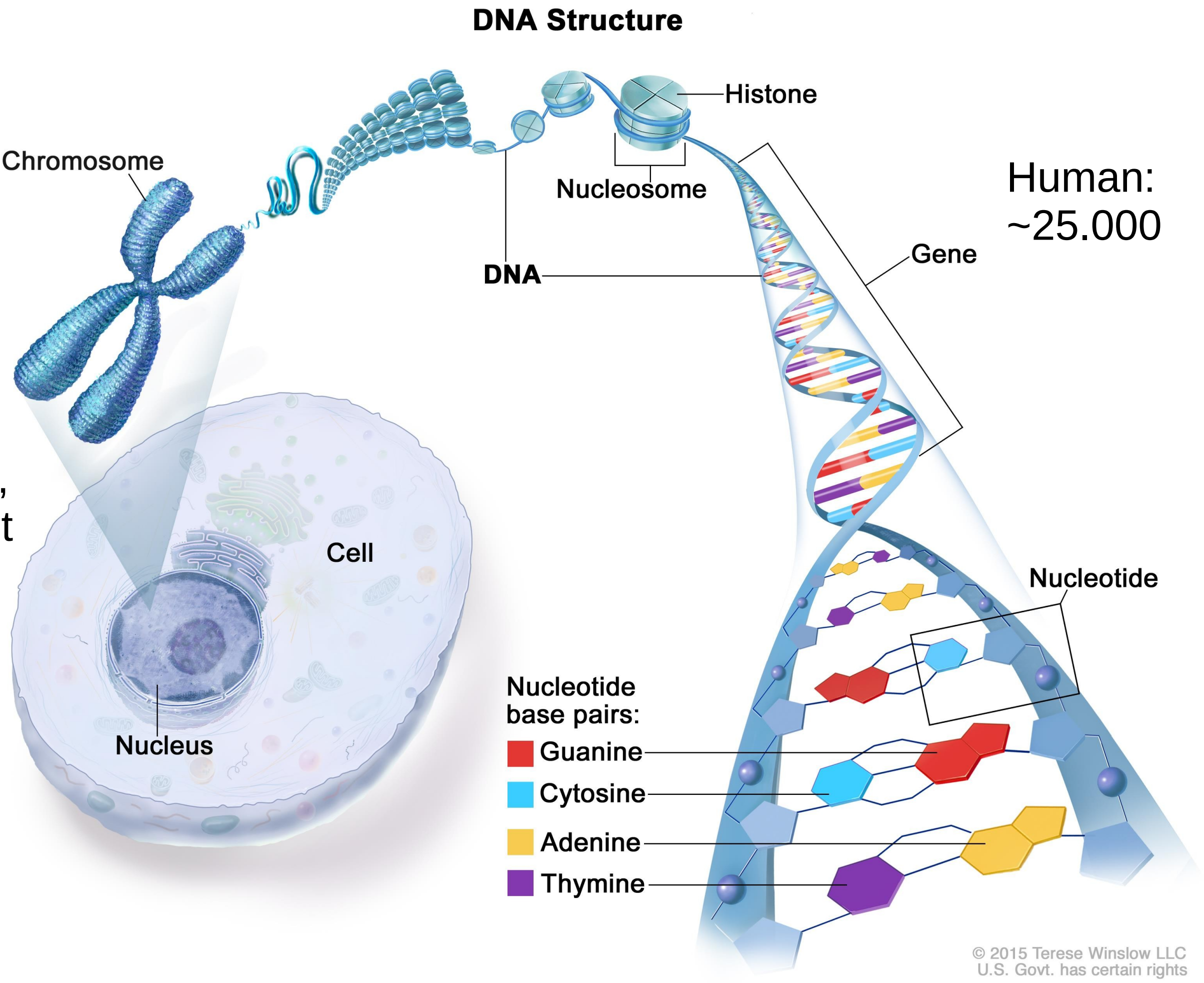
- Contain genetic information
- 23 pairs
- Made up of DNA (Desoxyribonucleic Acid)
- Composed of four nucleotides (A,C,G,T)
- DNA: double helix of complementary base pairs (bp)

Genes:

- „any discrete locus of heritable, genomic sequence which affect an organism's traits by being expressed as a functional product or by regulation of gene expression“

Genome:

- All genetic material (genes, non-coding DNA, mitochondria)
- Size: 3.2 Gbp



Human: ~25.000

© 2015 Terese Winslow LLC
U.S. Govt. has certain rights

Central Dogma of Molecular Biology

(a) Transcription

Genes on the DNA are read and transcribed to RNA (mRNA, miRNA, tRNA, …)

gene expression

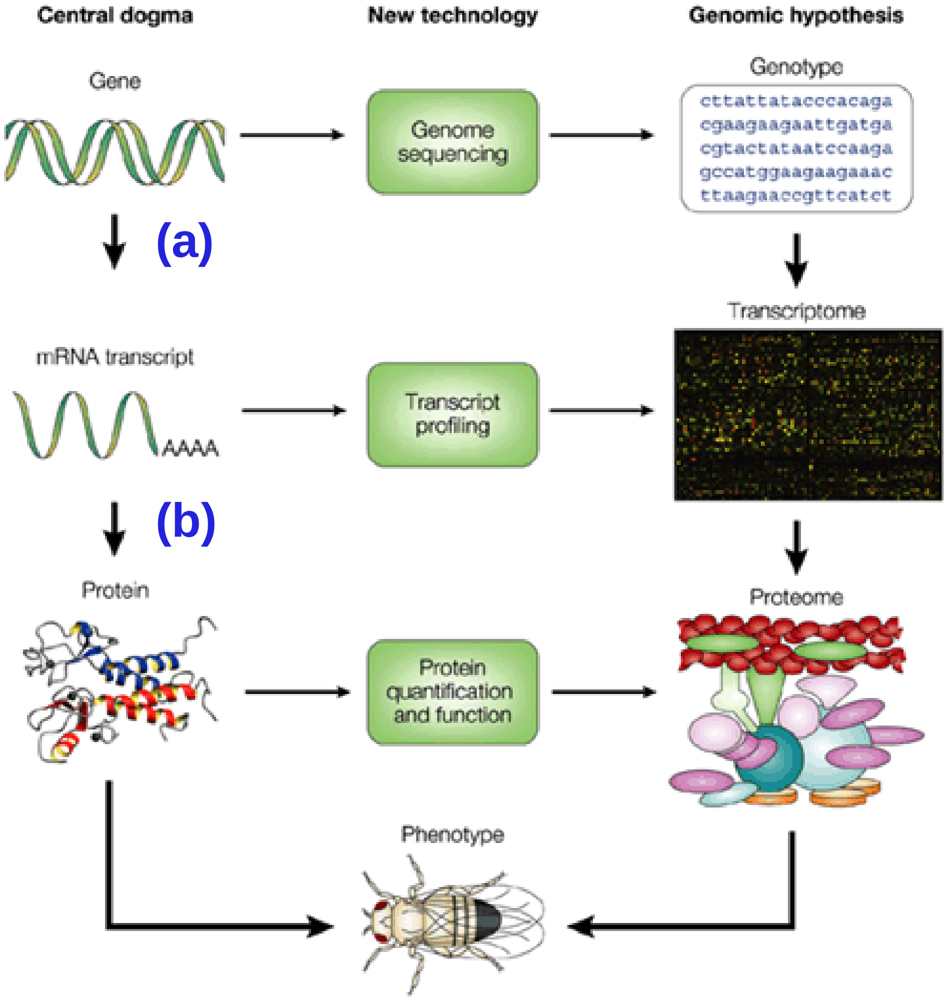
mRNA: messenger RNA,
single-stranded RNA
Transcript of a gene

(b) Translation

mRNA sequences are translated to proteins
Only 2% of the genome are protein-coding (exome)

Proteins have many functions:

- Enzymes
- Regulation
- Cell signaling and ligand binding (antibodies, receptors, …)
- Structural proteins (hair, nails, motor proteins, …)



Nature Reviews | Genetics

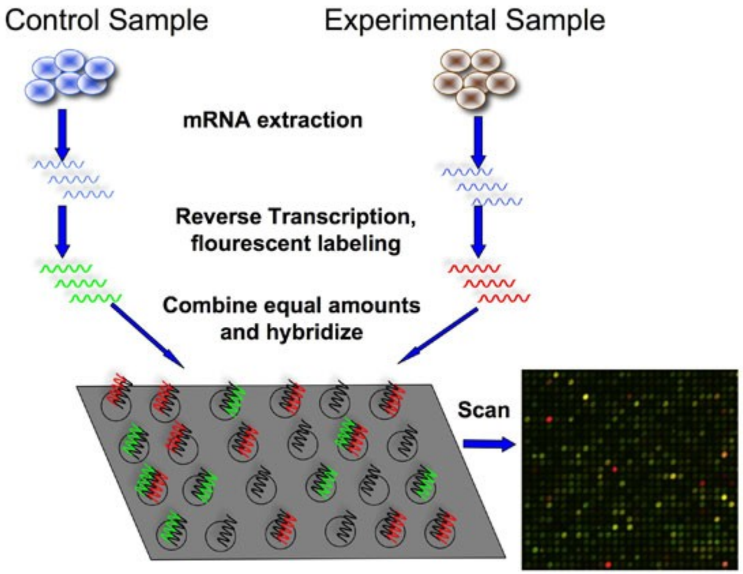
Transcription - Gene Expression

- Different techniques allow to count the number of transcripts to determine the activity of genes → **gene expression**

e.g. Northern Blot, Microarray, RNA-Seq

- **Every cell of an organism contains (nearly) the same set of genes but different cells show different patterns of gene expression!**

- Analysis and comparison of transcriptome of different types of cells:
  - What constitutes a specific cell type?
  - How works a specific type of cell?
  - How does changes in gene activity contribute/reflect to disease?


Control Sample
Experimental Sample
mRNA extraction
Reverse Transcription, flourescent labeling
Combine equal amounts and hybridize
Scan

Microarray experiment
(old but established technology)

Translation - Genetic Code

mRNA → protein

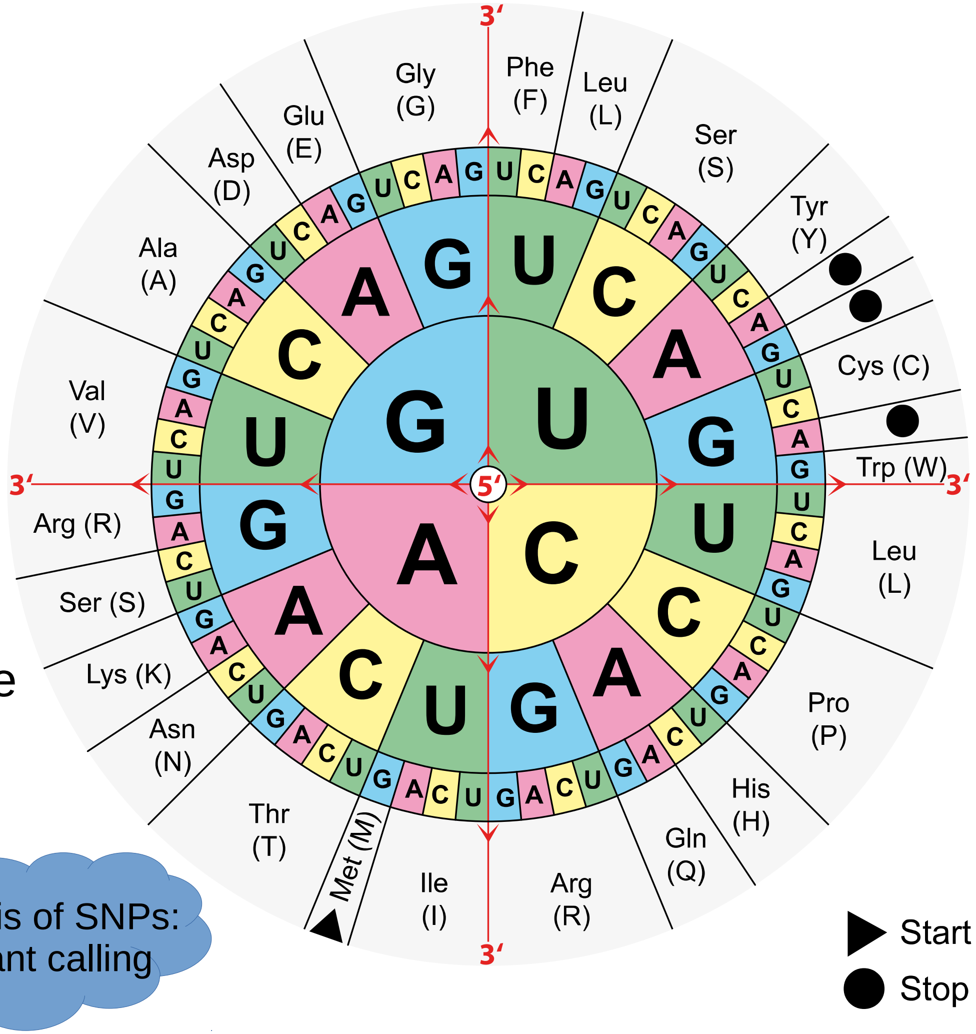
Nucleotide triplets (codons) code for one of 20 amino acids (aa, bricks of proteins)

Codon degeneracy

SNP (single nulceotide polymorphism): Variation in a single nucleotide within >1% of a population

→ may change aa and thus 3D structure and function of a protein

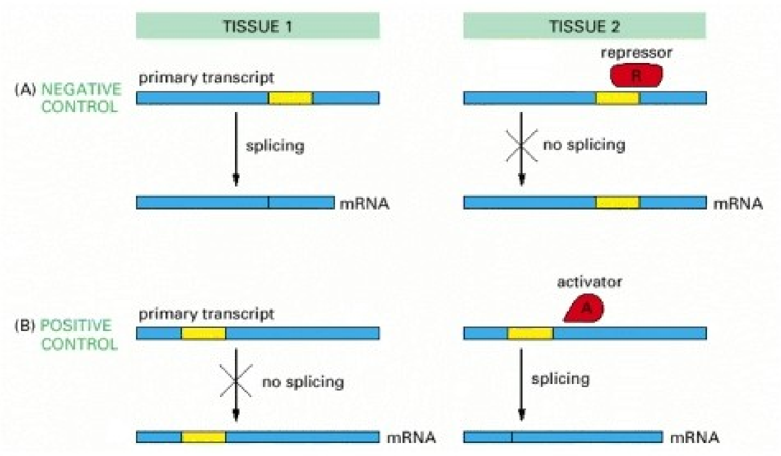
Examples: sickle cell anemia, cystic fibrosis



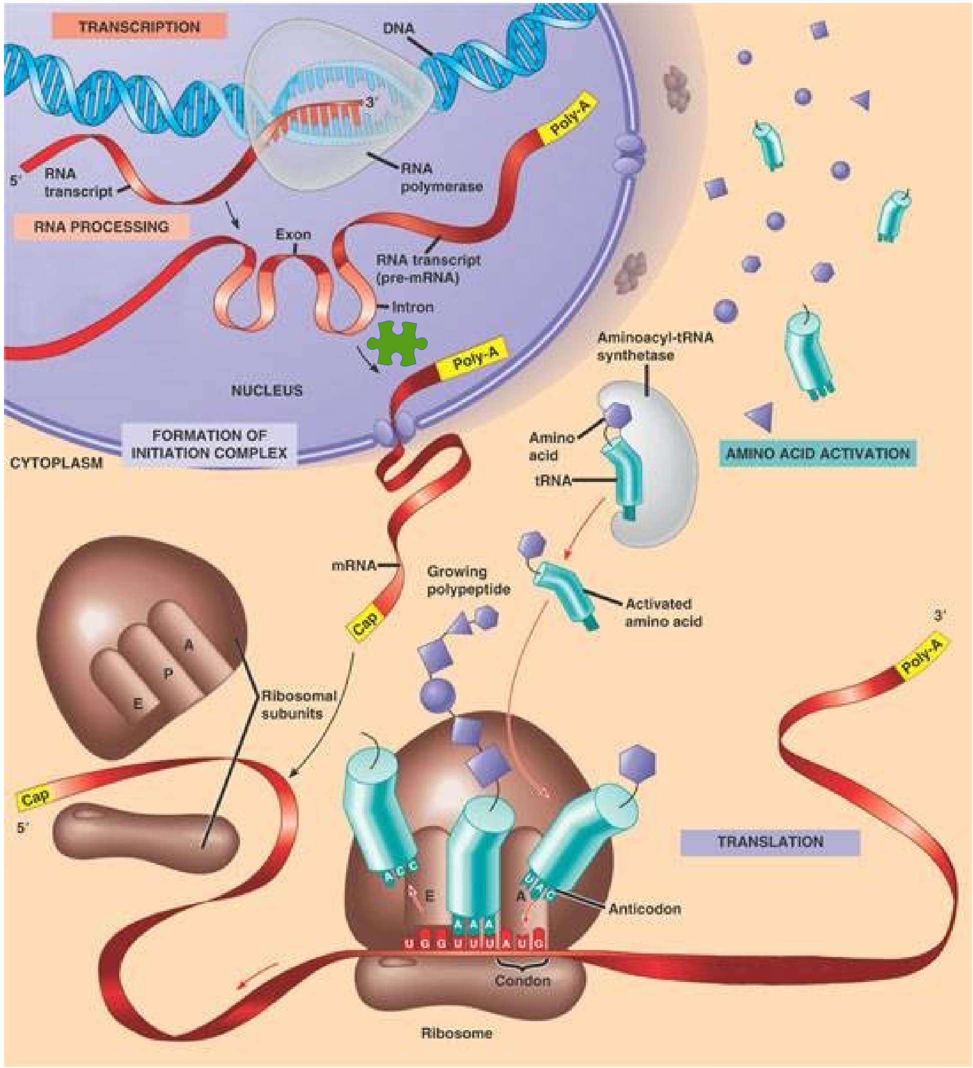
Analysis of SNPs:
Variant calling

Differential RNA Splicing

can produce different forms of a protein from the same gene:
Introns are removed from mRNA
Exons may be included or excluded from mRNA



Differential splicing from single gene → multiple proteins



http://www.proteinsynthesis.org/wp-content/uploads/2015/09/protein-synthesis-in-cytoplasm.jpg

Gene – Disease - Associations

Changes in DNA (SNP, insertion, deletion, …) → health problems

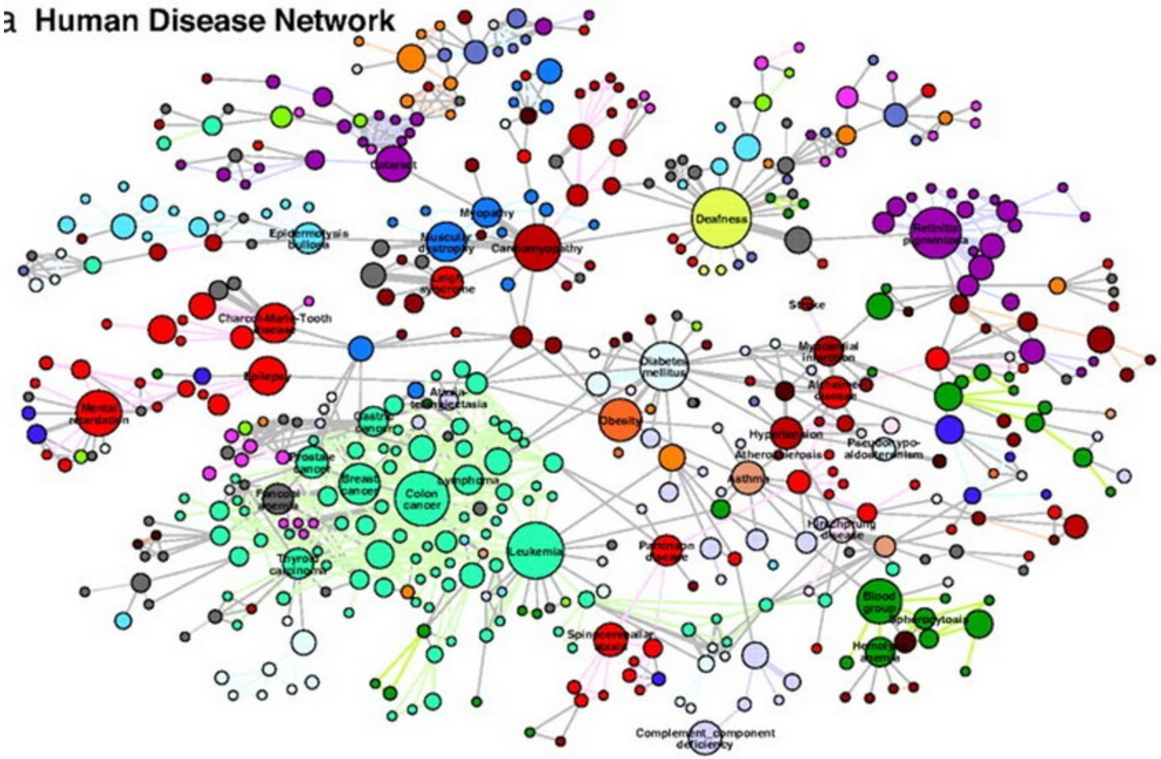
- Wrong amount of a certain protein is produced
- Proteins are misfolded → function is lost

<u>Example cancer:</u>

- changes cause cells to survive/grow out of control
- Three types of genes:
  - Proto-oncogenes (growth)
  - Tumorsuppressor genes (control cell growth)
  - DNA repair genes (DNA repair)

Large scale experimental analysis to find disease genes!

(Genomics, proteomics, transcriptomics, epigenomics)

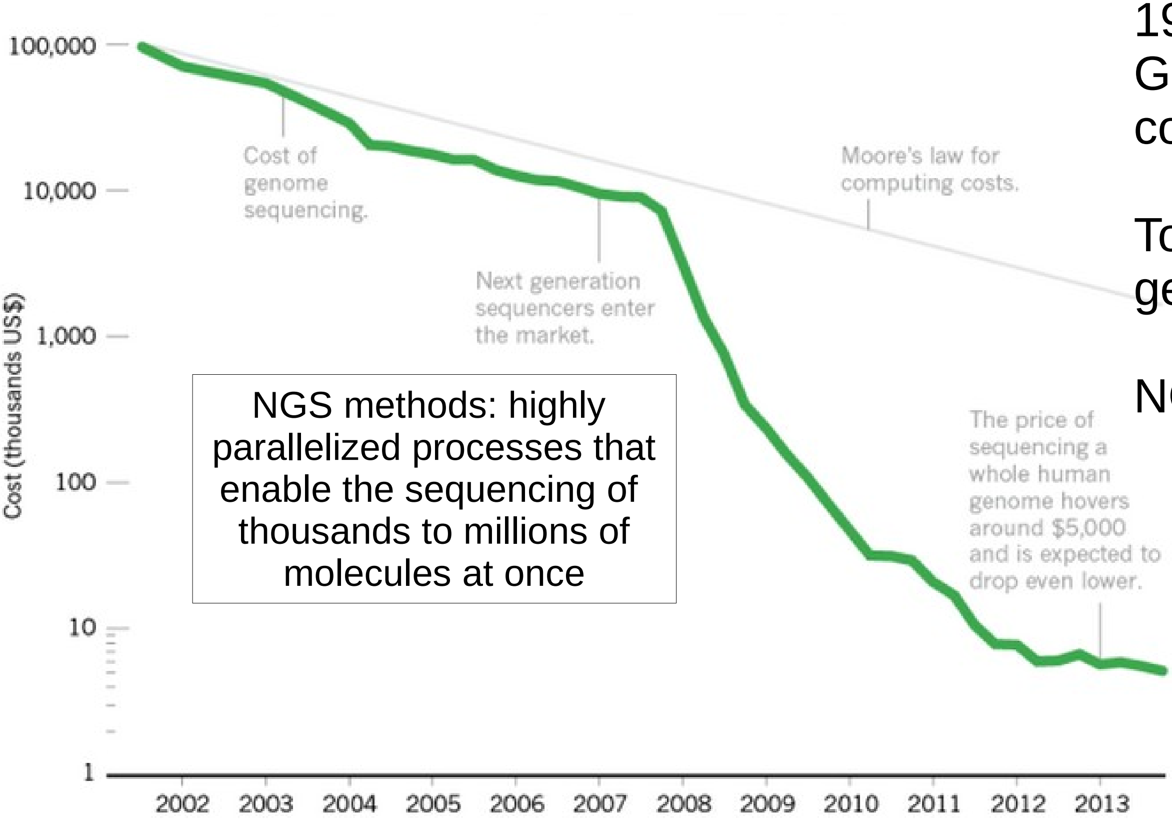


„The human disease network“, PNAS, 2007

data: OMIM database, 1,284 diseases and 1,777 disease genes

Sequencing Costs Decrease

Improvements in the area of high-throghput sequencing



NGS methods: highly parallelized processes that enable the sequencing of thousands to millions of molecules at once

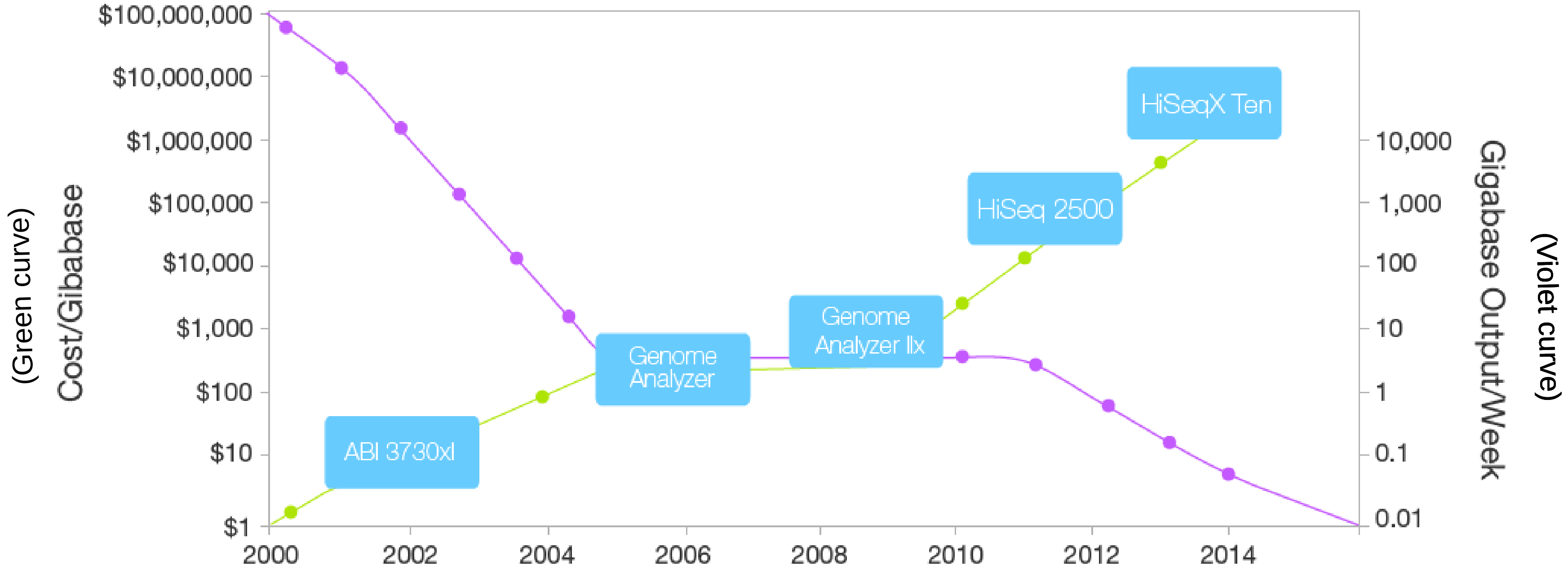
http://www.nature.com/news/technology-the-1-000-genome-1.14901

1990 – 2001: Human Genome Project (HGP), costs $3 billions

Today: $1000 per human genome, one day

NGS has outpaced Moore` law

The Sequencing Explosion

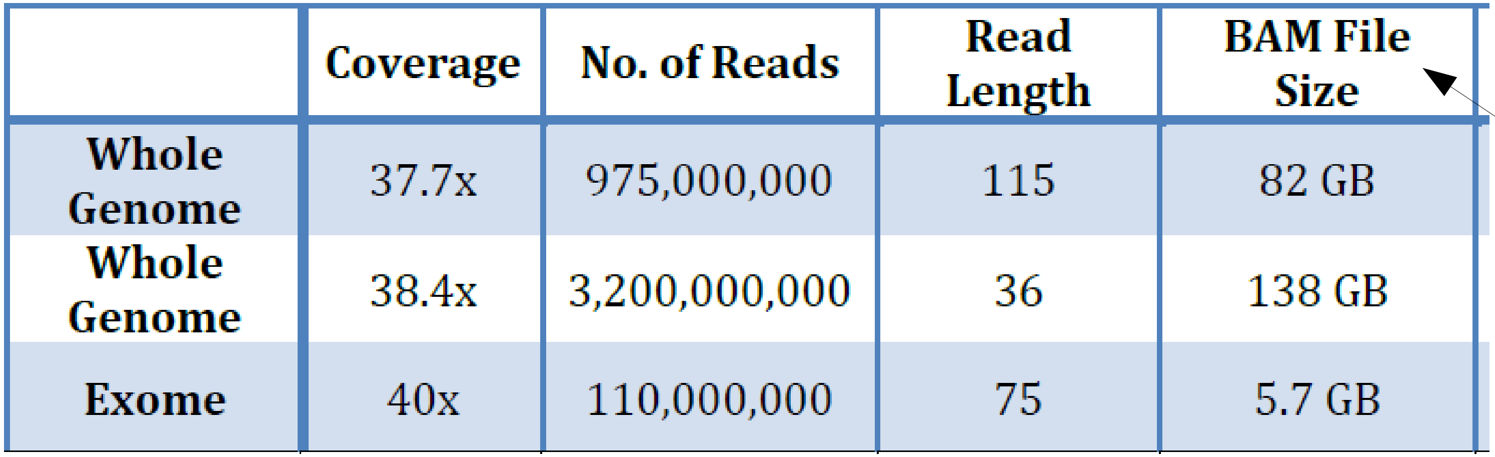


http://www.illumina.com

<u>Data output</u>
2005: GenomeAnalyzer 84kb 1 Gb per run
2014: 1.8 terabases per run (1000x increase)

➡ exponential growth of in the amount of (publicly available) sequence data, quality of data per sample is going up

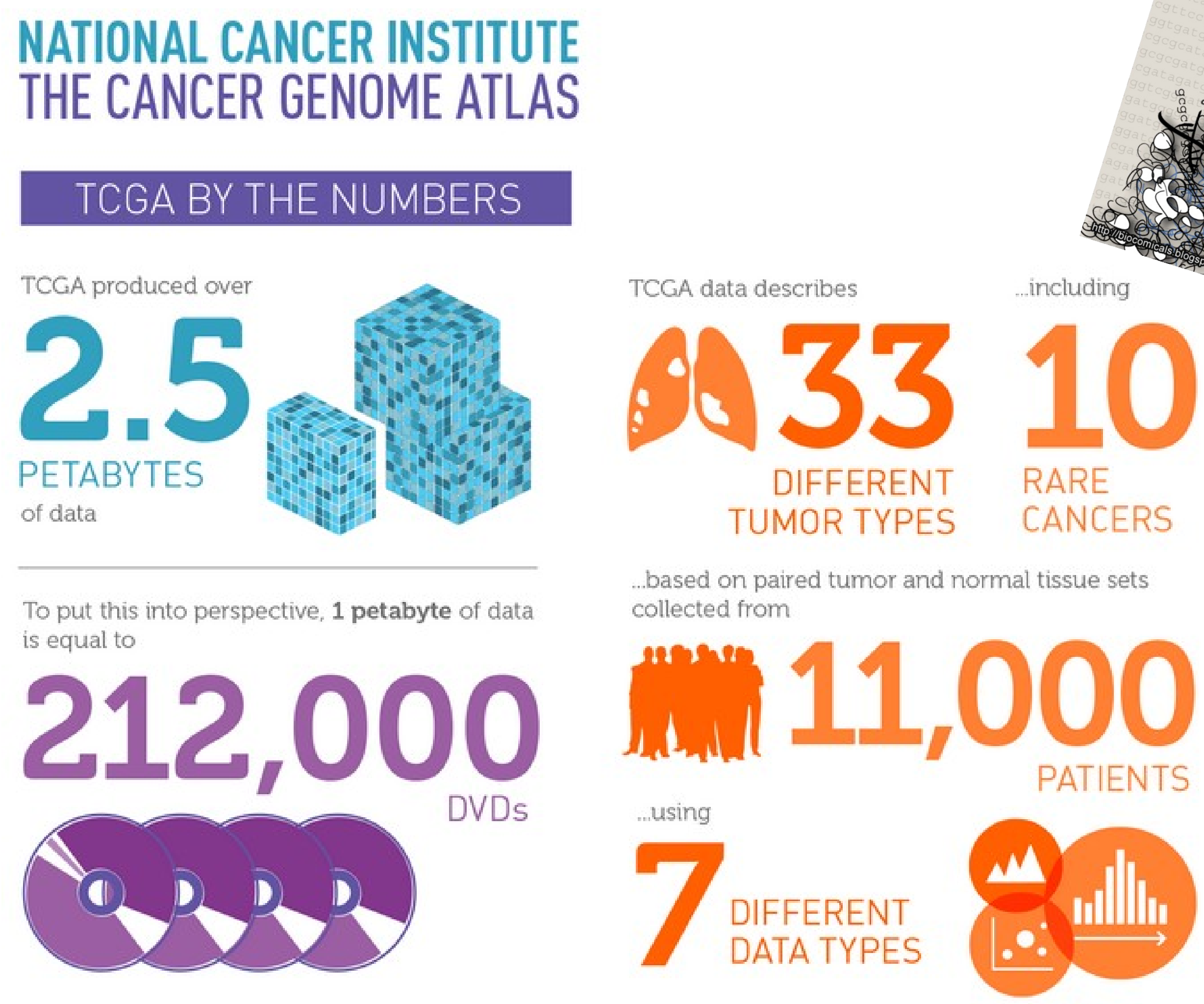
Example Data Sets



	Coverage	No. of Reads	Read Length	BAM File Size
Whole Genome	37.7x	975,000,000	115	82 GB
Whole Genome	38.4x	3,200,000,000	36	138 GB
Exome	40x	110,000,000	75	5.7 GB

Compressed file size

Example TCGA (1)



NATIONAL CANCER INSTITUTE
THE CANCER GENOME ATLAS

TCGA BY THE NUMBERS

TCGA produced over **2.5** PETABYTES of data

To put this into perspective, **1 petabyte** of data is equal to **212,000** DVDs

TCGA data describes **33** DIFFERENT TUMOR TYPES

...including **10** RARE CANCERS

...based on paired tumor and normal tissue sets collected from **11,000** PATIENTS

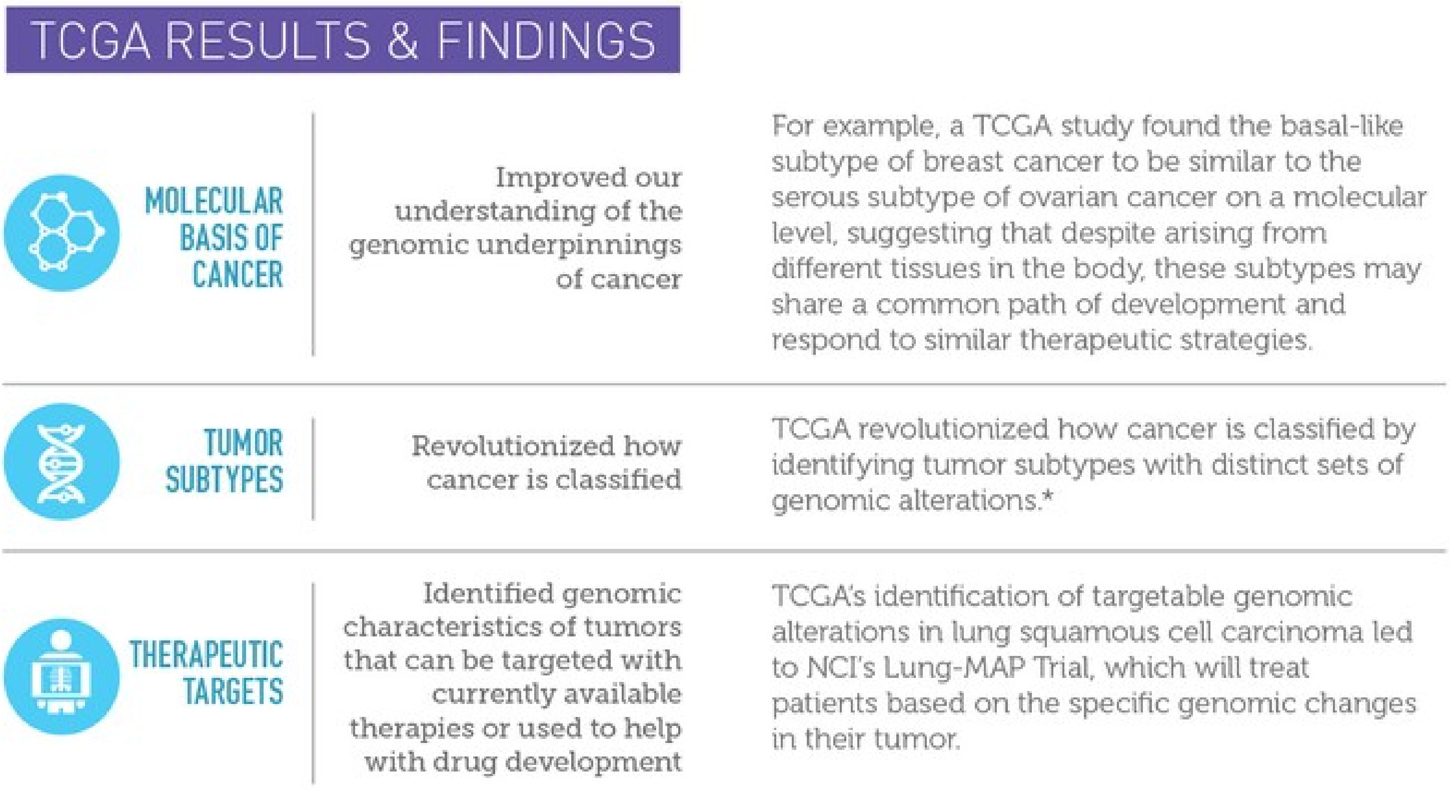
...using **7** DIFFERENT DATA TYPES

Drowned in next generation sequencing data
HELP!
http://biocomicals.blogspot.com

- collaboration between the NCI NHGRI
- >1000 studies of cancer by independent researchers
- improving cancer prevention, early detection and treatment

http://cancergenome.nih.gov

Example TCGA (2)

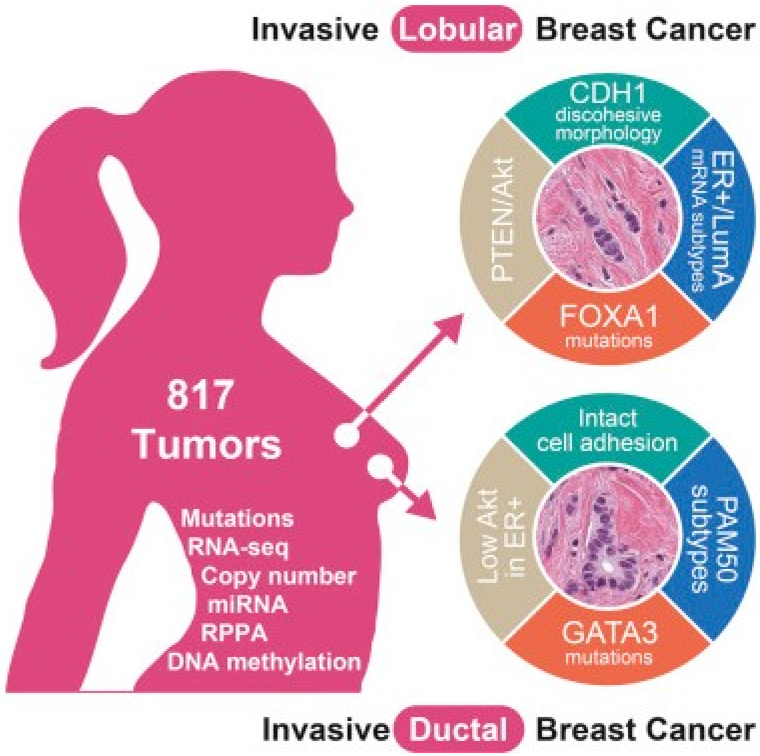


TCGA RESULTS & FINDINGS

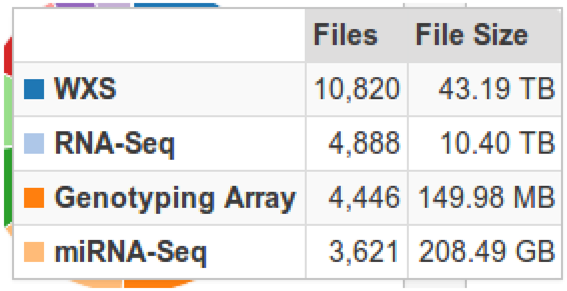
MOLECULAR BASIS OF CANCER	Improved our understanding of the genomic underpinnings of cancer	For example, a TCGA study found the basal-like subtype of breast cancer to be similar to the serous subtype of ovarian cancer on a molecular level, suggesting that despite arising from different tissues in the body, these subtypes may share a common path of development and respond to similar therapeutic strategies.
TUMOR SUBTYPES	Revolutionized how cancer is classified	TCGA revolutionized how cancer is classified by identifying tumor subtypes with distinct sets of genomic alterations.*
THERAPEUTIC TARGETS	Identified genomic characteristics of tumors that can be targeted with currently available therapies or used to help with drug development	TCGA's identification of targetable genomic alterations in lung squamous cell carcinoma led to NCI's Lung-MAP Trial, which will treat patients based on the specific genomic changes in their tumor.

http://cancergenome.nih.gov/researchhighlights

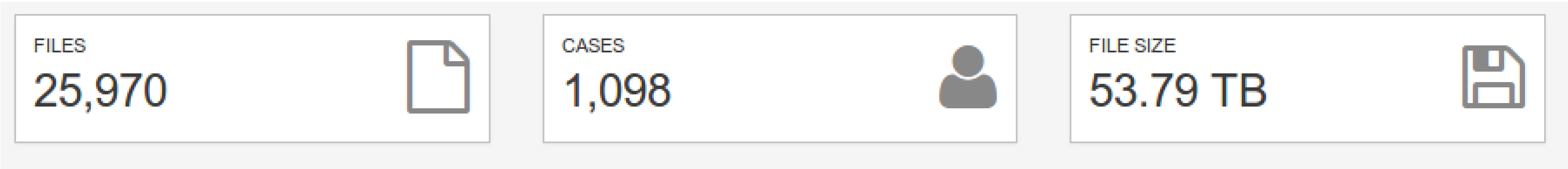
Example TCGA (3)



- Invasive lobular carcinoma (ILC) is a clinically and molecularly distinct disease
- ILCs show CDH1 and PTEN loss, AKT activation, and mutations in TBX3 and FOXA1



	Files	File Size
🟦 WXS	10,820	43.19 TB
🟦 RNA-Seq	4,888	10.40 TB
🟧 Genotyping Array	4,446	149.98 MB
🟧 miRNA-Seq	3,621	208.49 GB



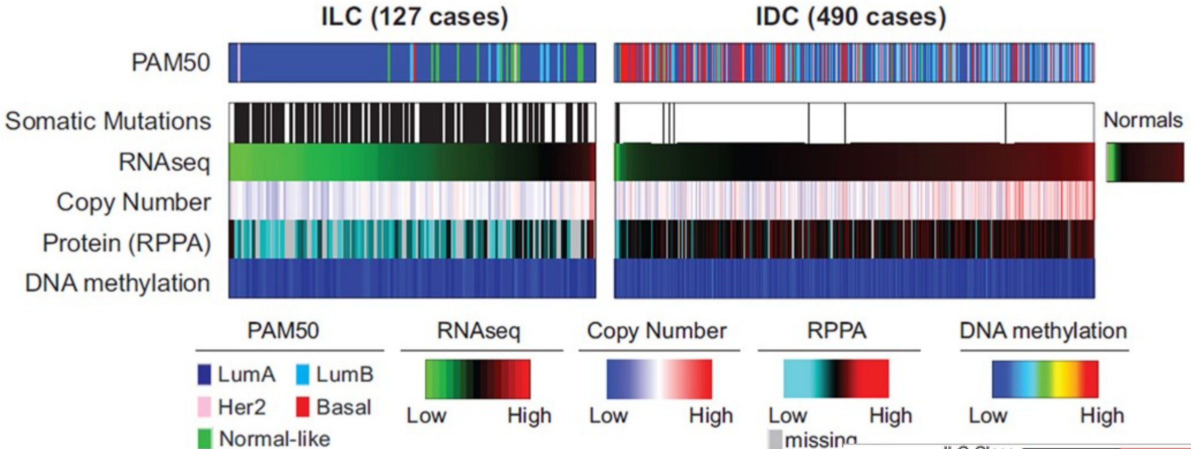
FILES
25,970

CASES
1,098

FILE SIZE
53.79 TB

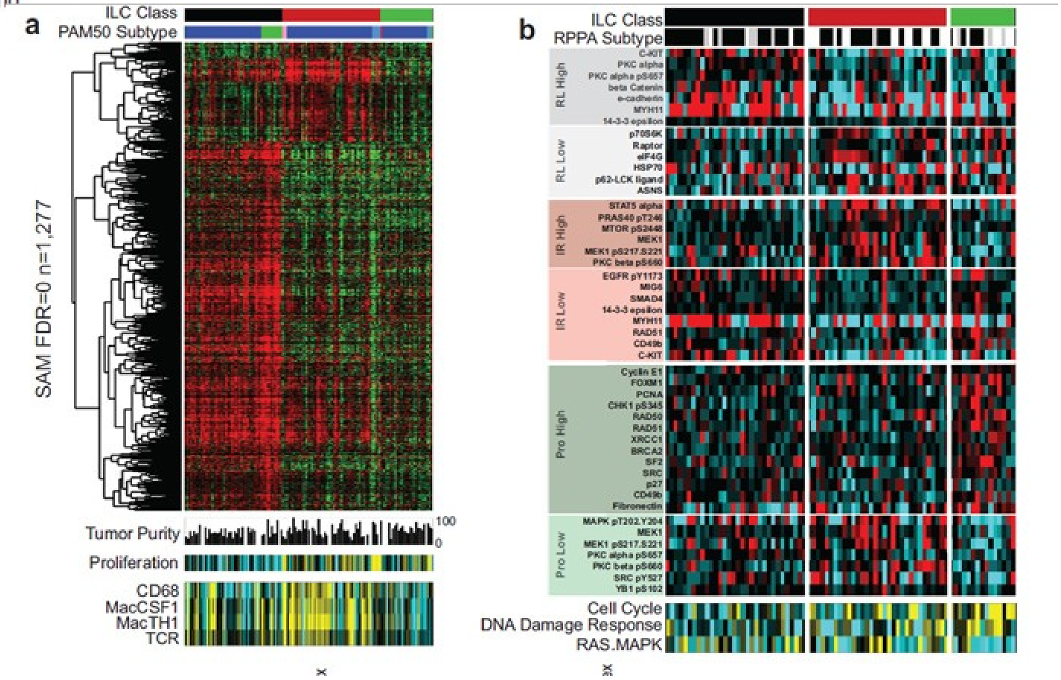
„Comprehensive Molecular Portraits of Invasive Lobular Breast Cancer“, Ciriello G. Et al., Cell 2015

Example TCGA (4)



- Proliferation and immune-related gene expression signatures define 3 ILC subtypes

- Genetic features classify mixed tumors into lobular-like and ductal-like subgroups



Agenda

- Introduction
- **Topics and assignment**
- Hints on presenting your topic and writing your thesis

Introducing Literature

- Berger et al.: **Computational solutions for omics data**, Nature Review Genetics, 2013
- Vogelstein et al.: **Cancer Genome Landscapes**, Science, 2013
- Biological Background:
  Alberts B, Johnson A, Lewis J, et al.: **Molecular Biology of the Cell**. 4th edition.
  https://www.ncbi.nlm.nih.gov/books/NBK21054/?term=molecular%20biology%20of%20the%20cell%20alberts
- More literature will be provided individually

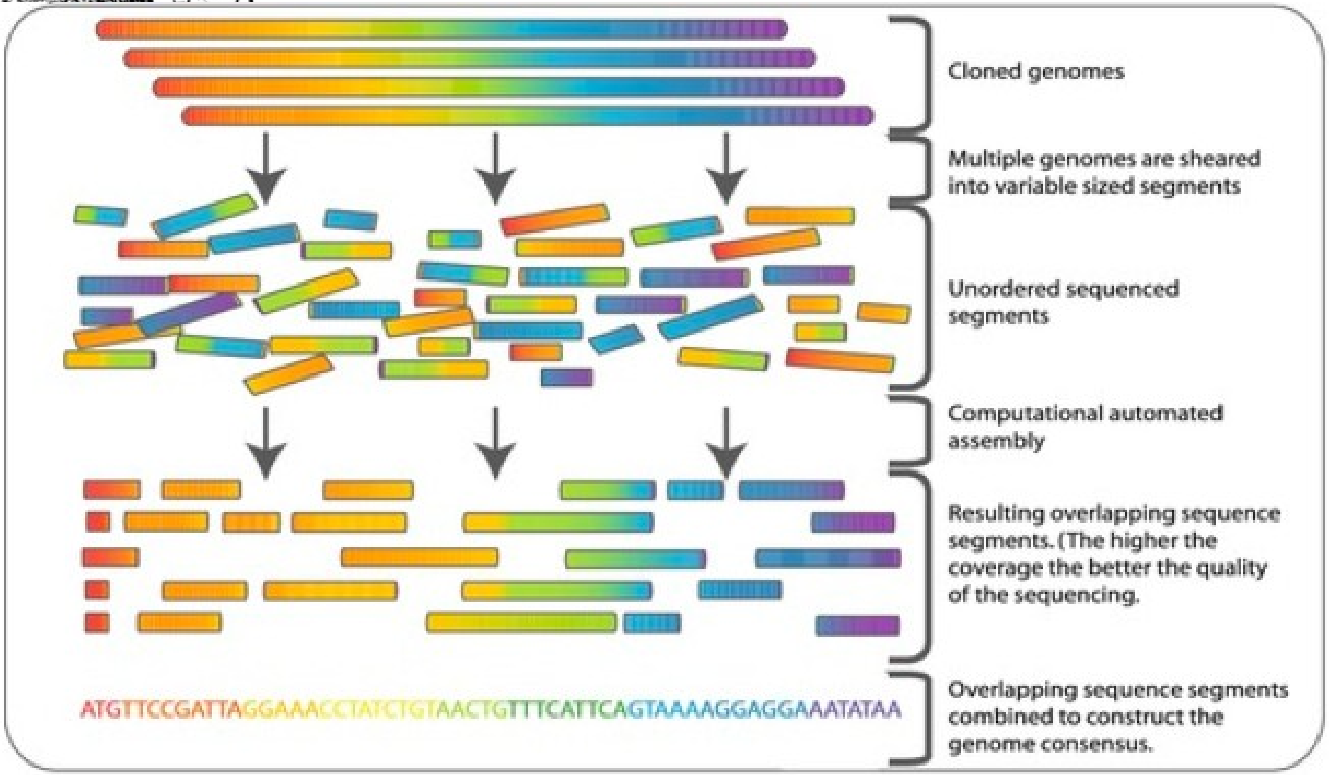
Overview Topics

Nr	Topic	Supervisor
1	Genome Assembly	Yvonne
2	Read Mapping in RNA-Seq	Yvonne
3	Compression Methods Genomics	Yvonne/Ulf
4	Variant Calling	Yvonne
5	Sequence Similarity	Yvonne
6	Co-Expression Networks	Saskia/Yvonne
7	Gene Regulatory Network Reconstruction/Regulatory activity	Saskia
8	Differential Expression Analysis of RNA-Seq Data	Saskia/Yvonne

Algorithmic efficiency to handle large datasets

Data analysis

Topic 1: Genome Assembly



Cloned genomes

Multiple genomes are sheared into variable sized segments

Unordered sequenced segments

Computational automated assembly

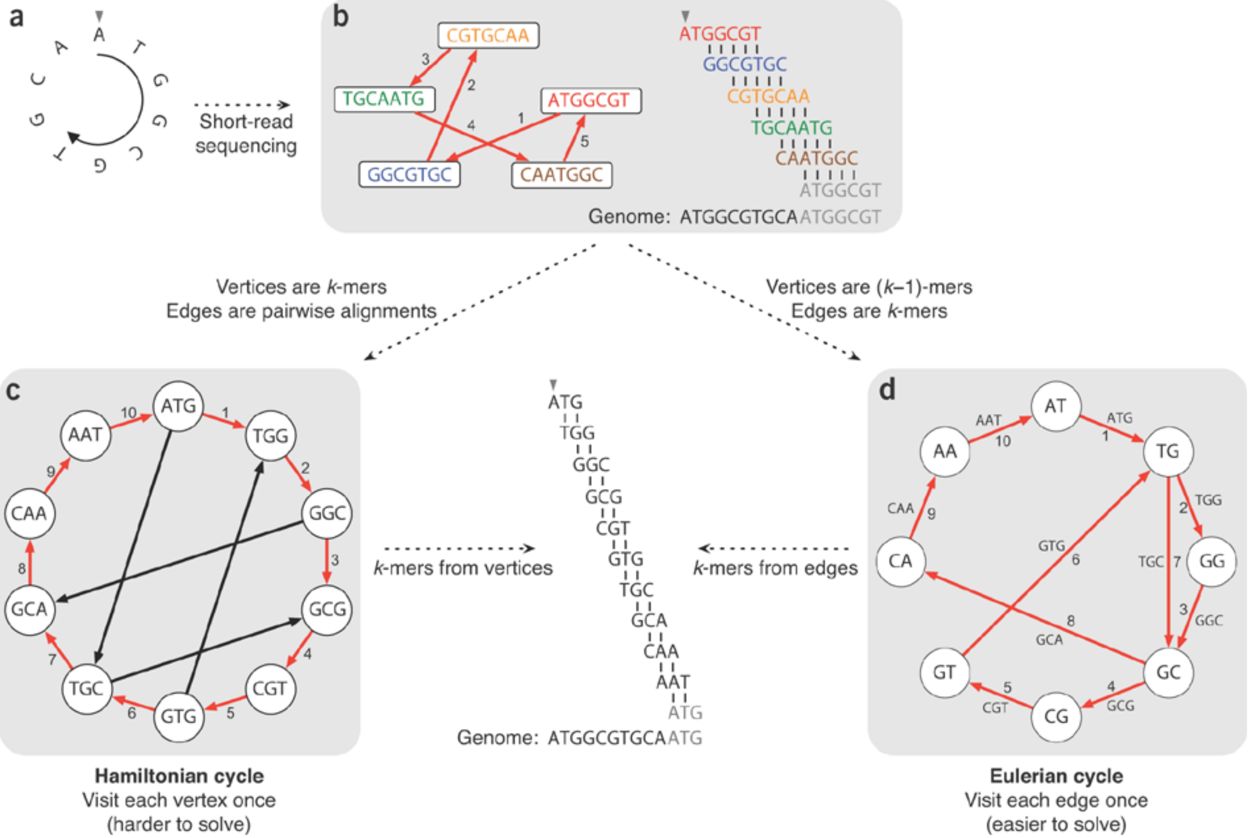
Resulting overlapping sequence segments. (The higher the coverage the better the quality of the sequencing.

ATGTTCCGATTAGGAAACCTATCTGTAACTGTTTCATTCAGTAAAAGGAGGAAATATAA

Overlapping sequence segments combined to construct the genome consensus.

Topic 1: Genome Assembly

Presentation of two different approaches:
(c) Overlap-Layout Consensus and (d) De Bruijn graphs



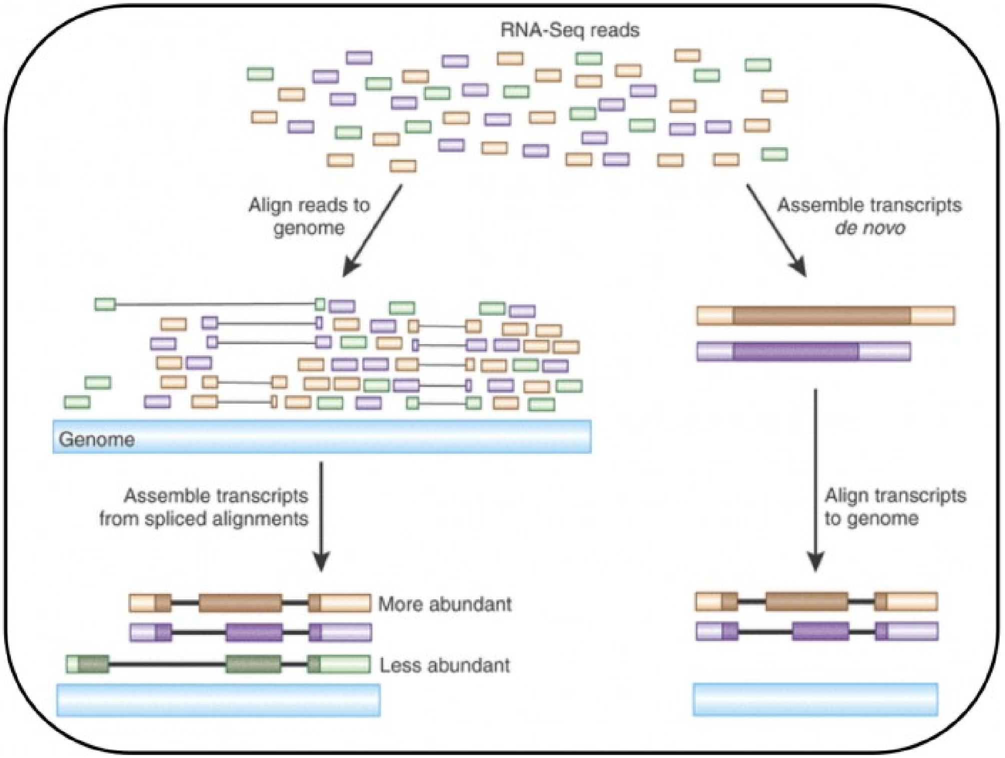
[1] "ARACHNE: A Whole-Genome Shotgun Assembler", Genome Research, 2002
[2] "How to apply de Bruijn graphs to genome assembly", Nature Biotechnology, 2011
[3] "Velvet: Algorithms for de novo short read assembly using de Bruijn graphs", Genome Research, 2008
[4] "Comparison of the two major classes of assembly algorithms: overlap-layout-consensus and de-bruijn-graph", Briefings in FG, 2011

Topic 2: Read Mapping in RNA Sequencing

- Presentation of two approaches: Bowtie and TopHat (CuffLinks)
- Memory usage / runtime
- Transcript assembly/workflow from biological sample to read counts

Challenges:

- Need to map millions of short reads to a genome
- NOT exact matching: sequencing erros, biological variants (substitutions, insertions, deletions, splicing)
- Advantage: discovery of new genes, transcripts, alternative splice isoforms


RNA-Seq reads
Align reads to genome
Assemble transcripts *de novo*
Genome
Assemble transcripts from spliced alignments
Align transcripts to genome
More abundant
Less abundant

TopHat

Topic 2: Read Mapping in RNA Sequencing

Bowtie
Extremely fast, general purpose short read aligner

TopHat
Aligns RNA-Seq reads to the genome using Bowtie
Discovers splice sites

Cufflinks package

Cufflinks
Assembles transcripts

Cuffcompare
Compares transcript assemblies to annotation

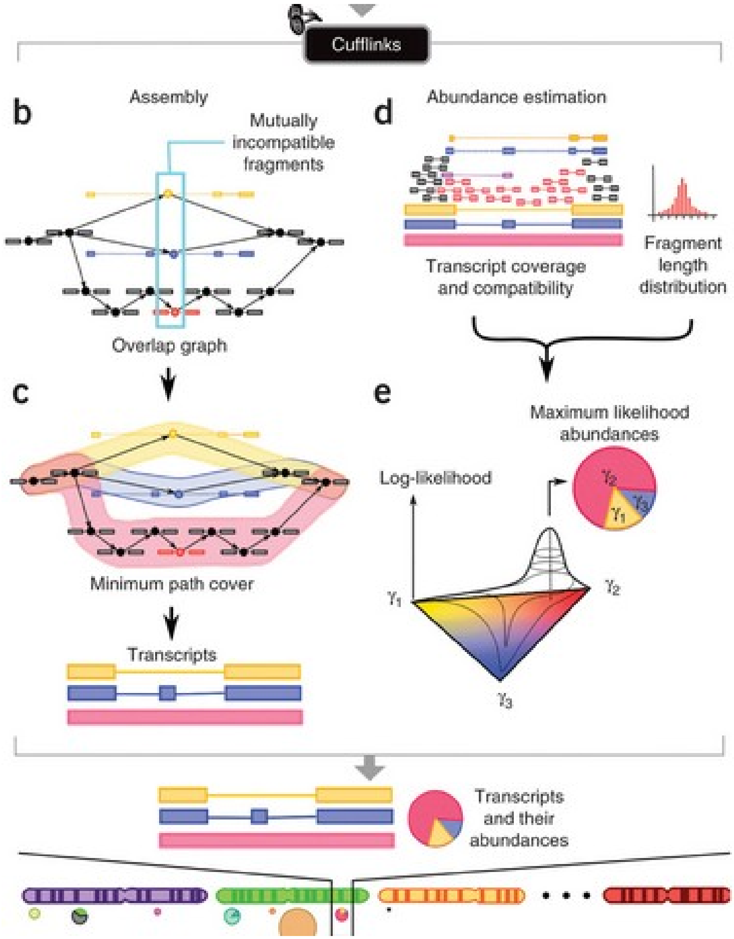
Cuffmerge
Merges two or more transcript assemblies

Cuffdiff
Finds differentially expressed genes and transcripts
Detects differential splicing and promoter use

CummeRbund
Plots abundance and differential expression results from Cuffdiff

Topic 2

Topic 8


Cufflinks
b
Assembly
Mutually incompatible fragments
Overlap graph
c
Minimum path cover
Transcripts
d
Abundance estimation
Transcript coverage and compatibility
Fragment length distribution
e
Maximum likelihood abundances
Log-likelihood
γ_1
γ_2
γ_3
Transcripts and their abundances

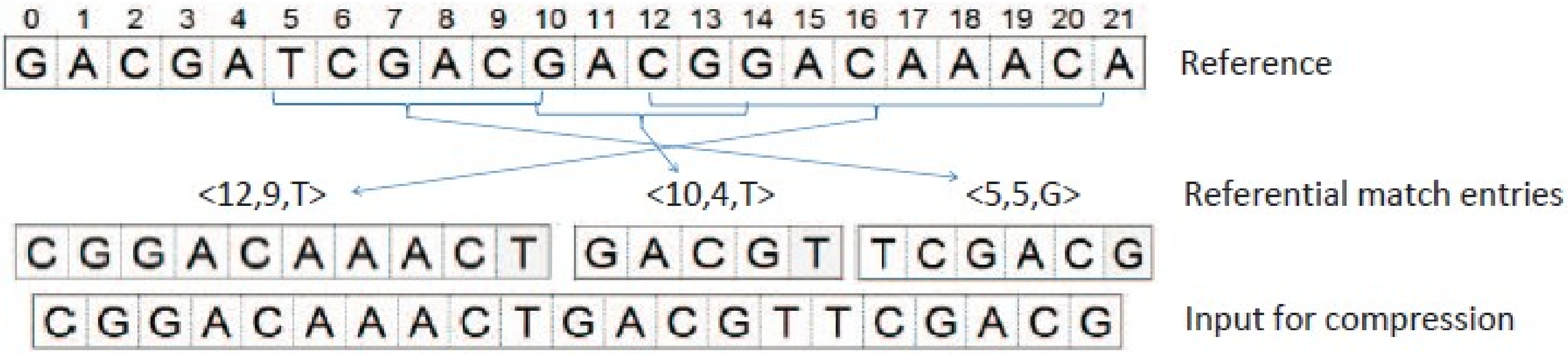
[1] "Fast gapped-read alignment with Bowtie 2", Nature Methods, 2011
[2] "TopHat: discovering splice junctions with RNA-Seq", Bioinformatics, 2009
[3] "Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks", Nature Protocols, 2014

Topic 3: Compression Methods Genomics

- Sequencing data accumulates
  → Reduction of data size for storage (and processing) necessary

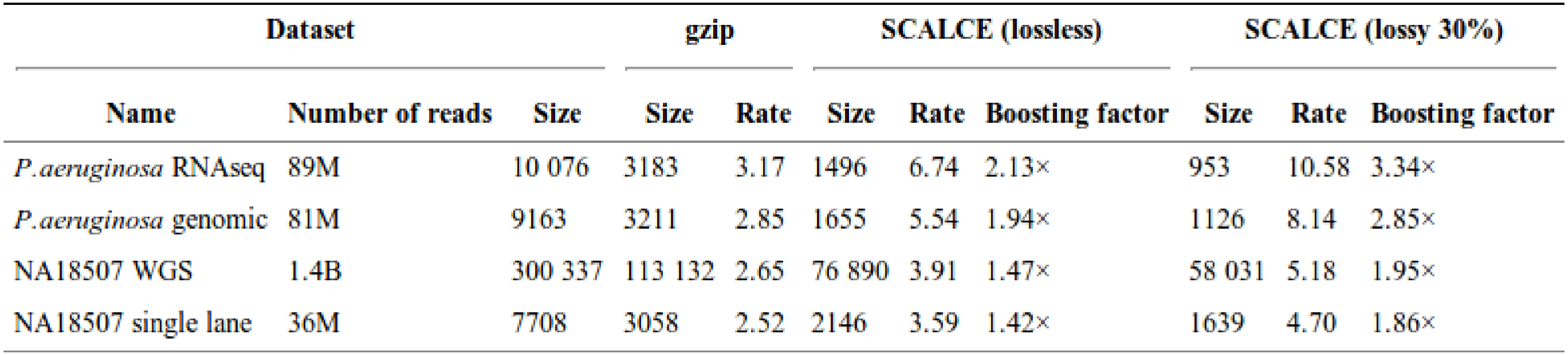
- Present reference-based and non-reference-based compression methods to achive high compression rate and speed

- Reference-based compression methods
  - Alignment of reads to a reference genome
  - Only differences are stored
  - Example: FRESCO (WBI)



Topic 3: Compression Methods Genomics

- Non-reference based methods
  - Rely on string compression algorithms
  - Exploits repetitive DNA segments
  - Use text compression algorithms (gzip, BWT, ..)
  - Example: SCALCE
    - Uses a locally parsing technique: combinatorial pattern matching technique that aims to identify ‘building blocks’
    - Reorganizes reads



Dataset			gzip		SCALCE (lossless)			SCALCE (lossy 30%)		
Name	Number of reads	Size	Size	Rate	Size	Rate	Boosting factor	Size	Rate	Boosting factor
P.aeruginosa RNAseq	89M	10 076	3183	3.17	1496	6.74	2.13×	953	10.58	3.34×
P.aeruginosa genomic	81M	9163	3211	2.85	1655	5.54	1.94×	1126	8.14	2.85×
NA18507 WGS	1.4B	300 337	113 132	2.65	76 890	3.91	1.47×	58 031	5.18	1.95×
NA18507 single lane	36M	7708	3058	2.52	2146	3.59	1.42×	1639	4.70	1.86×

[1] "Efficient storage of high throughput DNA sequencing data using reference-based compression", Genome Research, 2011

[2] "SCALCE: boosting sequence compression algorithms using locally consistent encoding", Bioinformatics, 2012

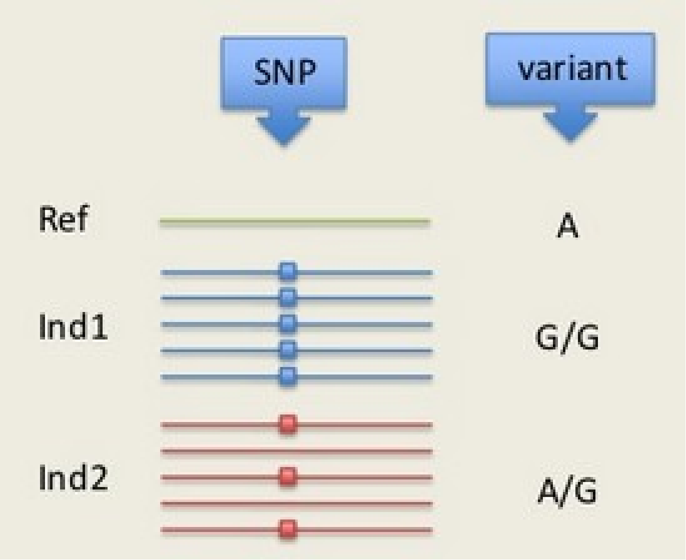
[3] "FRESCO: Referential Compression of Highly Similar Sequences", IEEE Transactions on Computational Biology and Bioinformatics, 2013

Topic 4: Variant Calling

Challenges, explain different methods, example project

- 1. Step: Read Mapping
- 2. Step: Different methods
  - Allele counting
  - Probabilistic models
    - To quantify statistical uncertainty
    - Assign priors e.g. by taking the observed allele frequency of multiple samples into account
  - Incorporating linkage disequilibrium
    - Specifically helpful for low coverage and common variants

- Examples (1000 / 100000 genomes project, african genome variation project, 23andMe, ...)



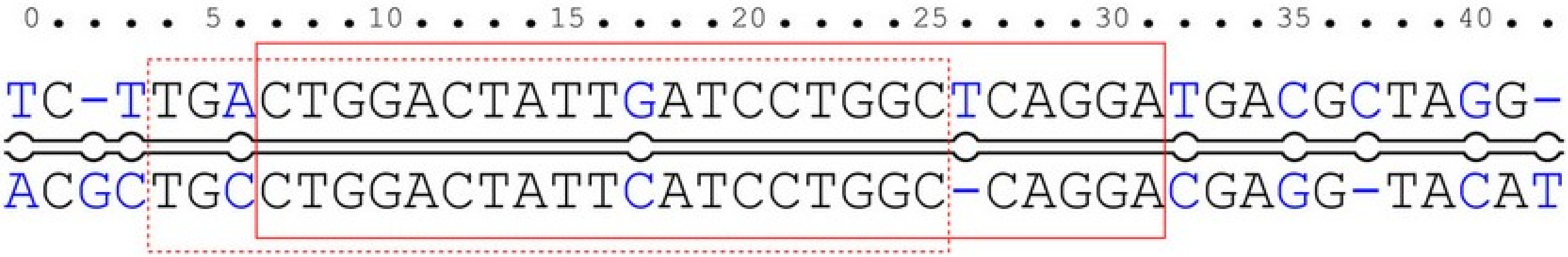
http://www.slideshare.net/allPowerde/variant-snpsindels-calling-in-dna-sequences-part-2

[1] „Best practices for evaluating single nucleotide variant calling methods for microbial genomics“, frontiers in genetics, 2015
[2] "Mapping short DNA sequencing reads and calling variants using mapping quality scores", Genome Research, 2008
[3] "Genotype and SNP calling from next-generation sequencing data", Nature Reviews Genetics, 2011
[4] "A map of human genome variation from population-scale sequencing", Nature, 2010

Topic 5: Sequence Similarity

- Requirements: BLAST
- Motivation: Identify homologous regions / Find functional similar sequences

- Presentation of two approaches
    - STELLAR (fast and exact local alignments)
        - Calculates only significant local alignments with high scores
        - Use a maximum error rate for alignments
        - Require minimal alignment length


0
5
10
15
20
25
30
35
40
TC-TTGACTGGACTATTGATCCTGGCTCAGGATGACGCTAGG-
ACGCTGCCTGGACTATTCATCCTGGC-CAGGACGAGG-TACAT

-
    - Alignment-free sequence comparison using spaces word frequencies
        - space words: defined by patterns of 'match' and 'don't care positions'
        - Fast implementation with recursive hashing and bit operations

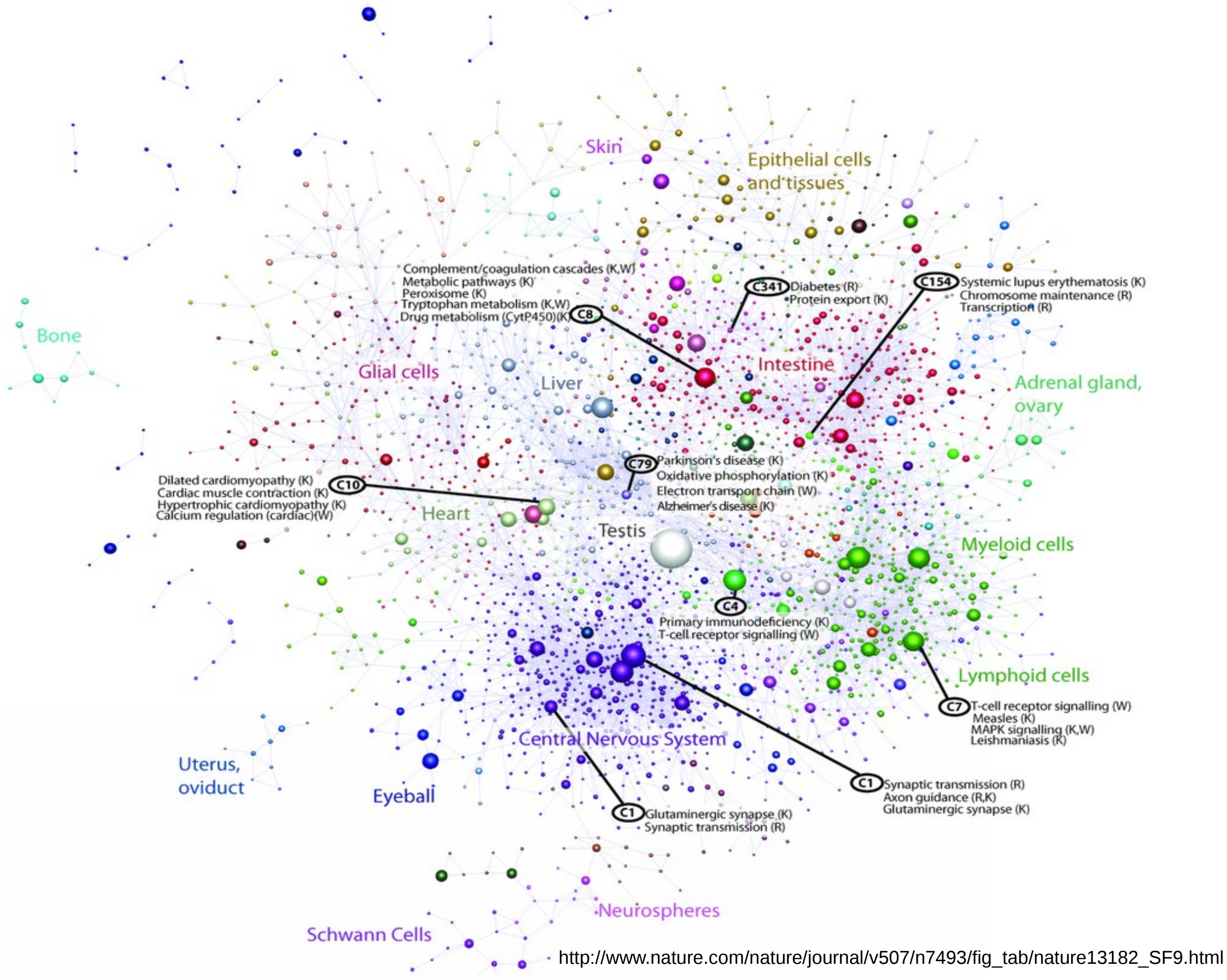
[1] "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.", Nucleic Acids Research, 1997

[2] "Fast alignment-free sequence comparison using spaced-word frequencies.", Bioinformatics, 2014

[3] "STELLAR: fast and exact local alignments.", BMC Bioinformatics, 2011

Topic 6: Co-Expression Networks

- Undirected graph, nodes: genes, edges: similar expression pattern across samples
- Co-expressed genes controlled by same regulatory program/ functionally related/ members of the same pathway

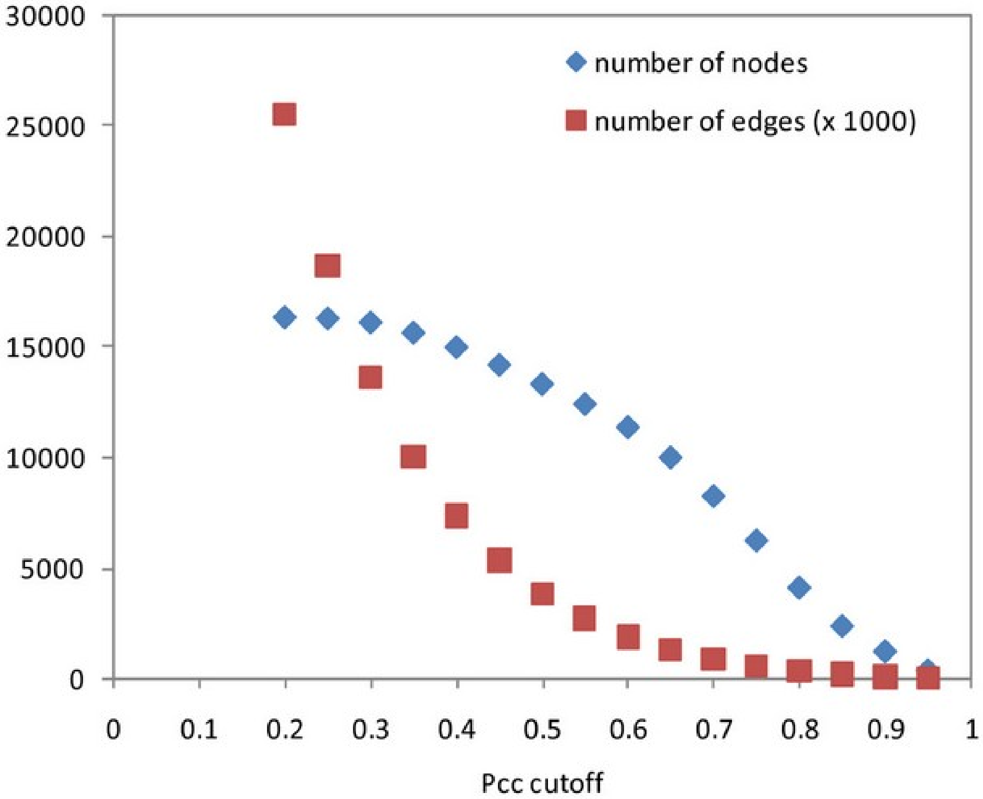


Skin
Epithelial cells and tissues
Bone
Complement/coagulation cascades (K,W)
Metabolic pathways (K)
Peroxisome (K)
Tryptophan metabolism (K,W)
Drug metabolism (CytP450)(K)
C8
C341
Diabetes (R)
Protein export (K)
C154
Systemic lupus erythematosis (K)
Chromosome maintenance (R)
Transcription (R)
Glial cells
Liver
Intestine
Adrenal gland, ovary
C79
Parkinson's disease (K)
Oxidative phosphorylation (K)
Electron transport chain (W)
Alzheimer's disease (K)
Dilated cardiomyopathy (K)
Cardiac muscle contraction (K)
Hypertrophic cardiomyopathy (K)
Calcium regulation (cardiac)(W)
C10
Heart
Testis
Myeloid cells
C4
Primary immunodeficiency (K)
T-cell receptor signalling (W)
Lymphoid cells
C7
T-cell receptor signalling (W)
Measles (K)
MAPK signalling (K,W)
Leishmaniasis (K)
Central Nervous System
Uterus, oviduct
Eyeball
C1
Synaptic transmission (R)
Axon guidance (R,K)
Glutaminergic synapse (K)
C1
Glutaminergic synapse (K)
Synaptic transmission (R)
Neurospheres
Schwann Cells

http://www.nature.com/nature/journal/v507/n7493/fig_tab/nature13182_SF9.html

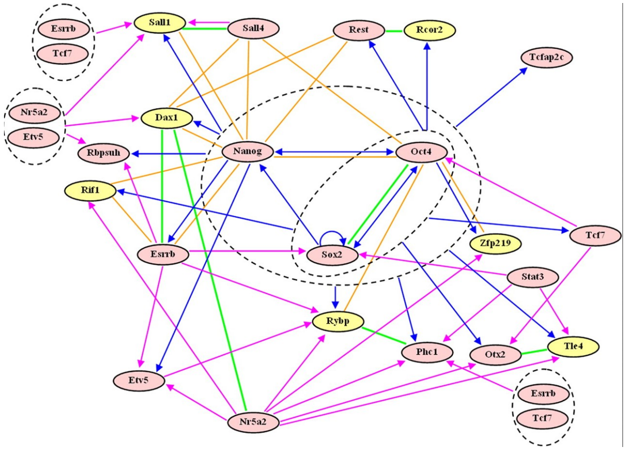
Topic 6: Co-Expression Networks

- Motivation, construction methods, measures, thresholds
- Network analysis (metrics e.g. betweenness centrality)
- Functional subnetworks



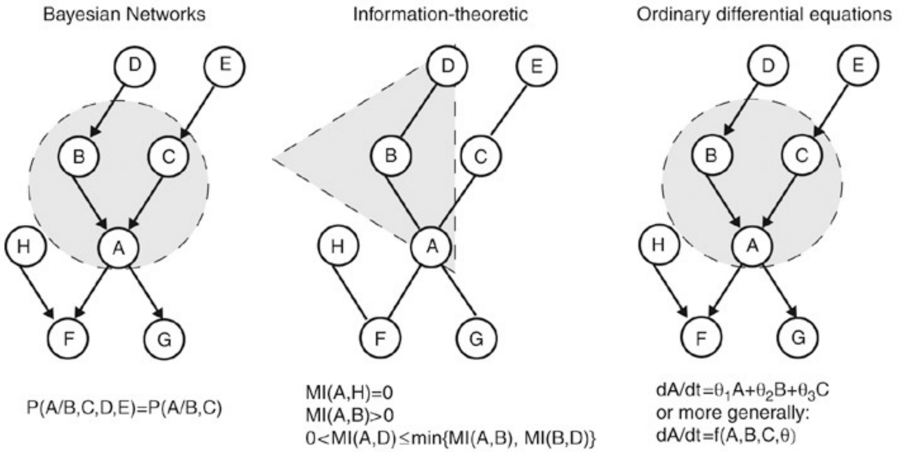
[1] "A General Framework for Weighted Gene Co-Expression Network Analysis", Statistical Applications in Genetics and Molecular Biology, 2005
[2] "Learning from Co-expression Networks: Possibilities and Challenges.", Frontiers in Plant Science, 2016
[3] "Arabidopsis gene co-expression network and its functional modules", BMC Bioinformatics, 2009

Topic 7: Gene Regulatory Network Reconstruction/ Regulatory Activity



http://www.pnas.org/content/104/42/16438/F3.expansion.html

- Regulators (e.g. transcription factors) govern gene expression levels
- GRN: directed graph, edges: represent biochemical process (activation/ inhibition)

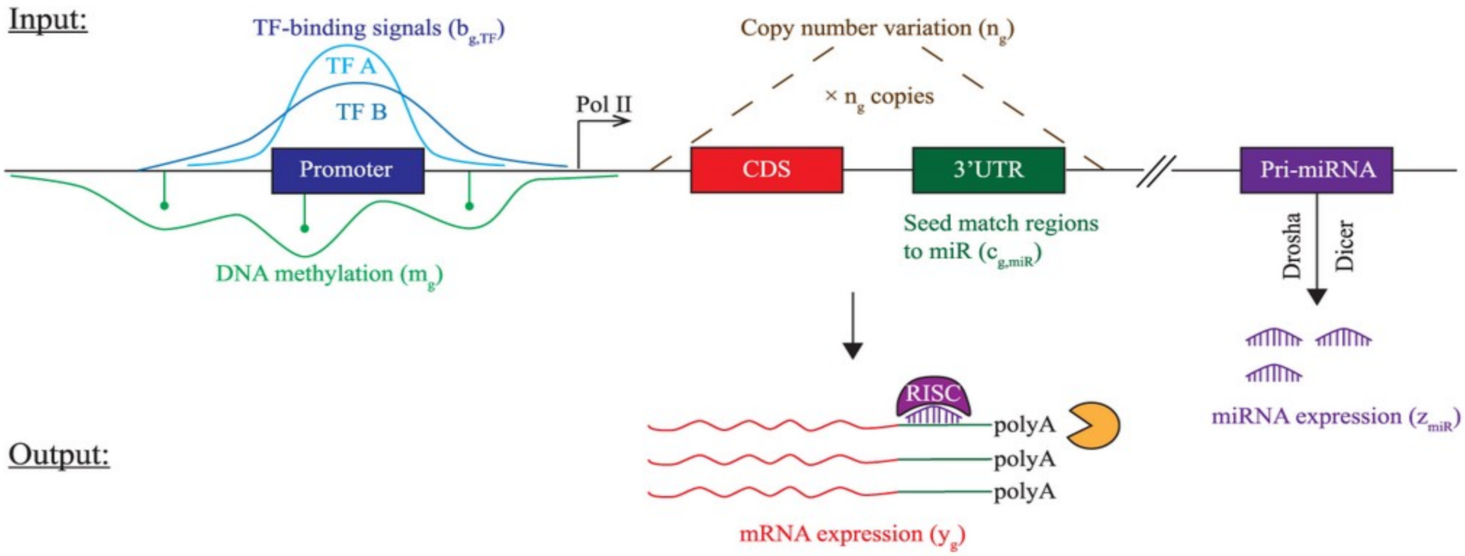


http://msb.embopress.org/content/3/1/78.figures-only

- Reconstruction of GRNs based on gene expression data using local inference
- Method „ARACNE“: mutual information, filtering, applications

[1] "ARACNE: An Algorithm for the Reconstruction of Gene Regulatory Networks in a Mammalian Cellular Context.", BMC Bioinformatics, 2006

Topic 7: Gene Regulatory Network Reconstruction/ Regulatory Activity



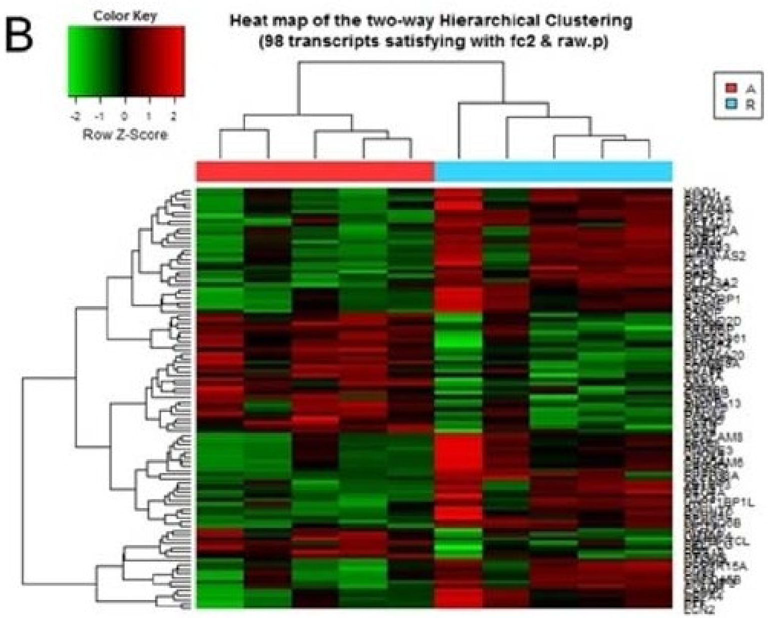
http://journals.plos.org/ploscompbiol/article?id=10.1371%2Fjournal.pcbi.1003908

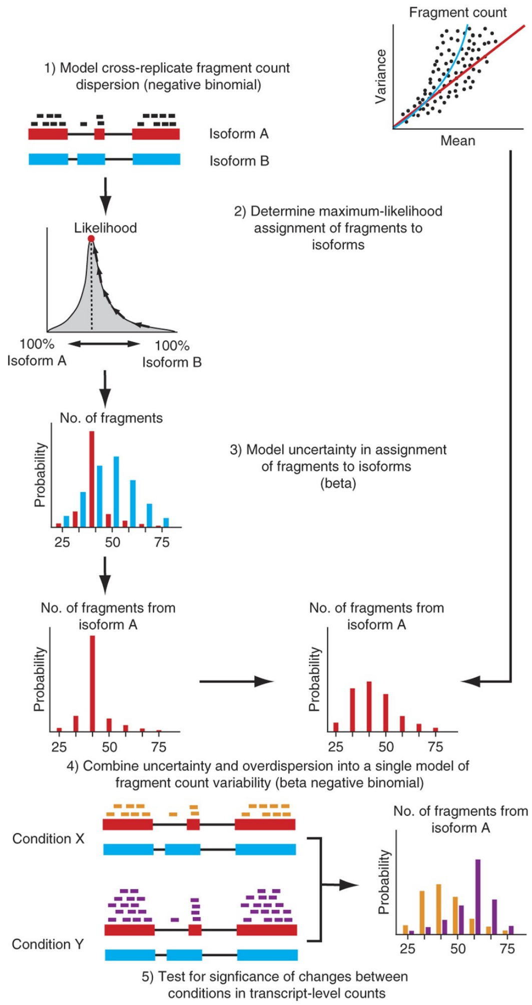
- GRN reconstruction: very complex problem → use additionally a priori knowledge about regulatory principles
- With network and expression data: estimate regulatory activity using mathematical optimization
- Methods: RACER, Rabit

[2] "Regression Analysis of Combined Gene Expression Regulation in Acute Myeloid Leukemia.", Plos Computational Biology, 2014
[3] "Inference of transcriptional regulation in cancers", PNAS, 2015

Topic 8: RNA-Sequencing: Differential Expression Analysis

- Differential expression: which transcripts from tumor sample are produced at significantly higher/ lower number than from healthy sample
- Tools: DESeq and CuffDiff
- Normalization of data
- Count uncertainty and overdispersion problem
- Statistical tests




1) Model cross-replicate fragment count dispersion (negative binomial)

2) Determine maximum-likelihood assignment of fragments to isoforms

3) Model uncertainty in assignment of fragments to isoforms (beta)

4) Combine uncertainty and overdispersion into a single model of fragment count variability (beta negative binomial)

5) Test for signficance of changes between conditions in transcript-level counts

[1] "Differential expression analysis for sequence count data.", BMC Genome Biology, 2010
[2] "Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks", Nature Protocols, 2014
[3] "Comparison of software packages for detecting differential expression in RNA-seq studies", Briefings in Bioinformatics, 2015

Potential Further Topics

- Genome Assembly FM-index based
- Small RNA Sequencing (miRNAs)
- Subnetwork analysis
- Proteomics

Overview Topics

Nr	Topic	Supervisor	Student
1	Genome Assembly	Yvonne	David
2	Read Mapping in RNA-Seq (*)	Yvonne	Lukas
3	Compression Methods Genomics	Yvonne/Prof.Leser	Jannes
4	Variant Calling (*)	Yvonne	
5	Sequence Similarity	Yvonne	Marti
6	Co-Expression Networks	Saskia/Yvonne	
7	Gene Regulatory Network Reconstruction/Regulatory activity	Saskia	Leon
8	Differential Expression Analysis of RNA-Seq Data (*)	Saskia/Yvonne	

Agenda

- Introduction
- Topics and assignment
- **Hints on presenting your topic and writing your thesis**

Allgemeine Hinweise

- Dozenten sind ansprechbar!
  - Vorbesprechung des Themas
  - Folien durchgehen
  - Abgrenzung der Ausarbeitung
- Diskussion erwünscht
  - Keine Angst vor Fragen: Fragen sind keine Kritik
  - Eine Frage nicht beantworten können ist in Ordnung
- Tiefe, nicht Breite
  - Lieber das Thema einengen und dafür Details erklären
- Bezug nehmen
  - Vergleich zu anderen Arbeiten (im Seminar)

Allgemeine Hinweise

- Werten und bewerten
  - Keine Angst vor nicht ganz zutreffenden Aussagen – solange gute Gründe vorhanden sind
  - Begründen und argumentieren
  - Kritikloses Abschreiben ist fehl am Platz
- Literaturrecherche ist notwendig
  - Die ausgegebenen Arbeiten sind Anker
  - Weiterführende Arbeiten müssen herangezogen werden
  - Auch Grundlagen nachlesen

- Auf der Homepage finden sie eine Liste zum Abhaken

Hinweise zum Vortrag

- ~30 Minuten inkl Diskussion
- Klare Gliederung
- Ab und an Hinweise geben, wo man sich befindet
- Bilder und Grafiken; Beispiele
- Font: mind. 16pt
- Eher Stichwörter als lange Sätze
- Vorträge können auch unterhaltend sein
  - Gimmicks, Rhythmuswechsel, Einbeziehen der Zuhörer, etc.
- Adressat sind alle Teilnehmer, nicht nur die Betreuer
- Technik: Laptop? Powerpoint?

Hinweise zur Ausarbeitung

- Eine gedruckte Version abgeben
  - Selbstständigkeitserklärung unterschreiben
- Eine elektronische Version schicken
- Referenzen: Alle verwendeten und nur die
  - Im Text referenzieren, Liste am Schluss
- Korrekt zitieren
  - Vorsicht vor Übernahme von kompletten Textpassagen; wenn, dann deutlich kennzeichnen
  - Aussagen mit Evidenz oder Verweis auf Literatur versehen
- Verwendung von gefundenen Arbeiten im Web
  - Möglich, aber VORSICHT
  - Eventuell Themenschwerpunkt verschieben – Betreuer fragen

Hinweise zur Ausarbeitung –2-

- Gezielt und sachlich schreiben
  - Ausführungen zur „Philosophische Überlegungen zu Vorzügen probabilistischer Verfahren im Vergleich zu Dempster‘s Theory of Evidence“ oder zur „Anmerkungen zur Trivialisierung des politischen Diskurs für soziale Netzwerke unter besonderer Berücksichtigung von Twitter“ möglichst kurz halten
  - Füllwörter vermeiden (dabei, hierbei, dann, …)
  - Knappe Darlegung, präzise Sprache
- Eine gute Gliederung ist die halbe Miete
- Kommen Sie zu Aussagen
  - Vorteile, Nachteile, verwandte Arbeiten, mögliche Erweiterungen, Anwendbarkeit, eigene Erfahrungen, …

Format

- Benutzung unserer Latex-Vorlage
- Nur eine Schriftart, wenig und konsistente Wechsel in Schriftgröße und –stärke
- Inhaltsverzeichnis
- Bilder: Nummerieren und darauf verweisen
- Referenzen:
  - [1] S. Wandelt and U. Leser (2013). "FRESCO: Referential Compression of Highly Similar Sequences". IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 10, no. 5, pp. 1275-1288.
  - [SWUL13] S. Wandelt and U. Leser (2013). "FRESCO: Referential Compression of Highly Similar Sequences". IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 10, no. 5, pp. 1275-1288.
- Darf man Wikipedia zitieren?
  - Ja, aber nicht dauernd

Wie halte ich einen Seminarvortrag

1. Wenn man nun so einen Seminarvortrag halten muss, dann empfiehlt es sich, möglichst lange Sätze auf die Folien zu schreiben, damit die Zuhörer nach dem Vortrag aus den Folienkopien noch wissen, was man eigentlich gesagt hat.

2. Während so einem Vortrag schaut sowieso jeder zum Projektor, also kann man das selbst ruhig auch tun - damit kontrolliert man gleichzeitig auch, ob der Beamer wirklich alles projeziert, was auf dem Laptop zu sehen ist. Ausserdem kann man so den Strom für das Laptop-Display sparen.

3. Übersichtsfolien am Anfang sind langweilig, enthalten keinen Inhalt und nehmen den Zuhörern die ganze Spannung. Schliesslich gibt‘s im Kino am Anfang auch keine Inhaltsangabe.

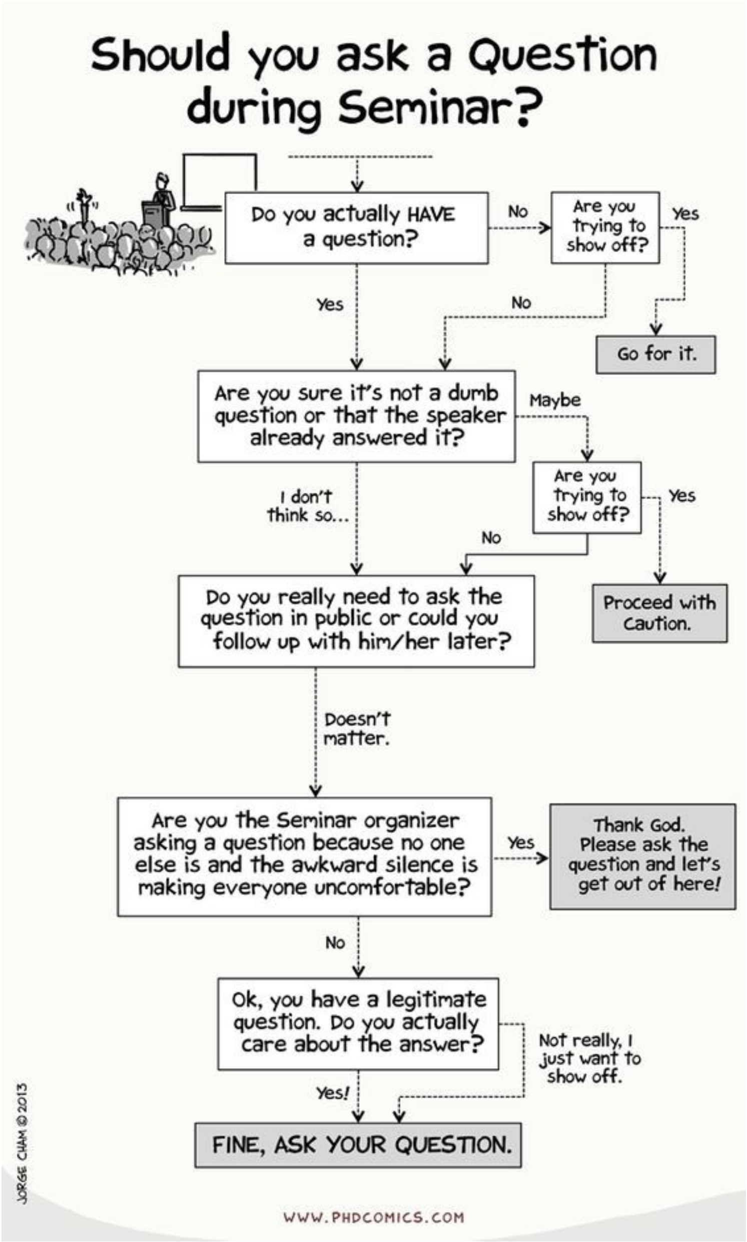
4. Powerpoint kann viele lustige Effekte, hat tolle Designs und Animationen. Die sollte man zur Auflockerung des Vortrags unbedingt alle benutzen, um zu zeigen, wie gut man das Tool im Griff hat.

5. Nicht zu wenig auf die Folien schreiben. Man weiß ja nie, ob man sie nicht doch ausdrucken muss, und man kann so wertvolle Zeit sparen, wenn man nicht weiterschalten muss.

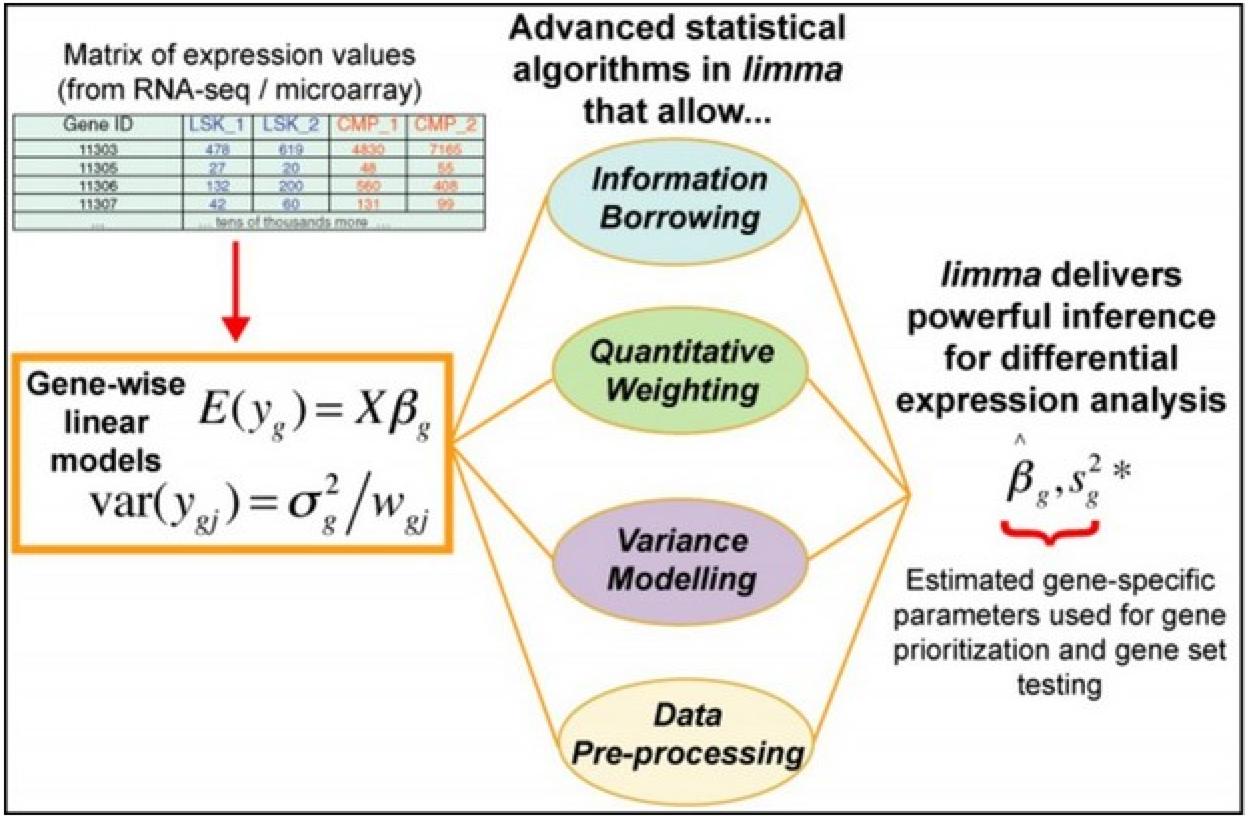
6. Man sollte versuchen, möglichst lange zu reden. Die Zeitvorgaben sind nur für die Leute, die nicht genug wissen - eigentlich will der Prüfer sehen, dass man sich auch darüber hinaus mit dem Thema beschäftigt hat.

Bloß keine Hervorhebungen im Text – sonst müssen die Zuhörer ja gar nicht mehr aufpassen!

Any Questions?


Should you ask a Question during Seminar?
Do you actually HAVE a question?
No
Are you trying to show off?
Yes
Go for it.
Yes
No
Are you sure it's not a dumb question or that the speaker already answered it?
Maybe
Are you trying to show off?
Yes
Proceed with Caution.
No
I don't think so...
Do you really need to ask the question in public or could you follow up with him/her later?
Doesn't matter.
Are you the Seminar organizer asking a question because no one else is and the awkward silence is making everyone uncomfortable?
Yes
Thank God. Please ask the question and let's get out of here!
No
Ok, you have a legitimate question. Do you actually care about the answer?
Not really, I just want to show off.
Yes!
FINE, ASK YOUR QUESTION.
JORGE CHAM ©2013
WWW.PHDCOMICS.COM

App1: Differential Expression Analysis of Transcriptome Data based on Microarray Data

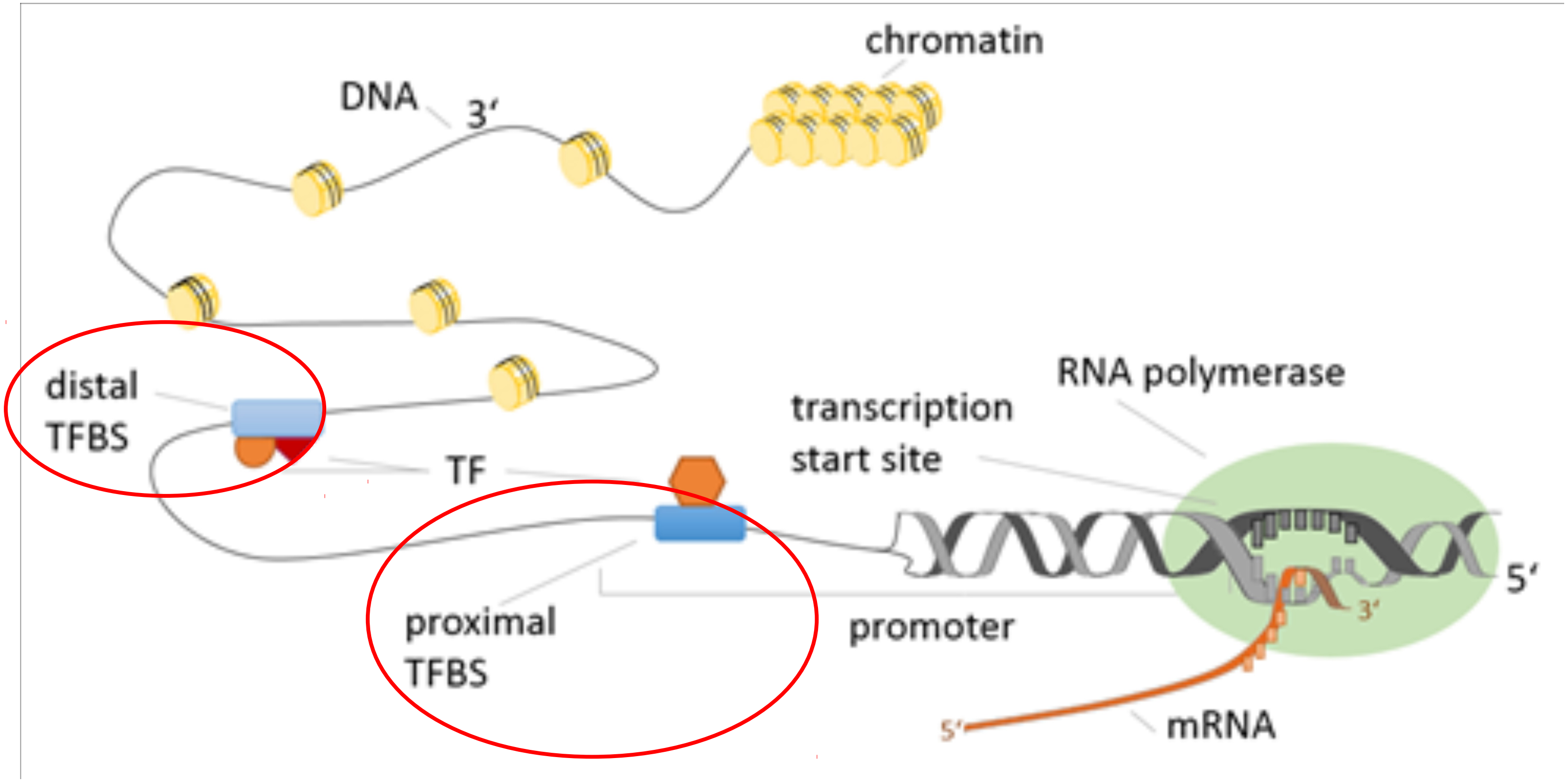

Matrix of expression values (from RNA-seq / microarray)
Advanced statistical algorithms in *limma* that allow...
Information Borrowing
Quantitative Weighting
Variance Modelling
Data Pre-processing
Gene-wise linear models
$E(y_g) = X\beta_g$
$\mathrm{var}(y_{gj}) = \sigma_g^2 / w_{gj}$
limma delivers powerful inference for differential expression analysis
$\underbrace{\hat{\beta}_g, s_g^2}$ *
Estimated gene-specific parameters used for gene prioritization and gene set testing

Gene ID	LSK_1	LSK_2	CMP_1	CMP_2
11303	478	619	4830	7165
11305	27	20	48	55
11306	132	200	560	408
11307	42	60	131	99
...	... tens of thousands more ...			

http://www.rna-seqblog.com/microarray-analysis-workhorse-limma-now-capable-of-differential-expression-and-differential-splicing-analyses-of-rna-seq-data/

- Linear models (limma), tests, multiple testing correction, prefiltering, ... [DOI: 10.1007/0-387-29362-0_23]

App2: Annotation of Regulatory Sites



- Overview of computational methods and databases [DOI: 10.1038/nbt1053], examples