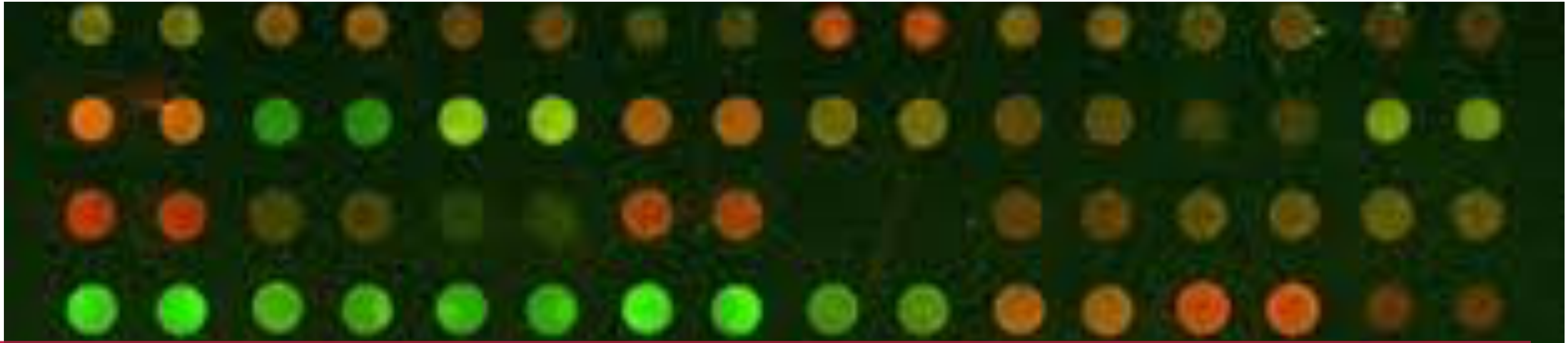
HPI Hasso Plattner Institut
IT Systems Engineering | Universität Potsdam



Introduction to RNAseq Analysis

Milena Kraus

Apr 18, 2016

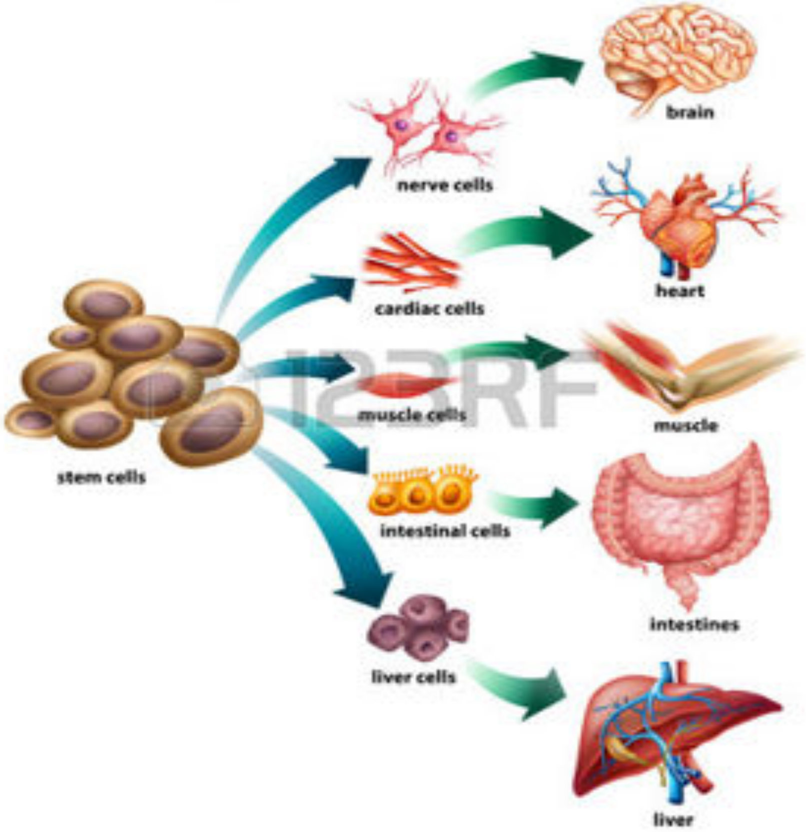
Agenda

What is RNA sequencing used for?

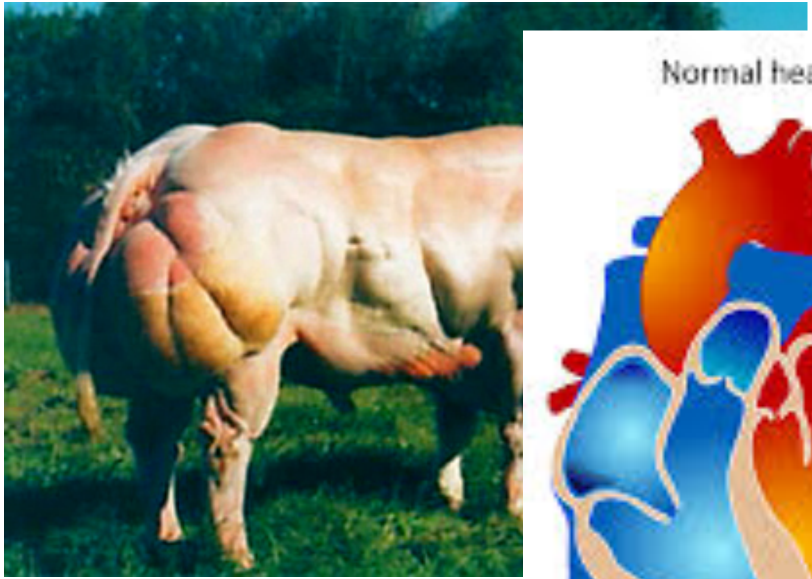
1. Biological background
2. From wet lab sample to transcriptome
   a. Experimental procedure
   b. Raw data
   c. Processing pipeline(s)
   d. Downstream analysis

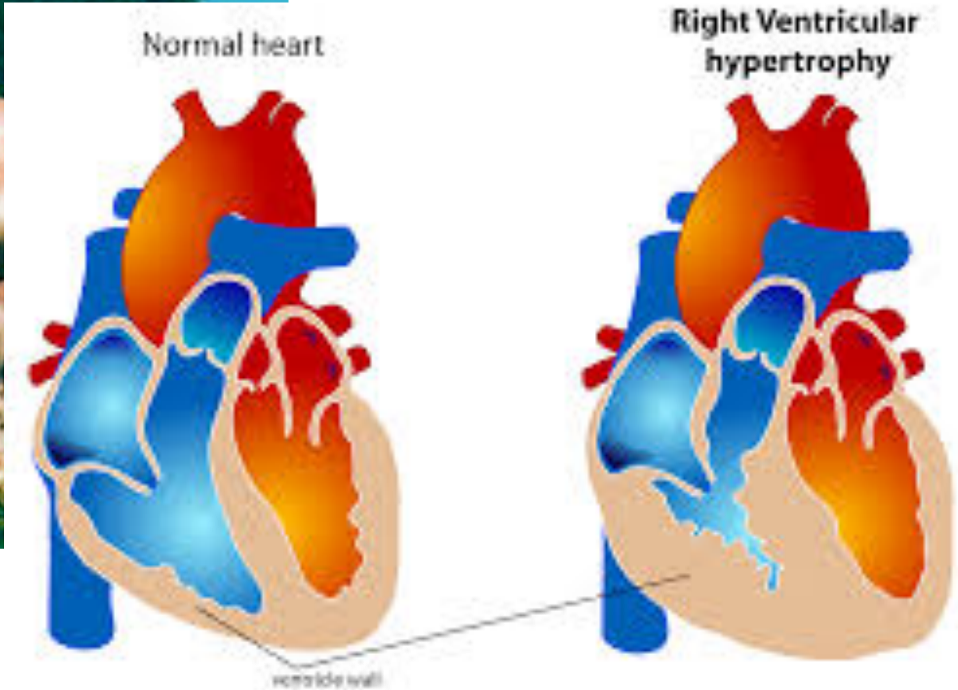
How is a muscle cell different from a liver cell?

- Every cell in your body contains the same DNA as every other cell
- The DNA codes for every process in the cell

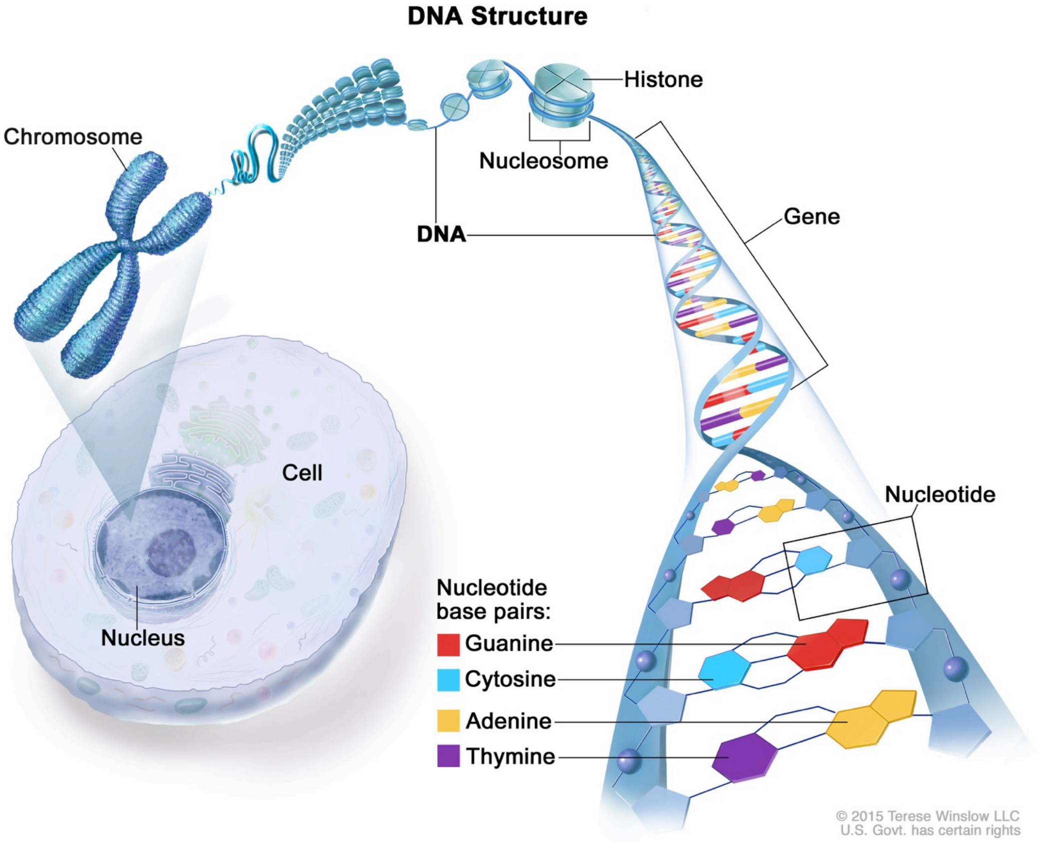


What is the difference between a healthy heart and a sick heart?



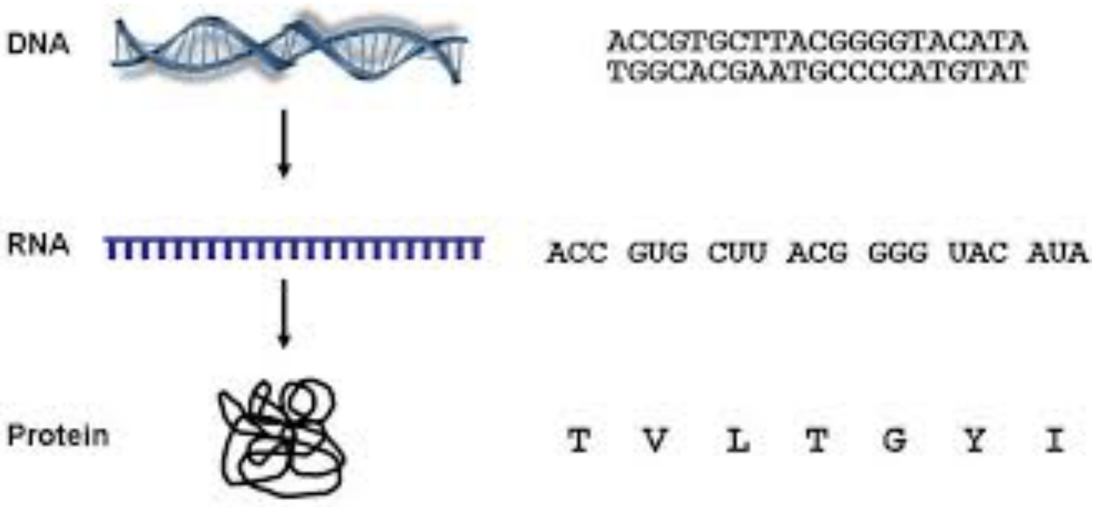


Biological Background

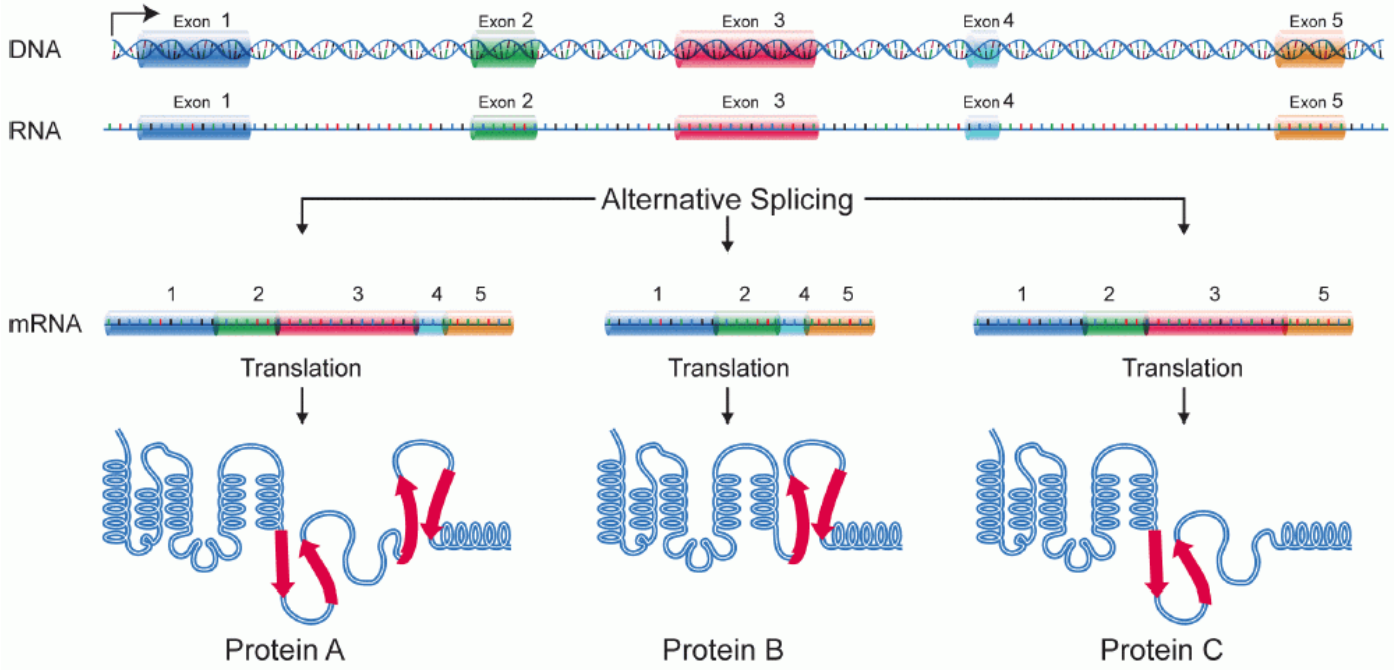


Central Dogma of Molecular Biology
From DNA to RNA to protein

- Interesting information from RNA:
  - Sequence
  - Quantity



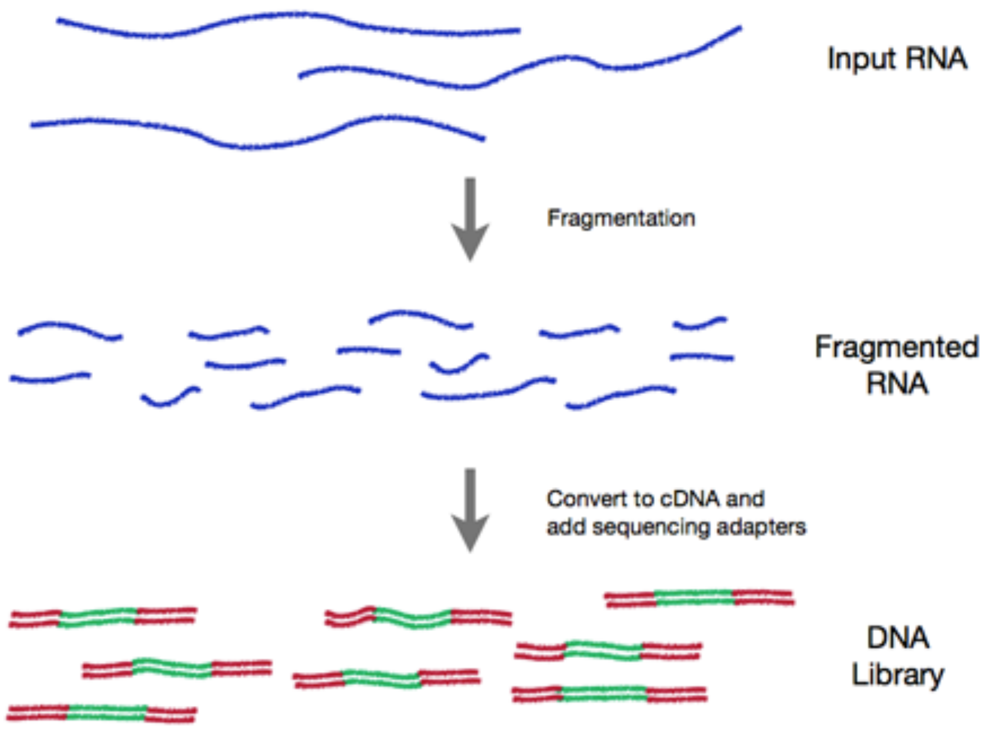
One more bio fact before we start: Alternative Splicing



- Von National Human Genome Research Institute - http://www.genome.gov/Images/EdKit/bio2j_large.gif, Gemeinfrei, https://commons.wikimedia.org/w/index.php?curid=2132737

From Wet Lab Experiment to Transcriptome

Experimental Procedure



- Over 20.000 RNAs in variable abundance (1-k times)
- 1.500-2.000 nt

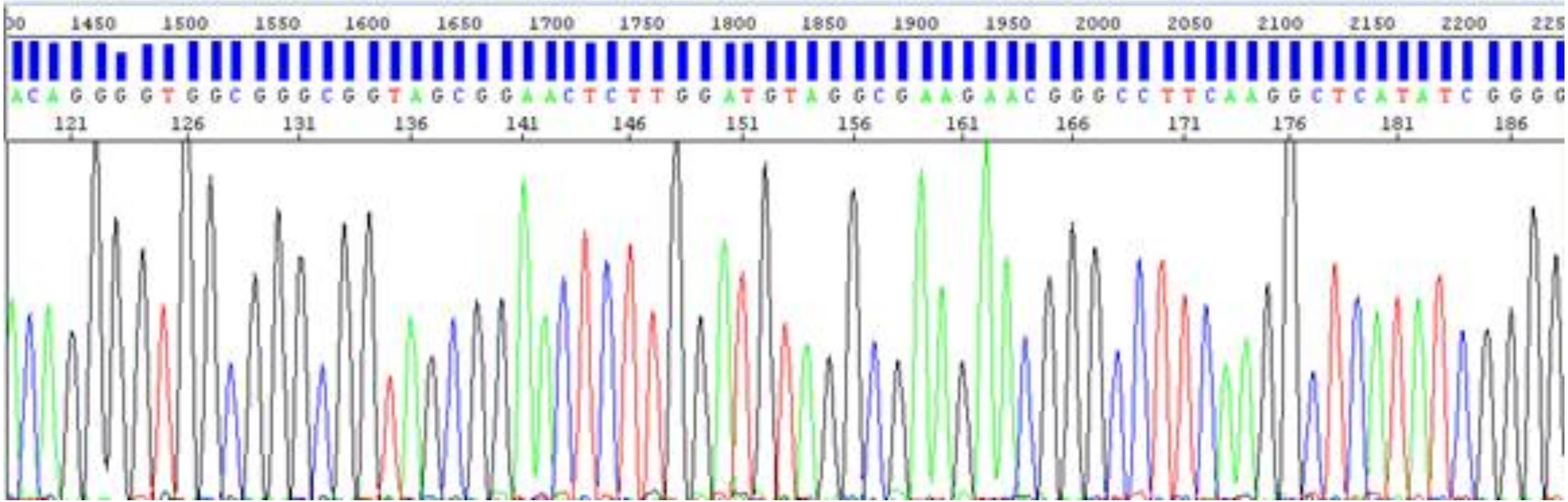
- 30-200 nt

Sequencer

SOLID 5500

Illumina HiSeq2000

Sequencing Signal



Raw data
FASTQ files

@SRR831012.1 HWI-ST155_0742:7:1101:1284:1981/1
NGAGATGAAGCACTGTAGCTTGGAATTCTCGGGTGCCAAGGAACTCCAGT
+
%1=DDDFFHHHGFIHHIIIIIIIIIIIIIIIIIEHIIIIIIIFIIIIIII

@SRR831012.2 HWI-ST155_0742:7:1101:2777:1998/1
NGAGATGAAGCACTGTAGCTCTTTGGAATTCTCGGGTGCCAAGGAACTCC
+
%1=DFFFFHHHHHIIIIIIIIIIIIIIIIIIIIGIIIIIIIIIIIIIIIG

Quality score (increasing from worst to best):

!"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstuvwxyz{|}~

@SampleID.ReadNr

Experimental Setup

In our setting:

- ~1.4 GB per file
- ~8 Mio reads per file
- 80 files

Raw data
Reference genome

- FASTA-file

>Sequenz 1

;comment A

ACAAGATGCCATTGTCCCCCGGCCTCCTGCTGCTGCTGCTCTCCGGGGCCACGGC
CACCGCTGCCCTGCCCCTGGAGGGTGGCCCCACCGGCCGAGACAGCGAGCATAT
GCAGGAAGCGGCAGGAATAAGGAAAAGCAGCCTCCTGACTTTCCTCGCTTGGTGG
TTTGAGTGGACCTCCCAGGCCAGTGCCGGGCCCCTCATAGGAGAGGAAGCTCGG
GAGGTGGCCAGGCGGCAGGAAGGCGCACCCCCCCAGCAATCCGCGCGCCGGGA
CAGAATGCCCTGCAGGAACTTCTTCTGGAAGACCTTCTCCTCCTGCAAATAAAACCT
CACCCATGAATGCTCACGCAAGTTTAATTACAGACCTGAA...

In our setting:

- Indexed HG19 (Humane Genome V19)
- HG consists of approx 3.2B nucleotides

Raw data
Gene library

- 20k-25k protein coding genes representing small part of the genome
- Using the annotation to speed up processing
- If the discovery of new genes in a sample is expected, a custom annotation can be calculated from the reads

Col 1	Col 2	Col 3	Col 4	Col 5	Col 6	Col 7	Col 8	Col 9
chr21	HAVANA	transcript	10862622	10863067	.	+	.	gene_id "ENSG00000169..
chr21	HAVANA	exon	10862622	10862667	.	+	.	gene_id "ENSG00000169..
chr21	HAVANA	CDS	10862622	10862667	.	+	0	gene_id "ENSG00000169..
chr21	HAVANA	start_codon	10862622	10862624	.	+	0	gene_id "ENSG00000169..
chr21	HAVANA	exon	10862751	10863067	.	+	.	gene_id "ENSG00000169..
chr21	HAVANA	CDS	10862751	10863064	.	+	2	gene_id "ENSG00000169..
chr21	HAVANA	stop_codon	10863065	10863067	.	+	0	gene_id "ENSG00000169..
chr21	HAVANA	UTR	10863065	10863067	.	+	.	gene_id "ENSG00000169..

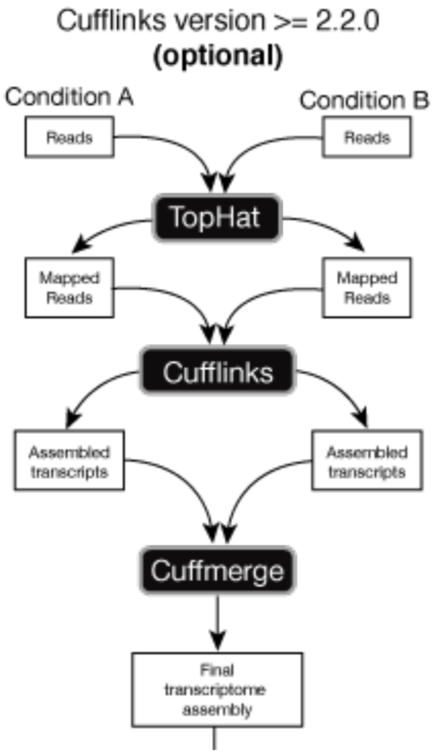
- In our setting: geneshg19.gtf

Processing pipeline
Gold standard – tophat/cufflinks

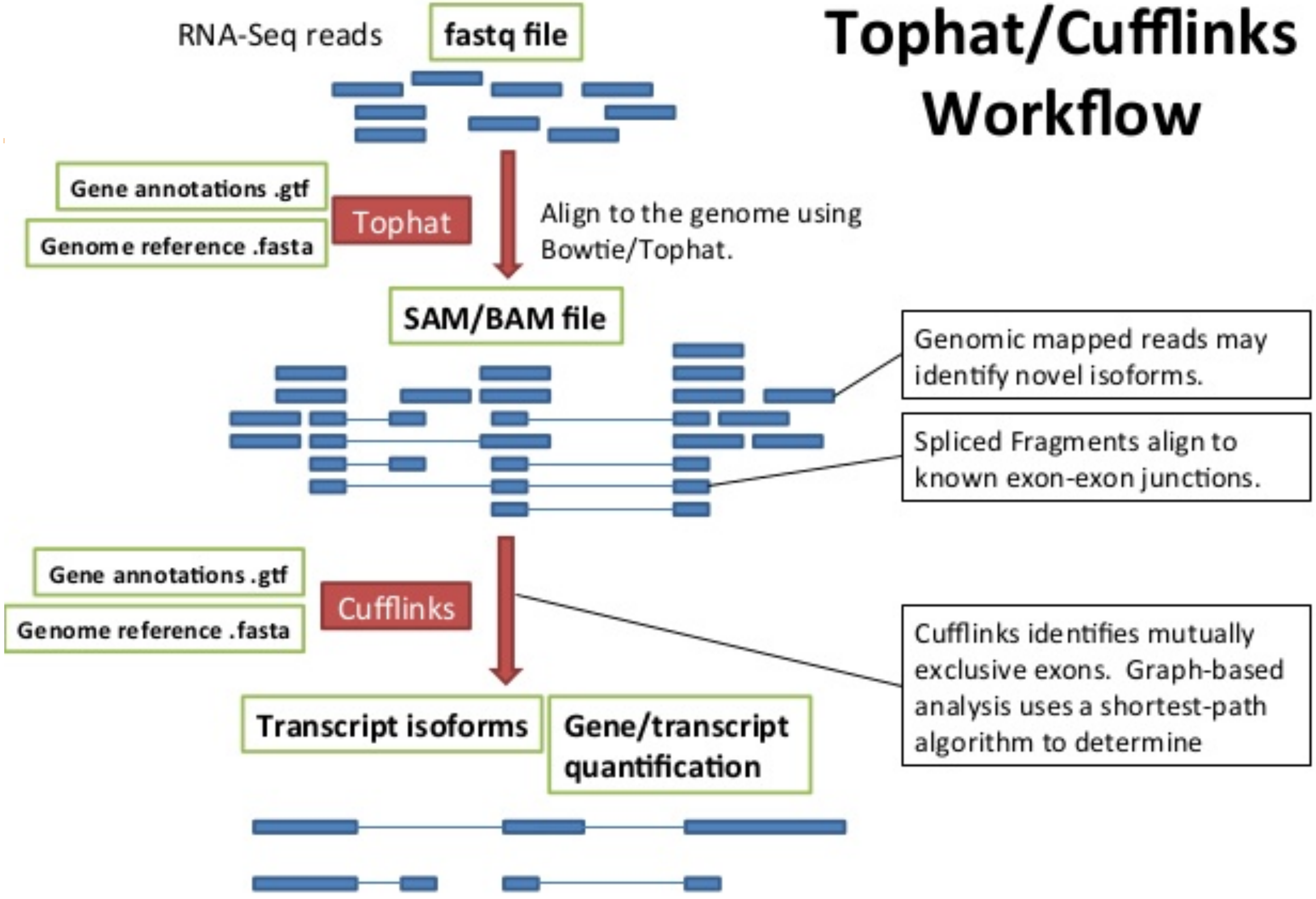
- **TopHat** aligns RNA-Seq reads to mammalian-sized genomes using the ultra high-throughput short read aligner BOWTIE, and then analyzes the mapping results to identify splice junctions between exons.

- **Cufflinks** assembles mapped RNA-Seq reads into transcripts.

- **Cuffmerge** creates an assembly of all transcripts to build the transcriptome (ocurrence transcripts).



Tophat/Cufflinks Workflow



Cuffmerge

- Input: Transcript library condition-wise (.gtf)
- Algorithm: Counts/Assembles all transcripts found in the different conditions
- Output: Library of all transcripts over all conditions (.gtf)

- The transcriptome …
  - Serves as a reference for further analysis,
  - Contains all found transcripts over all conditions, and
  - Resembles a rough profile of the studied tissue.

Processing Pipeline
New approach: DESeq and DEXseq

- Preprocessing to generate count tables from .bam files with htseq-count

DESeq

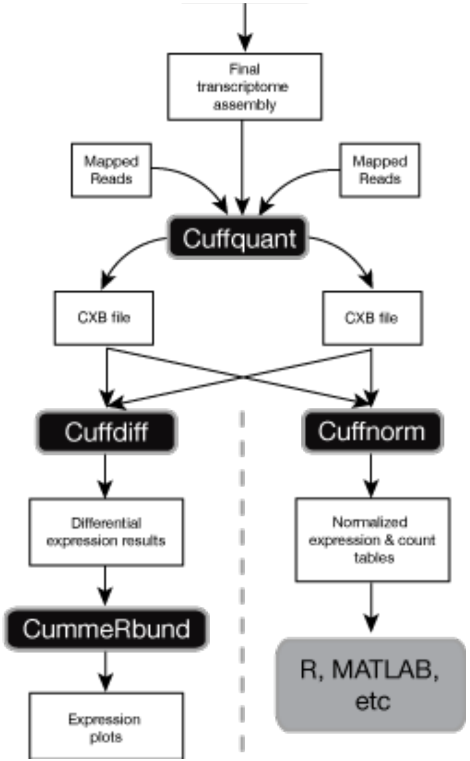
- Input: count table including all conditions
- Algorithm: Estimates variance-mean dependence in count data using a negative binomial distribution instead of maximum likelihood.
- Output: table containing gene identifiers and their normalized counts

DEXseq

- Same stat. method as DESeq but the output shows differentially expressed exons

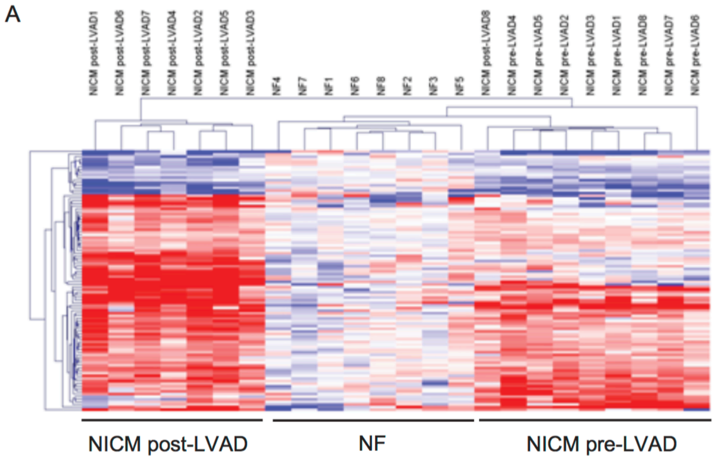
Downstream Processing
Statistical analysis and visualization

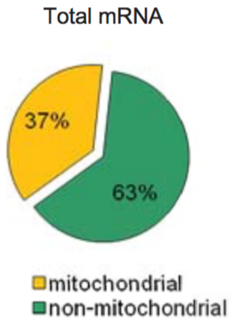
- **Cuffquant** is an intermediate step that helps to serialize and parallelize analysis.
- **Cuffdiff** compares expression levels of transcripts and shows differentials spliced genes and isoforms.
- **Cuffnorm** normalizes expression levels for exact comparison (usually optional).
- **CummeRbund** is an R package that provides various methods to visualize the data.

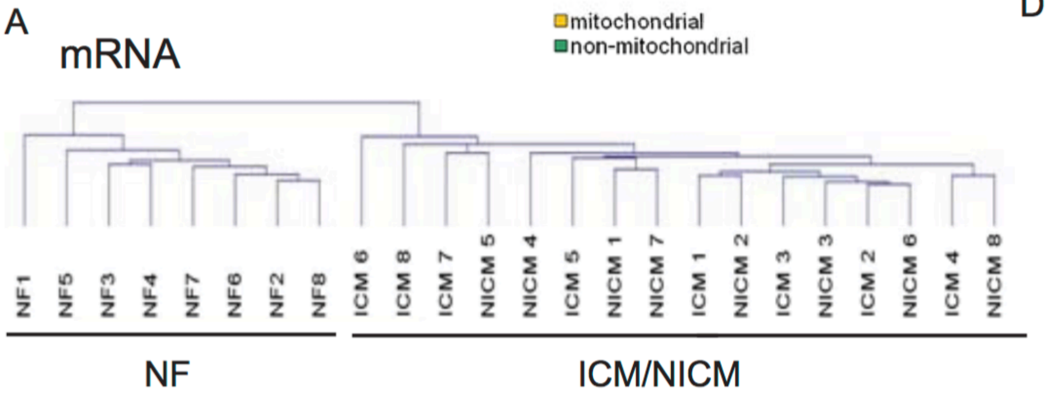


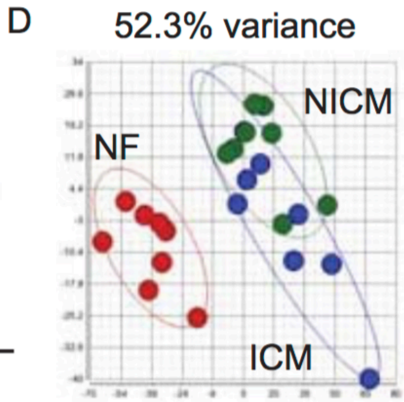
CummeRbund

- R package with common methods for
  - Statistical analysis
  - Visualization

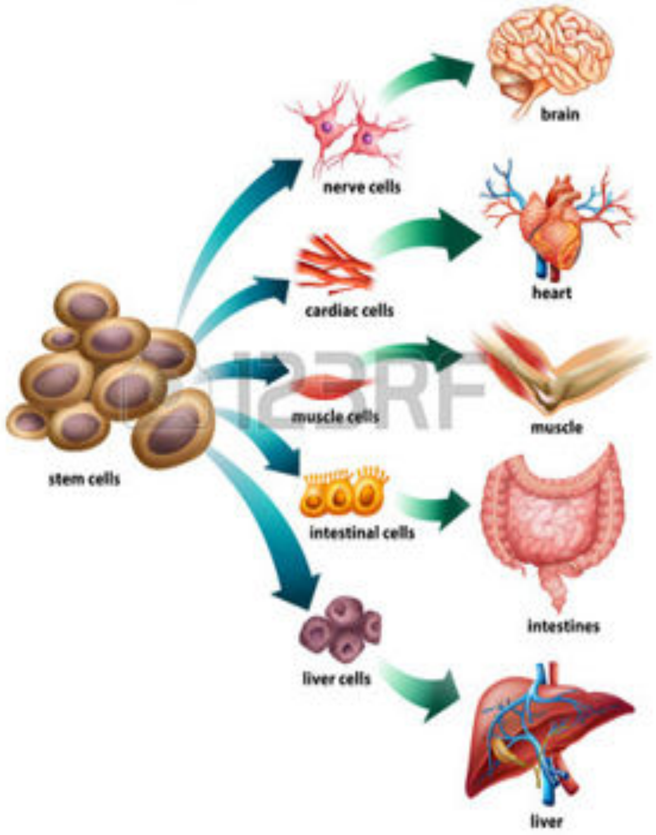


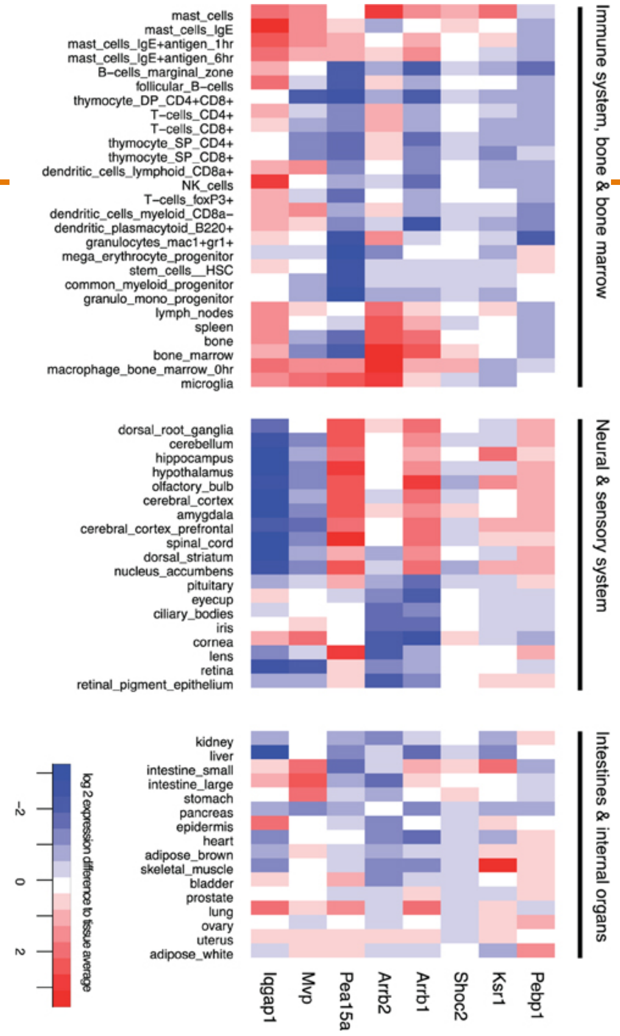






Tissue specificity





RNAseq Intro

Milena Kraus, 19.04.2016

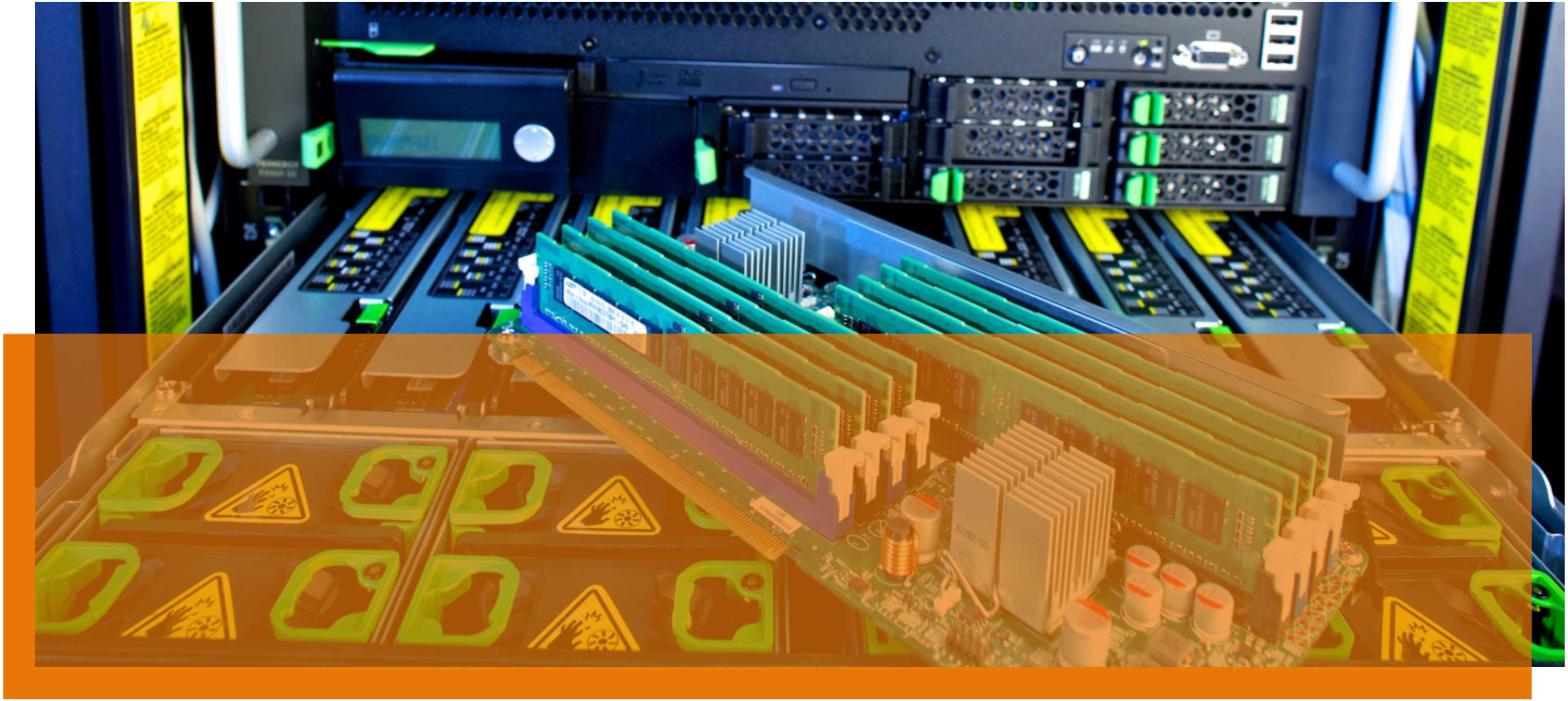
Chart **21**

Variant calling from RNAseq data

Thank you
for your attention!

Speaker
Job Description
Institute

HPI Hasso Plattner Institut
IT Systems Engineering | Universität Potsdam



HPI Hasso Plattner Institut

IT Systems Engineering | Universität Potsdam

HPI Hasso Plattner Institut
IT Systems Engineering | Universität Potsdam

Explanations
- Text layers

First text layer for running text.

- Second level for bullet points
  - Third level for bullet points
    - Fourth level for bullet points

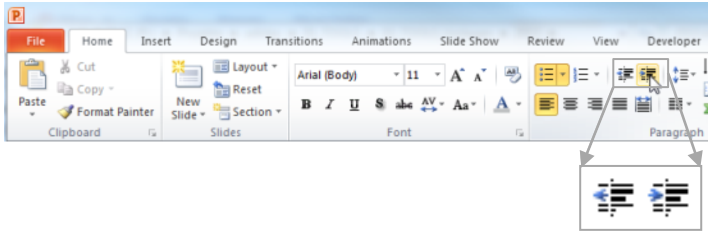
1. Fifth level for numberings
   a) Sixth level for listings

SEVENTH TEXT LAYER
FOR CORE MESSAGES

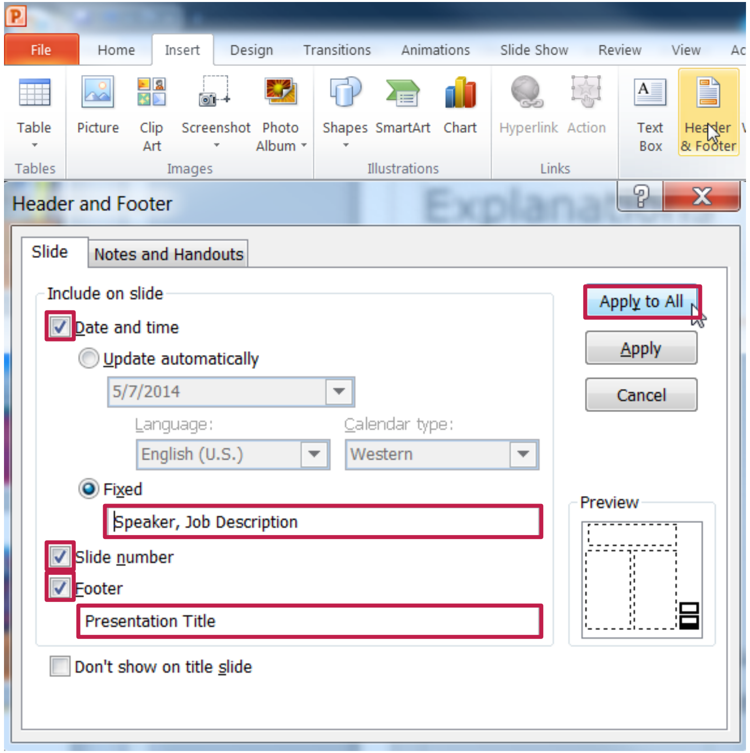
In this template, we pre-formatted different text layers
(as you can see on the right side).

You don’t have to generate bullet points manually.
By the way: Please avoid this!

To change from one text layer to the next, use the Increase / Decrease List Level buttons:



Explanations
- Footer



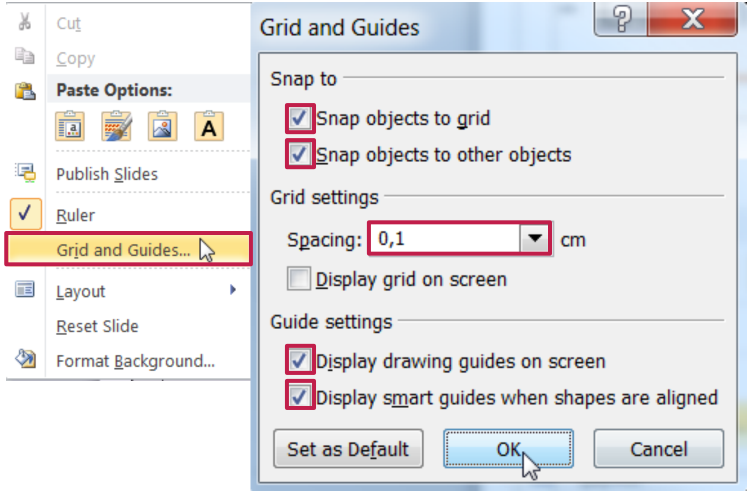
You can insert or change your presentation's footer. Click on the Insert-tab | Header and Footer | After filling in your descriptions click on **Apply to All.**

Descriptions:

- Activate date and time and write in:*Speaker, Job Description*
- Activate the slide number.
- Activate the footer and write in: *Presentation Title*

Don't use the template without the complete footer.

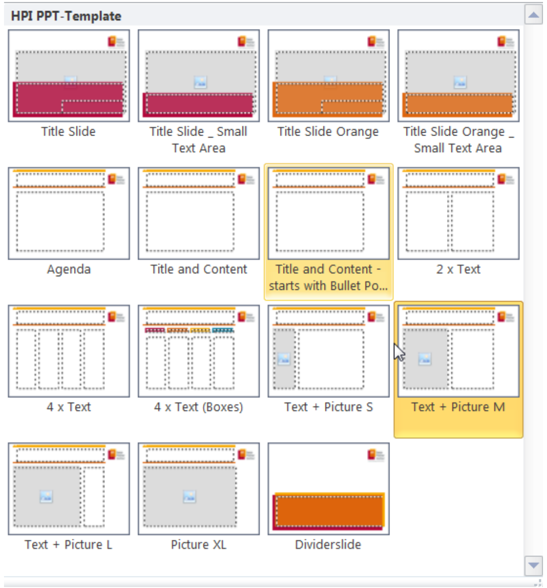
Explanations
- Drawing guides

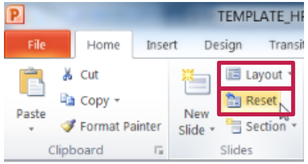


You can enable your guide-lines to align objects on the slide **(View | Show | Select the option „Guides“)**

Or hit the right mouse button outside
the slide and go at
„Grid and Guides…“

Explanations
- Slide layouts





You can choose between different slide layouts.
These pre-defined layouts gives you the oportunity to use text and visualisations just the right way.

To use these layouts:

Click on the Home-tab | New Slide or Layout | and choose one out of the layouts

Click „Reset“ to reset to the predefined slide layout.