Hasso Plattner Institut
IT Systems Engineering | Universität Potsdam

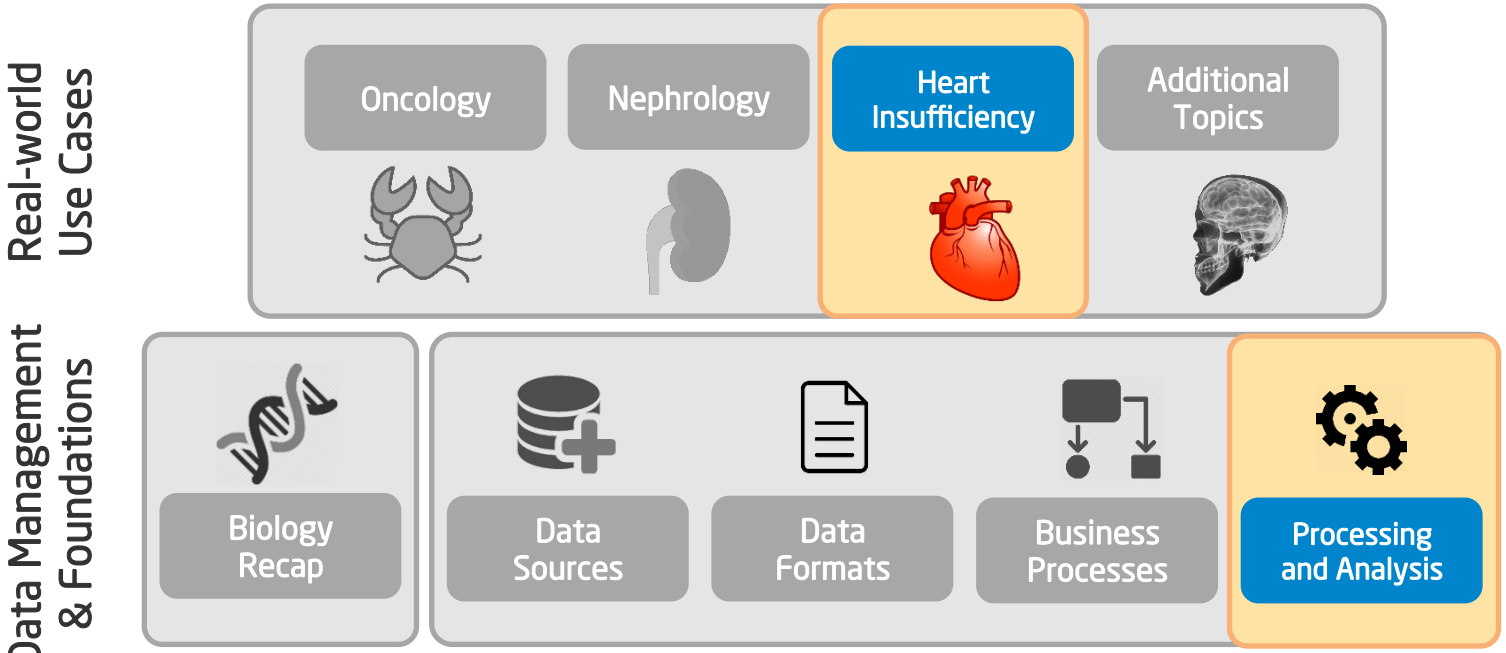
Evaluation Exercise III

Milena Kraus, Harry Cruz

Data Management for Digital Health

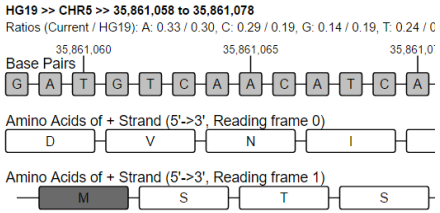
Summer 2017

Exercise III



Exercise III
Topics

- Genome Browser
- Medical Knowledge Cockpit
- Use Case Heart Failure (Systems Medicine)



AG Genome Browser

Jane Dough

female, 51 years, non-smoker

Markers

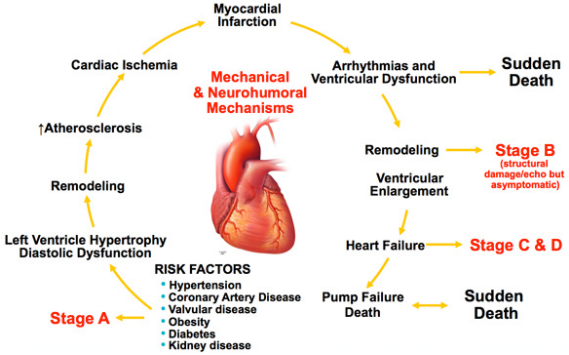
KRAS EGFR BRAF NRAS

Diagnosis

non-small cell lung cancer, stage IV

Charité

Medical Knowledge Cockpit



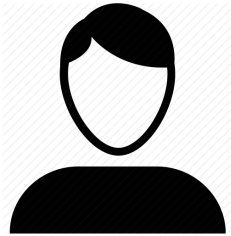
Evaluation Exercise III

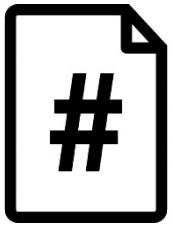
Data Management for Digital Health, Summer 2017

Exercise III
Key Stats

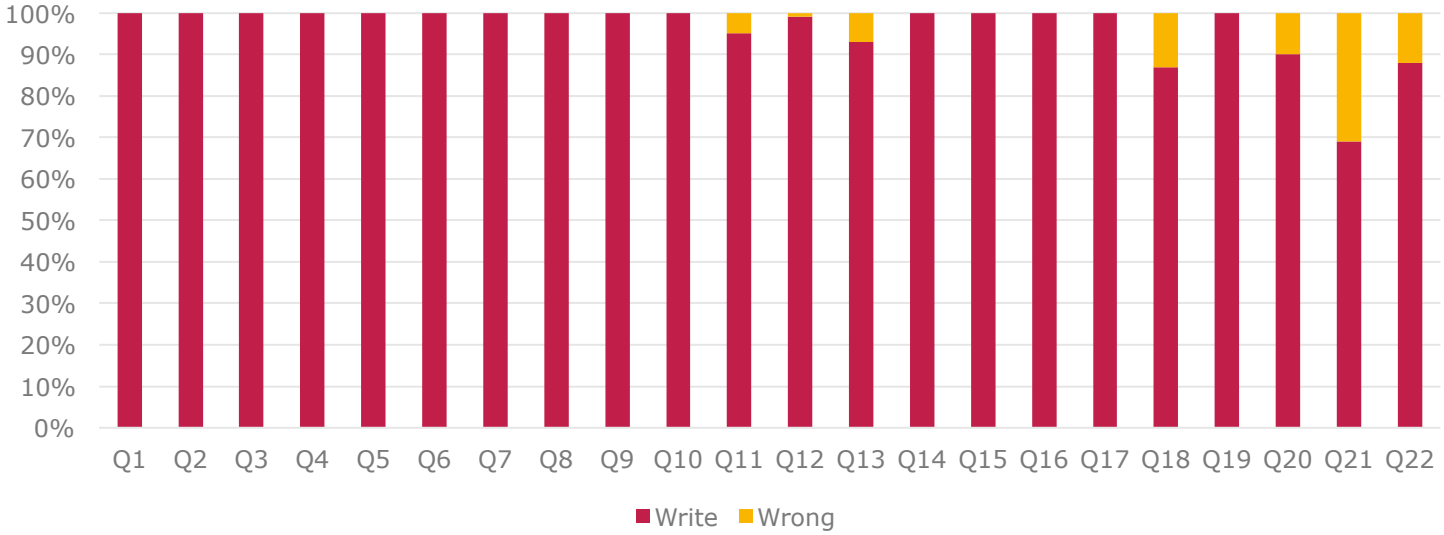
Hasso Plattner Institut

Part 1 – 5pts Part 2 – 35 pts	31 Students 29 Passed	Average score 5 / 100% 33.69 / 96%	Average time 68min





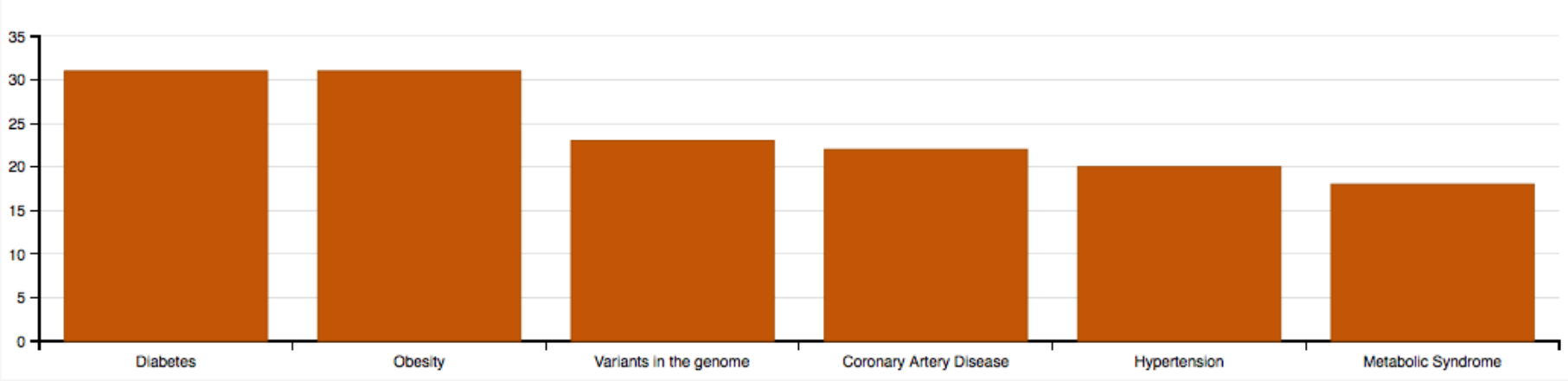
Exercise III
Key Stats



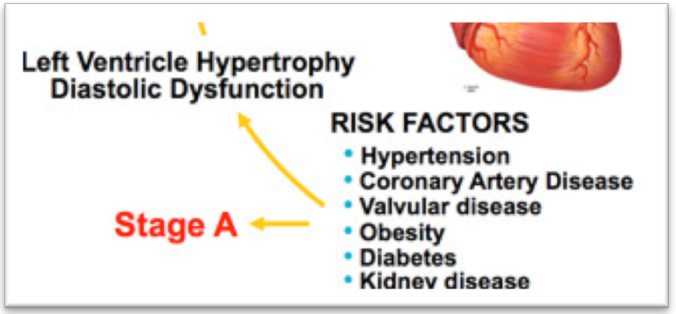
AG Genome Browser
Medical Knowledge Cockpit

Flawless!

Q 14: Which risk factors might affect the onset and progression of heart failure?



- Slide 20 (HF_SystemsMedicine): “Many genes have been found to be related to inherited forms of heart failure”
- Slide 51 (Mixed-type data analysis): “Metabolic syndrome is associated with the risk of developing cardiovascular disease and type 2 diabetes. “



Be aware of all right answers in multiple choice questions!

- In some cases, OpenHPI gave full credit if you picked at least one right answer (e.g. Q14, Q15, Q19)
- Please double check with the actual results and reconsider your answer before the exam
- In the exam, you will not get the full credit for only one correct answer!

Q 19: What are typical challenges that need consideration, when biomedical/ systems medicine data sets are analyzed, e.g., in clustering algorithms?

a) Data is of mixed-type (numerical, categorical etc.)

b) Missing values

c) Weight of variable subset within complete data set

d) Patient age

- Consider tables/variables from different origins, e.g.,:
  - 30 m variants x 200 patients
  - 20 k gene expression values x 200 patients
  - 150 clinical parameters x 200 patients
- Per default, all variables are accounted for in the same amount → variants would clearly dominate the complete analysis

Q 20: Principal Component Analysis is a technique applied to...

a) ... extract components of highest variation from a data set.

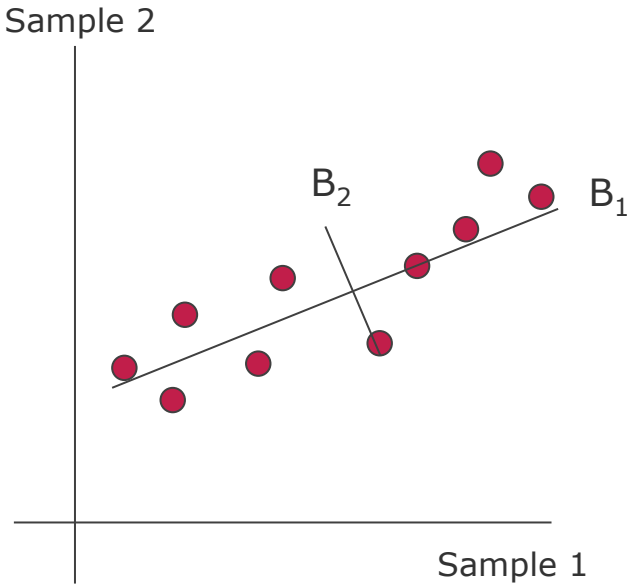
b) ... preprocess raw RNA sequencing reads.

c) ... reduce the dimensionality of large data sets.

d) ... model observed variables through a statistical distribution.

Principle Components Analysis

- Consider the mean of all points m, and a vector B going through the mean
- The vector B (PC1) is stretched along the path of most variation
- Vector B_2 (PC2) is stretched along the path of second most variation and orthogonal to B_1
- Length and orientation of the B vectors are most influenced by the outer points



Q21: Which of the following statements are correct regarding DNAseq and RNAseq?

Hasso Plattner Institut

a) Their primary goal is to provide sequence information
b) After the preparation phase, sequencing is performed on DNA molecules
c) Sequencing output is usually a FASTQ file
d) There is no difference between algorithms used for the alignment of reads.

Information Content in RNA

- Information retrieved from RNA:
  - Quantity (primary, How many RNAs are transcribed from a specific gene?)
  - Sequence (secondary, as sequence information can be inferred more precisely from DNA)

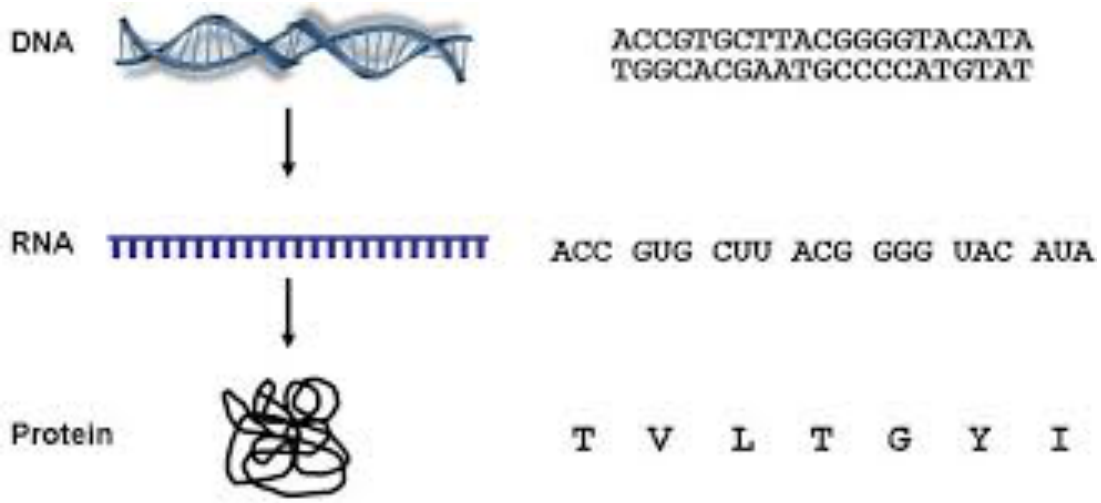


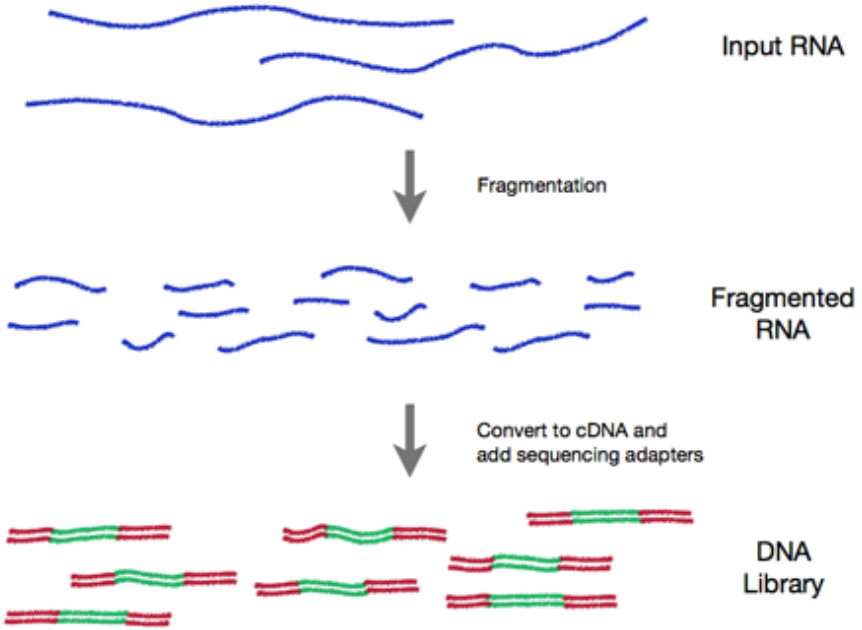
Image source: http://cureangelman.org/understanding-angelman/testing-101/

Experimental Procedure of RNAseq

- Generally similar to DNA sequencing
- Over 20.000 single stranded RNAs in variable abundance (1-k times) of 1.500-2.000 nt
- Fragmented into 30-200 nt

Differences:

- RNA is single stranded and needs to be revers-transcribed to DNA for sequencing
- Coverage is dependent on expression value of gene



Q22: Differential gene expression analysis is:

a) Used to compare different sample or patient groups
b) Performed on count data from RNAseq experiments
c) A method to find SNP's in RNAseq data
d) A model-based clustering method

- What caused the confusion to pick answer d?