RNA-SEQUENCING DATA ANALYSIS

Project Workflow

GeneVia Technologies

CONTENTS

Background and aim..3

Project outline..4

1. Kick-off...5
2. Preprocessing...5
3. Differential expression analysis 7
4. Sequence analyses 8
5. Predictive modeling...9
6. Reporting and documentation..10

BACKGROUND AND AIM

This is a generic project plan for analyzing RNA-sequencing data. Such a data set may comprise samples from a setting such as treated vs. untreated, disease vs. healthy control, gene-edited model vs. wild type, or different tissues or time points. The aim of this project is to study transcriptome-wide differences between the samples on the level of genes and pathways.

In the following section, we walk you through the main steps of the workflow. While this project plan is not an exhaustive list of all possible analyses for RNA-seq data, it does adapt to most research questions that we have encountered, and many of the analyses also apply to other expression data modalities such as expression arrays, Nanostring and proteomics. We hope this gives you ideas on how gene expression analyses could drive your research forward!

PROJECT OUTLINE

An overview of the analysis is given in ***Figure 1***. This project plan is adapted to your experimental setting, used library preparation protocol and research questions. It is likely that some of the listed analyses are impossible or do not make sense with your data. For most steps in the workflow, there are multiple available tools, of which the most suitable one is selected.

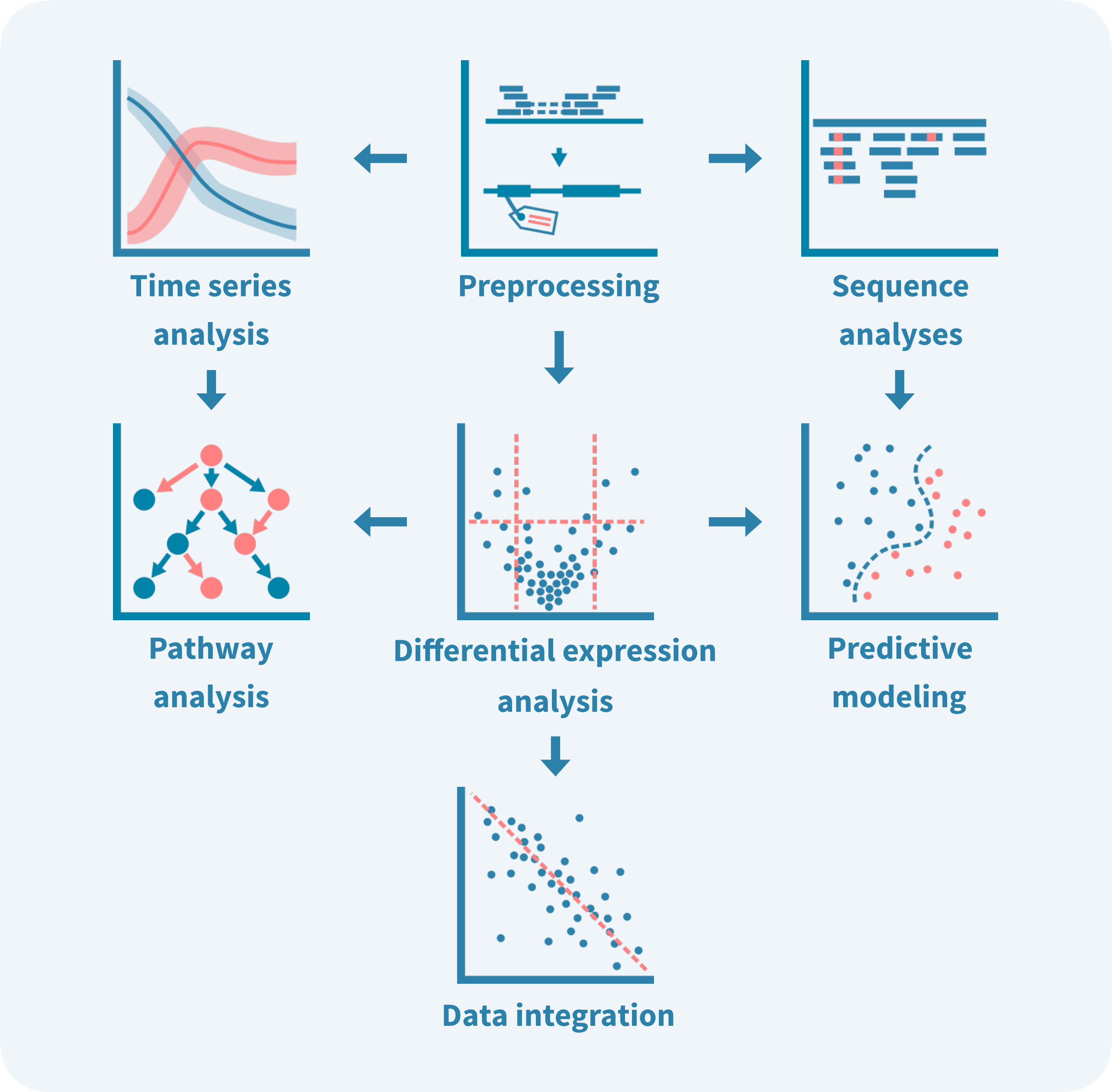

Figure 1. *The RNA-seq data analysis workflow*

1. KICK-OFF

The aim of this phase is to ensure that we agree on the overall goal of the project, that we have all required material to implement the analysis plan, and to sort out practicalities such as the schedule and preferred modes of communication.

1) Sequencing data files and a sample sheet with all pertinent metadata are delivered to us via
    - a download link to a data repository, or
    - a hard drive (Genevia Technologies Oy, Hämeenkatu 14 C 33, 33100 Tampere, Finland)
2) The project team reviews the delivered files and relevant background information
3) A kick-off teleconference is held with your contact person and our project manager to discuss
    - the overall goals of the project,
    - any technical or practical details and preferences, and
    - the schedule and deliverables of the first batch of work

2. PREPROCESSING

The aim of preprocessing is to ensure that the data is of sufficient quality for the planned analyses, to mitigate potential quality issues and confounding factors, and to render the data into a format required by downstream analyses. If data quality is deemed insufficient, we agree with the customer on 1) discarding low-quality data from further analysis, 2) putting the project on hold while new, better-quality data is generated, or 3) terminating the project.

Note that some data quality issues may become obvious only after downstream analyses, requiring revisiting this phase and readjusting quality thresholds. Some of these preprocessing steps may be carried out only for specific downstream analyses. The workflow is shown in ***Figure 2***.

4) Read quality control (QC) metrics are computed for each sample, reads are quality-trimmed and adapter sequences are removed
5) If reference genome is not available:
    - *de novo* transcriptome assembly is performed using e.g, Trinity
    - quality of the transcriptome is assessed using, e.g., BUSCO and TransRate
    - the transcriptome is annotated using e.g., Trinotate
6) Reads are aligned to the reference using, e.g. STAR
7) Alignment QC metrics are computed for each sample (e.g., alignment rates, duplication rates, rRNA contamination)
8) A gene expression table is formed with normalized expression levels for all genes in all samples, using annotations from e.g., GENCODE or RefSeq
9) Expression levels of individual copies of all annotated repeat elements (or transposable elements, TE) can be quantified using e.g., RepeatMasker
10) Downstream QC analyses are conducted to identify outliers or potential batch effects (e.g., sample-wise correlation, hierarchical clustering, principal component analysis (PCA))

11) If samples represent tissues with a mixture of cell types, cell type deconvolution can be performed
12) A teleconference is held to
    - report data quality and discuss potential quality issues
    - agree on the next batch of analyses and their schedule

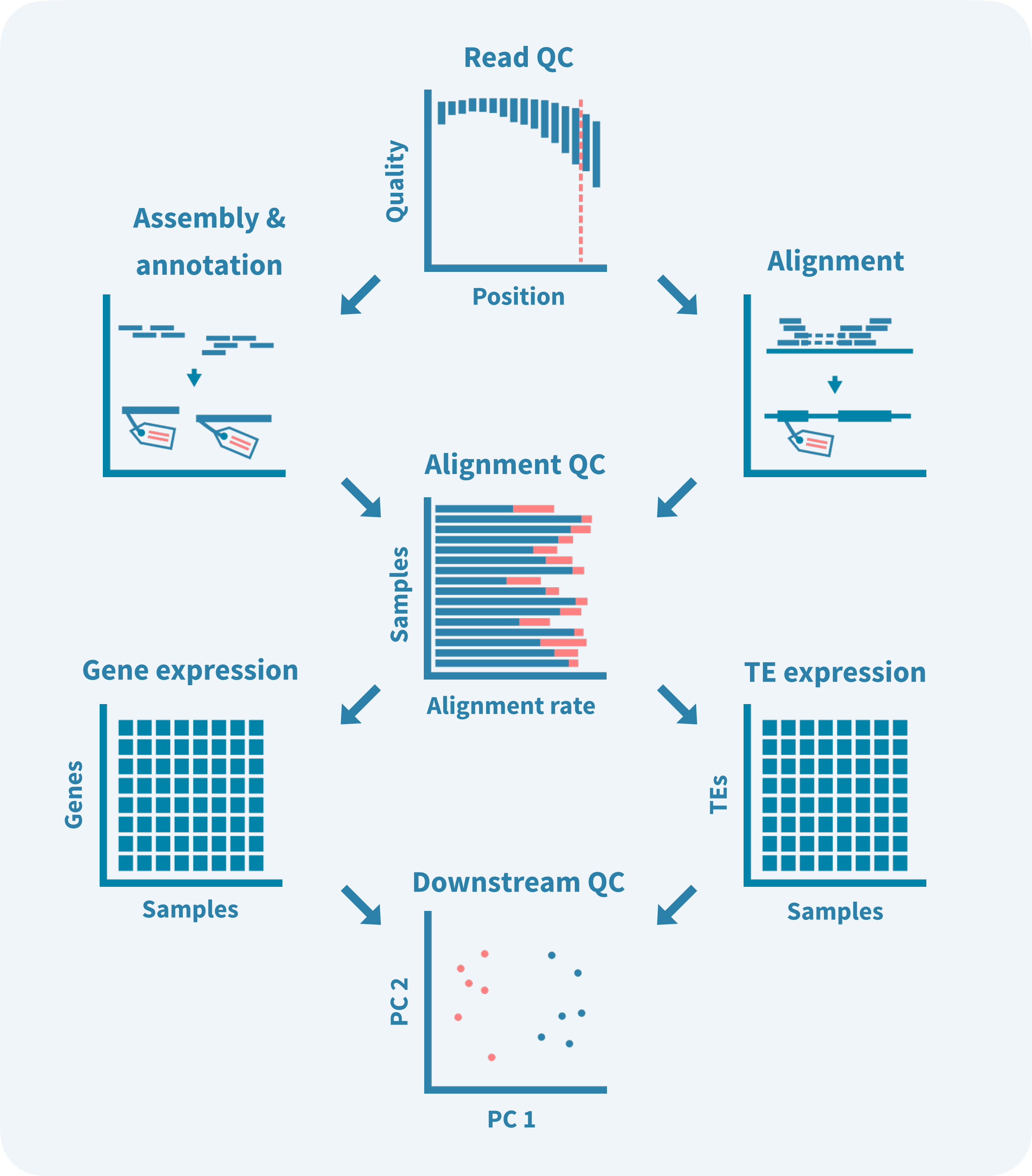

Figure 2. The preprocessing workflow

3. DIFFERENTIAL EXPRESSION ANALYSIS

The aim of differential expression (DE) analysis is to associate differences between sample groups to genes, gene sets and pathways. ***Figure 3*** describes the workflow.

1) Adjusted *p*-values for differential expression are computed for genes between any two sample groups of interest
2) DE genes are visualized with volcano plots, box plots, heatmaps and Venn diagrams
3) Biological interpretation to sets of DE genes are given by computing enriched processes and pathways with Gene Ontology enrichment analysis
4) Further regulation analyses are run and visualized using Ingenuity Pathway Analysis (IPA):
   - enrichment analysis of canonical pathways, diseases or functions
   - analysis of upstream regulators and downstream effects
   - comparison of pathway activation profiles to those of public data sets
5) A teleconference is held to
   - present and discuss the results of DE and pathway analysis
   - agree on the next batch of analyses and their schedule

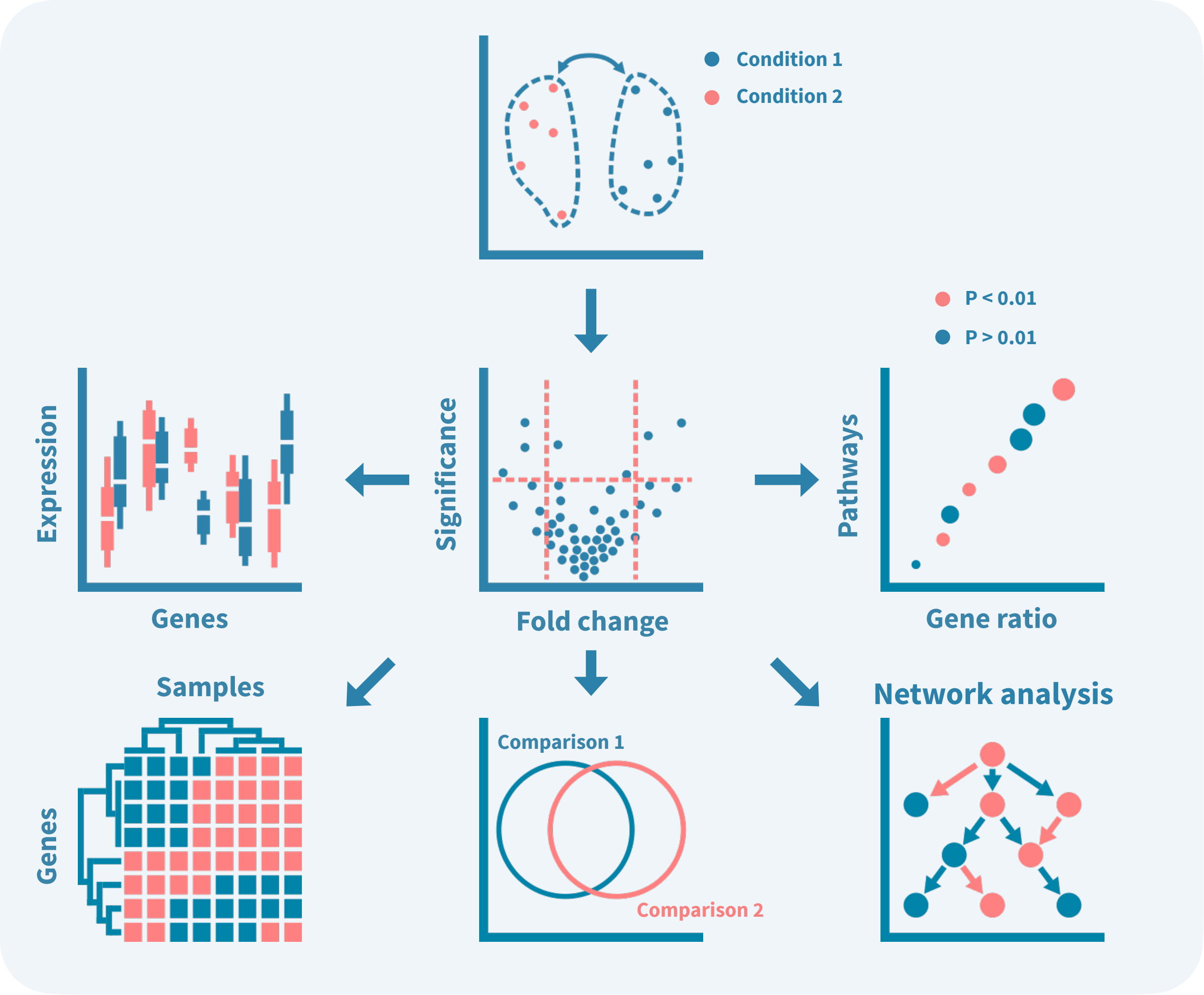

Figure 3. *The differential expression analysis workflow*

4. SEQUENCE ANALYSES

If the amount and quality of RNA-seq reads are sufficient, sequence-level analyses may be performed in addition to those based on gene-level expression estimates. Confidence in the identified sequence-level events depends heavily on the read coverage, and thus, abundance of the corresponding transcript.

1) Alternative splicing events can be called between samples
2) Fusion genes may be identified by studying discordantly mapping read pairs
3) Variants or RNA-editing events can be called from sufficiently abundant transcripts and compared between samples
4) Allelic expression can be studied by phasing reads by haplotype
5) For T cells and B cells, the TCR and antibody repertoires, respectively, can be studied
6) A teleconference is held to
   - present and discuss the results of the sequence analyses
   - agree on the next batch of analyses and their schedule

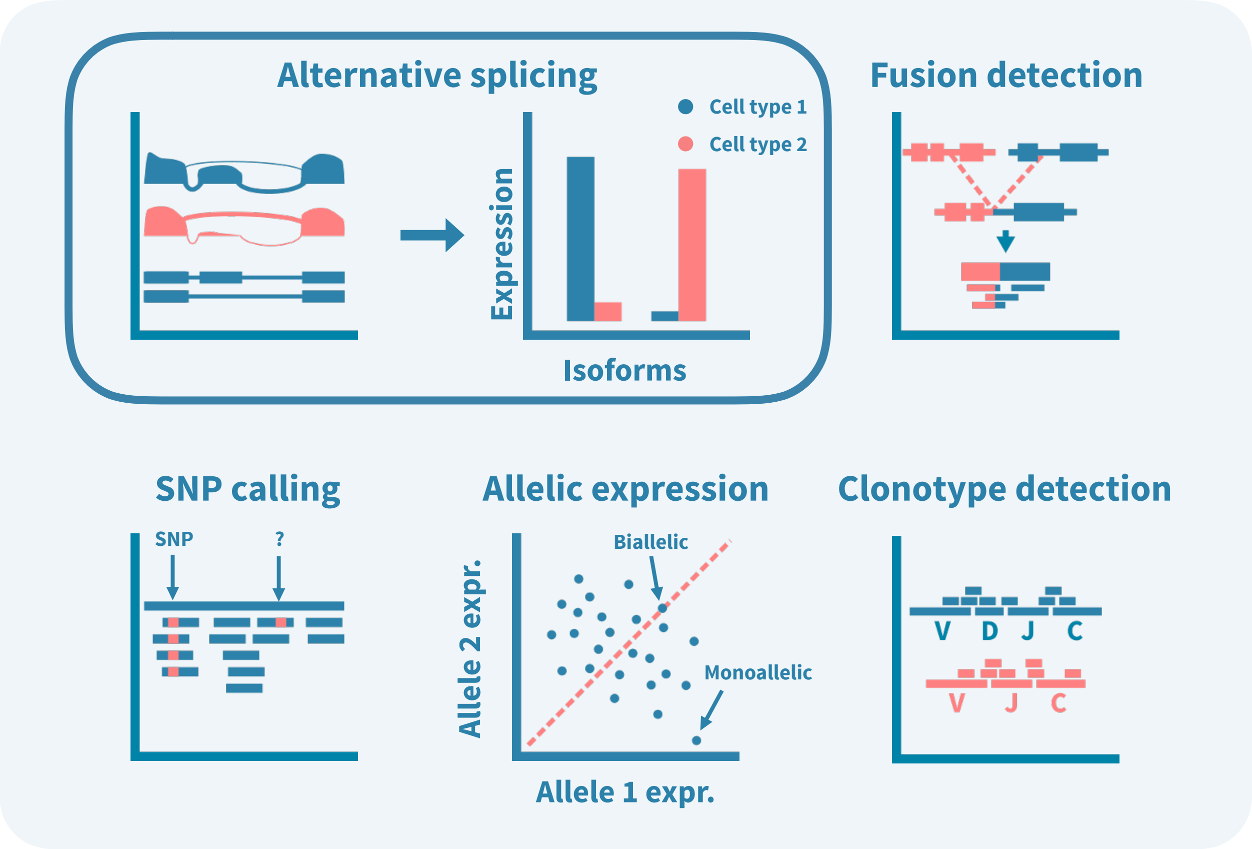

__Figure 4.__ Sequence-level transcriptomic analyses

5. PREDICTIVE MODELING

Often times, RNA-seq data sets come with additional data, such as other ‘omics measurements or phenotypic or clinical sample data. While approaches to correlate different data modalities are next to limitless in number, many of them fall under the umbrella category of predictive modeling. The aim here is to train a computational model to predict a variable of interest from the high number of molecular markers generated in the previous steps.

1) The prediction problem is agreed upon — it depends on the available data modalities and research questions, but could be
   - predicting sample phenotype or clinical variable (e.g., survival or treatment response) from expression levels or sequence-level events (***Fig 5***)
   - predicting protein expression (as quantified by mass spectrometry from the same samples) from expression levels or sequence-level events
   - predicting gene expression from epigenetic markers (as measured by ATAC-seq or BS-seq from same samples) or vice versa
2) A panel of predictive models are trained
   - training may be preceded by feature selection to ease computational burden
   - model types may range from, e.g, regression and decision-tree based ones to deep learning
   - both hyper parameters, i.e. model structures, and training parameters are optimized in a cross-validation scheme to avoid overfitting the model to training data
   - in addition to cross-validation, an independent validation data set, if available, is used to assess performance of the models
3) A teleconference is held to
   - present and discuss the results of predictive analysis
   - agree on the next batch of analyses and their schedule

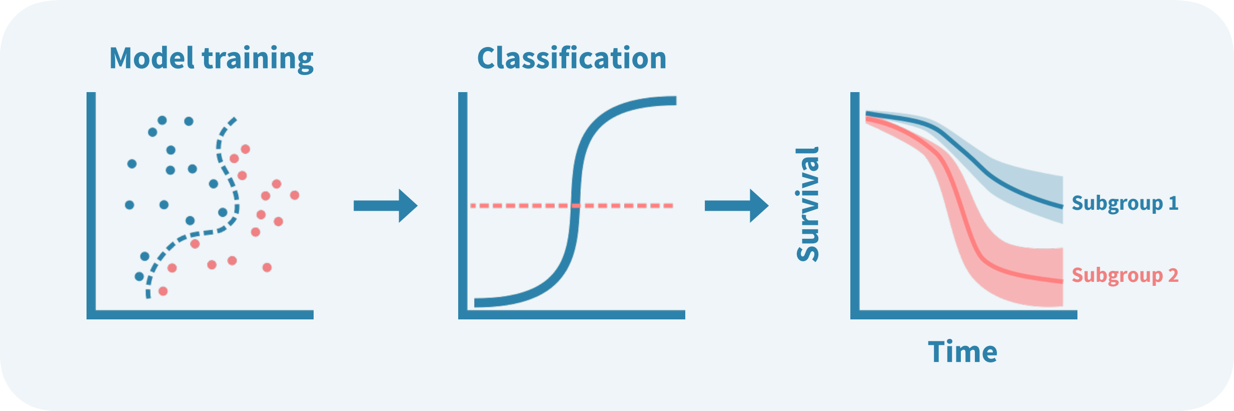

Figure 5. *Biomarkers and predictive signatures can stratify patients into prognostic subgroups*

6. REPORTING AND DOCUMENTATION

None of the above-listed analyses make much sense unless the delivered figures, tables and analysis methodology is described in sufficient detail to allow further communication of the results in, for instance, a peer-reviewed research article.

1) All deliverables are presented and discussed in teleconferences, and any analysis can be re-run using alternative approaches or different parameters
2) All computational methods are described at the level required by high-quality scientific publishing
3) The work is summarized in a project report, allowing a researcher not involved with the project to understand and use the results later on
4) After the project, we can review the ensuing manuscript to ensure that the bioinformatics part is correct, and we can reply to bioinformatics-related reviewer comments
5) If requested by the journal upon submitting a manuscript, we can deliver any code written during the project