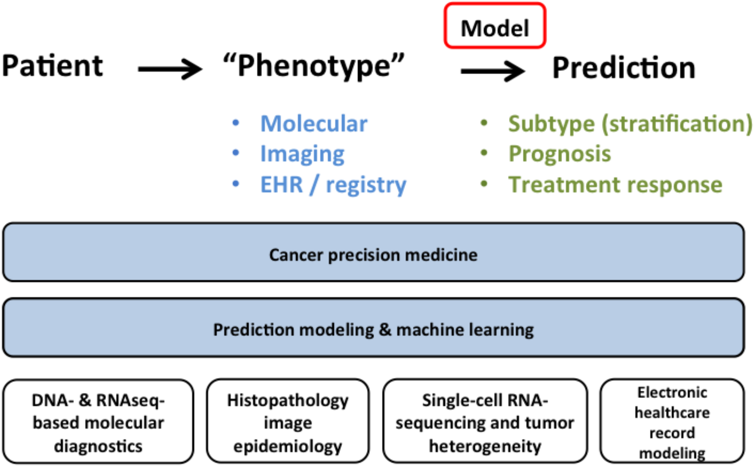
Biomarkers and disease subtyping from an epidemiological perspective

Mattias Rantalainen, PhD
Department of Medical Epidemiology and Biostatistics
Karolinska Institutet

EFSPI / FMS meeting, 20th of March 2019

Predictive medicine group @ MEB



Model

Patient → "Phenotype" → Prediction

- Molecular
- Imaging
- EHR / registry

- Subtype (stratification)
- Prognosis
- Treatment response

Cancer precision medicine

Prediction modeling & machine learning

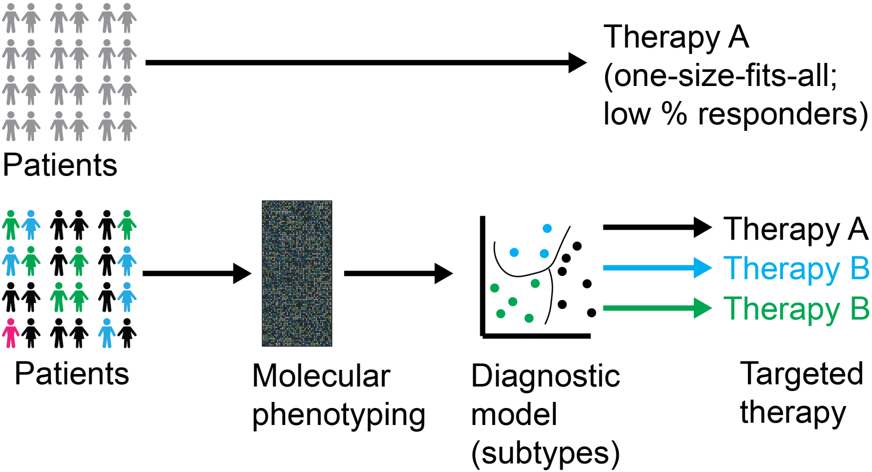
DNA- & RNAseq-based molecular diagnostics

Histopathology image epidemiology

Single-cell RNA-sequencing and tumor heterogeneity

Electronic healthcare record modeling

Precision medicine & patient stratification



Patients

Therapy A
(one-size-fits-all;
low % responders)

Patients

Molecular
phenotyping

Diagnostic
model
(subtypes)

Therapy A

Therapy B

Therapy B

Targeted
therapy

Biomarkers and subtypes

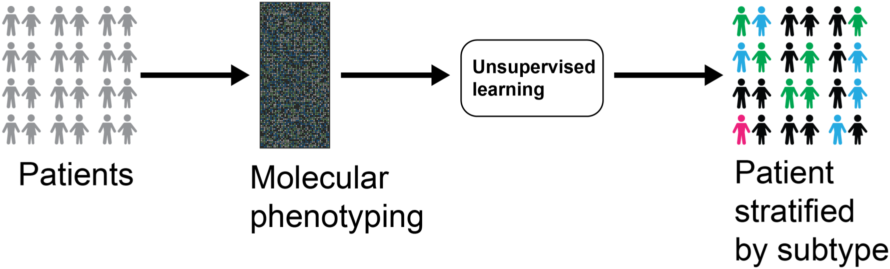
Definition: Biomarker

A measurable entity that is related to a biological state

Definition: Subtype

Distinct group of the patient population that can be defined by phenotypic data (e.g. molecular phenotypes)

Two main modes of subtype discovery:
1) Data-driven unsupervised subtype discovery



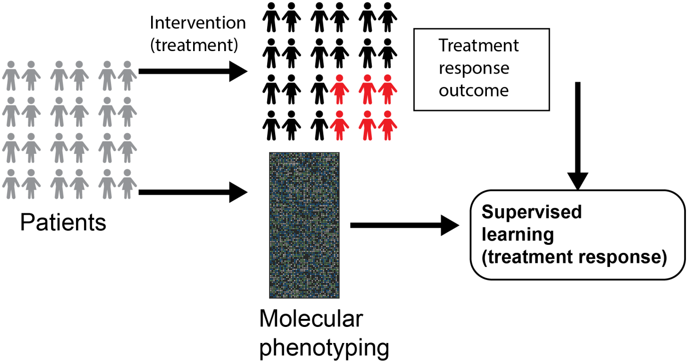
Patients

Molecular phenotyping

Unsupervised learning

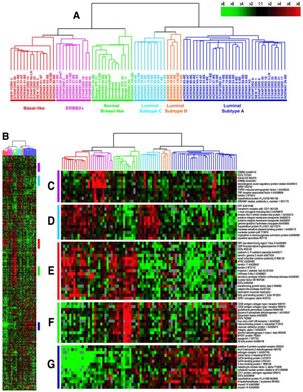
Patient stratified by subtype

Two main modes of subtype discovery:
2) Outcome-driven supervised "subtype" discovery

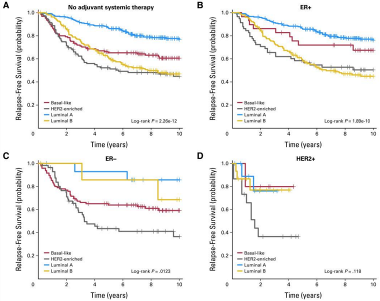


Intervention (treatment)
Treatment response outcome
Patients
Molecular phenotyping
Supervised learning (treatment response)

Example: Early success story - Intrinsic subtypes of breast cancer; data-driven discovery; subtypes associated with treatment response and prognosis

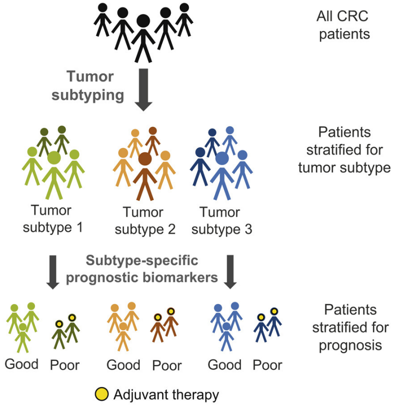


Srlie T, Perou CM, Tibshirani R, et al. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. Proc Natl Acad Sci U S A. 2001;98(19):10869-74.



Parker JS, Mullins M, Cheang MC, et al. Supervised risk predictor of breast cancer based on intrinsic subtypes. J Clin Oncol. 2009;27(8):1160-7.

Example: Subtype-specific prognostic models to guide adjuvant therapy in colorectal cancer



All CRC patients

Tumor subtyping

Patients stratified for tumor subtype

Tumor subtype 1

Tumor subtype 2

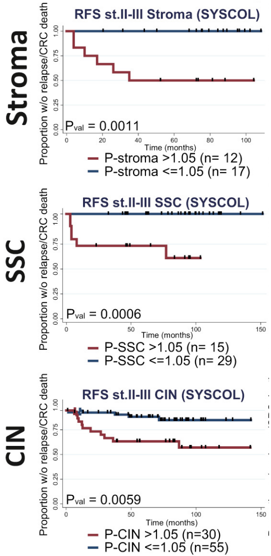
Tumor subtype 3

Subtype-specific prognostic biomarkers

Good Poor Good Poor Good Poor

Patients stratified for prognosis

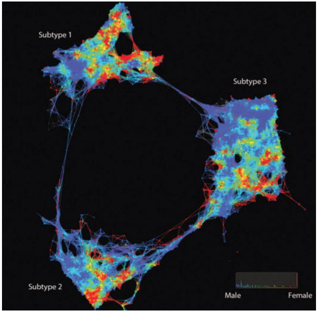
Adjuvant therapy



Bramsen, J.B. et al., 2017. Molecular-Subtype-Specific Biomarkers Improve Prediction of Prognosis in Colorectal Cancer. Cell reports, 19(6), pp.12681280.

Example: Identification of type 2 diabetes subgroups through analysis of patient similarity in electronic medical records

- Subtype discovery based on 73 clinical features from EMRs
- Subtypes associated with SNPs and with distinct co-morbidities



Li, Li, et al. Identification of type 2 diabetes subgroups through topological analysis of patient similarity. Science translational medicine 7.311 (2015): 311ra174-311ra174.

Common limitations in subtype studies

- Poor study design (study do not reflect the actual patient population; poorly characterised patients)
- Too small study size
- Mainly descriptive analysis
- Lack of proper validation (internal or external)

What can epidemiological approaches contribute with?

- **Study design** (ensuring population representative studies)
- Opportunity to ascertain prevalence of subtypes (cohort studies)
- Opportunity to estimate RR or HR associated with biomarker or subtype status (cohort studies)
- + Comprehensive baseline characteristics and (registry-based) outcomes

Subtype discovery and validation process in the CLINSEQ-AML study (Acute myeloid leukemia)

1. Mer, Arvind Singh, et al. "Expression levels of long non-coding RNAs are prognostic for AML outcome." Journal of hematology & oncology 11.1 (2018): 52.
2. Wang, M., et al. "Validation of risk stratification models in acute myeloid leukemia using sequencing-based molecular profiling." Leukemia 31.10 (2017): 2029.
3. Wang, Mei, et al. "Development and Validation of a Novel RNA SequencingBased Prognostic Score for Acute Myeloid Leukemia." JNCI: Journal of the National Cancer Institute 110.10 (2018): 1094-1101.

Sequencing-based cancer diagnostics

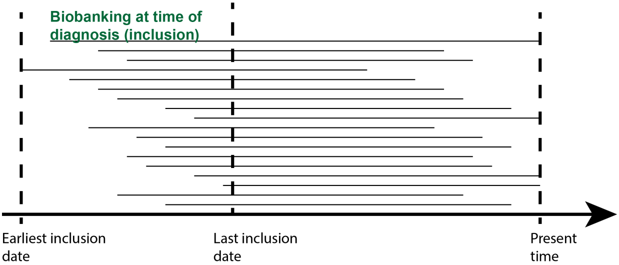
- **RNAseq (25M reads per tumour)**
  - **Expression profiles (for subtyping)** (including lncRNA)
  - Validation of point mutations
- Low-pass whole genome sequencing (0.5-1x coverage)
  - CNV profile
- Panel DNA sequencing (650 genes, >300x average coverage)
  - Point mutations and indels
  - Pharmacogenomic loci
  - Germline risk variants

(Molecular phenotyping is based on biobanked material taken prior to treatment. Comprehensive clinical information and registry-based outcomes are available.)

Study design

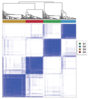
Retrospective cohort of AML patients (N=274).

- Population representative and well-characterised cohort
- Possible to estimate subtype prevalence
- Possible to estimate subtype prognosis (HR)
- Reduced risk for selection bias (improved generalisability of models)

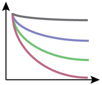


Overview of our discovery and validation process

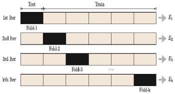
1) Robust subtype discovery



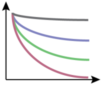
2) Assessment of prognostic stratification



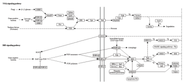
3) Internal validation of the consistency
in subtype assignments



4) External validation (time-to-event)



5) Descriptive analysis
(molecular pathways)



Subtype discovery through clustering

Relative importance of different factors in the discovery process (from experience):

study design $>$ features $>$ dissimilarity metric $>$ clustering algorithm

Objectives:

- Estimate the number of clusters, K, from data, $(\mathbf{X})$
- Learn the subgroups, from data $(\mathbf{X})$

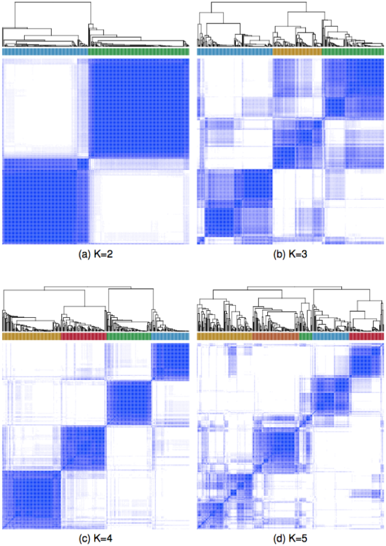
Consensus clustering offer improved robustness

- Ensemble-based approach to clustering
- Improved stability of clusters (robustness against noise as well as starting conditions)

Algorithm 1 Pseudo-code: Consensus clustering

```
for $k = 1 : k_{max}$ do
    while $i < i_{max}$ do
        $a_{sub} =$ Draw subsample of observations
        $b_{sub} =$ Draw subsample of features
        Clustering: $f(X_{a_{sub},b_{sub}}, k)$ (e.g. k-medoids or iCluster)
        Save results into the consensus matrix $[N \times N]$
    end while
    Save final consensus matrix for $k$
end for
Determine optimal $k$ (model selection)
```

1. Consensus clustering ($k = 2 \ldots 5$)

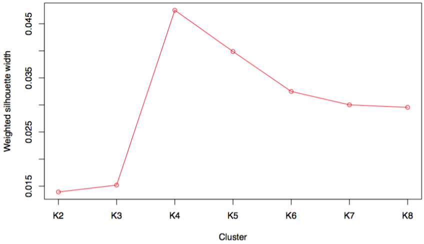


(a) K=2

(b) K=3

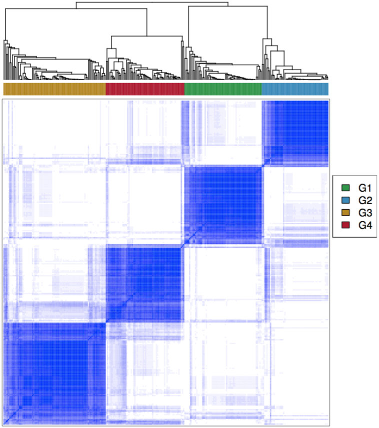
(c) K=4

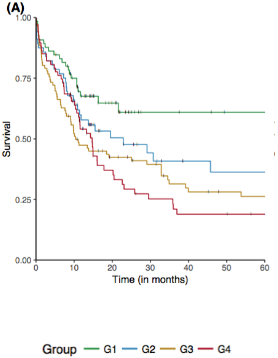
(d) K=5

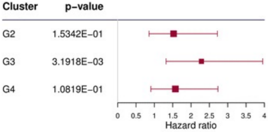


The average silhouette index was used to determine number of clusters (K) (the metric is a function of distance difference within allocated cluster compared to points in the other clusters)

2. Prognostic stratification by subtypes (OS and EFS)



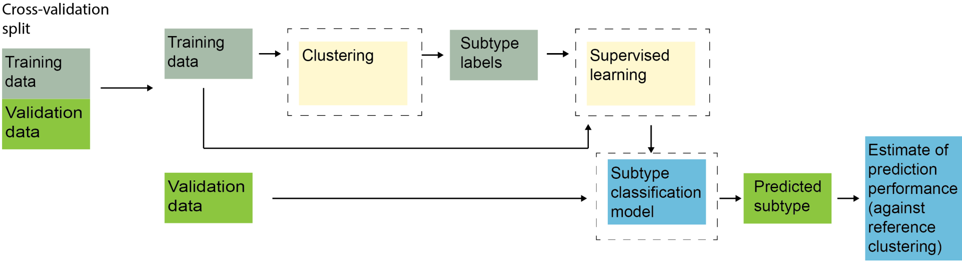




Cluster	p–value
G2	1.5342E–01
G3	3.1918E–03
G4	1.0819E–01

(Cox-PH model adjuste for age, gender, aetiology, ELN

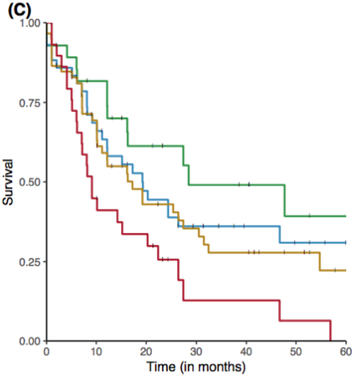
3. Internal validation of consistency in subtype prediction provide some ascertainment of the reproducibility of the subtypes



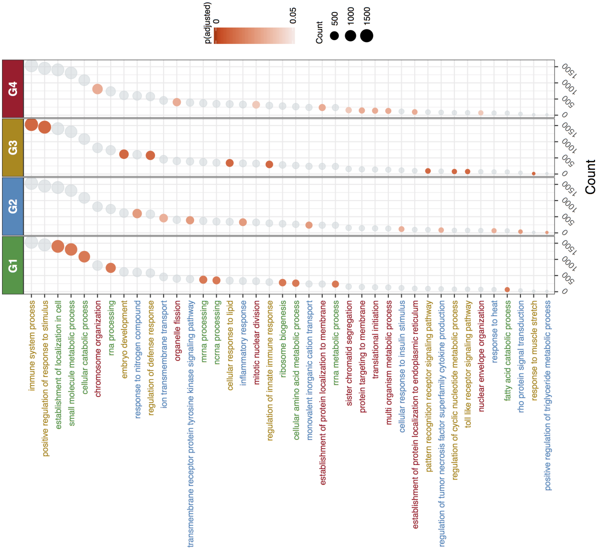
Subtype classification accuracy of 0.85 (cross-validation of the whole discovery process)

4. External validation of prognostic stratification (TCGA-AML)

1. Subtype discovery in primary study (CLINSEQ-AML)
2. Supervised subtype classifier optimised on primary study (CLINSEQ-AML)
3. Subtype classifier applied to external study (TCGA-AML)
4. Validation of prognostic stratification by predicted subtype in external study (TCGA-AML)



5. Qualitative interpretation - pathway analysis by subtype



Examples of commonly observed limitations in biomarker and subtype focused studies

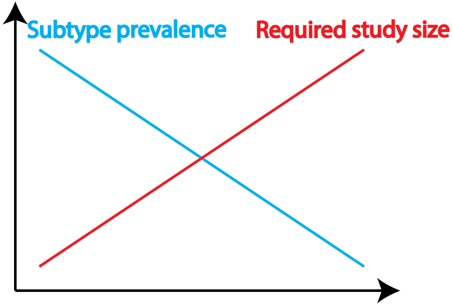
Many reported subtype discovery studies (and outcome associations) may not be reproducible

There are many published studies that simply has **no validation results**, or only have some descriptive or qualitative analysis.

Proper validation of reported subtypes would almost surely help the field forward:

1. Internal empirical assessment of whether cluster labels can be predicted (supervised learning)
2. External retrospective validation
3. Prospective validation

In data-driven subtype discovery study size will limit detection of subtypes with low prevalence



In diseases with potentially many, but low prevalence subtypes, this will limit efficient data-driven subtype discovery

In cancer focused studies, the "bulk" level average subtype often masks intra-tumour subtype heterogeneity

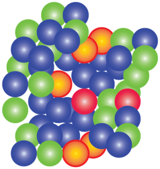
Genomic instability

Tumour evolution

Mutations

Stochastic variability

Cellular hierachies



Intra-tumour heterogeneity

↓

Bulk-average subtype dominated by the most common subtype

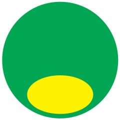
Generalisability of subtype stratification models can be limited by several factors

PROBLEM

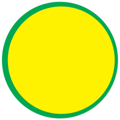
SOLUTION

Selection bias
in discovery data

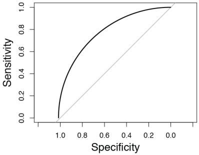
Biased sample

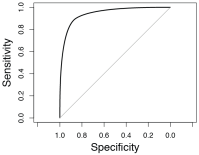


Representative sample



Poor or
overfitted
model





Lack of
robustness

Too limited training data or
noise / outlier sensitive model

Larger study and
more robust models

Sensitivity to
assay-related
effects

Assay related batch effects
unaccounted for

Improved pre-processing
or standardization of assays

Can generalisability of subtype models improve by translation to rule-based classifiers?

A classifier based on a set of K rules, of the type $x_i < x_j$, reduce the dependency on perfect normalization of data.

Algorithm 2 Pseudo-code: training of a rule-based model using data, $\mathbf{X}$, and subtype labels, $\mathbf{y}$

```
$\theta = f(\mathbf{X}, \mathbf{y})$ (Heuristic search to establish a set of decision rules, $\theta$, based on pairs $(i, j)$ of features in $\mathbf{X}$ to maximise class separation)
$\mathbf{I}_x = f(\mathbf{X}, \theta)$ (Apply discovered rules to generate, $\mathbf{I}_x$)
Train a classification model on $\mathbf{I}_x$ and labels, $\mathbf{y}$ (e.g. Naive Bayes)
```

Eric R. Paquet, Michael T. Hallett; Absolute Assignment of Breast Cancer Intrinsic Molecular Subtype, JNCI: Journal of the National Cancer Institute, Volume 107, Issue 1, 1 January 2015,

Concluding remarks

1. Study design is of central importance in the discovery phase (improve generalisability, minimize selection bias, enable time-to-event analyses, etc.)
2. Data-driven subtype discovery should probably be performed in well-defined cohorts (followed by potential trial post-hoc analyses)
3. Validation and demonstration of generalisability necessary

Acknowledgements

Current team members:

Yinxi Wang
Bojing Liu
Youcheng Zhang
Boxi Zhang
Balasz Acs
Charlotte von Heijne Widlund
Abhinav Sharma
Alva Sandin

KAROLINSKA INSTITUTET ANNO 1810
Karolinska Institutet

Vetenskapsrådet

SeRC
Swedish e-Science Research Centre

ecpc
e-Science for Cancer Prevention and Control

Cancerfonden