Bayesian Clustering of Multi-Omics

for Cardiovascular Diseases

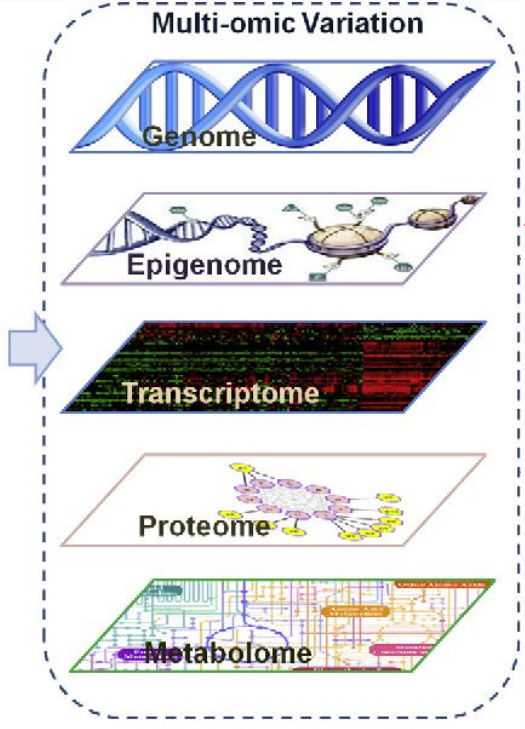
Nils Strelow

22./23.01.2019
Final Presentation Trends in Bioinformatics WS18/19

Recap

Intermediate presentation

Precision Medicine



High-throughput multi-omic information

Multi-Omics
Machine Learning

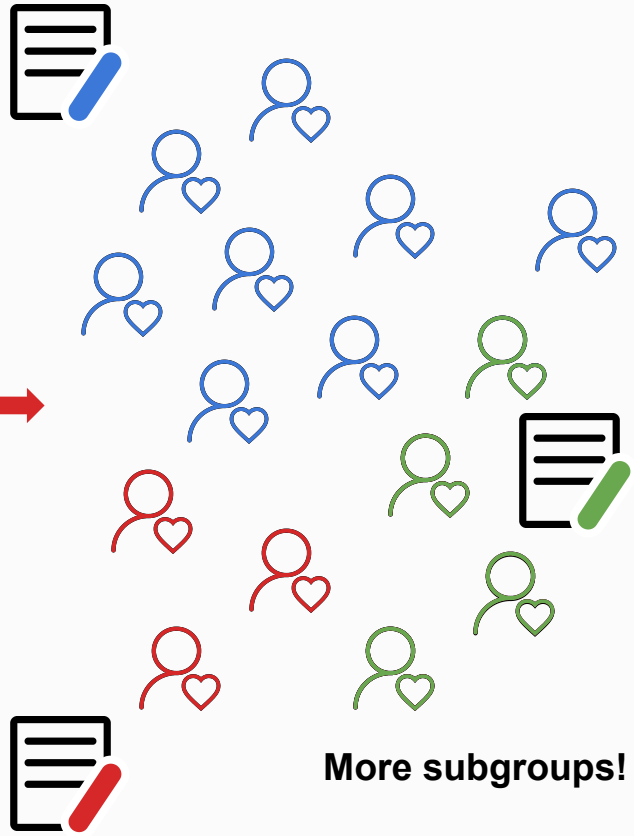


Image: Yan V. Sun, Yi-Juan Hu (2016)

iClusterBayes

- joint integrative clustering framework
- uses Bayesian latent variable regression models

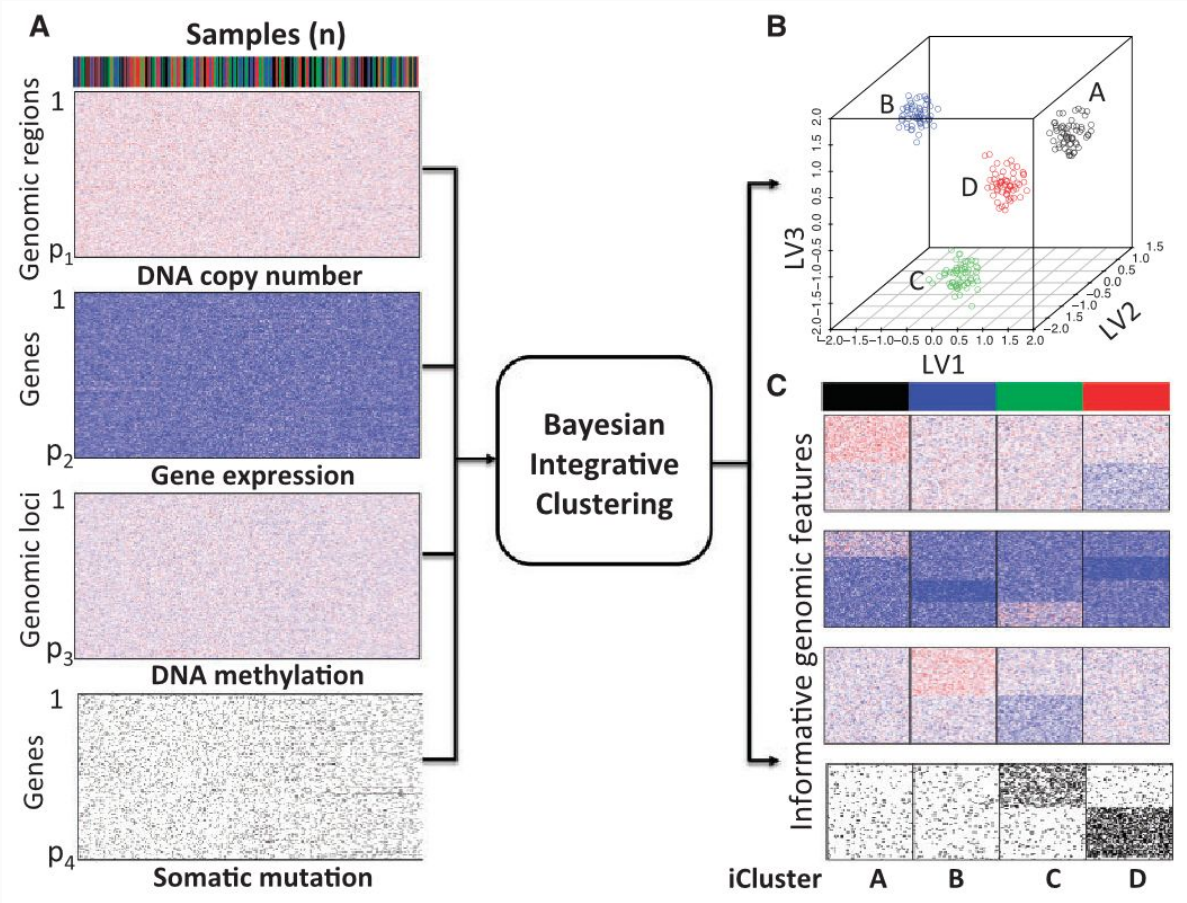
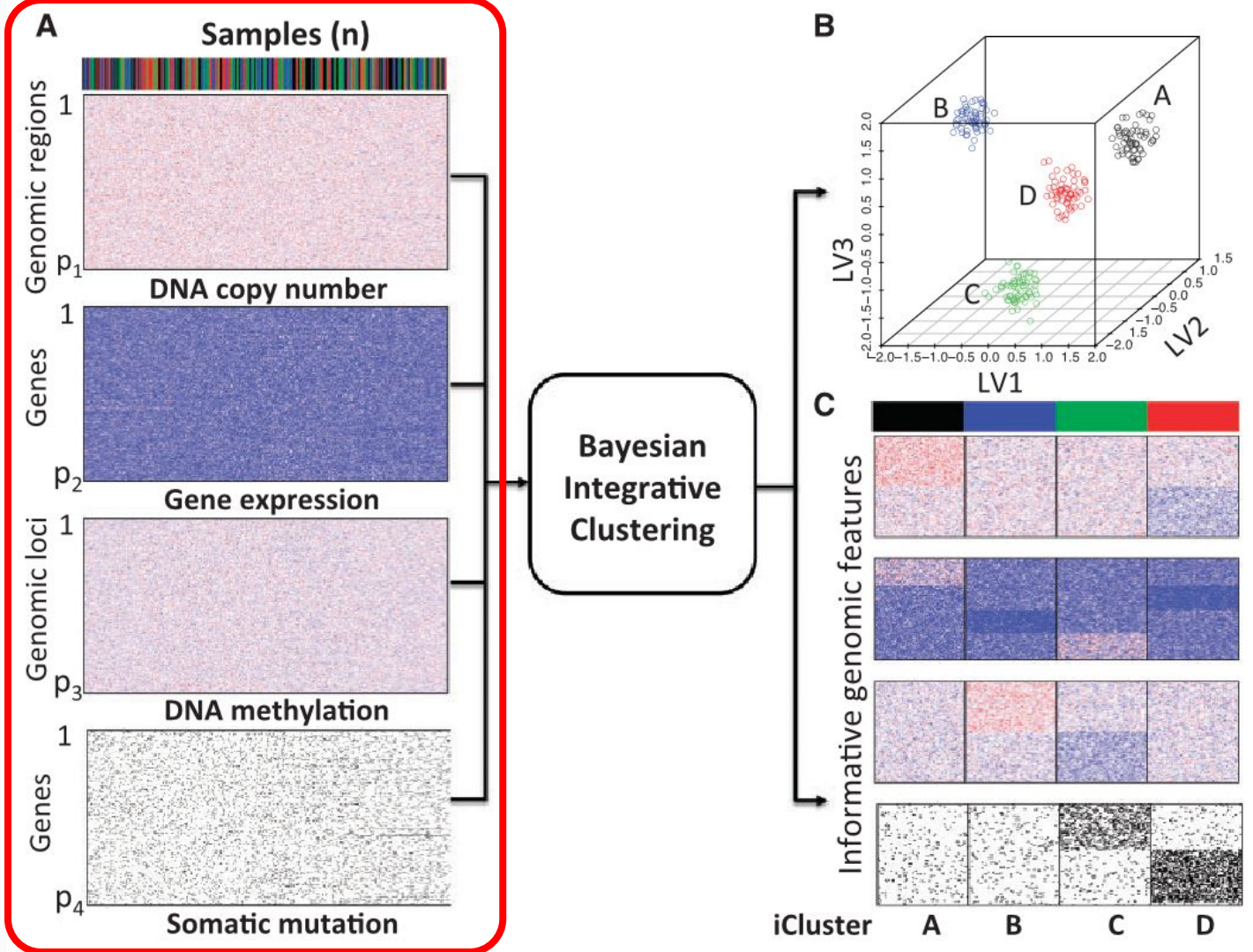


Image: iClusterBayes

Agenda

- Omics Data: Origin and significance
- Data Preparation
- Bayesian Inference
- iClusterBayes: Deep dive
- What’s next

Omics data


A
Samples (n)
Genomic regions
1
p1
DNA copy number
Genes
1
p2
Gene expression
Genomic loci
1
p3
DNA methylation
Genes
1
p4
Somatic mutation
Bayesian Integrative Clustering
B
A
B
C
D
LV1
LV2
LV3
C
Informative genomic features
iCluster
A
B
C
D

Genomics

"The sequence, structure and function of the entire genome in a cell."

- Blueprint for transcripts (RNA)
- Sequencing
  - methods such as Shotgun, Next-Gen, Illumina
  - Differences: cost, coverage, time, number of base pairs in one read
- Alignment
  - assembled using a reference genome
  - Difference to reference genome can show
    - mutations, insertion/deletion of gene fragments
    - genetic variants (SNPs)
    - copy number variations

Not the whole genome is transcribed

Transcriptomics

"The complete set of transcripts (RNA) in a cell & their quantity, for a specific developmental stage or physiological condition"

- Used to translate the genetic code into proteins (by ribosomes)
- In contrary to genome: varies under conditions
  - disease
  - drugs

Not everything is translated into proteins,
but it still has functionality

Transcriptomics Analysis

Methods:

- Microarray: Most popular
  - Benefits: coverage, cost, high-throughput, uncomplicated analysis
  - Limits: amount of RNA required, dynamic range, semi-quantitative approach, detection of predefined transcripts
- RNA-seq: Newcomer
  - Benefits: absolute quantification of transcripts, includes variants, unknown, very short RNAs
  - Limits: cost, data storage, computational resources, complex

Led to the discovery of novel biomarkers e.g. GDF15

Proteomics

“A complex dynamic system formed by all proteins encoded by more than 20.000 genes encode proteins (3% of the DNA)”

Varies in:

- abundance
- isoform expression
- subcellular localization
- interactions
- turnover rate
- posttranslational modifications (PTMs)

More variants through splicing

Proteomics Analysis

Steps to analyze

- Proteins separated using gel
- Analyzed by mass spectrometry

Label free quantification: Newcomer

- Faster, higher flexibility in analysis and studies, less comparability of samples

Field of proteomics is still in early stages

- Only approx. 10.000 proteins can be mapped today

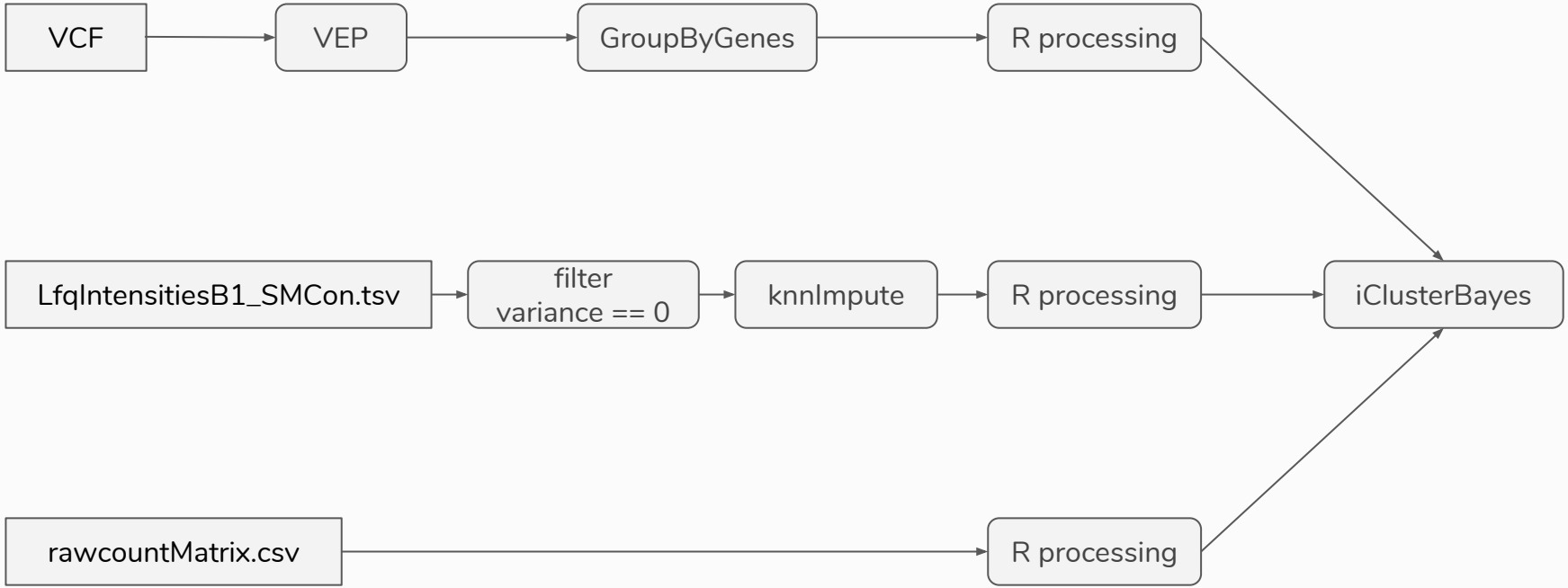
Data preparation

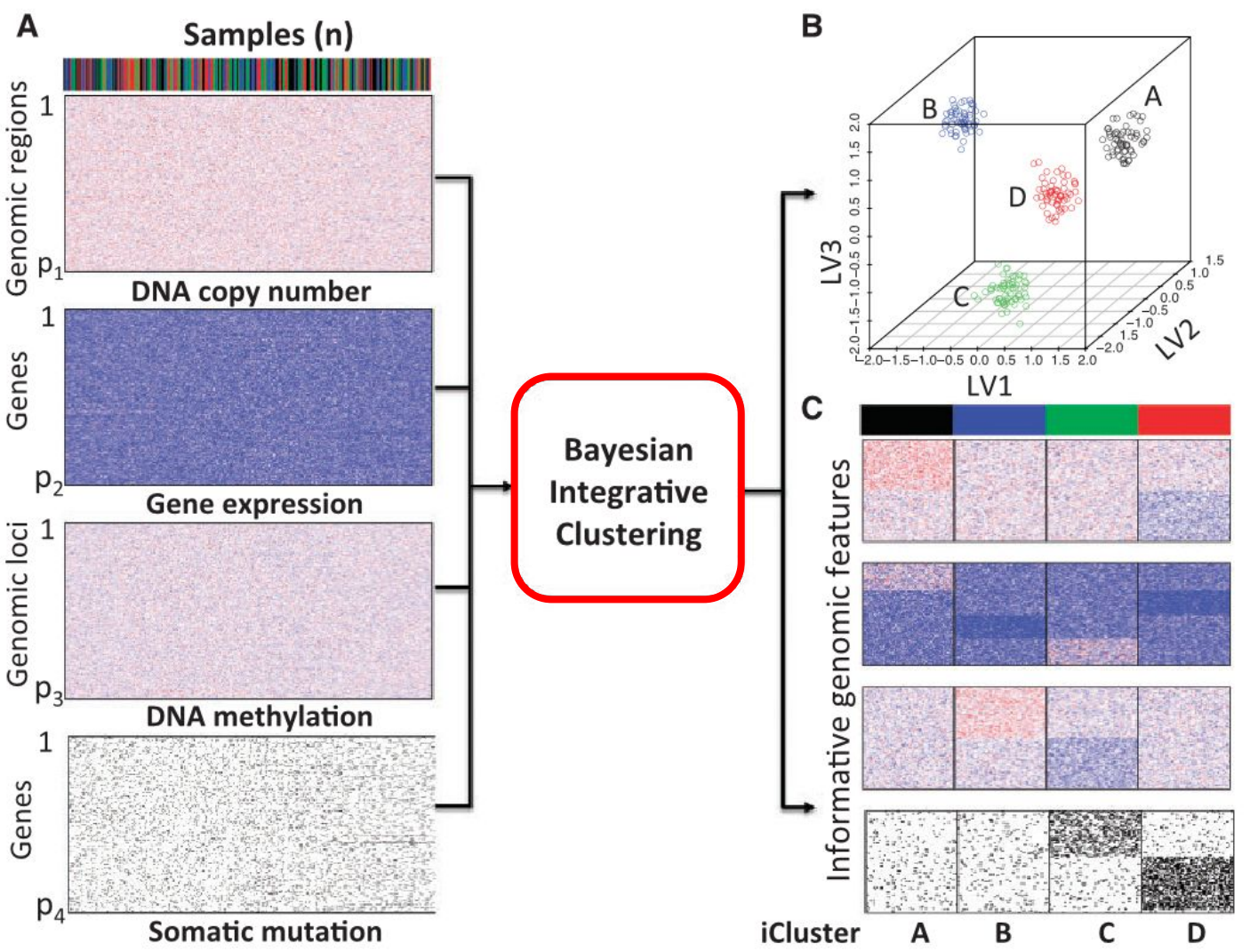
80% time of data analytics

Data sets

- Genome
  - Binary: Mutation Binary Matrix
  - 350GB VCF file (Variant Call format)
- Transcriptome
  - Count: RNA-seq gene expression data
  - 5MB
- Proteome
  - Continuous: Label free quantification measurements
  - 2MB

Data preparation pipeline



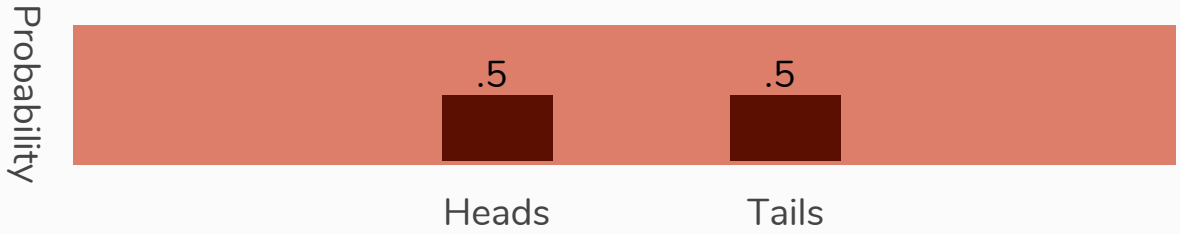

A
Samples (n)
Genomic regions
1
p1
DNA copy number
Genes
1
p2
Gene expression
Genomic loci
1
p3
DNA methylation
Genes
1
p4
Somatic mutation
Bayesian Integrative Clustering
B
A
B
C
D
LV1
LV2
LV3
C
Informative genomic features
iCluster
A
B
C
D

Bayesian Inference

Guessing in the style of Bayes

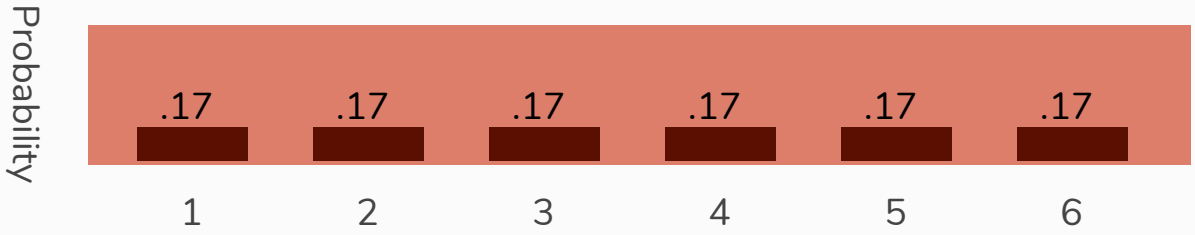
Probability distributions

Tossing a fair coin



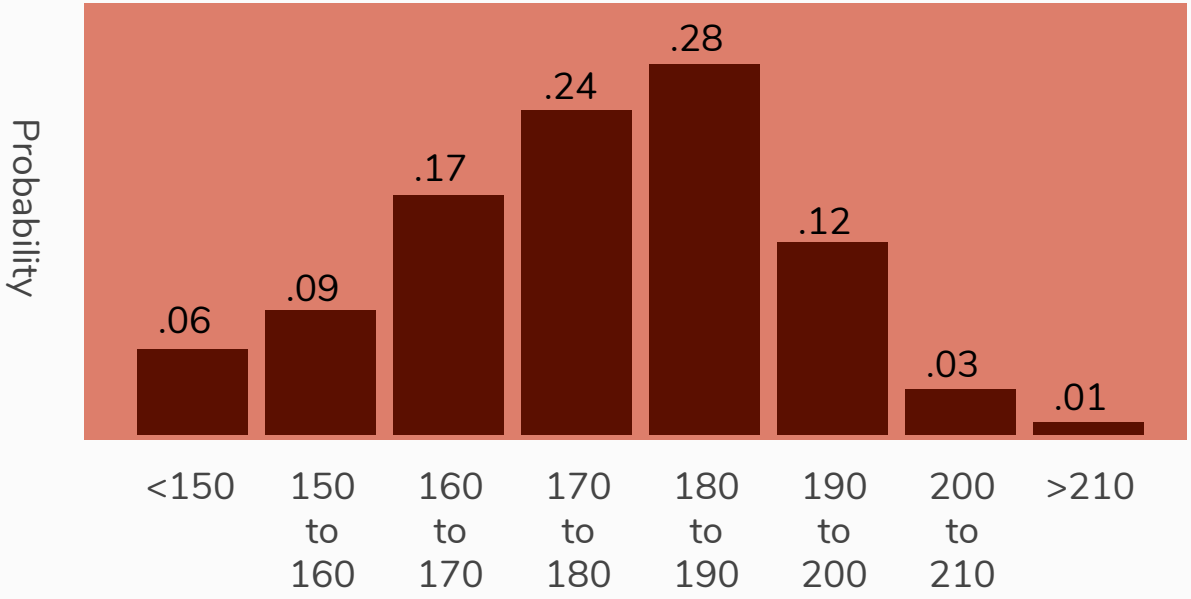
Probability distributions

Rolling a fair die



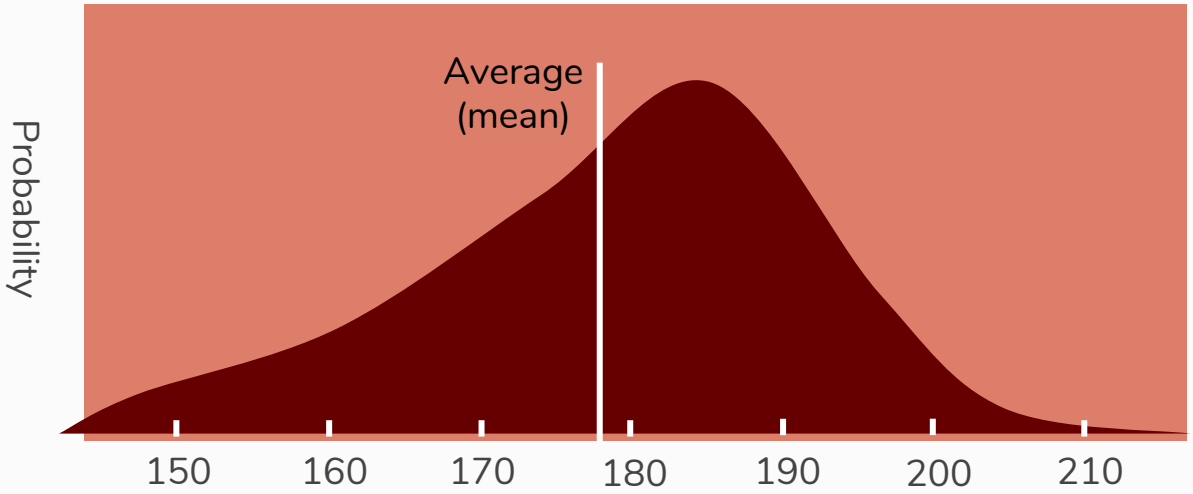
Probability distributions

Height of adults in cm



Probability distributions

Height of adults in cm



Temperature sensors in my room

2x ESP8266 with DHT22

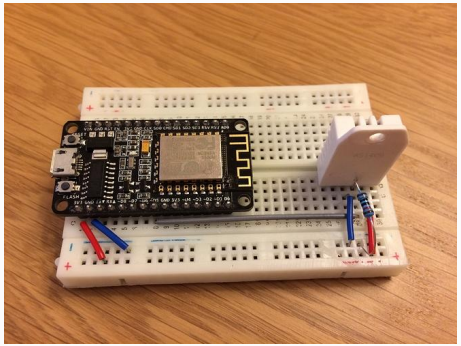
1x Bluetooth temperature sensor with display

I measure:

- 13.9
- 14.1
- 17.5

What is the actual temperature given those measurements?

⇒ P(temperature | measurements)

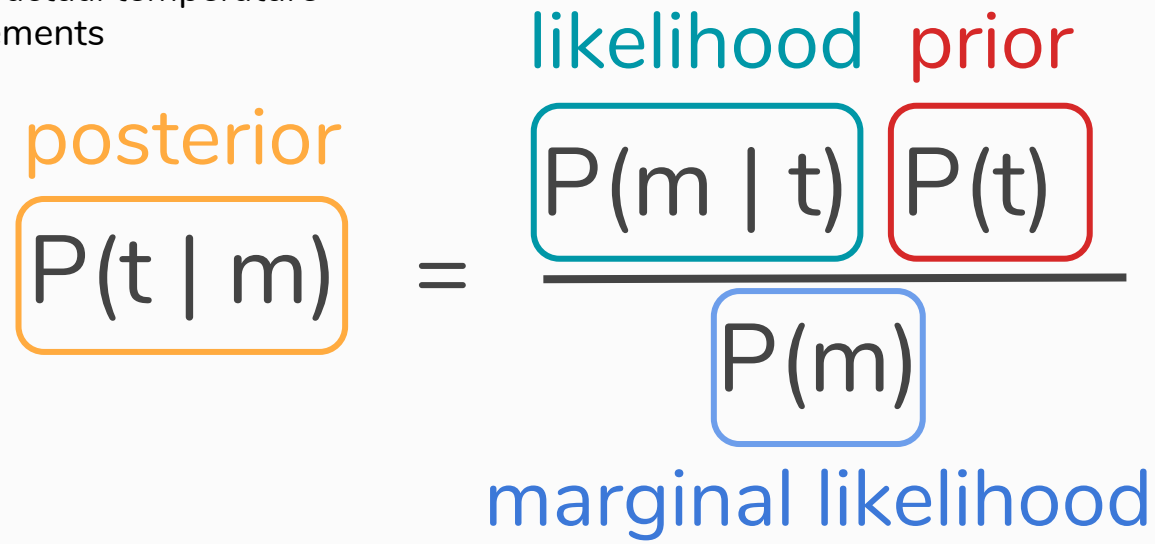


Bayes' Theorem

$$P(A \mid B) = \frac{P(B \mid A)\ P(A)}{P(B)}$$

Bayes' Theorem

t = unknown actual temperature
m = measurements



Bayesian Inference

"process of deducing properties (parameters) about a population or probability distribution from data using Bayes' theorem."

1. Generate random t according to the distribution of prior P(t)

2. Calculate the posterior distribution by using the generated t and our measurements: P(m | t) * P(t)
   = P(m = [13.9, 14.1, 17.5] | t = 17)
   = P(m=13.9|w=17) *P(m=14.1|w=17) *P(m=17.5|w=17) *$P(t)^3$
   ⇒ returns one value of the posterior distribution

3. P(m) can be neglected, since it is constant
   P(t | m) ∝ P(m | t) * P(t)

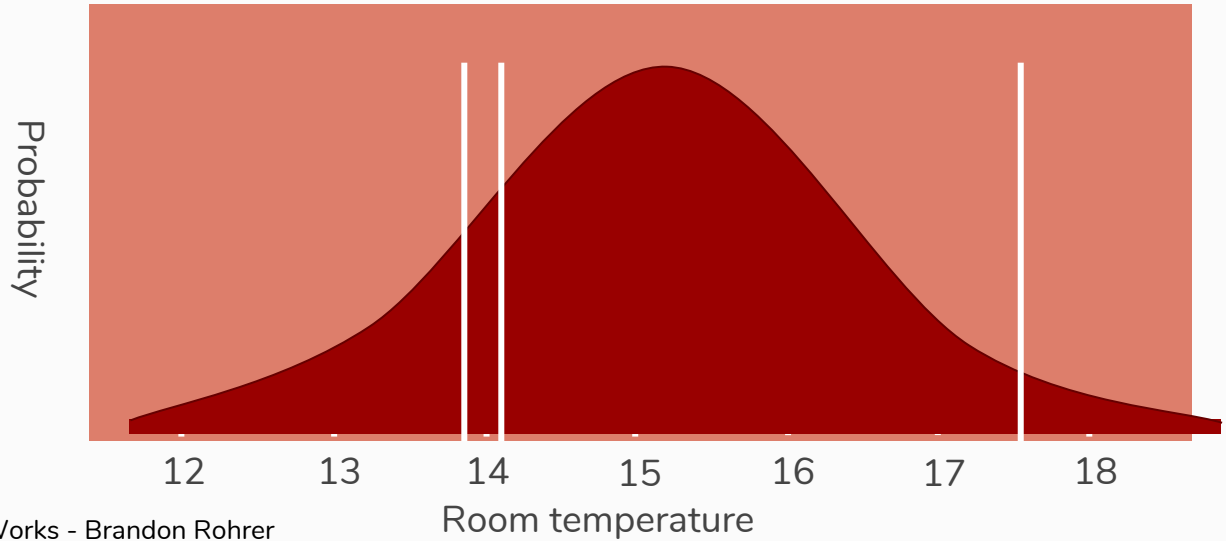
$$P(t \mid m) = \frac{P(m \mid t)\ P(t)}{P(m)}$$

t = unknown actual temperature
m = measurements

Uniform prior

No assumption about the distribution before the measurements
Uniform distribution: Every value has the same probability

mean = 15.2°
Posterior distribution uniform prior: Also known as a Maximum Likelihood Estimate (MLE)

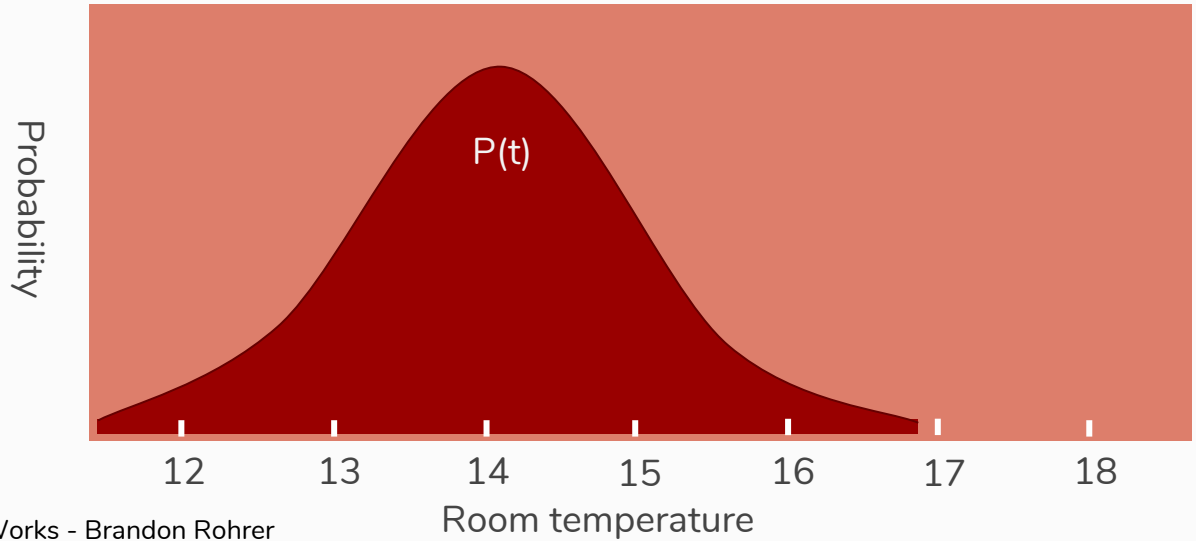


Informative prior

Last time I measured: 14.2°

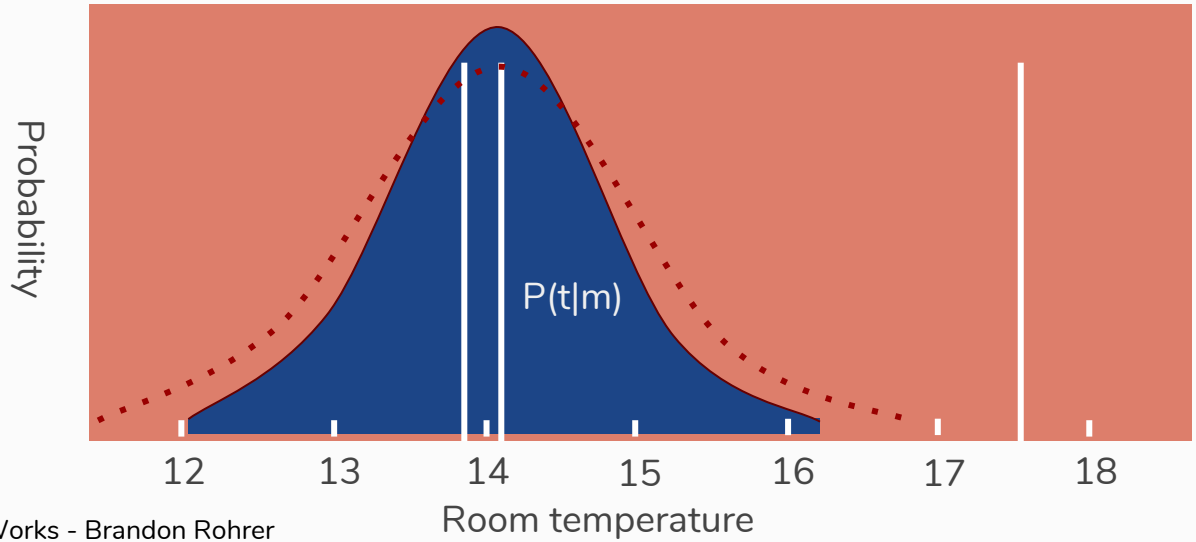
Prior = normal distribution

- mean = 14.2°
- standard error = 0.5°



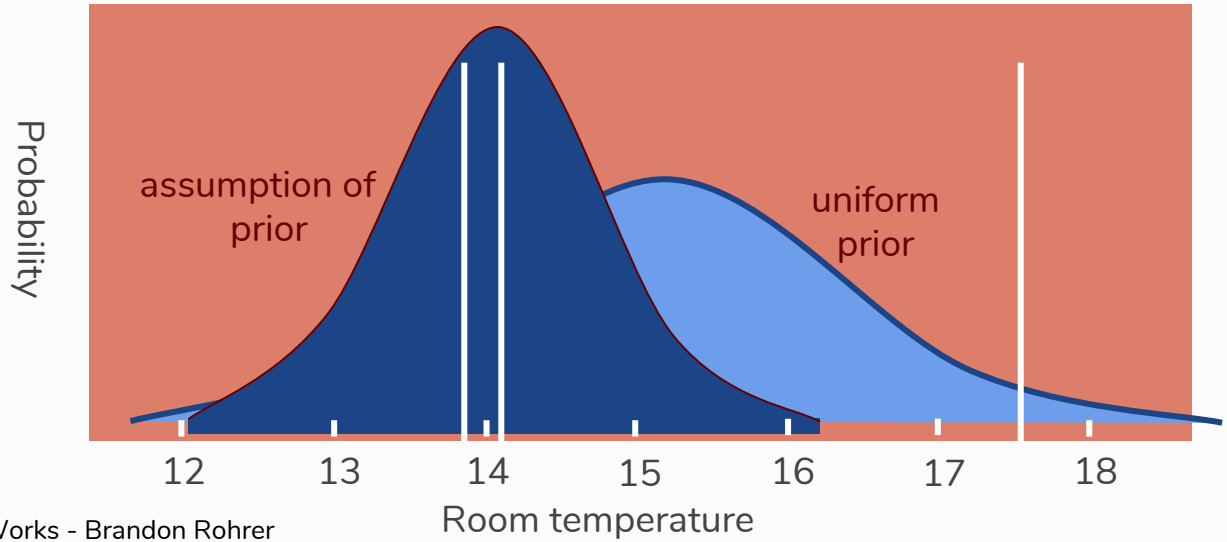
Posterior with informative prior

mode (most common) = 14.1°
Also known as Maximum A Posteriori (MAP)



Uniform vs. A prior belief

- With an assumptions of the prior
  - ignores the 17.5° like an outlier
  - greater confidence



How Bayesian inference Works - Brandon Rohrer

Takeaway

Assumptions of the data/distribution

(e.g. temperature > -273° Celsius)

enable us to use bayesian inference (with an informative prior)

which helps us to get sharper estimates with fewer measurements

iClusterBayes

Integrative clustering of multi-omics data

Variables

- i = sample (1, ..., n)
- j = genomic feature (1,...,p_t)
- t = data set (1, ..., m)
- y_{ijt} = matrix with samples, features and data sets
- **z_i = latent variable, used for clustering**

i

	SM_12	SM_32	SM_10
DDX11L1	0	0	0
WASH7P	35	0	82
MIR6859-1	0	0	1

j

Transcriptome: rawCountMatrix.csv

t = 1

i

	SM.10	SM.11	SM.13
A30	667090	1166500	223830
A4GALT	0	0	0
AAAS	1799900	1443300	1611700

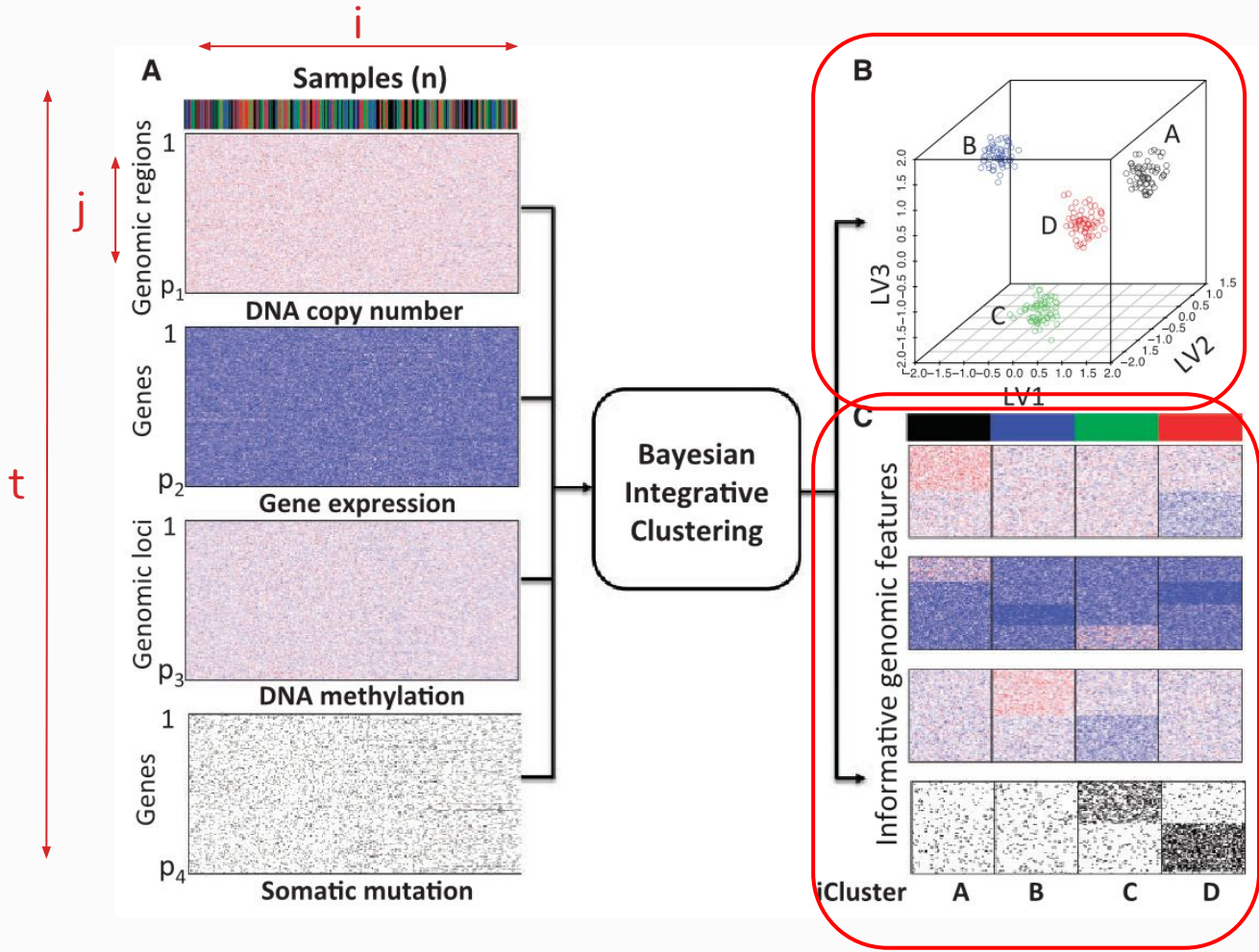
Proteome: LfqIntensitiesB1_SMCon.tsv

t = 2

Core concept:
Dimensionality reduction

identify latent variables to cluster samples in a lower dimensional subspace

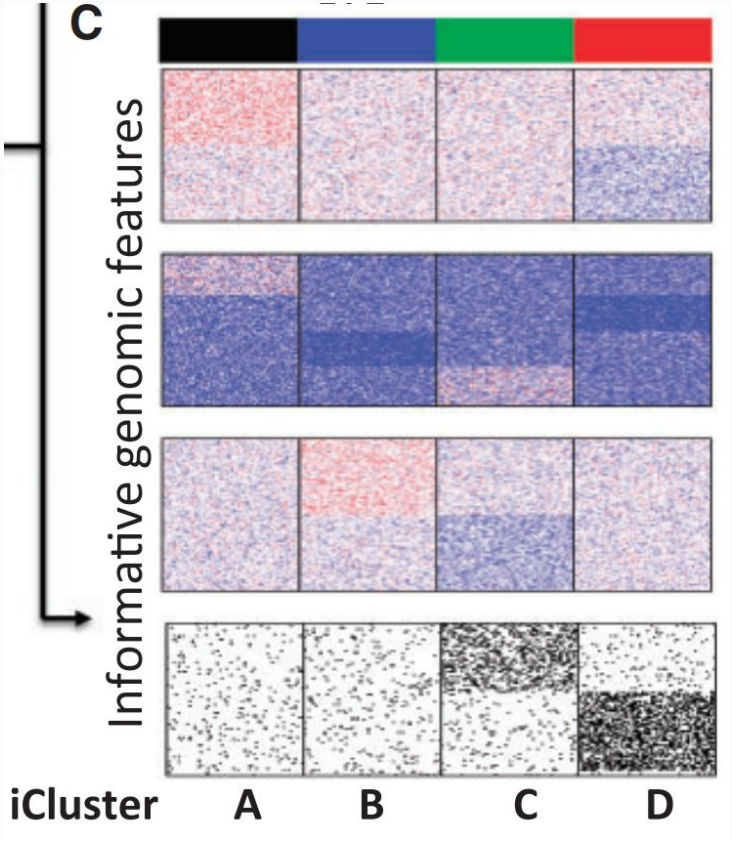
driving features that contribute to clustering



Model parameters

omics feature ***j*** in data set ***t***

- β_{jt} = coefficient vector
- γ_{jt} = indicator variable
  - 0: β_{jt} small ⇒ does not contribute to clustering
  - 1: β_{jt} big ⇒ contributes to clustering



Continuous model

Statistical framework:

x_i includes latent variable z_i, Γ_{jt} includes γ_{jt}

$$y_{ijt} = \mathbf{x}_i \boldsymbol{\Gamma}_{jt} \boldsymbol{\beta}_{jt} + \varepsilon_{ijt}$$

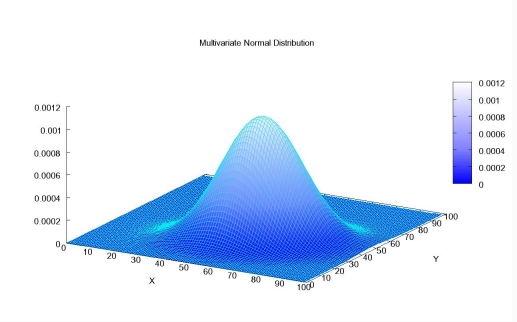
Model for omics feature j in data set t:

X includes latent variable z_i, Γjt includes γjt

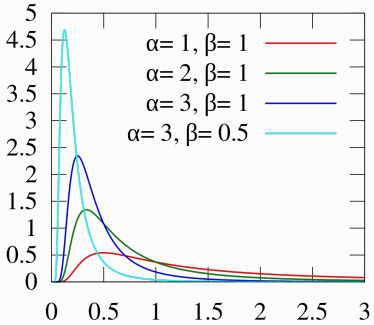
$$\mathbf{y}_{jt} = \mathbf{X} \boldsymbol{\Gamma}_{jt} \boldsymbol{\beta}_{jt} + \boldsymbol{\varepsilon}_{jt}$$

Priors

- $\beta_{jt} \sim MVN(\beta_{0t}, \sum_{0t})$
  - Multi-variant Normal distribution with ***mean*** β_{0t} and ***covariance*** $\sum_{0t}$



- $\sigma^2_{jt} \sim IG(v_0/2, v_0\sigma^2_0/2)$
  - Inverse Gamma distribution with ***shape*** $v_0/2$ and ***scale*** $v_0\sigma^2_0/2$
  - only for continuous model



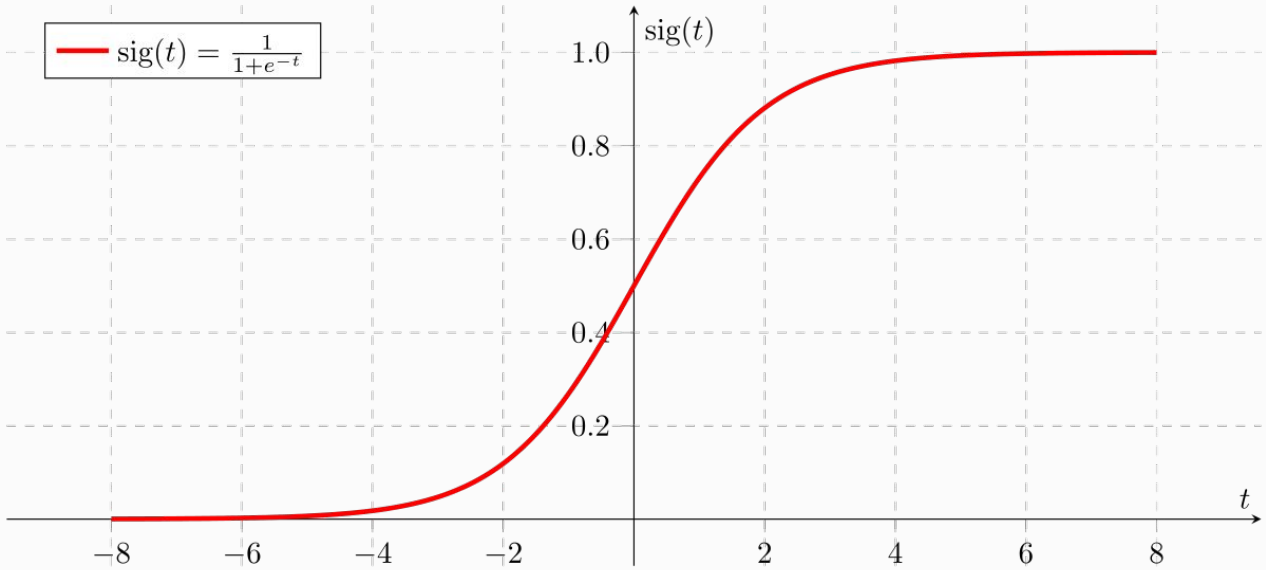
- $\gamma_{jt} \sim Bernoulli(q_t)$
  - q_t : probability of omics feature being a driving factor for clustering

Image: Wikipedia

Binary model

$$\log \frac{P(y_{ijt} = 1 \mid \mathbf{z}_i)}{1 - P(y_{ijt} = 1 \mid \mathbf{z}_i)} = \mathbf{x}_i \boldsymbol{\Gamma}_{jt} \boldsymbol{\beta}_{jt}$$

use a logistic regression



example: Sigmoid Activation Function

Image: towardsdatascience.com

Count model

$$\log\left(\lambda(y_{ijt} \mid \mathbf{z}_i)\right) = \mathbf{x}_i \boldsymbol{\Gamma}_{jt} \boldsymbol{\beta}_{jt}$$

uses poisson regression

commonly used for count data

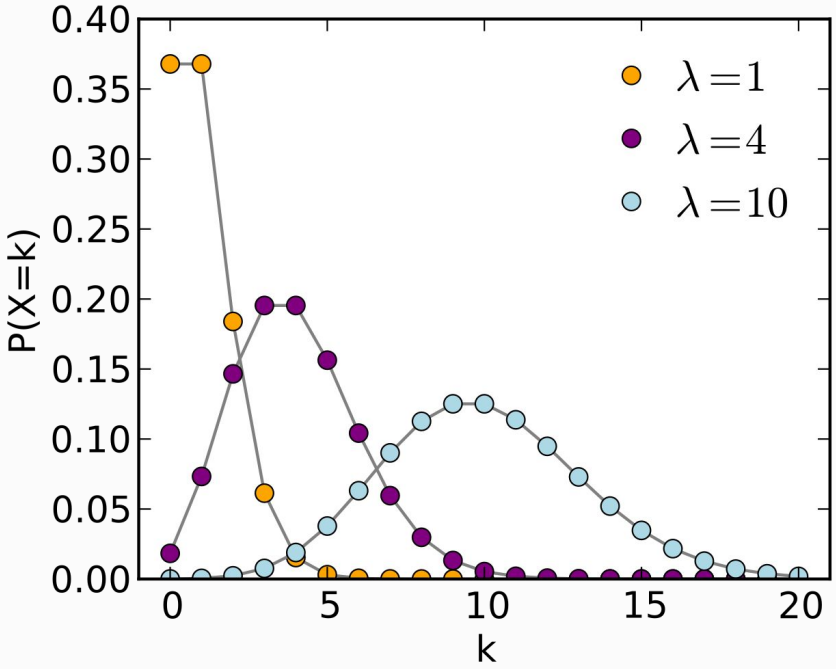


Image: Wikipedia

Joint modeling: Model latent variable z_i

posterior prior likelihood

$$P\left(\mathbf{z}_i \mid \mathbf{y}_{jt}, \boldsymbol{\beta}_{jt}, \gamma_{jt}\right) \propto P(\mathbf{z}_i) \prod_t^m \prod_j^{p_t} P\left(y_{ijt} \mid \mathbf{z}_i, \boldsymbol{\beta}_{jt}, \gamma_{jt}\right)$$

$P(\mathbf{z}_i)$: MVN(0, I_k)

y_{ijt}: data

$$P\left(y_{ijt} \mid \mathbf{z}_i, \boldsymbol{\beta}_{jt}, \gamma_{jt}\right) \propto \begin{cases} \sigma_{jt}^{-1} \exp\left(-\left(y_{ijt} - \mathbf{x}_i \boldsymbol{\Gamma}_{jt} \boldsymbol{\beta}_{jt}\right)^2 / (2\sigma_{jt}^2)\right), & \text{normal,} \\ \left(\exp(\mathbf{x}_i \boldsymbol{\Gamma}_{jt} \boldsymbol{\beta}_{jt})\right)^{y_{ijt}} \left(1 + \exp(\mathbf{x}_i \boldsymbol{\Gamma}_{jt} \boldsymbol{\beta}_{jt})\right)^{-1}, & \text{binomial,} \\ \left(\exp(\mathbf{x}_i \boldsymbol{\Gamma}_{jt} \boldsymbol{\beta}_{jt})\right)^{y_{ijt}} \exp\left(-\exp(\mathbf{x}_i \boldsymbol{\Gamma}_{jt} \boldsymbol{\beta}_{jt})\right), & \text{Poisson.} \end{cases}$$

β_{jt} (binary, count), γ_{jt} and z_i cannot be calculated in a finite number of steps

⇒ How to sample from their distribution?

Posterior distribution through sampling

Metropolis–Hastings algorithm (β_{jt} (binary, count), γ_{jt} and z_i):

- Markov chain Monte Carlo (MCMC) method
- generates random samples from a probability distribution

Gibbs Sampling (β_{jt} , σ^2_{jt} for continuous):

- also generates samples using MCMC
- specific: approximates from a specified multivariate probability distribution

Used when calculating the theoretical distribution is too complex (e.g. multi-dim Integrals)

What’s next

- Run iClusterBayes on the complete data set
- Visualize the data
- Evaluate findings
  - Do the cluster and their driving features make sense?
  - Did we find known or novel driver genes and molecular subtypes?
- Evaluate which data sets drive clustering by using only pairs of data sets
  - Which data set influence the clustering most and why?
  - Is there a data set that can be left out?

Omicum:
Building of the Estonian Biocentre
in Tartu

Image: Wikipedia

Sources of images used

- http://simpleicon.com/wp-content/uploads/note-4.png
- https://pngtree.com/free-icon/patient_1257502
- https://www.semanticscholar.org/paper/Integrative-Analysis-of-Multi-omics-Data-for-and-of-Sun-Hu/bc9cf73b72be9c1769ccb60f3f3d24f0c22cf1ab
- https://www.simula.no/sites/default/files/styles/original_dimension_image/public/articles/images/01_icon_software_engineering_rgb_black.png?itok=HNDDcPzS
- A fully Bayesian latent variable model for integrative clustering analysis of multi-type omics data. - Mo Q. et al

Backup slides

Break glass in case of emergency

Transcriptomics in CVD

Led to the discovery of novel biomarkers

- GDF15
  - Acute coronary syndromes
  - angina pectoris
  - heart failure
- And other circulating microRNAs
  - coronary heart disease
  - myocardial infarction

Proteomics Analysis

- Initial protein separation methods
  - 2-D - two dimensional gel electrophoresis
  - DIGE - differential in-gel electrophoresis
- After separation
  - Protein spots are picked and digested with proteolytic enzymes
  - analyzed by tandem mass spectrometry (MS/MS)
- Label free quantification
- Field of proteomics is still in early stages
  - Only approx. 10.000 proteins can be mapped today