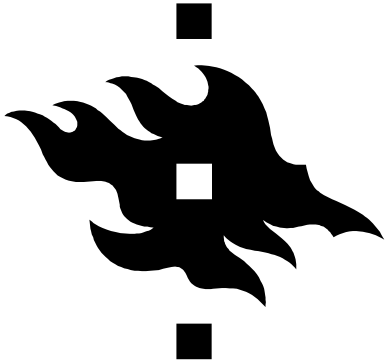


Latent stochastic models for comparing tumor samples of unknown purity

Antti Häkkinen

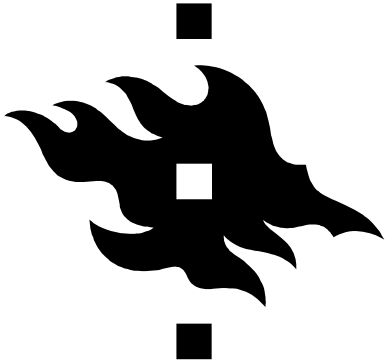
Hautaniemi lab, Faculty of Medicine, University of Helsinki

Sep 11, 2017

HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

Introduction

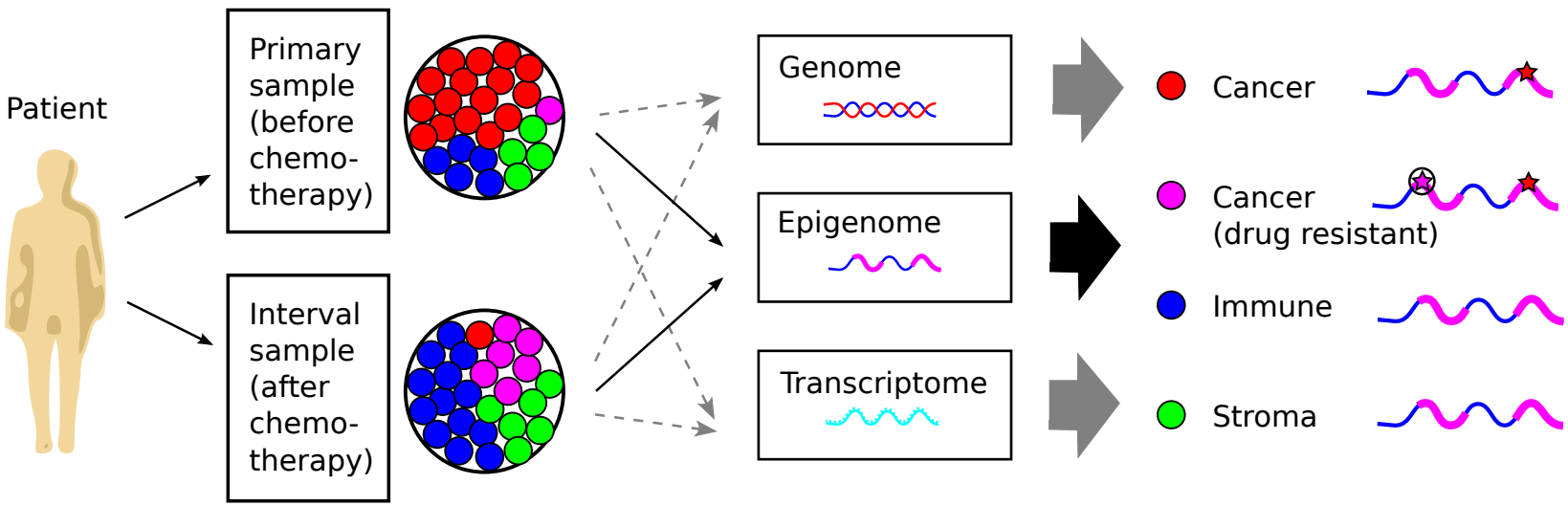
- We work on understanding the development of chemoresistance in ovarian cancer using a systems biology approach
  - $>$ 40,000 OVCA-caused deaths per year in Europe alone [1]
  - $>$ 50% of patients die within 5 years from diagnosis [1, 2]
  - Alterations driving it are heterogeneous [2]
- As part of their treatment, tumor samples are collected from the OVCA patients at multiple stages of the disease progression
  - Primary sample (at diagnosis)
  - Interval samples (during treatment)
  - Typically the patient response is good, but the cancer recurs
  - Which aberrations drive the development of chemoresistant cancer cell subpopulations?

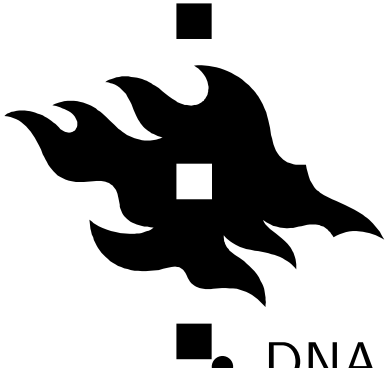


Introduction

- Clinical tumor samples contain various cell types in addition to the cancer cells: immune cells, stroma (blood cells, nerves), etc.
- Samples vary greatly in purity, some as low as 30% [3]; public resources like TCGA discard low purity samples ($< 60\%$) [4]
- Dilutes comparison at best–i.e. differences represent those between normal cells and not between the cancer cells
- Injects systematic bias at worst
- Problematic when comparing primary vs. interval as the treatment affects the tumor composition
- The mixed signal confounds subsequent analysis of active biological processes in cancer cells (DNA repair, apoptosis, etc.)

Introduction

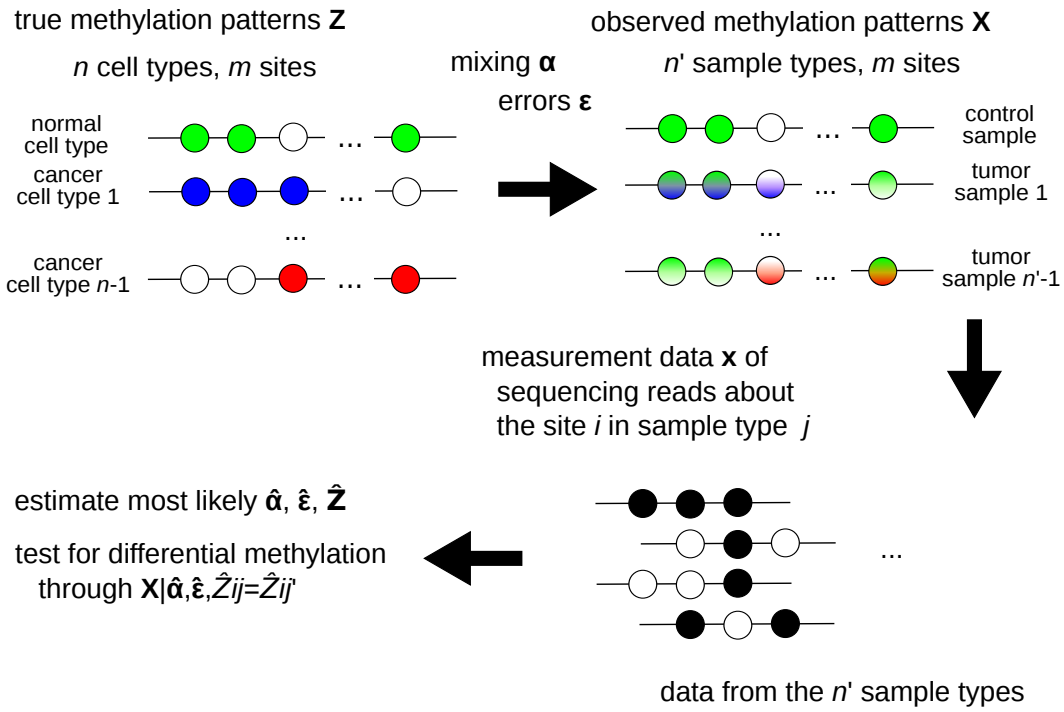




Methylation: background

- DNA methylation is an epigenetic modification [5]
  - Aberrations observed in most cancer types
  - Modulates expression at regulatory areas e.g. silencing of tumor suppressor genes
  - General lack of methylation causes genomic instability
- Standard methods assume pure samples, only few recent ones control but require normal cell profiles
  - Most techniques use linear regression [6, 7]
    - Suitable only for large number of replicates
    - Purity must be obtained by other means (e.g. IHC)
  - Can we do this without control and purity estimates?
    - Reliable information cannot be obtained from all patients

Methylation: models



$$\mathbb{P}\left[X_{i,1} | Z_{i,0}, Z_{i,1}, \epsilon_1, \alpha_1 \right]$$

	$(Z_{i,0}, Z_{i,1}) = (0,0)$
$X_{i,1} = 0$	$1 - \epsilon_1$
$X_{i,1} = 1$	ϵ_1

	$(Z_{i,0}, Z_{i,1}) = (1,0)$
$X_{i,1} = 0$	$(1-\alpha_1)\,\epsilon_1 + \alpha_1\,(1-\epsilon_1)$
$X_{i,1} = 1$	$(1-\alpha_1)\,(1-\epsilon_1) + \alpha_1\,\epsilon_1$

$$\ldots$$

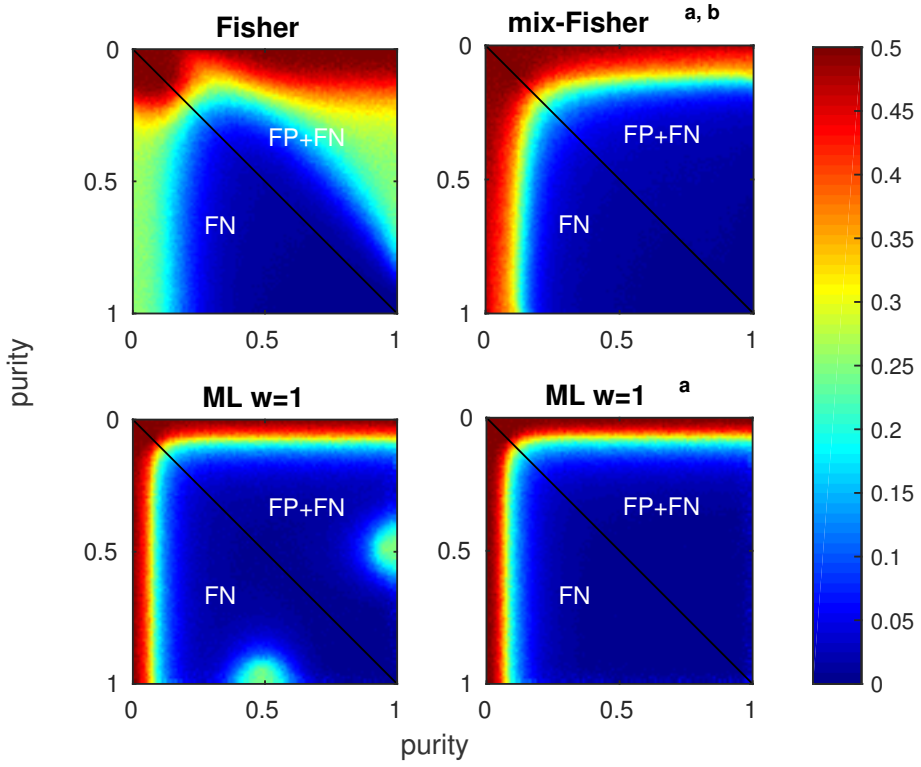
EM-algorithm:

$$\mathbb{P}[\mathbf{Z}|\mathbf{X},\,\theta] = \frac{\mathbb{P}[\mathbf{X},\,\mathbf{Z}|\,\theta]}{\mathbb{P}[\mathbf{X}|\,\theta]} = \frac{\mathbb{P}[\mathbf{X},\,\mathbf{Z}|\,\theta]}{\sum_{\mathbf{Z}} \mathbb{P}[\mathbf{X},\,\mathbf{Z}|\,\theta]}$$

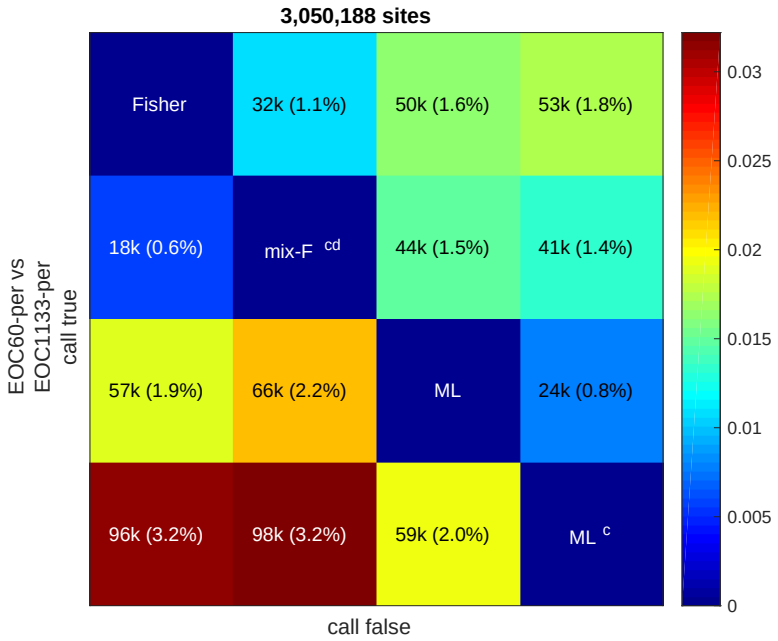
$$\theta^{(r+1)} = \underset{\theta}{\operatorname{argmax}}\, \mathbb{E}\left[\log \mathbb{P}[\mathbf{X}|\mathbf{Z},\,\theta] \,\middle|\, \mathbf{Z} = \mathbf{Z}^{(r)} | \mathbf{X},\,\theta^{(r)} \right]$$

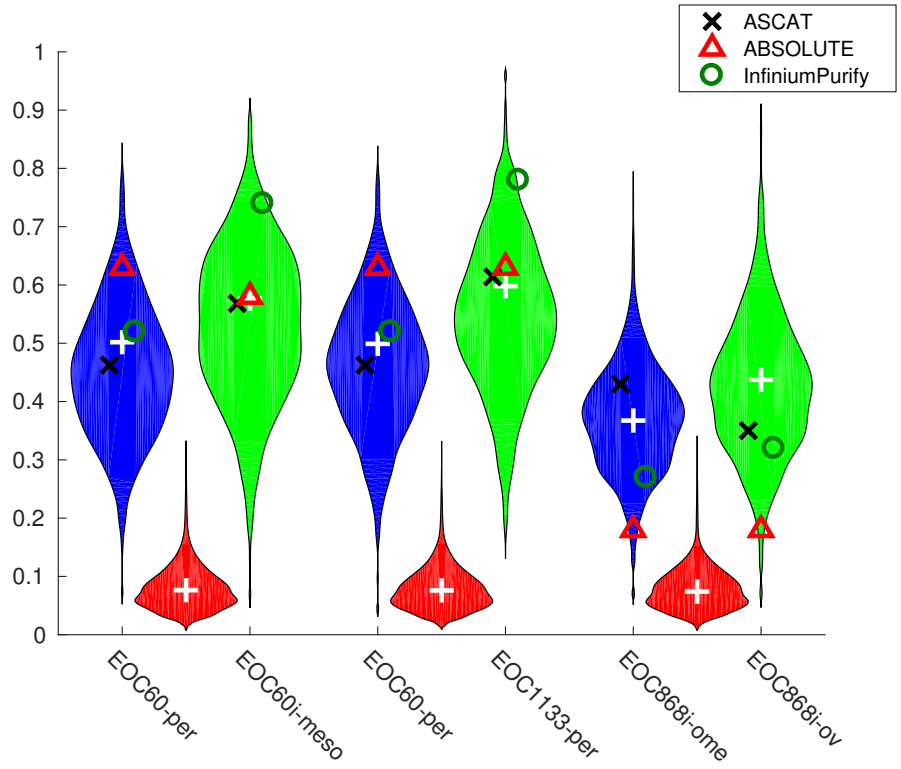
$$= \underset{\theta}{\operatorname{argmax}} \sum_{\mathbf{Z}} \log \mathbb{P}[\mathbf{X}|\mathbf{Z},\,\theta]\; \mathbb{P}\left[\mathbf{Z} = \mathbf{Z}^{(r)} | \mathbf{X},\,\theta^{(r)} \right]$$

Methylation: results

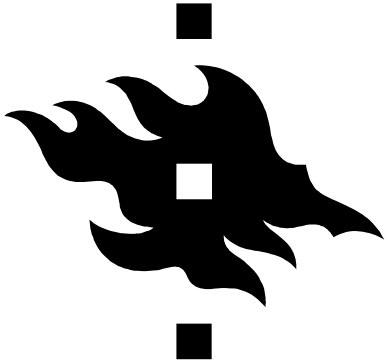


Methylation: results



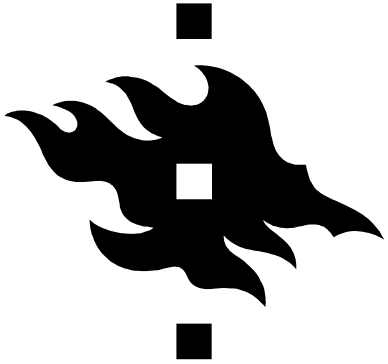


Sample A	Sample B	Sites	Corr -	$P_{-,0}$	Corr ML	$P_{ML,0}$	$P_{-,ML}$
EOC60-per	EOC60i-meso	F ∨ ML	−3.57%	0.04	−8.55%	8.44×10^{-7}	0.04
EOC60-per	EOC1133-per	F ∨ ML	−6.63%	9.74×10^{-27}	−10.48%	1.52×10^{-64}	9.23×10^{-6}
EOC868i-ome	EOC868i-ov	F ∨ ML	−1.88%	0.21	−12.04%	4.73×10^{-16}	1.22×10^{-6}



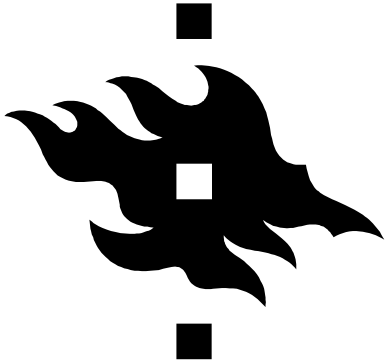
Methylation: results

- Reveals methylomes of cancer cells or compares them from two (or more) samples
- Generally outperforms other methods in all settings
- Suggests 5–10% new findings and false positives
- Predicts RNA expression better
- Purity estimates comparable to WGS methods
- Does not require normal cell control nor prior purity estimates
  - If these are available, used for increased accuracy



RNA-seq: background

- Cells harbor various alterations, only few drive tumor progression
- Alterations can affect the expression of the gene or its product
- RNA sequencing is widely employed: detect expression chages, discover splice variants, gene fusions, etc.
- There are methods of varying quality for decomposing the RNA-seq signal (e.g. from $\hat{\mathbf{W}} = \mathbf{X}^{\dagger}\,\mathbf{Y}$ to [8]):
  - Most are for microarray data (not NGS)
  - Based on libraries of known expression patterns for known cell types and deconvolution
    - No attempt to model heterogeneity
    - Cannot adapt to unknown cell types



RNA-seq: models

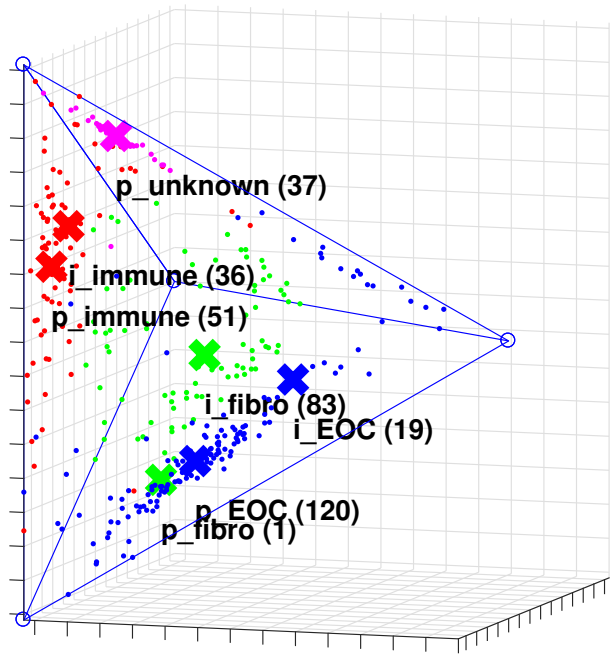
$\mathbf{X} \sim \mathcal{P}(\bar{\mathbf{U}}\,\mathbf{D}_U\,\bar{\mathbf{W}}\,\mathbf{D}_W)$
$[\mathbf{W}]_{:,j} \sim \mathcal{D}(\boldsymbol{\alpha}_j)$

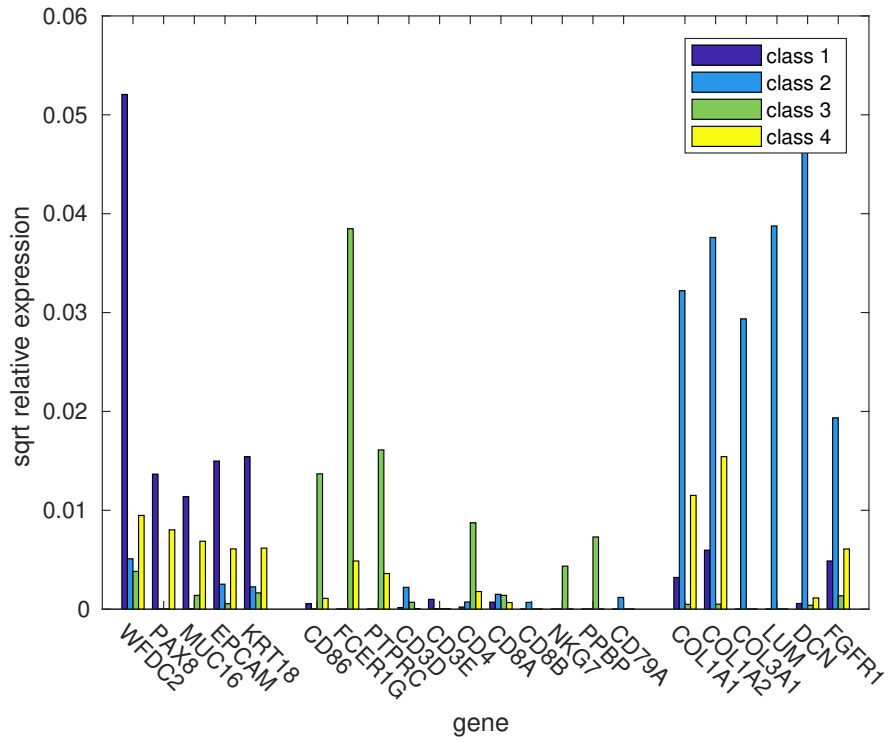
$\mathbf{X}$ are the responses, can be bulk or single-cell data
$\bar{\mathbf{U}}$ are the normalized expression profiles
$\mathbf{D}_U$ contains the expression level of each cell type (e.g. if cancer has double expression level in all genes)
$\bar{\mathbf{W}}$ is the normalized mixing matrix
$\mathbf{D}_W$ is the batch gain of the j sample (e.g. different amounts of RNA sequenced, amplification differences)

estimate $\bar{\mathbf{U}}\,\mathbf{D}_U$ and $\bar{\mathbf{W}}\,\mathbf{D}_W$ simulataneously
EM-algorithm with $Z_{i,j,k} \sim \mathcal{P}(U_{i,j}\,[\mathbf{D}_U]_{j,j}\,W_{j,k}\,[\mathbf{D}_W]_{k,k})$.

substitute $\mathcal{NB}$ for $\mathcal{P}$ for heterogeneity: mixture of $\mathcal{P}$

RNA-seq: results

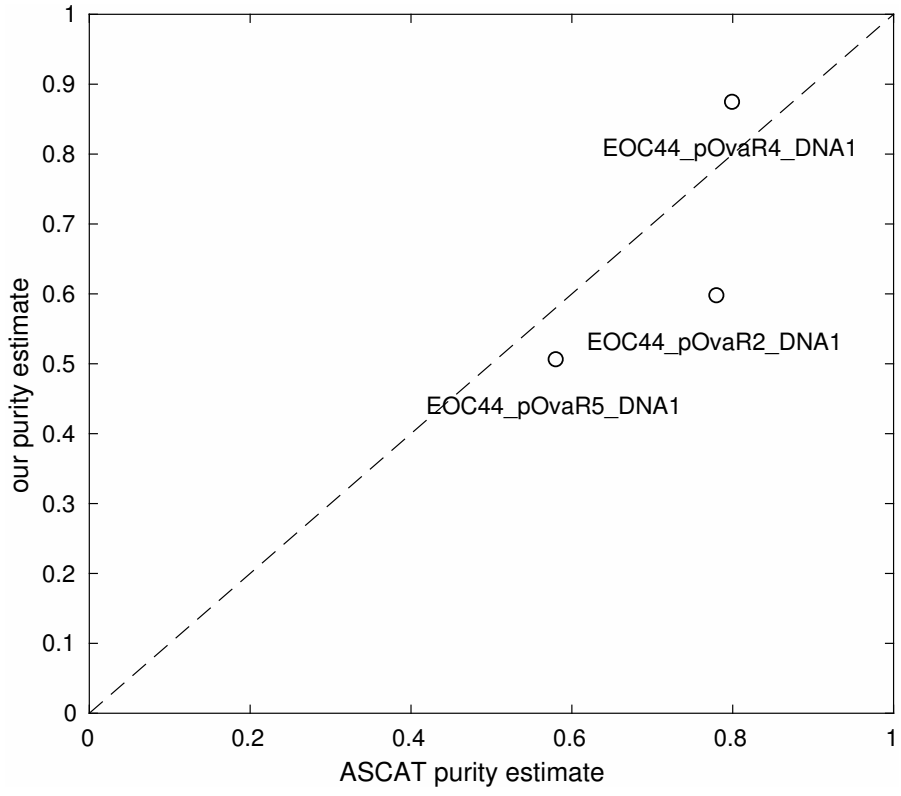


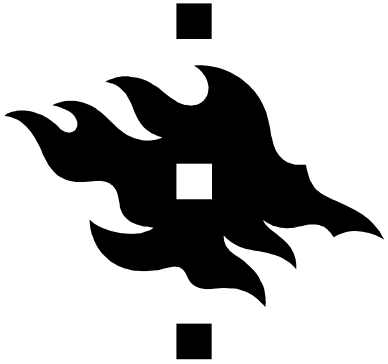


	p_EOC	p_fibro	p_immune	p_unknown	i_EOC	i_fibro	i_immune
class 1	0.0605	0	0.0029	0	0.0346	0	0
class 2	0.0058	0	0	0	0	0.2277	0.0058
class 3	0.0029	0	0.1095	0.1066	0.0058	0	0.0951
class 4	0.2767	0.0029	0.0346	0	0.0144	0.0115	0.0029

$\sim$85.3% accuracy of discovering hand-labeled cell types in 347 single-cell profiles

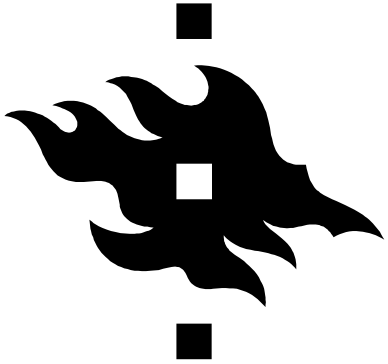
RNA-seq: results





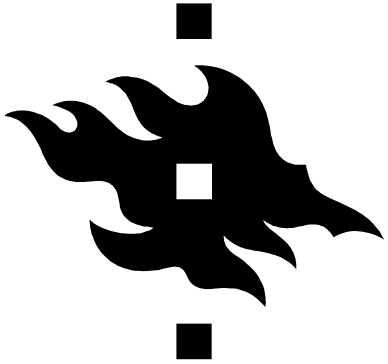
RNA-seq: results

- Our decomposition allows:
  - Blind clustering of single-cell data to identify candidate cell types
  - Decomposition of bulk samples into their components
    - Using single-cell + 1 unknown for 1 bulk sample
    - "Clean" a bulk sample using a bulk normal
- Correct batch effects etc. in the process
- The development & evaluation are still ongoing



Conclusion

- Remove biases that result from sample impurities
- Adapt to patient-specific profiles
- Adapt to missing or suspect controls
- All parameters estimated from the data–allow more flexible experimental settings
- Better accuracy–less prominent differences can be detected
- Accurate and unbiased quantification is required for integrative analyses to identify malfunctioning the biological processes
- Allows NGS of patient-derived samples to be used as personalized diagnostic and prognostic biomarkers and aids discovering therapeutic targets



References

[1] Siegel RL et al., CA Cancer J Clin 66: 7, 2016

[2] Berns EMJJ & Bowtell DD, Cancer Res 72: 2701, 2012

[3] Aran D et al., Nat Commun 6: 8971, 2015

[4] The Cancer Genome Atlas Research Network, Nature 474: 609, 2011

[5] Esteller M, N Engl J Med 358: 1148, 2008

[6] Feng H et al., Nucl Acids Res 42: e69, 2014

[7] Zheng X et al., Genome Biol 18: 17, 2017

[8] Yoshihara K et al., Nat Commun 4: 2612, 2013