



Statistical inference and
learning in genomics, with
application to sepsis

Pål Vegard Johnsen, PhD SINTEF DIGITAL

Courses

- Autumn 2018
 - MA8704 - Probability Theory and Asymptotic Methods
 - TMA4285 – Time series models
 - MOL8008 - Bioinformatics Methods for next Generation Sequencing Analysis

- Spring 2019
 - Individual study syllabus (MA8701, deep learning, statistical learning).
 - Oslo Health Hackaton (Glioblastoma).



Facts

30 000 000

O_2

6 000 000





GEMINISENTER SEPSISFORSKNING

Check out sepsis.no for more information
about sepsis.

Genetic predisposition to sepsis

- Higher risk for getting sepsis when reduced immune response:
- Old, chronic diseases, cancer etc.

- Big question: Can some people have more genetic tendency to get sepsis than others?

- Key: Eventually would want to look at SNP-SNP interactions.



DATA FROM HUNT

- HelseUndersøkelsen i Nord-Trøndelag
- Health data from over 135 000 participants living in Nord-Trøndelag. Used world-wide.

70 000 participants

2900 with sepsis

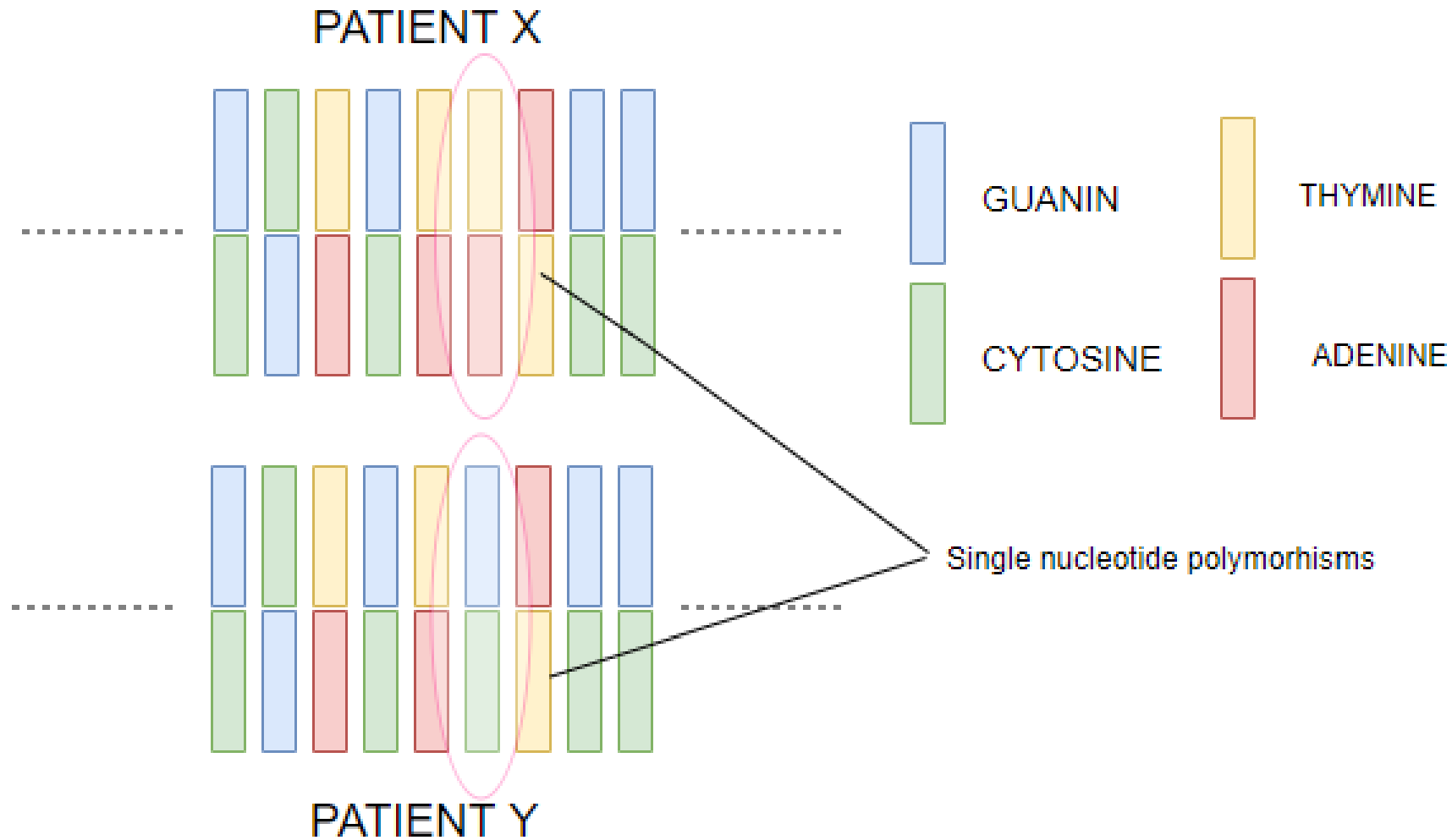
300 000 SNPs

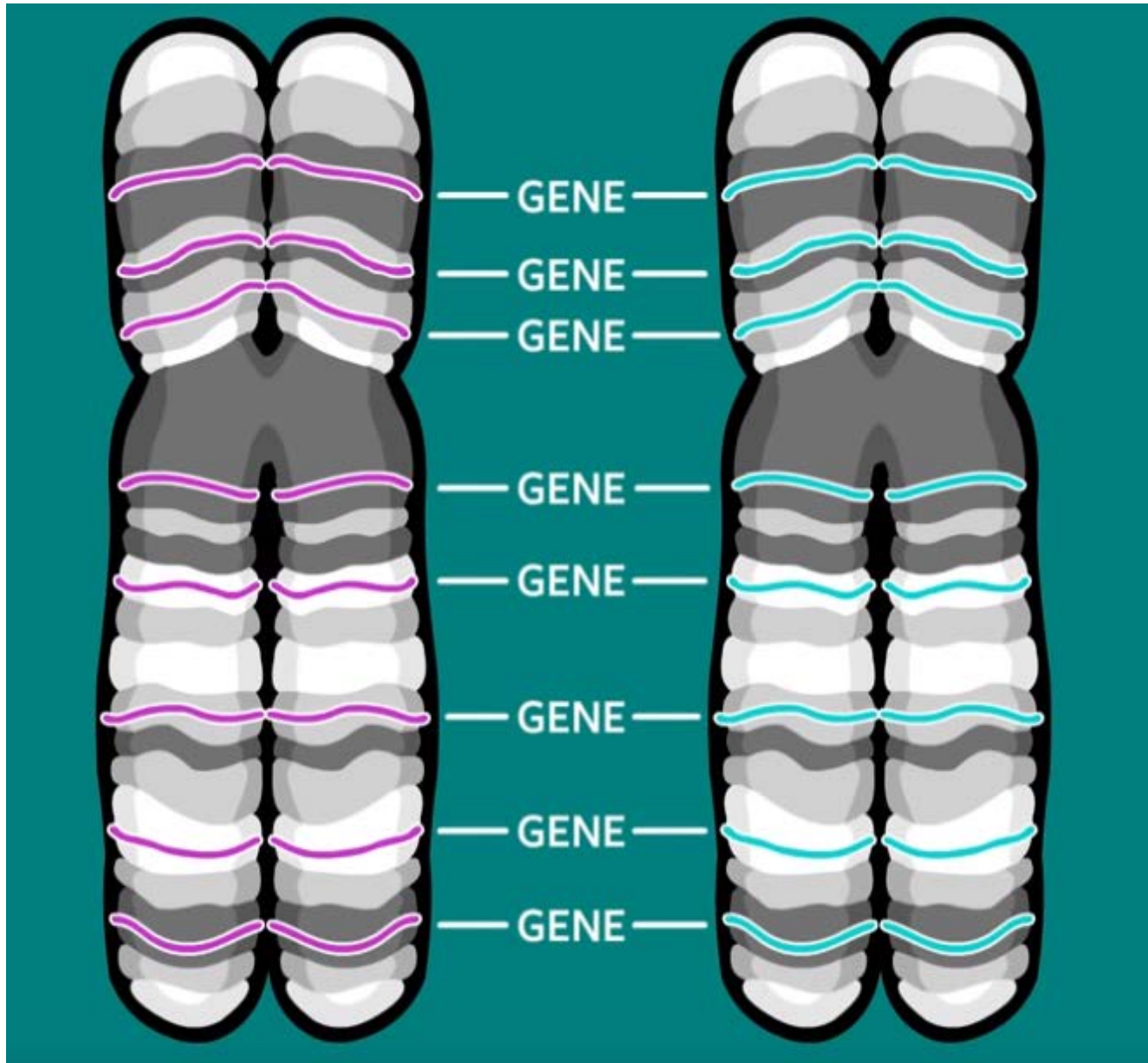


CLEANING THE DATA

- Hardy Weinberg – Hmm...
- Cluster separation score – I guess is OK!

What is a SNP?





	22.16855618.G	22.17057138.G.	22.17073066.A.	X22.17075353.C.	22.17094749.A.C
ID 1	0	1	0	2	2
ID 2	0	0	1	1	2
ID 3	1	1	NA	2	2
ID 4	0	2	0	2	NA

...

•
•
•
•

Additional data and lack of data

- Sex and age.
- bacterias found in blood, gram-test.
- Problem: Should have had even more clinical data, protocols etc.
- Do not have information about kinship (relatives).



IMPORTANT:
DATA MUST BE STORED
SECURELY

HUNT Cloud



Statistical methods

- Genome-wide association studies (GWAS)
- Multiple testing (statistical inference)

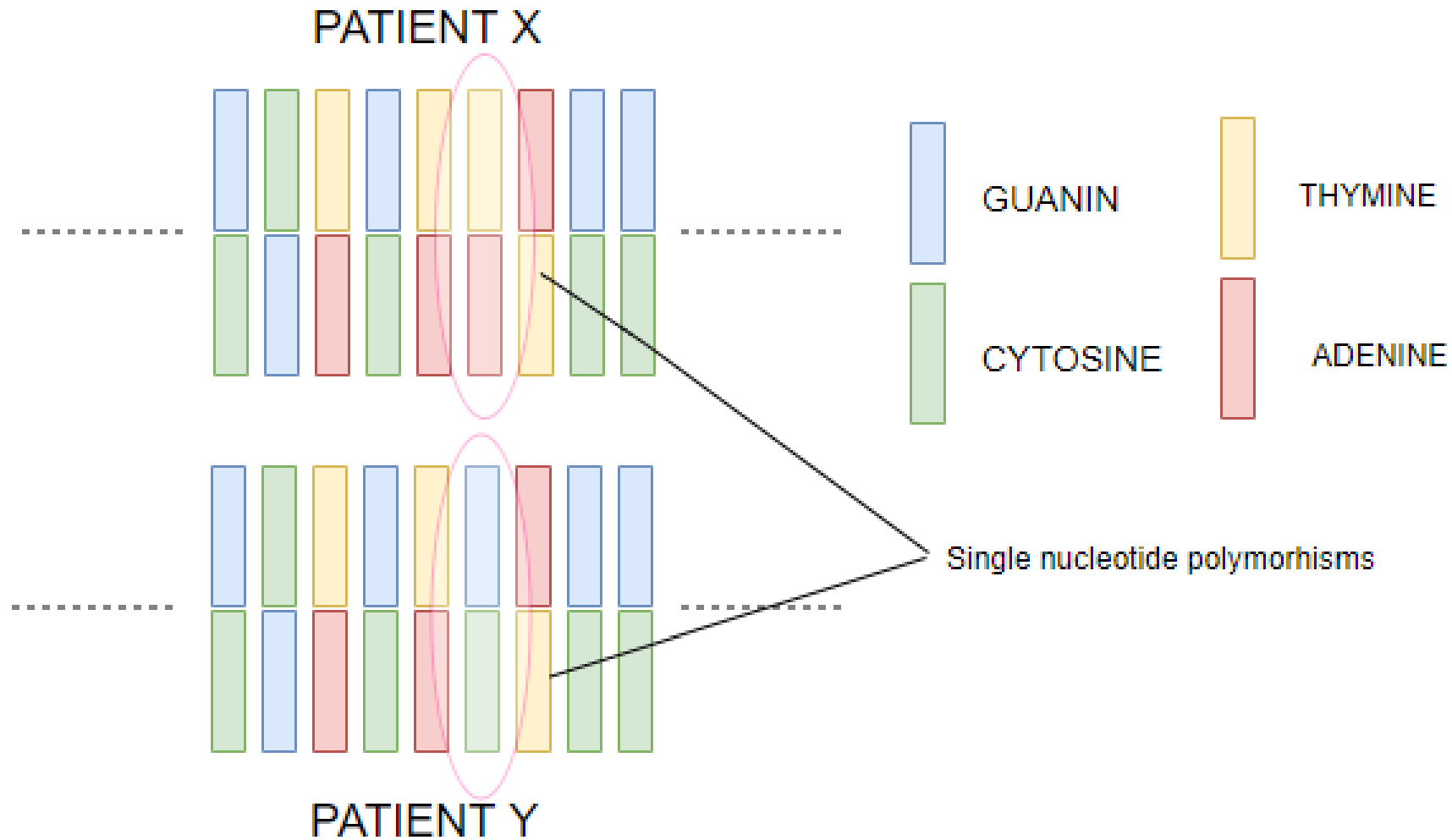
One single hypothesis:

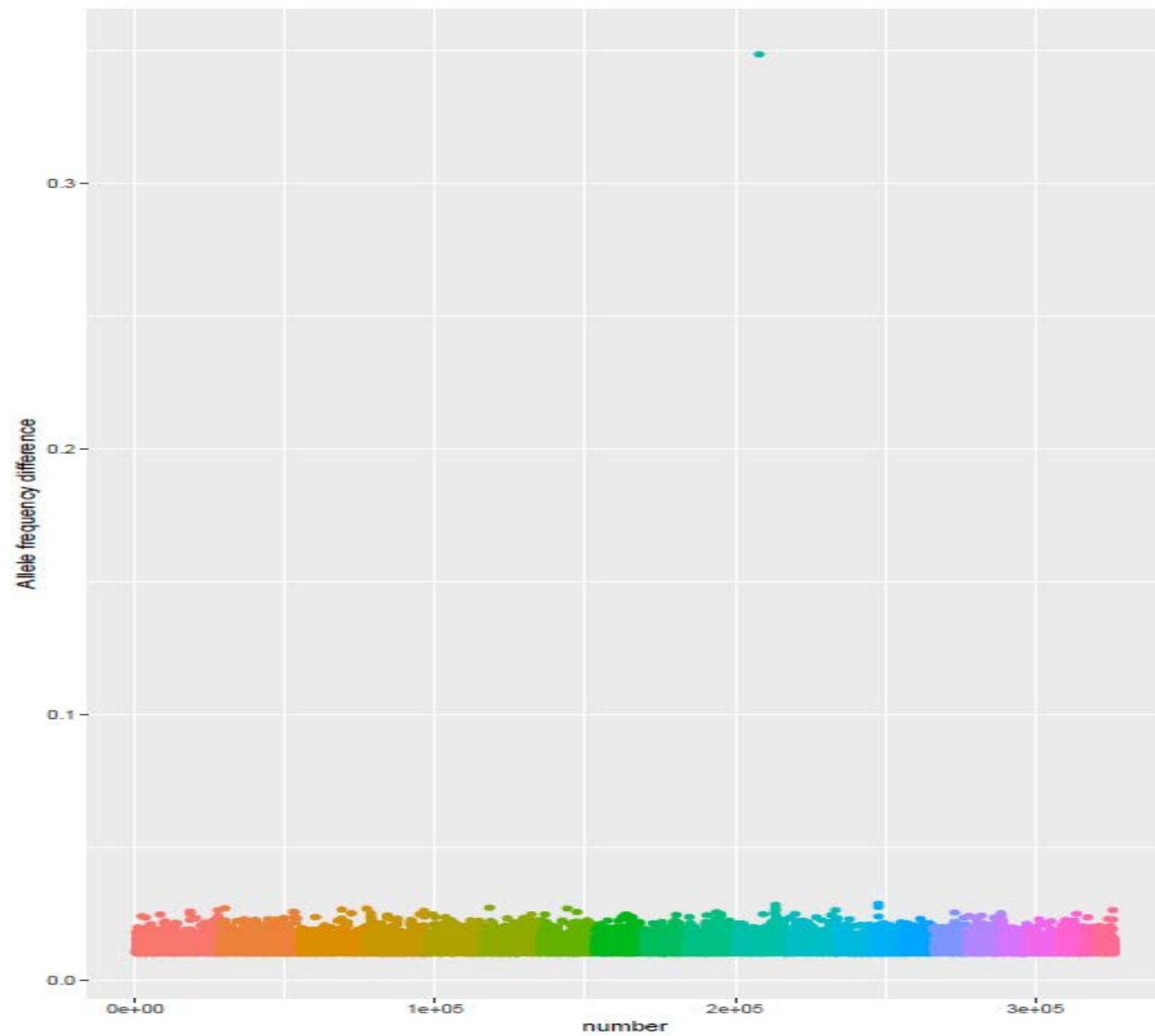
H_0 : SNP x is not associated with sepsis.

H_1 : SNP x is associated with sepsis.



What is a SNP?





Association tests:

- Tea lady test:




The lady in question eventually answered correctly six out of the eight trials. The results can be assembled in a 2 by 2 contingency table:

		TRUE ORDER:		TOTAL (MARGIN)
		TEA FIRST	MILK FIRST	
LADY'S GUESSES:	TEA FIRST	$a = 3$	$b = 1$	$a + b = 4$
	MILK FIRST	$c = 1$	$d = 3$	$c + d = 4$
TOTAL (MARGIN)		$a + c = 4$	$b + d = 4$	$n = 8$

$$P(X = a) = \frac{\binom{a+b}{a} \binom{c+d}{c}}{\binom{n}{a+c}}$$

$$P(X \geq a) = \sum_{j=a}^J \frac{\binom{a+b}{j} \binom{c+d}{a+c-j}}{\binom{n}{a+c}} :$$




$$\log \left(\frac{p(x)}{1 - p(x)} \right) = \beta_{SNP} x_{genotype} + \boldsymbol{\beta}^T \mathbf{x}$$

$p(x)$ is risk of getting sepsis given value of SNP (0,1 or 2)
+ other covariates in x .

What should the other covariates be?

Rejection of null hypothesis based on p-value computed
for β_{SNP} (asymptotically).

IMPORTANT ASSUMPTION IN LOGISTIC REGRESSION MODELLING: INDEPENDENT DATA



If we attempt to model the probability that Y is 0 or 1 with the function $Pr(y | X; \theta) = h_{\theta}(X)^y(1 - h_{\theta}(X))^{(1-y)}$, we take our likelihood function assuming that all the observations in the sample are independently Bernoulli distributed,

$$\begin{aligned} L(\theta | x) &= Pr(Y | X; \theta) \\ &= \prod_i Pr(y_i | x_i; \theta) \\ &= \prod_i h_{\theta}(x_i)^{y_i} (1 - h_{\theta}(x_i))^{(1-y_i)} \end{aligned}$$

Multiple testing

- End up with hundreds of thousands of hypotheses
- How about type I errors?
- *Familywise error rate (FWER)*: Given m hypotheses with m_0 being true null hypotheses, denote q_1, q_2, \dots, q_{m_0} the corresponding p-values, then by controlling:

$$FWER = P(\text{One or more type I errors}) \leq \alpha$$

, means that each hypothesis can be rejected if $q_i \leq \alpha/m$, since:

$$P(\cup_{i=1}^{m_0} \{q_i \leq \alpha/m\}) \leq \sum_{i=1}^{m_0} P(q_i \leq \frac{\alpha}{m}) \leq \sum_{i=1}^{m_0} \frac{\alpha}{m} \leq \frac{m_0}{m} \alpha \leq \alpha$$



Statistical learning

- Eventually, association between several SNPs (SNP-SNP-interactions) and risk of sepsis.
- Might require ideas from statistical learning.



XGBoost – To find interactions

SCIENTIFIC REPORTS

OPEN

Machine learning identifies interacting genetic variants contributing to breast cancer risk: A case study in Finnish cases and controls

Received: 9 February 2018

Accepted: 22 August 2018

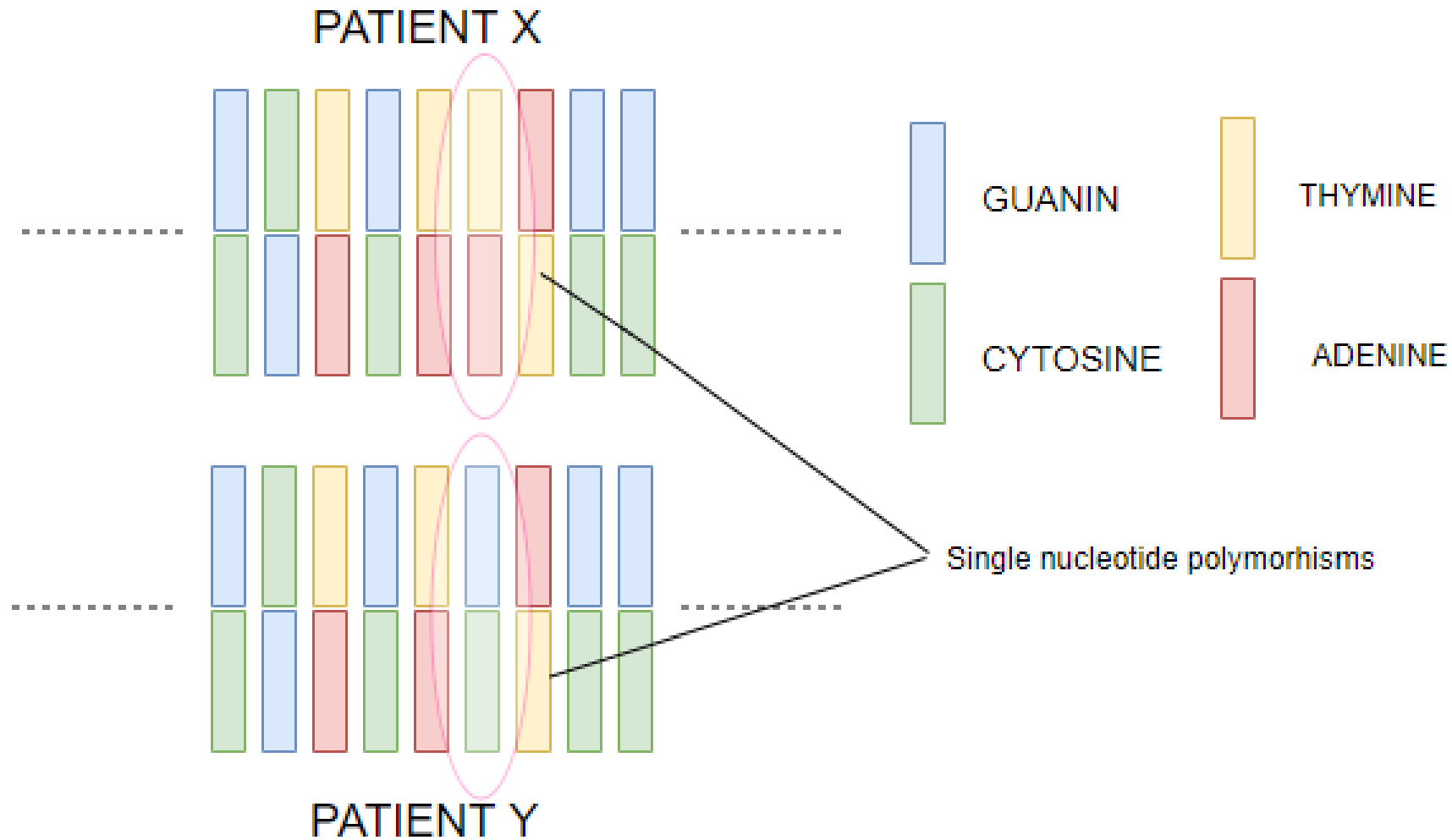
Published online: 03 September 2018

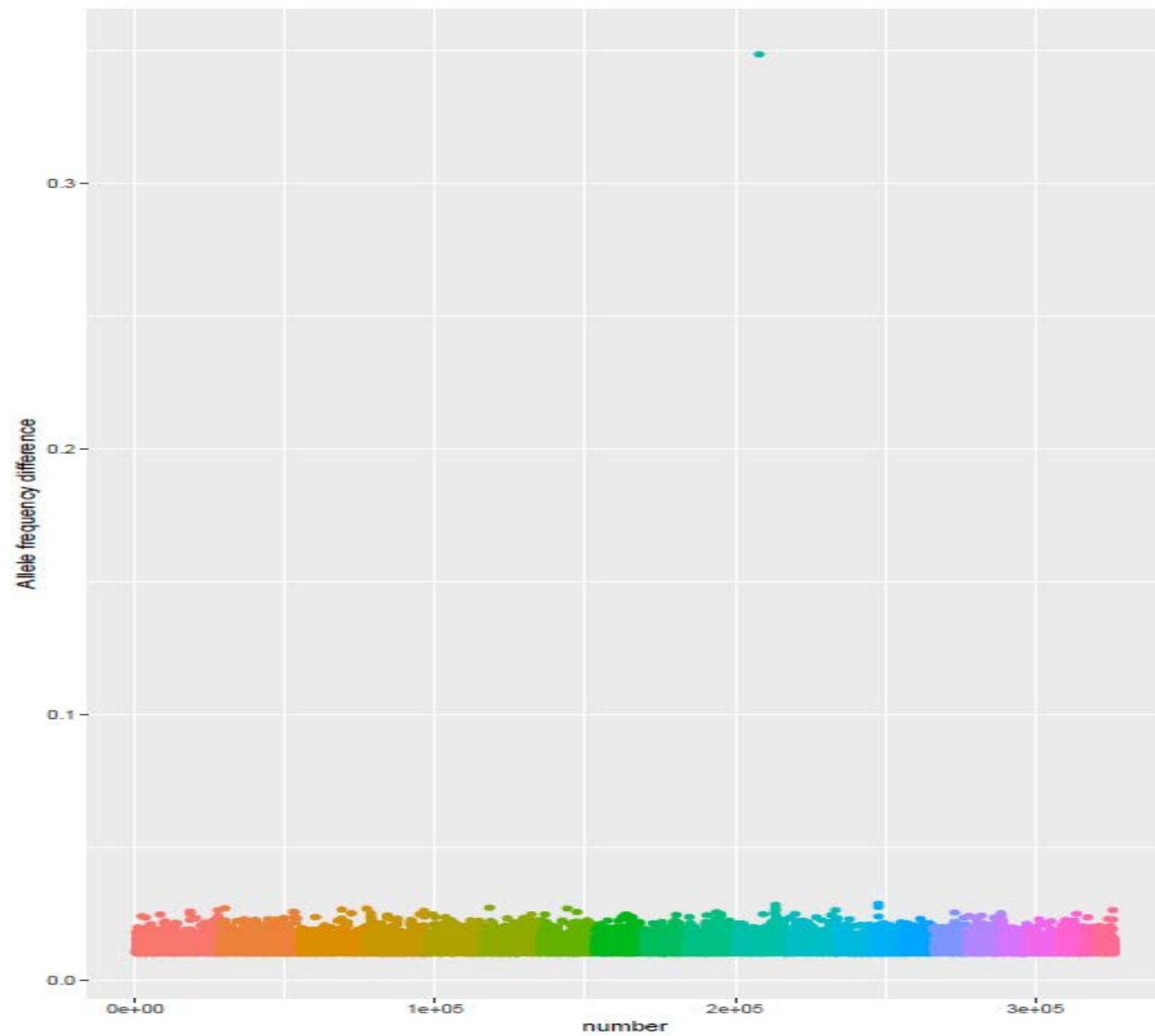
Hamid Behravan¹, Jaana M. Hartikainen¹, Maria Tengström^{2,3}, Katri Pylkäs⁴, Robert Winqvist⁴, Veli-Matti Kosma^{1,5} & Arto Mannermaa^{1,5}

We propose an effective machine learning approach to identify group of interacting single nucleotide polymorphisms (SNPs), which contribute most to the breast cancer (BC) risk by assuming dependencies among BCAC iCOGS SNPs. We adopt a gradient tree boosting method followed by an adaptive iterative SNP search to capture complex non-linear SNP-SNP interactions and consequently, obtain group of interacting SNPs with high BC risk-predictive potential. We also propose a support vector machine formed by the identified SNPs to classify BC cases and controls. Our approach achieves mean average precision (mAP) of 72.66, 67.24 and 69.25 in discriminating BC cases and controls in KBCP, OBCS and merged KBCP-OBCS sample sets, respectively. These results are better than the mAP of 70.08, 63.61 and 66.41 obtained by using a polygenic risk score model derived from 51 known BC-associated SNPs, respectively, in KBCP, OBCS and merged KBCP-OBCS sample sets. BC subtype analysis further reveals that the 200 identified KBCP SNPs from the proposed method performs favorably in classifying estrogen receptor positive (ER+) and negative (ER-) BC cases both in KBCP and OBCS data. Further, a biological analysis of the identified SNPs reveals genes related to important BC-related mechanisms, estrogen metabolism and apoptosis.



What is a SNP?





MOST IMPORTANT TO IMPROVE
MODEL:

MORE DATA



Challenges

- The number of covariates much larger than the number of data points.
- FWER is conservative => Reduces power.
- Construction of logistic regression. What covariates to add in order to increase power? How to deal with confounders?
- Asymptotic computation of p-values for imbalanced data sets.
- Dependence in data (relatives).
- Handling large amount of data.



Papers in the future

- One paper in a statistical journal regarding improvements within multiple testing (Independent Hypothesis Weighting).
- One paper in a biology/medicine journal with multiple testing applied to the sepsis data.
- One additional paper in a statistical journal – Possibly about computation of valid p-values for imbalanced sets or SNP-SNP interactions.



Research stay abroad: Yale School of Public Health

