

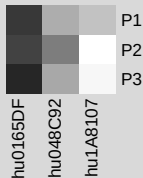
CAGI slide pack guidelines

I agree to my slides being used according to the following terms.

Tokyo, March 25th, 2015 *M. Stanke*

- To disseminate the CAGI results, I encourage all CAGI participants to remix these CAGI slides for inclusion in their oral and poster presentations
- CAGI participants may use or publish any content of these slides only in compliance with the CAGI Data Use Agreement
- CAGI participants may not cite or publish any content of these slides except with the written permission of the originating author or in the primary CAGI publications
- CAGI participants must acknowledge the original authors of these slides and CAGI 4. Include the acknowledgement banner across the bottom
- CAGI participants must include the CAGI credits slide in their presentations
- Slides with a red slash (no remix) may not be included in a presentation unless all attendees are registered CAGI participants who signed the data use agreement
- Slides with a red X over the entire slide may not be used in any circumstances
- Slides with a blue slash (embargoed) may not be used unless the embargo is explicitly lifted
- Slides without slash or X may be tweeted unless flagged with the “No tweet” icon.
- CAGI participants using these slides must explain the “No tweet” icon, which means that the slide content should not be disseminated outside the presenting conference hall. If the presentation venue permits pervasive tweeting, you may not include these slides in your talk 
- Note for slide authors: we expect all CAGI participants to abide by these restrictions, and will make best effort to ensure they are followed. However, but bear in mind that we have limited means of enforcing them, and therefore the restrictions cannot be guaranteed.

Mario Stanke



Predicting Phenotypes

Probabilities of
Matchings

A Markov Chain Monte Carlo Approach for Matching Genomes to Phenotype Profiles

Critical Assessment of Genome Interpretation
San Francisco, March 26th, 2016

Mario Stanke
University of Greifswald, Germany

Phenotypes and classification methods

I used only

- the training data of 145 genomes+phenotype profiles and
- the GWAS catalog to select SNPs associated with a phenotype in a previous study.

Only $K = 9$ phenotypes were used:

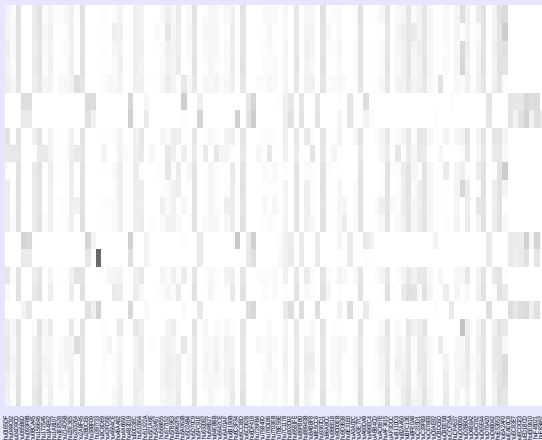
k	phenotype	prediction method	SNPs used	covariates
1	sex	ad hoc	presence of any SNP on chr Y	sex, age age sex, age
2	'baldness'	logistic regression	rs2497938	
3	astigmatism	logistic regression	rs10496034, rs1658442, rs7534824	
4	osteoarthritis	logistic regression	rs1417066, rs3757837	
5	blood type rh	logistic regression	rs590787	
	blood type ABO	conditional multinomial	rs8176719, rs8176747	
6	eye color blue	logistic regression	rs12913832	
7	floaters	6 nearest neighbors	random 1000 SNPs not on chr Y	
8	high triglycerides	6 nearest neighbors	random 1000 SNPs not on chr Y	
9	gingivitis	6 nearest neighbors	random 1000 SNPs not on chr Y	

Example 'baldness'

$$P(\text{bald}) = \frac{1}{1 + e^{-(-1.461 + 1.74 \cdot (\text{male} \ \& \ \text{age} \geq 50) - 2.74 \cdot \text{rs2497938})}}$$



Color Key



1.3

Made M quadratic

- Introduced 101 – 23 **dummy (missing) genomes** $\mathbf{x}_{n+1}, \dots, \mathbf{x}_N$:

$$M = \left(P(\mathbf{y}_j | \mathbf{x}_i) \right)_{1 \leq i, j \leq N}$$

- M was now a 101×101 stochastic matrix.
- 'Dummy' rows 24, ..., 101 of M constitute a **prior on phenotypes** and are important!

Matching example

$M =$	0.691	0.187	0.122	P1
	0.633	0.365	0.002	P2
	0.787	0.202	0.011	P3
	hu0165DF	hu048C92	hu1A8107	

- problem is **symmetric**:

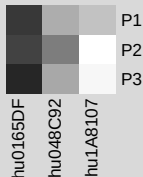
each genome phenotype profile assigned to exactly one phenotype profile genome

- seek a bijective mapping

$$\pi : \{1, \dots, 101\} \rightarrow \{1, \dots, 101\}$$

- π **depends on whole matrix M**

Mario Stanke



Predicting Phenotypes

Probabilities of
Matchings

Matching/permutation Π

The unknown matching is a permutation and is modeled as a random variable

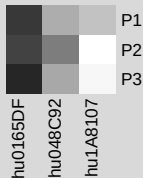
$$\Pi : \{1, \dots, N\} \rightarrow \{1, \dots, N\}.$$

Joint distribution of matching, genomes and phenotype profiles:

$$P(\Pi = \pi, \mathbf{x}, \mathbf{y}) = \underbrace{P(\Pi = \pi)}_{=1/N!} \cdot P(\mathbf{x}) \cdot \prod_{i=1}^N P(\mathbf{y}_{\pi(i)} \mid \mathbf{x}_i),$$

where $\mathbf{x} := \mathbf{x}_1, \dots, \mathbf{x}_N$ and $\mathbf{y} := \mathbf{y}_1, \dots, \mathbf{y}_N$.

Mario Stanke



Predicting Phenotypes

Probabilities of
Matchings

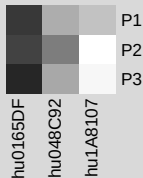
I submitted an estimate of the doubly stochastic matrix S

$$S := (s_{i,j}) := \left(P(\Pi(i) = j | \mathbf{x}, \mathbf{y}) \right)_{1 \leq i, j \leq N}$$

- Entry $s_{i,j}$ of S holds the probability that genotype i is matched with phenotype profile j , **given all the genomes and profiles**.
- The **rows and columns of S add up to 1**:
Each phenotype is matched to exactly 1 genotype and vice versa.
- Computing $s_{i,j}$ exactly from M is equivalent to computing the **permanent of the matrix** (e.g. Volkovs and Zemel, NIPS 2012) obtained from M by deleting the i -th row and j -th column.

$$\text{perm}(A) = \sum_{\pi \in \Sigma_N} \prod_{i=1}^N a_{i, \pi(i)}$$

Mario Stanke



Predicting Phenotypes

Probabilities of
Matchings

Approximating S using Markov Chain Monte Carlo (MCMC)

Sample from the posterior distribution on matchings

$$P(\pi | \mathbf{x}, \mathbf{y}) \propto \prod_{i=1}^N P(\mathbf{y}_{\pi(i)} | \mathbf{x}_i).$$

- Markov chain defined on the set of all permutations with $P(\pi | \mathbf{x}, \mathbf{y})$ as limiting distribution
- implemented `sample-matchings`, Metropolis algorithm, C++, chain was run for 10^7 iterations, < 3 min
- <http://bioinf.uni-greifswald.de/bioinf/cagi/sample-matchings-0.1.tar.gz>

Example: computing S from M

$M =$	0.691	0.187	0.122	P1
	0.633	0.365	0.002	P2
	0.787	0.202	0.011	P3
	hu0165DF	hu048C92	hu1A8107	

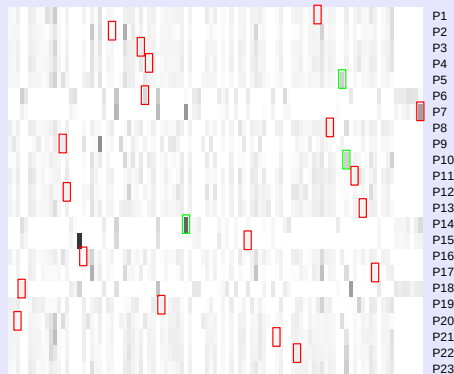
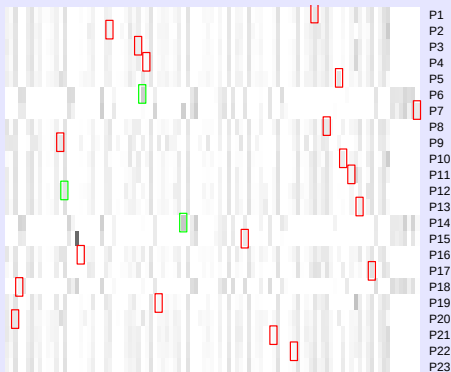
$S =$	0.053	0.027	0.92	P1
	0.311	0.681	0.008	P2
	0.637	0.292	0.071	P3
	hu0165DF	hu048C92	hu1A8107	

Evaluation: same number of hits, but better likelihood

M



S

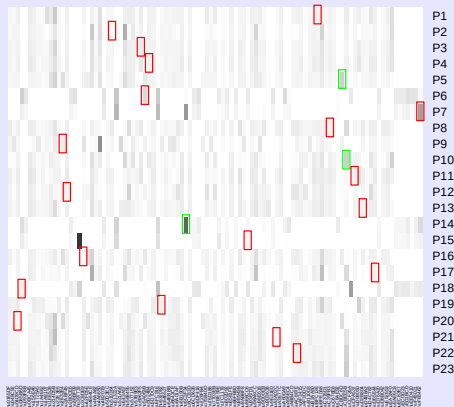
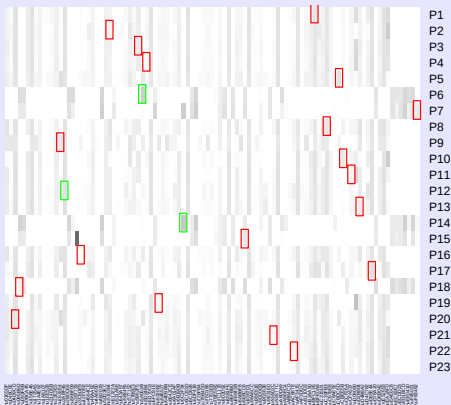


Evaluation: same number of hits, but better likelihood

M



S



$\pi :=$ true matching (rectangles)

$$\text{loglik}(A) := \ln \left(\prod_{i=1}^{23} A[i, \pi(i)] \right)$$

$$\text{loglik}(M) \approx -93.1$$

$$\text{loglik}(S) \approx -89.3$$