


SNP analysis linking between genotype and phenotype

Susana Vinga, Jonas Almeida

Symposium of the PNEUMOPATH project
ITQB
Oeiras, 27 June 2012



Instituto de Engenharia de Sistemas e Computadores Investigação e Desenvolvimento em Lisboa

Problem

Given:

- Set of genomes from bacteria
- Set of (replicate) phenotype information

- How to find the SNPs associated with given phenotypes

Data – case-study

- Whole **genome sequences** of 9 bacteria strains: BHN418 ,BHN191 ,CBR206 ,LgtSt215 ,SP14-BS69 ,SP3-BS71 ,BHN100 ,TIGR4 ,D39
- Phenotype information for 3 mice strains:
 - MF1, BALB/c, CBA/Ca
 - 10 replicates: Survival , Blood24, Blood, Lungs

Workflow Genomic analysis

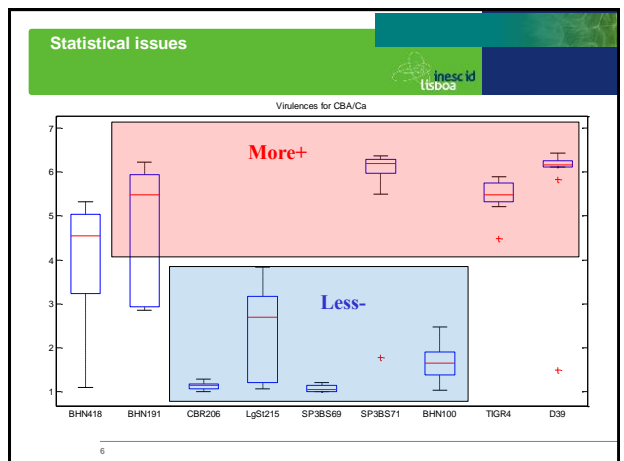
- Multiple alignment of homologous genes
 - 10 strains: BHN418, BHN191, CBR206, LgtSt215, SP14-BS69, SP3-BS71, BHN100, TIGR4, D39
 - Threshold: 70% overlap (80% identity)
- Filter redundant files
 - 2525 homologous genes found
- Get SNP information for proteins
 - 33796 positions where at least one difference is found in the aminoacids

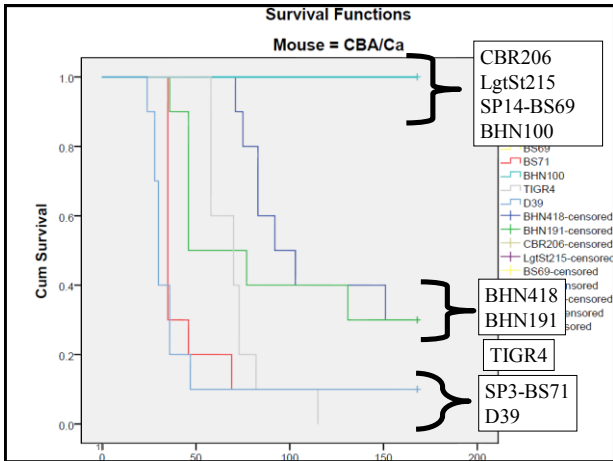
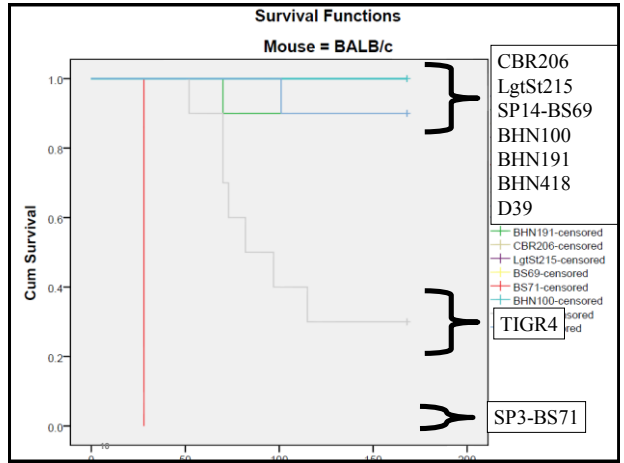
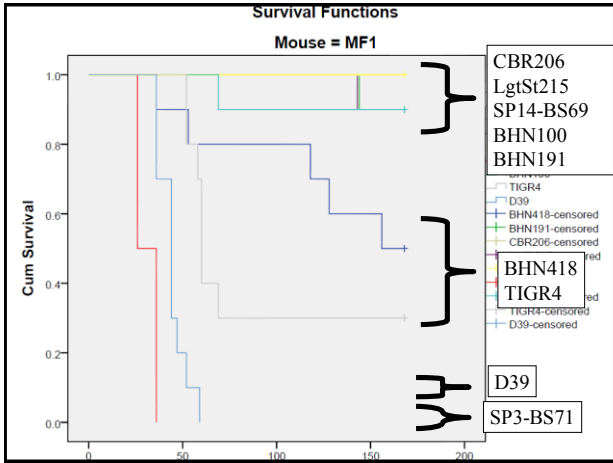
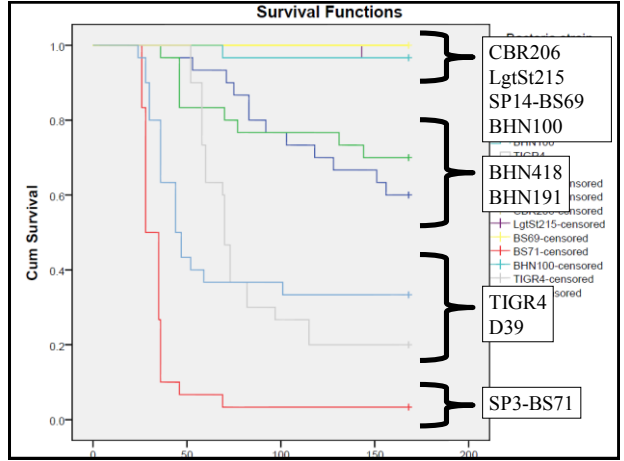
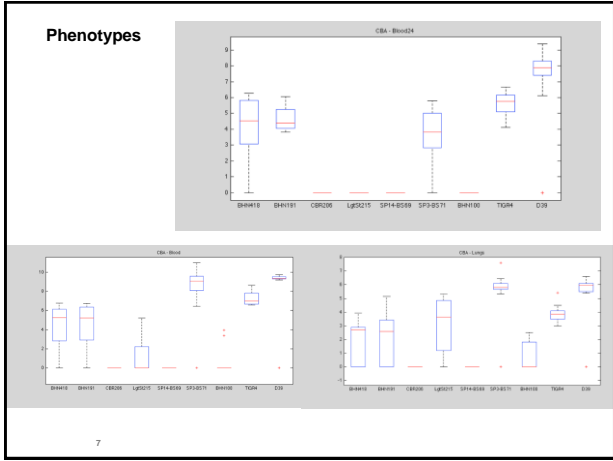
protein ecsB

BHN100_04521	BHN100_04521	G11A1HRLA9PIVA
>BHN100_04521 1464339 146558 11050bp	1	BHN100_04521
MKDLFLKRRQAFRKECLGIVRYVNDHFVFLVLLGFLA	2	BHN100_0452
YQVSQLQHFFHNHPILLFVGTISVLLGGIATYMEA	3	BHN191_0495
PKKLLVGGVEEIKLHKRGTIGISLVFVFLVQTLFLLFA	4	BHN418_0423
FLFLAMVGLFVLLVLLVGGVGYFHRQKSKFTEETG	5	CBR206_0523
LNDWYLIQSSEKRRQVLLRFFALFTQVKGISNSVKRRAYL	6	DCCF215_0461
DFLTKAVQVFGKIWQNLVRSYLRGDLFALSRLLLLS	7	SP3_0524
VLAQVFEQAWIATAVVLFRVYLLFPQLLALYHAFDYQYL	8	SPD_0465
TQLFLKRSQKESLQEVVR	9	SP_0523-G11ALHCVAPGLLA
>BHN191_0495 1466099 1467148 11050bp	10	SP_0460
MKDLFLKRRQAFRKECLGIVRYVNDHFVFLVLLGFLA		G11ALHRVAPGIVA
YQVSQLQHFFHNHPILLFVGTISVLLGGIATYMEA		
PKKLLVGGVEEIKLHKRGTIGISLVFVFLVQTLFLLFA		
FLFLAMVGLFVLLVLLVGGVGYFHRQKSKFTEETG		
LNDWYLIQSSEKRRQVLLRFFALFTQVKGISNSVKRRAYL		

matrix_c.txt

33796 SNPs





Question

inesc id
tsboa

- How to relate any continuous variable/phenotype with the genome/SNPs?
 - Compare bacteria segregation/partitioning generated by each SNPs
 - Test for statistical differences in the generated groups
 - Sort results from most (statistically) significant to least (statistically) significant differences (better: define p-value threshold)

12

SNP → Phenotypes

- SNP's → check strain partition defined by *each* SNP

```

1 BHN100.....RNASYGLALRLAPGLVA.....PS.....GAVLLE
2 BHN191..TAVELV...SYGILALRCLAPSLVAVAGCAH...SFT.....GAVLII
3 BHN418..TAVELV...SYGILALRCLAPSLVAVAGCAH...SFT.....GAVLII
4 CBR206.....SDGILALHRVAPGIVAVERYVY...GSVYA.....AVLII
5 DCCN215.....SDGILALHRVAPGIVAVERYVY...GSVYA.....AVLII
6 SP7.....RD...GTLALRCVAPGLLAVAGCAH...SFTV...SVCVHTIRQDAVLII
7 SP14.....CPTSD.....GTFY.....
8 SP3.....RHANYCIIVLVRVYSGLLS.....TDKPS.....TMFLE
9 SPD.....IVIDEICH...GTLALHRVAPGIVAVERYVY...GSVYA.....AVLII
10 spz.....IVIDEICH...GTLALHRVAPGIVAVERYVY...GSVYA.....AVLII
11 H
M
M
T
T
M
T
T
M
T

```

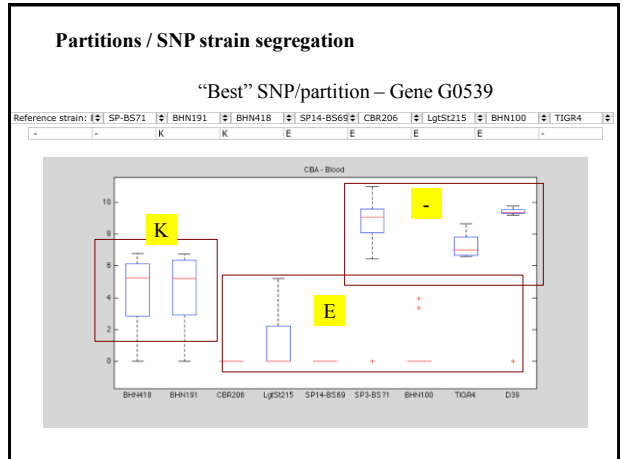
Compare phenotypes of the groups → statistically significant differences?
 H_0 : There are no differences → p-values

Classification

- 1575 distinct partitions defined by all SNPs
 - define 2 to 7 clusters of the strains
- Compare clusters
 - Kruskal-Wallis statistical tests (non-parametric, medians)
- Sort SNP's according to p-values
 - csv files with classification

csv files

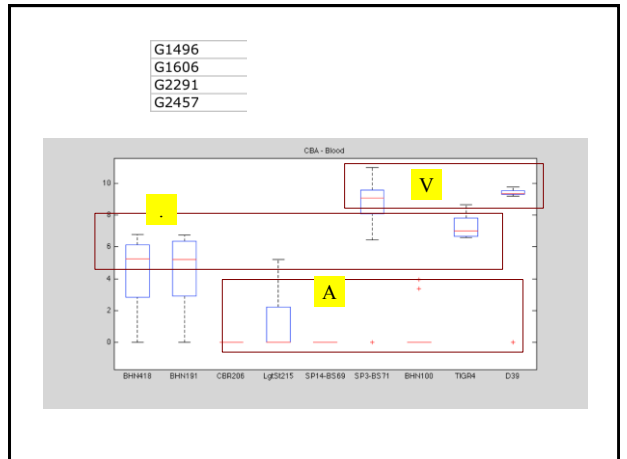
Gene name	Gene description	SNP_ID	SNP position	log(p-value)	Reference	SP-BS71	BHN191	BHN418	SP14-BS69	CBR206	LgtS215	BHN100	TIGR4	SNP name
G0539	(hypothetical protein)	12402	57	22.808858	T	C	C	C	C	C	C	C	C	SP-BS71
G1214	(hypothetical protein)	12402	57	22.808858	T	C	C	C	C	C	C	C	C	BHN191
G1214	(hypothetical protein)	12402	57	22.808858	T	C	C	C	C	C	C	C	C	BHN418
G1214	(hypothetical protein)	12402	57	22.808858	T	C	C	C	C	C	C	C	C	SP14-BS69
G1214	(hypothetical protein)	12402	57	22.808858	T	C	C	C	C	C	C	C	C	CBR206
G1214	(hypothetical protein)	12402	57	22.808858	T	C	C	C	C	C	C	C	C	LgtS215
G1214	(hypothetical protein)	12402	57	22.808858	T	C	C	C	C	C	C	C	C	BHN100
G1214	(hypothetical protein)	12402	57	22.808858	T	C	C	C	C	C	C	C	C	TIGR4

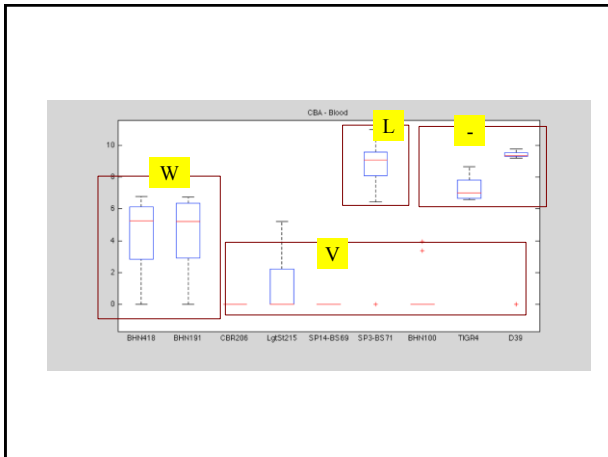


Same partition: 13 SNPs / 5 genes

- G0539; G1214; G2349; G2867; G3025

No missing orthologous





Consensus combined p-values Phenotype meta-analysis



- How to combine results from k statistical tests
→ pool individual p-values:
 - Fisher's combined probability test (2k degrees of freedom)

$$\chi^2 = -2 \sum_{i=1}^k \ln(p_i)$$

- Stouffer's Z transform test
 - $p_i \rightarrow Z_i$ (normal distribution) $Z_s = \frac{\sum_i Z_i}{\sqrt{k}}$
- Psum or Pmean
- Pmax and Pmin (are NOT consensus)

20

Combine p-values



- GOAL: find a reasonable number of SNP's for further analysis by combining several experiments (meta-analysis) → hypothesis generation
 - side effect: phenotype correlations
- Use rank ordering

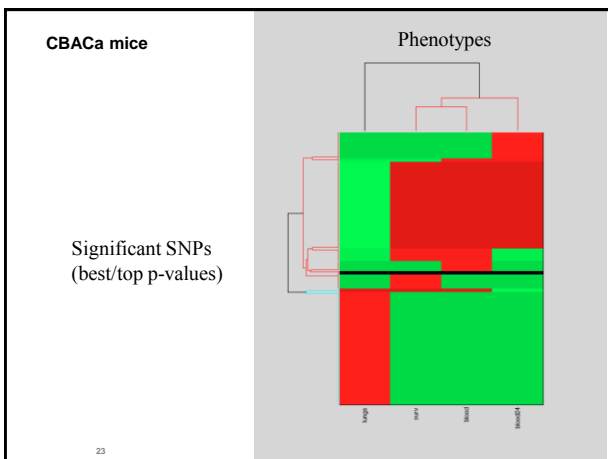
21

Rank combination and Biclustering

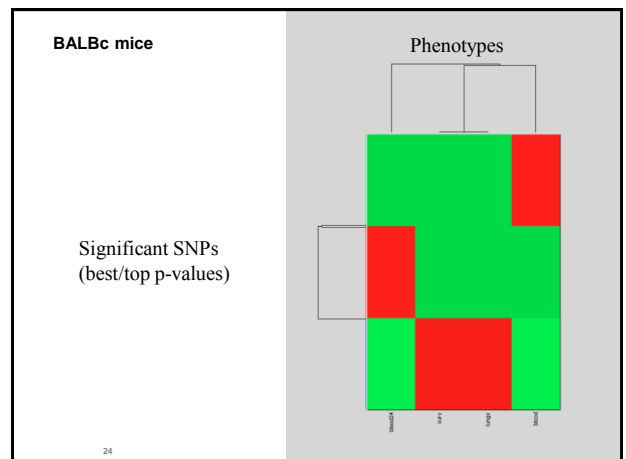


- Biclustering - simultaneous clustering of the rows and columns of a matrix
 - Find clusters constituted by **subsets** of rows and **subsets** of columns
- Select top SNPs for each phenotype
 - include SNPs that are in that set in *at least* one phenotype

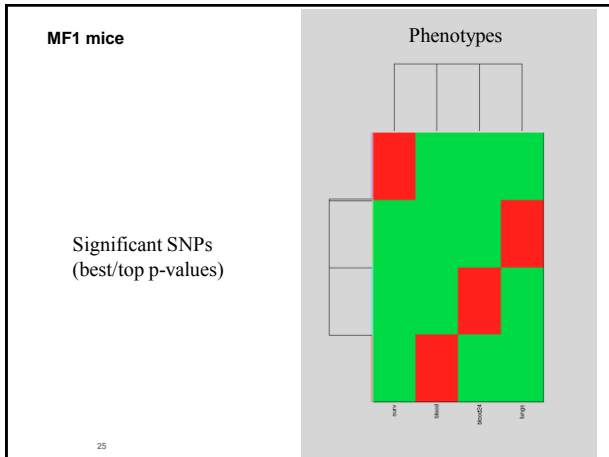
22



23



24



Web-based application

- Analyses any phenotype
- Submit SNPs (or pre-inserted)
- For each strain, insert phenotypic measurement (virulences, survival rates, parameters,...), with or without replicates
- Get most significant SNPs
- Links to external information

26

Discussion

- Consider SNP location and abundance per gene
- Improve statistical tests, permutation and multiple testing corrections
- Define p-value threshold for significance (or is ordering sufficient?)
- Validate generated hypotheses (several SNPs correspond to same gene)

27

Acknowledgments

- KDBIO group INESC-ID
 - Alexandre Francisco
- Univ. Leicester
 - Magdalena Jonczyk
 - Peter Andrew
- Karolinska Institutet
 - Birgitta Henriques-Normark
- ITQB-UNL
 - Helen Andrade Arcuri
 - Raquel Sá-Leão
 - Hermínia de Lencastre
- Univ. Siena
 - Marco Oggioni

HEALTH-F3-2009-222983

28