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

1

Problem

Given:

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

- How to find the SNPs *associated* with given phenotypes

2

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

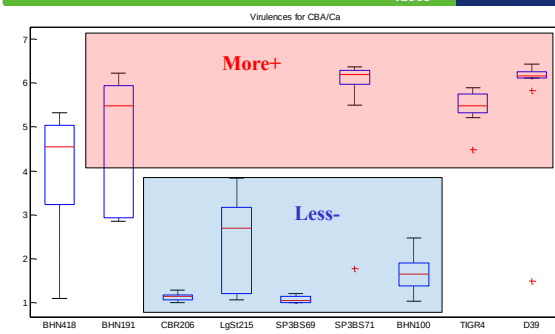
3

Workflow
Genomic analysis

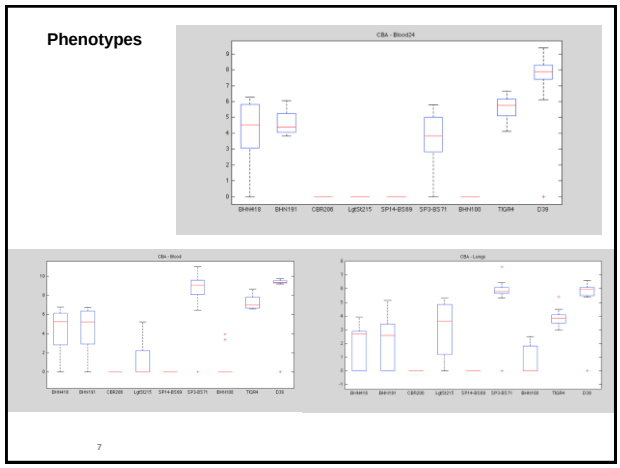
- Multiple alignment of homologous genes
  - 10 strains: BHN418, BHN191, CBR206, LgtSt215, SP14-BS69, SP3-BS71, BHN100, TIGR4, D39
  - Threshold: 70% overlap (80% identity)
- Filter redundant files
  - 2525 homologous genes found
- Get SNP information for proteins
  - 33796 positions where at least one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

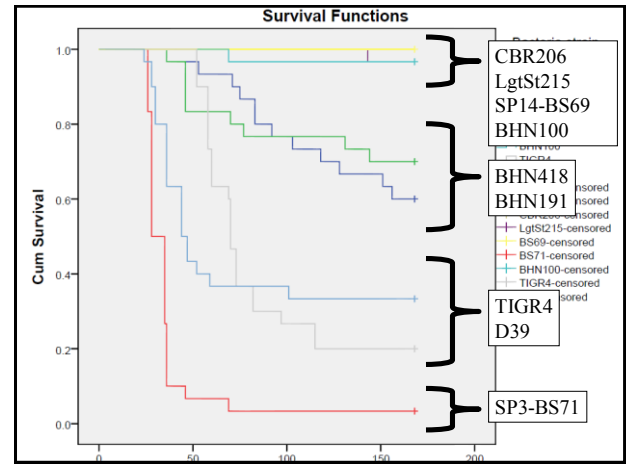
4

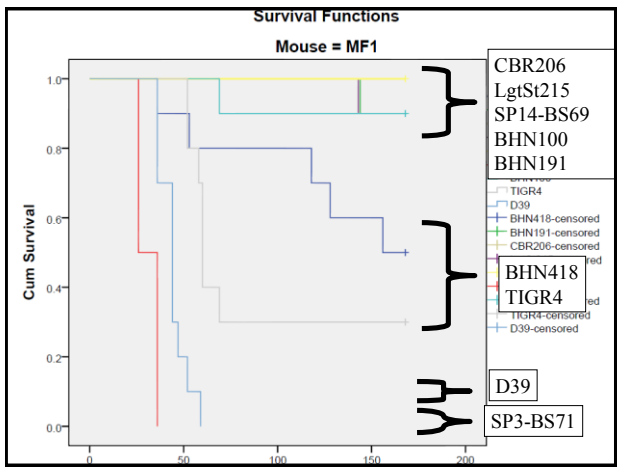
Statistical issues

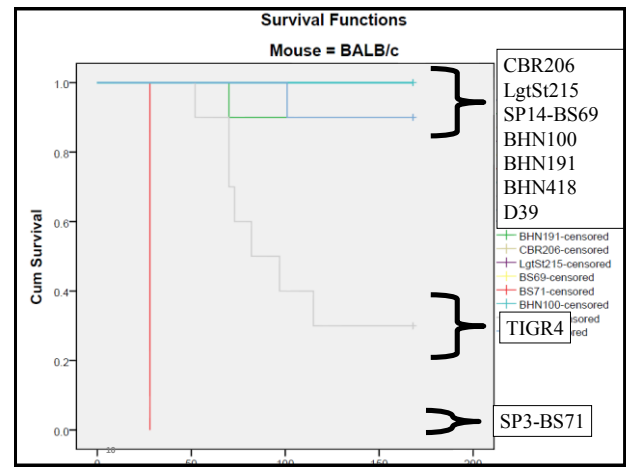


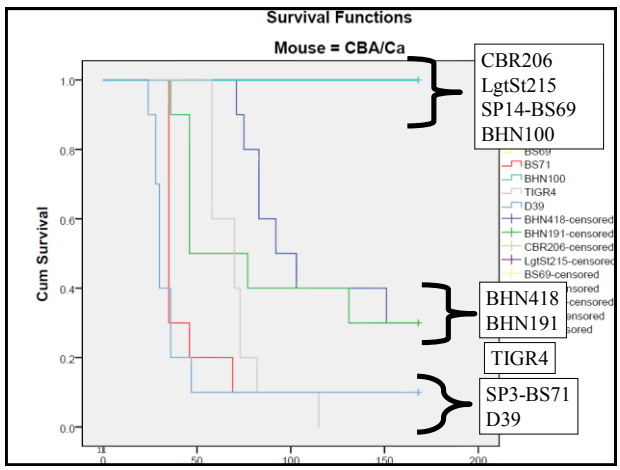
6









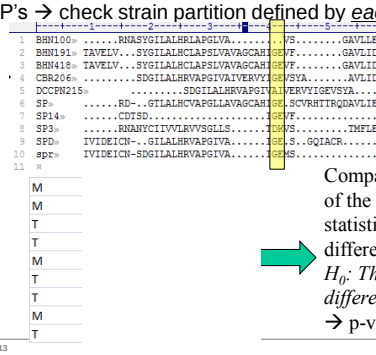


Question

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

SNP → Phenotypes

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



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

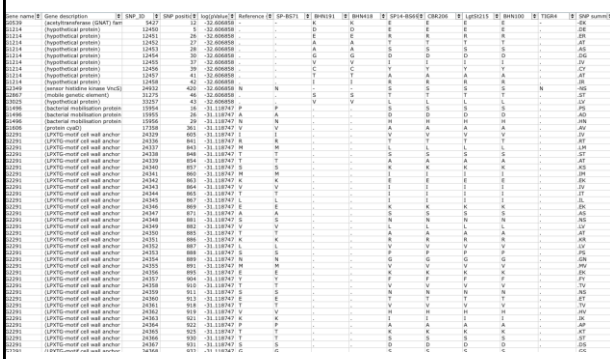
13

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

14

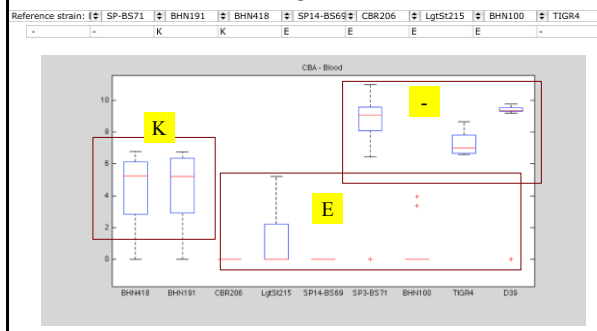
csv files



15

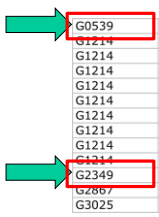
Partitions / SNP strain segregation

"Best" SNP/partition – Gene G0539



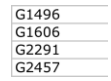
Same partition: 13 SNPs / 5 genes

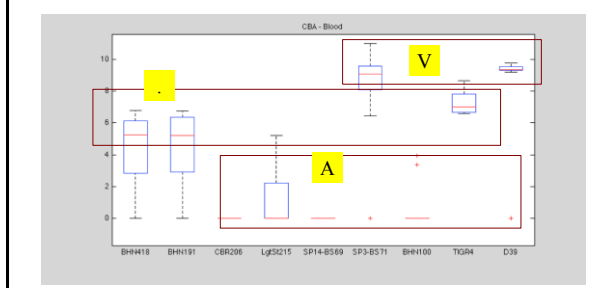
- G0539; G1214; G2349; G2867; G3025

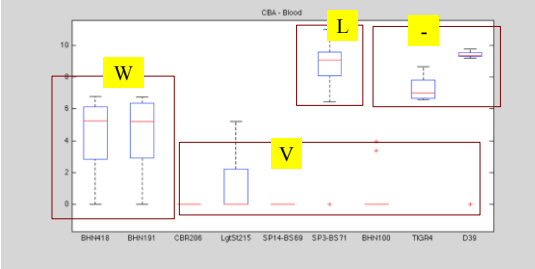


No missing orthologous

17







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)

$$\chi^2 = -2\sum_1^k \ln(p_i)$$

  - Stouffer's Z trasform test
    - pi → Zi (normal distribution) $Z_s = \frac{\Sigma_i^k Z_i}{\sqrt{k}}$
  - 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

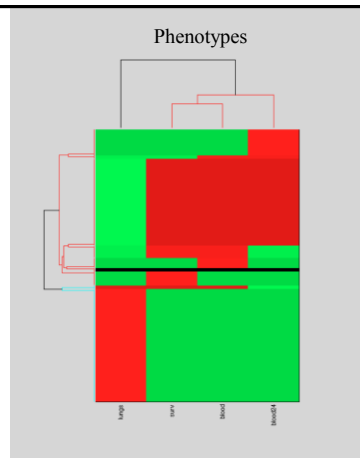
Rank combination and Biclustering

- 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

CBACa mice

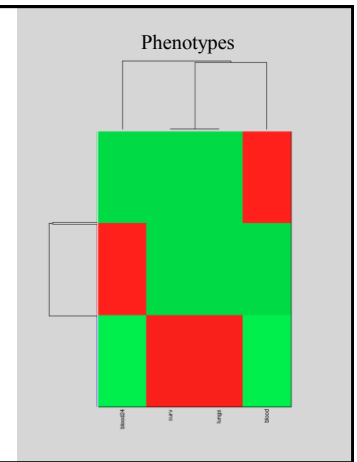
Significant SNPs (best/top p-values)



23

BALBc mice

Significant SNPs (best/top p-values)

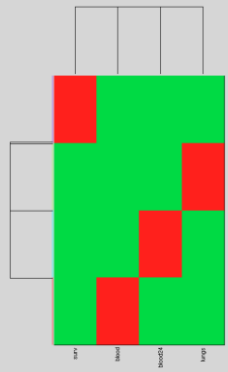


24

MF1 mice

Significant SNPs
(best/top p-values)

Phenotypes



25

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