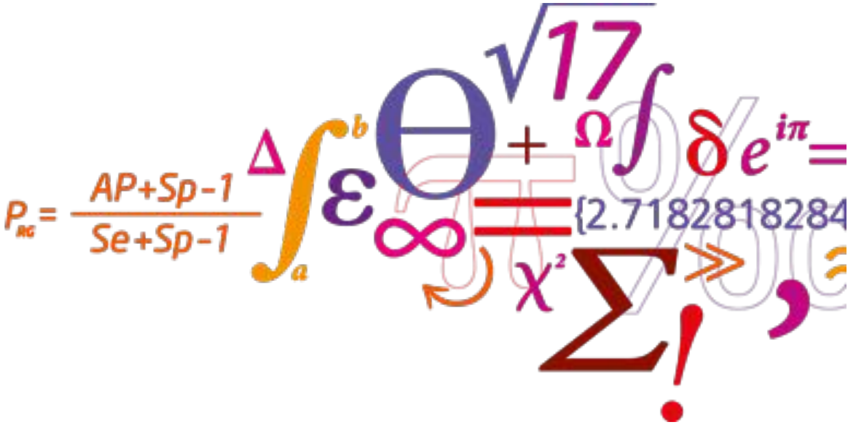
Statistical Genotype-phenotype Correlation

Leon Eyrich Jessen

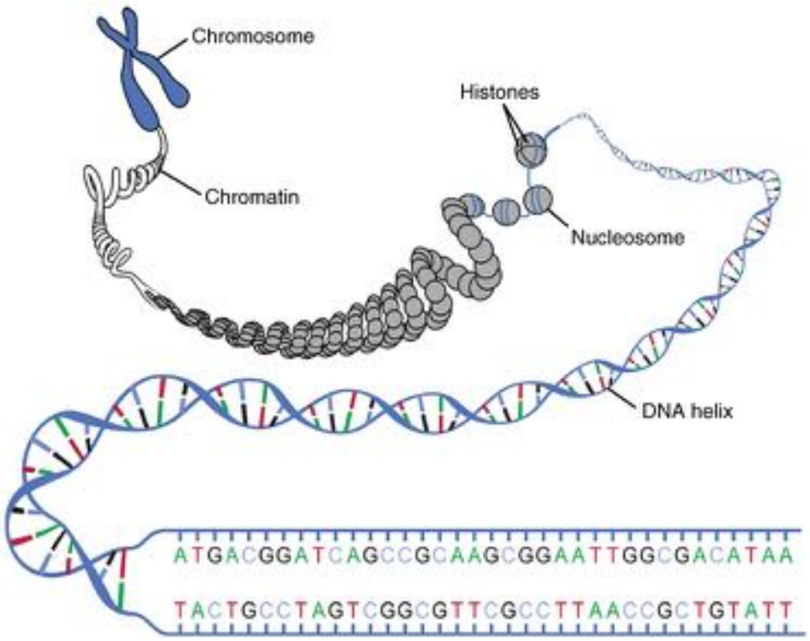
36685 - Immunological Bioinformatics January 2018



Immunoinformatics and Machine Learning
Department of Bio and Health Informatics

Phenotype – What you “observe”

Genotype – Determinants of what you “observe”


Chromosome
Histones
Chromatin
Nucleosome
DNA helix
ATGACGGATCAGCCGCAAGCGGAATTGGCGACATAA
TACTGCCTAGTCGGCGTTCGCCTTAACCGCTGTATT

Example of genotype-phenotype correlation

- GWAS (Genome Wide Association Study)

Height

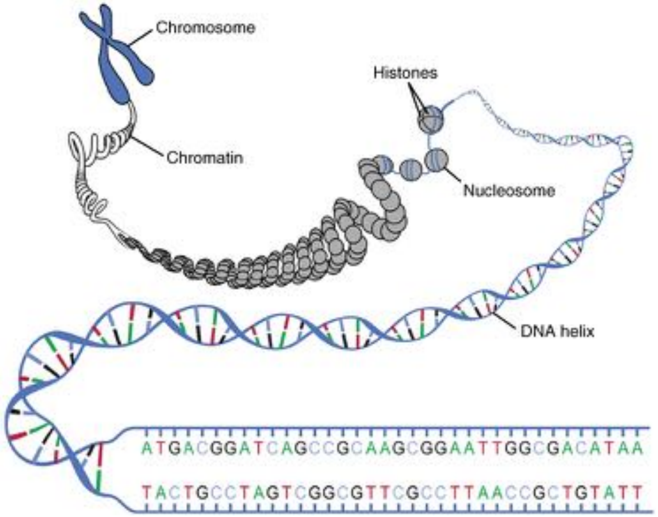
Eye colour

Hair colour

Facial hair

Disease

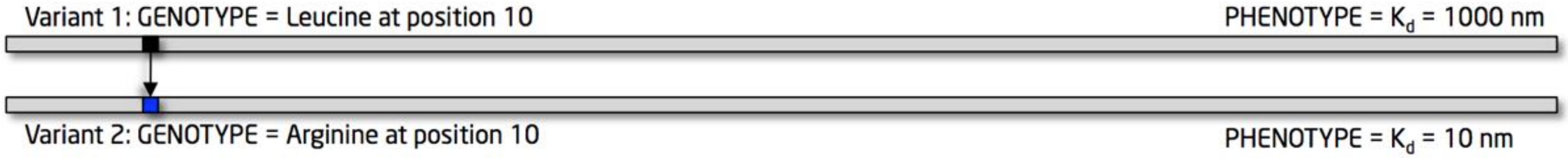
Etc.



Chromosome
Histones
Chromatin
Nucleosome
DNA helix
ATGACGGATCAGCCGCAAGCGGAATTGGCGACATAA
TACTGCCTAGTCGGCGTTCGCCTTAACCGCTGTATT

Molecular level genotype-phenotype correlation

- Two variants of the same protein with two different phenotypes

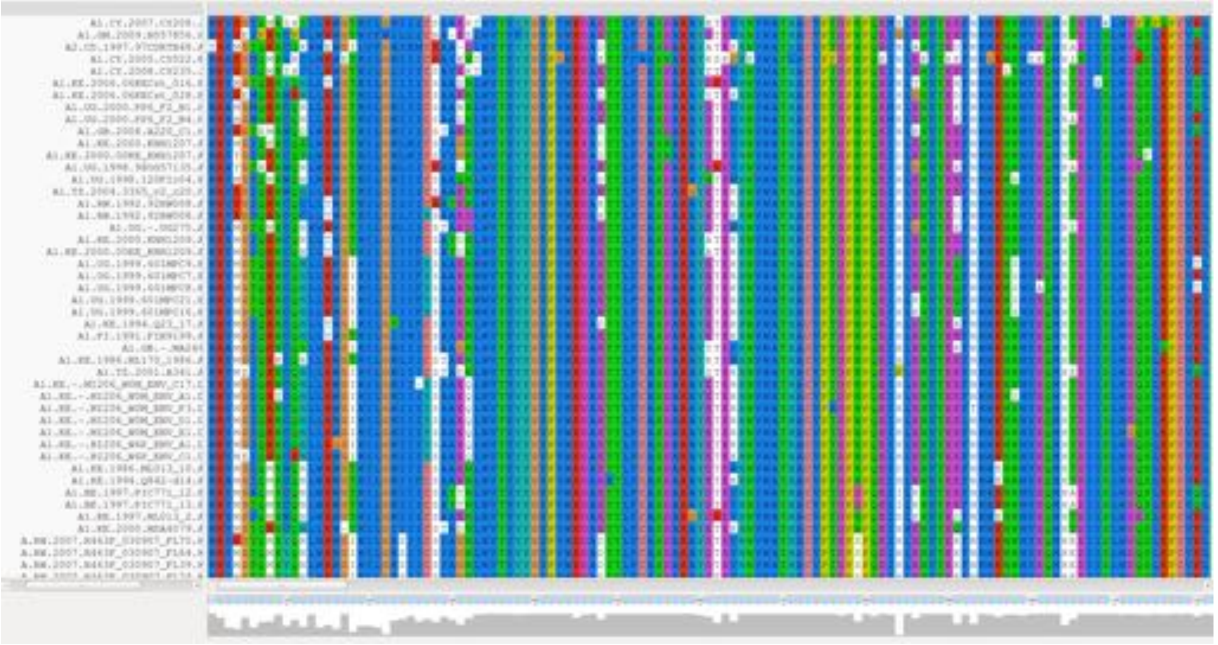


Variant 1: GENOTYPE = Leucine at position 10
PHENOTYPE = K_d = 1000 nm
Variant 2: GENOTYPE = Arginine at position 10
PHENOTYPE = K_d = 10 nm

- Clearly L10R has a significant impact on the phenotype
- But…

Molecular level genotype-phenotype correlation

- What if you have 300 variants each with 10-20 mutations?
- How to figure out what’s what in terms of genotypes and phenotypes?



Enter: SigniSite

- Input: A multiple sequence alignment (MSA) with a numerical value to each sequence
- Algorithm:
  1. Form a MSA of different variants of the same protein
  2. Rank the associated numerical values
  3. Sort the sequences based on the ranks
  4. For each amino acid residue at each position:
     1. Calculate the mean observed rank
     2. Compare with the mean expected rank
     3. Derive z-score from comparison
  5. Form z-score matrix with number of rows corresponding to the number of positions in the MSA and number of columns corresponding to the number of proteogenic amino acids (20)
  6. Adjust z-scores for multiple comparisons (each z-score calculated is one test)
  7. The adjusted z-score matrix can now be viewed as a Position Specific Scoring Matrix and we can create sequence logo based on calculated adjusted z-scores

Let us take a closer look

- After forming an MSA and sorting the sequences, we have this observed amino acid distribution at p_{13}
- Do genotype arginine at p_{13} significantly impact the phenotype?

R	1
R	2
K	3
K	4
R	5
R	6
R	7
H	8
N	9
S	10
M	11
W	12
N	13
N	14
H	15
D	16
D	17
E	18
P	19
D	20

Do the GT R at p_{13} significantly impact the PT?

- The test is rank based, so we want to test the observed mean rank with the expected mean rank and then test the hypothesis:

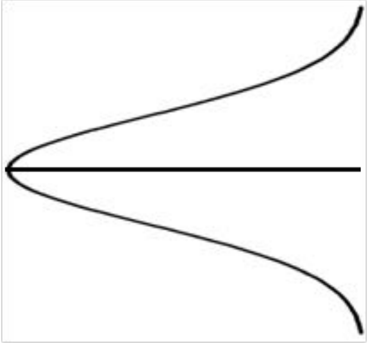
$$H_0: \quad \mu_{p,a}^{exp} = \overline{x}_{p,a}^{obs} \quad H_1: \quad \mu_{p,a}^{exp} \neq \overline{x}_{p,a}^{obs}$$

R	1
R	2
K	3
K	4
R	5
R	6
R	7
H	8
N	9
S	10
M	11
W	12
N	13
N	14
H	15
D	16
D	17
E	18
P	19
D	20

Do the GT R at p_{13} significantly impact the PT?

- The test is rank based, so we want to test the observed mean rank with the expected mean rank and then test the hypothesis:

$$H_0: \quad \mu^{exp}_{p,a} = \overline{x}^{obs}_{p,a} \qquad H_1: \quad \mu^{exp}_{p,a} \neq \overline{x}^{obs}_{p,a}$$



R	1
R	2
K	3
K	4
R	5
R	6
R	7
H	8
N	9
S	10
M	11
W	12
N	13
N	14
H	15
D	16
D	17
E	18
P	19
D	20

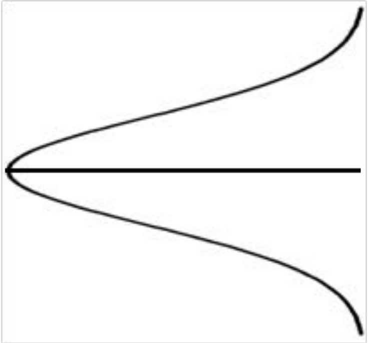
Do the GT R at p_{13} significantly impact the PT?

- The expected mean rank is

$$\mu_{p,a}^{exp} = \frac{N+1}{2}$$

- I.e. “the middle” of all amino acids, where N is the number of sequences in the MSA

- In this example we get (20 + 1)/2 = 10.5



R	1
R	2
K	3
K	4
R	5
R	6
R	7
H	8
N	9
S	10
M	11
W	12
N	13
N	14
H	15
D	16
D	17
E	18
P	19
D	20

$$H_0: \quad \mu_{p,a}^{exp} = \overline{x}_{p,a}^{obs} \qquad H_1: \quad \mu_{p,a}^{exp} \neq \overline{x}_{p,a}^{obs}$$

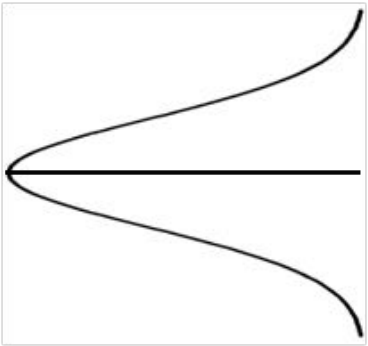
Do the GT R at p_{13} significantly impact the PT?

- The observed mean rank is

$$\overline{x}^{obs}_{p,a} = \frac{1}{n_{p,a}} \sum_{i=1}^{N} rank_{p,b_i} \cdot \delta(b_i,a)$$

- I.e. “the middle” of the observed amino acid, where n is the number of the particular amino acid were testing

- In this example we get (1 + 2 + 5 + 6 + 7) / 5 = 4.2



R	1
R	2
K	3
K	4
R	5
R	6
R	7
H	8
N	9
S	10
M	11
W	12
N	13
N	14
H	15
D	16
D	17
E	18
P	19
D	20

$$H_0: \quad \mu^{exp}_{p,a} = \overline{x}^{obs}_{p,a} \qquad H_1: \quad \mu^{exp}_{p,a} \neq \overline{x}^{obs}_{p,a}$$