

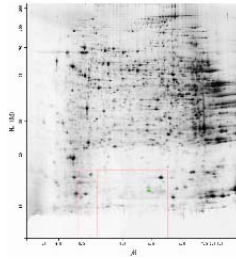
P – Proteomics

The Size of the Proteome:

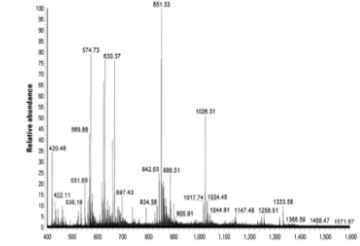
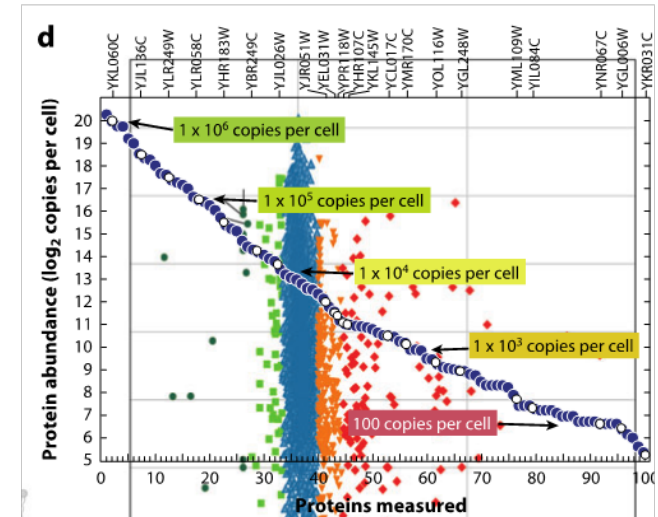
- 24.000 genes
- Alternative Splicing
- Post-translational modifications
 - Phosphorylation of especially serine and threonine
 - Glycolysation
 - Ubiquitination

Experimental techniques:

- 2D electrophoresis



- Mass Spectroscopy



Analysis Techniques:

Segments of proteins have known weights, modifications create known weight changes.

148.2 261.3 376.4 491.5 606.6 719.8 820.9 936.0 1051.1 1164.2 1311.4 1472.6 1571.7
Phe [Leu [Asp [Asp [Asp [Leu [Thr [Asp [Asp [Ile [Met [Cys [Val [Lys

Properties of Data:

- Noisy
- Hard to make dynamic
- Quality improving quickly
- Qualitative
- Average over an ensemble of cells

M – Metabonomics

The Size of the Metabolome:

- *Set of small molecules*
 - *Combinatorial techniques allow exhaustive listing – extremely large numbers*
 - *Databases exists (eg Beilstein) with all empirically known – millions.*
 - *Standard textbook – maximally thousands. Observed tens of thousands*

Experimental techniques:

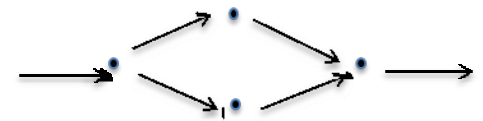
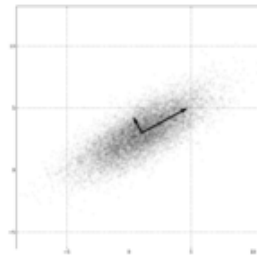
- *Gas chromatography*
- *Mass Spectroscopy*
- *Nuclear Magnetic Resonance (NMR)*

Objectives:

- *Mapping of Metabolism*
- *Disease Markers*

Analysis Techniques:

- *Principal Component Analysis*
- *Metabolic Network Analysis*



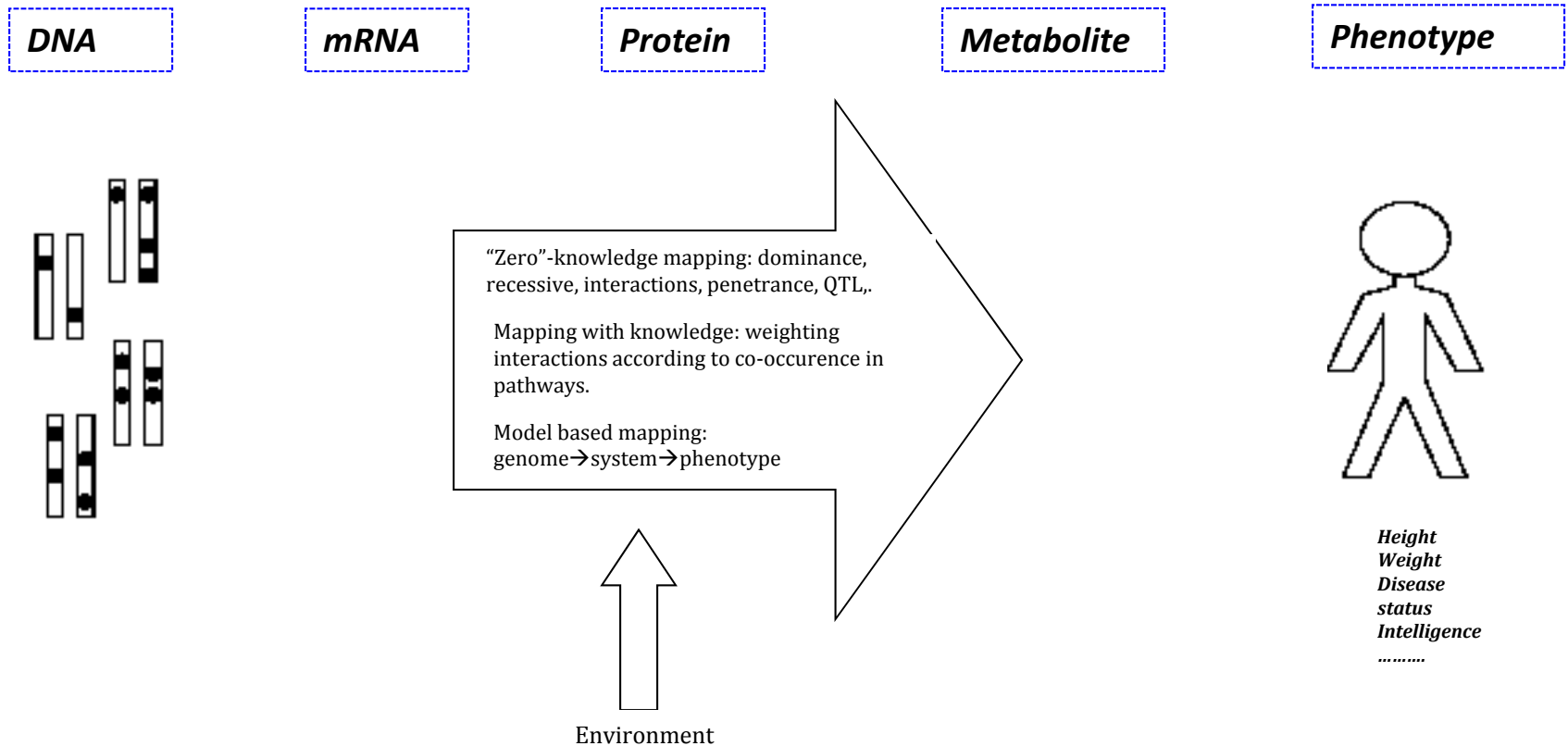
Properties of Data:

- *Noisy*
- *Hard to make dynamic*
- *Quality improving quickly*
- *Qualitative*
- *Average over an ensemble of cells*

$G \rightarrow F$

- *Mechanistically predicting relationships between different data types is very difficult*
- *Empirical mappings are important*
- *Functions from Genome to Phenotype stands out in importance*

G is the most abundant data form - heritable and precise. F is of greatest interest.



Phenomics

Why care about phenotypes ??

They define an individual [including disease status]

Evolution/Selection works on G via F:

How to define a phenotype??

Major issues:

Measurement error

Correlation between traits

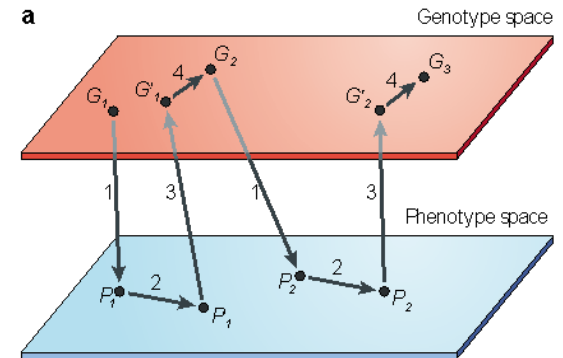
Life stage and environmental dependencies?

Special Traits:

Fitness

Disease susceptibility

Molecular Phenotypes



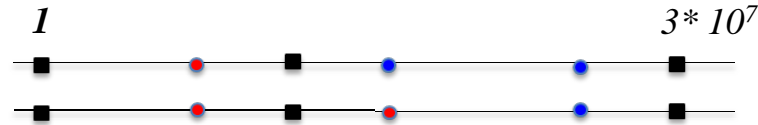
International effort in systematic collection of phenotypes especially in model organisms

<i>Arabidopsis</i>	Collaboration between groups working on <i>A. thaliana</i> , resulting in phenotyping and GWA studies on an overlapping set of 191 inbred lines	USNSF and NIH funding for initial GWA studies. Phenotyping supported by diverse grants to individual investigators	107 mostly quantitative phenotypes included in initial GWA studies ¹²⁶ , including resistance to pathogens, flowering traits, ionome and life history traits. No intensive phenotypes	250,000 SNPs genotyped using a chip	• http://walnut.usc.edu/2010/GWA
--------------------	---	--	--	-------------------------------------	---

The General Problem is Enormous

Set of Genotypes:

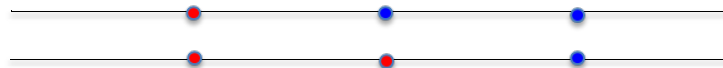
- Diploid Genome**



- In 1 individual, $3 * 10^7$ positions could segregate.
- In the complete human population $5 * 10^8$ might segregate.
- Thus there could be $2^{500,000,000}$ possible genotypes

Partial Solution: Only consider functions dependent on few positions

- Causative for the trait**



Classical Definitions:

- Single Locus**

Dominance

Recessive

Additive

Heterotic

- Multiple Loci**

Epistasis: The effect of one locus depends on the state of another

Quantitative Trait Loci (QTL). For instance sum of functions for positions plus error term.

$$\sum_{i \text{ causative positions}} X_i(G_i) + \varepsilon$$

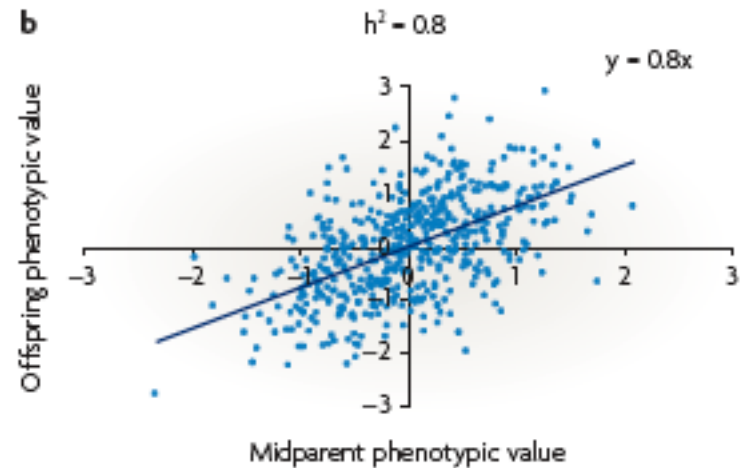
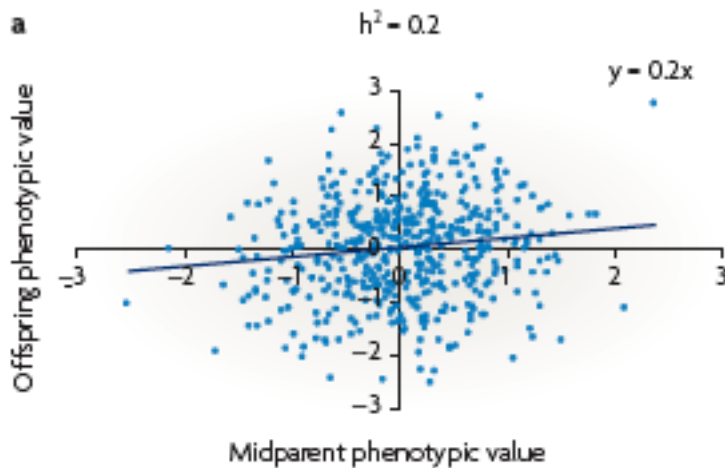
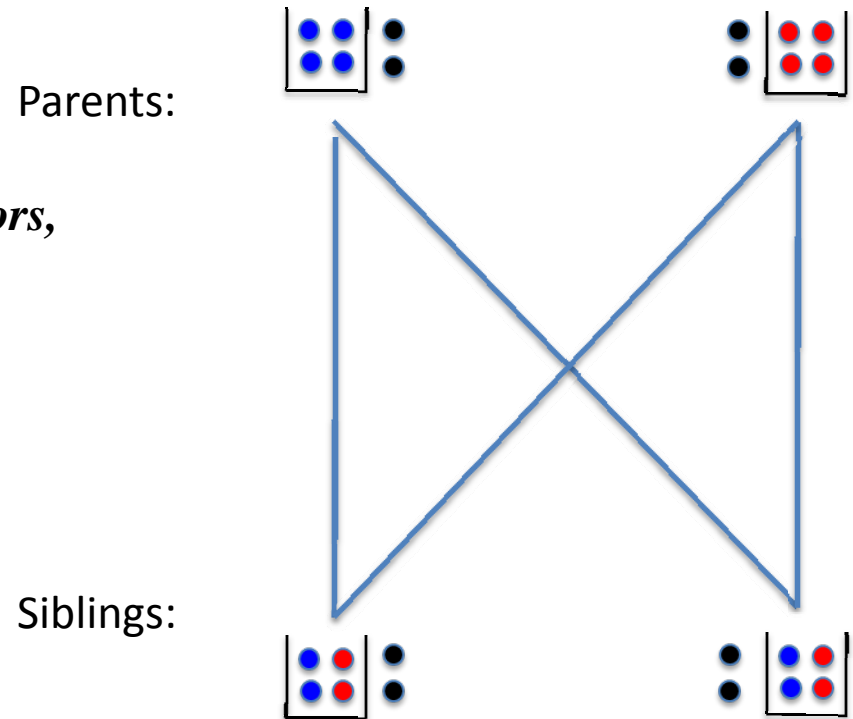
Heritability: Inheritance in bags, not strings.

The Phenotype is the sum of a series of factors, simplest independently genetic and environmental factors: $F = G + E$

Relatives share a calculatable fraction of factors, the rest is drawn from the background population.

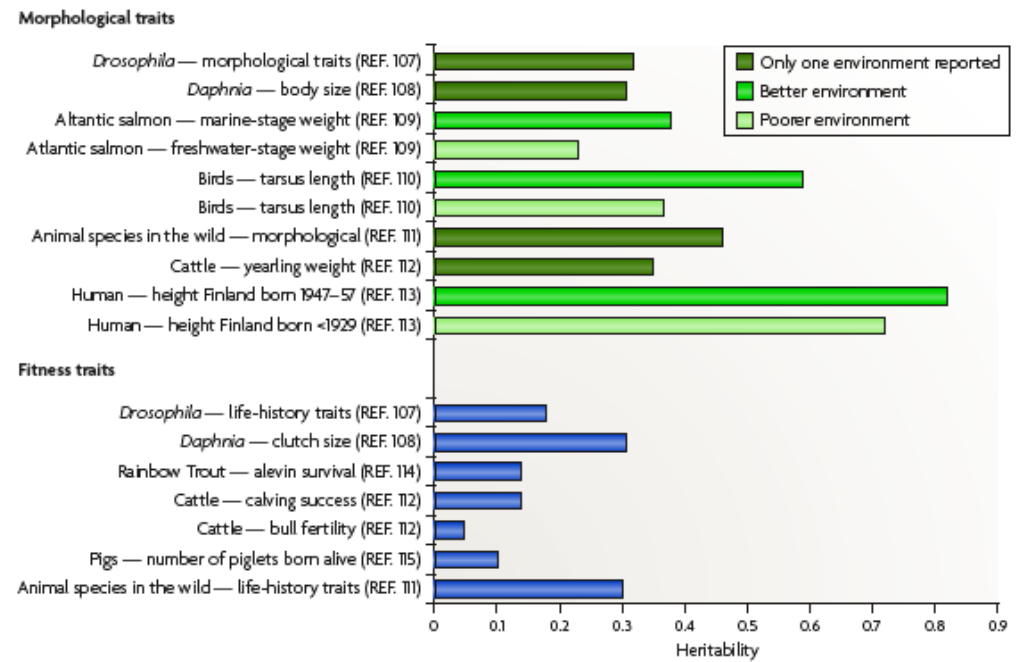
This allows calculation of relative effect of genetics and environment

Heritability is defined as the relative contribution to the variance of the genetic factors: σ_G^2 / σ_F^2

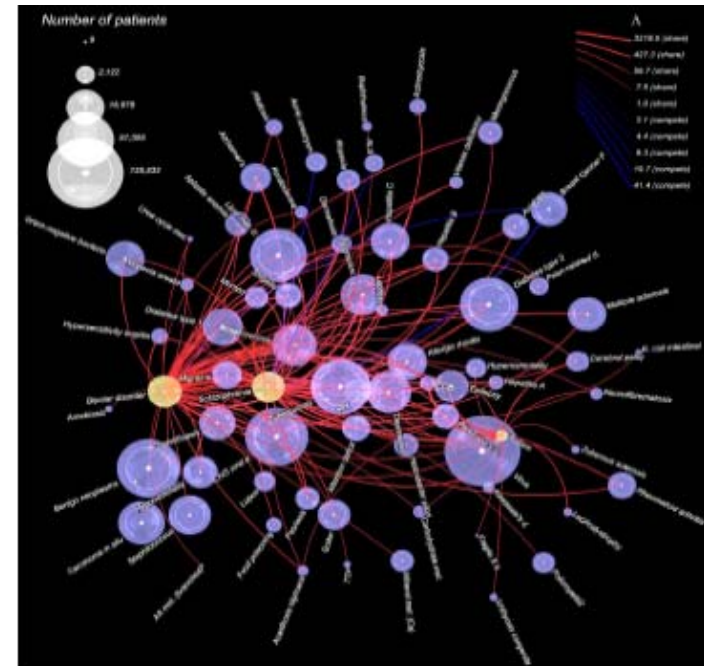
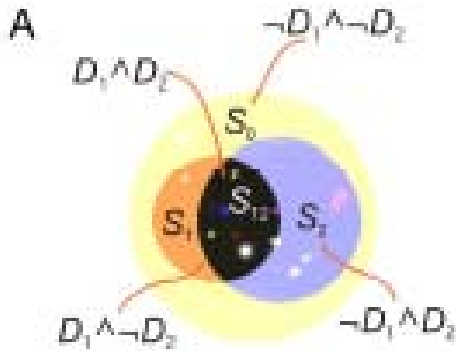


Heritability

Examples of heritability

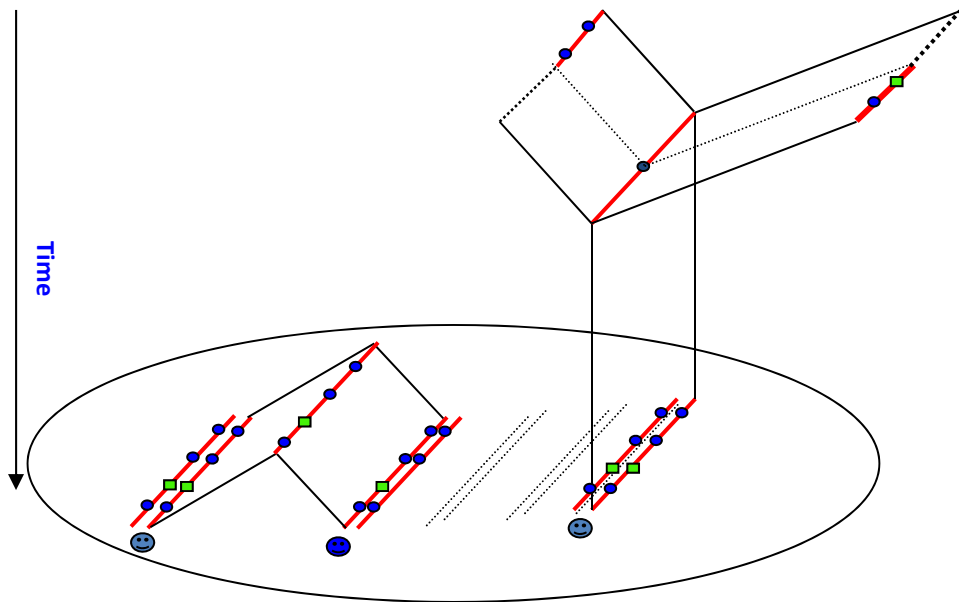


Co-Heritability of multiple characters:

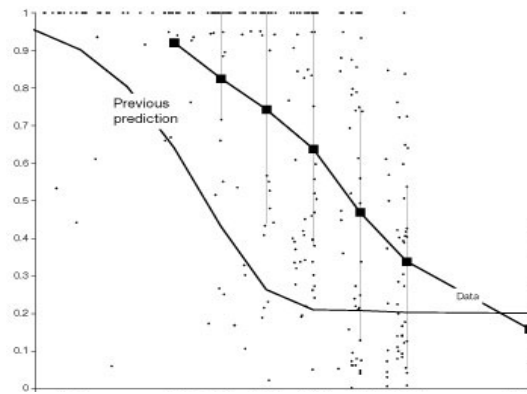


Genotype and Phenotype Co-variation: Gene Mapping

Sampling Genotypes and Phenotypes

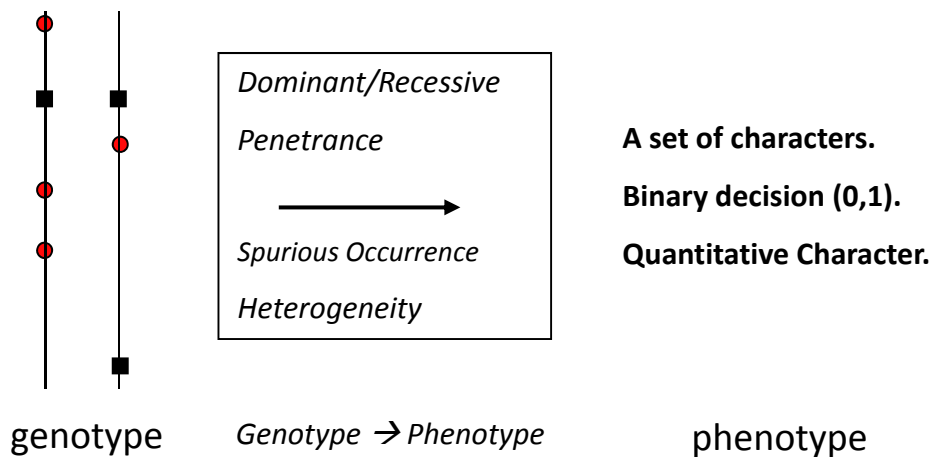


Decay of local dependency

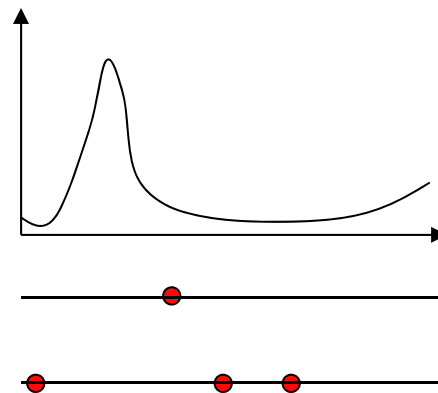


Reich *et al.* (2001)

Genotype --> Phenotype Function

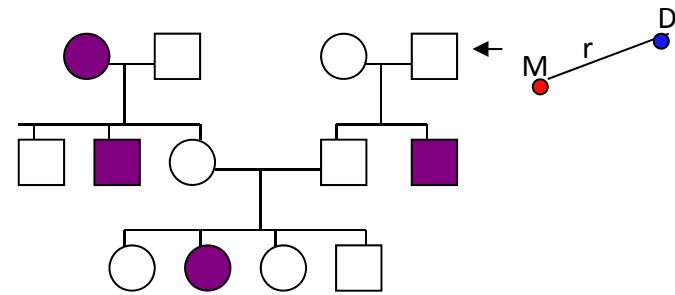


Result: The Mapping Function



Pedigree Analysis & Association Mapping

Pedigree Analysis:

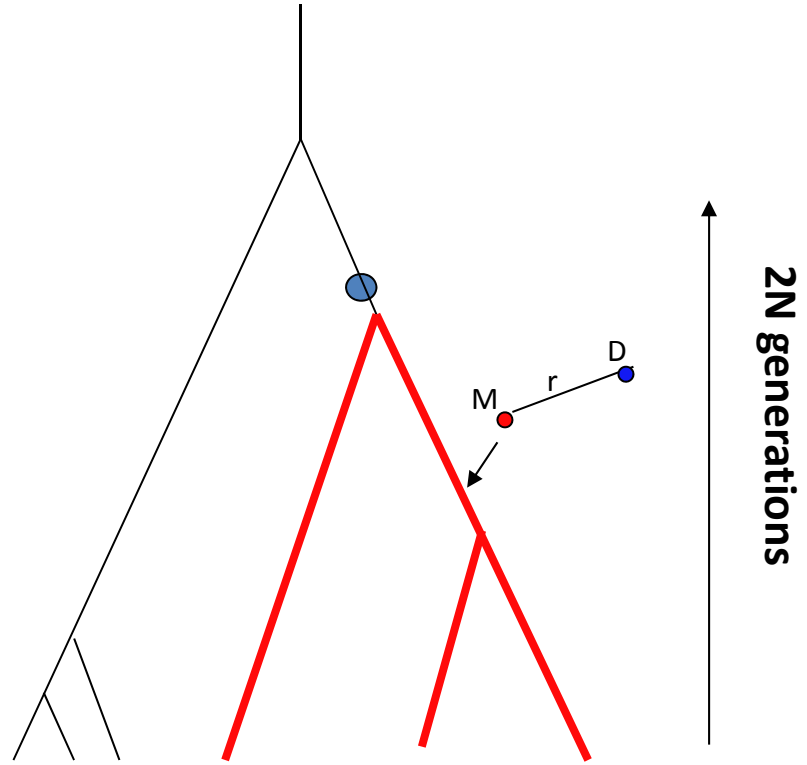


Pedigree known

Few meiosis (max 100s)

Resolution: cMorgans (Mbases)

Association Mapping:



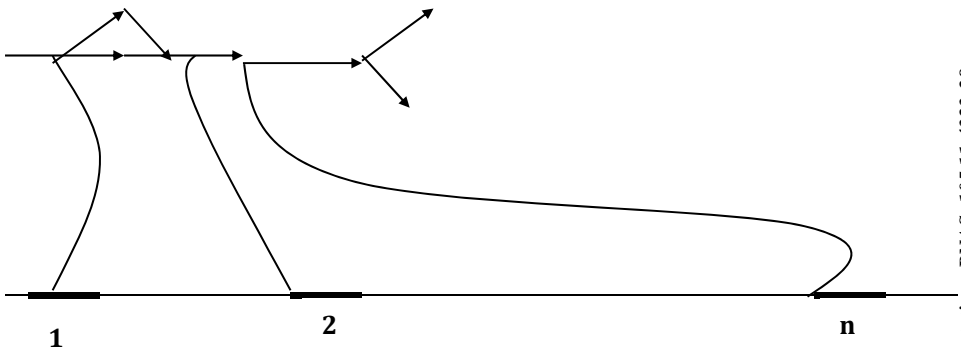
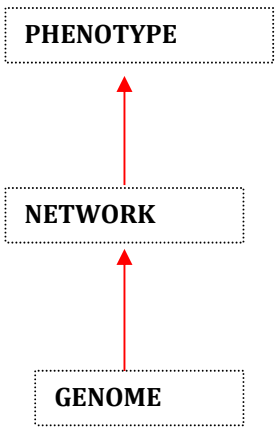
Pedigree unknown

Many meiosis ($>10^4$)

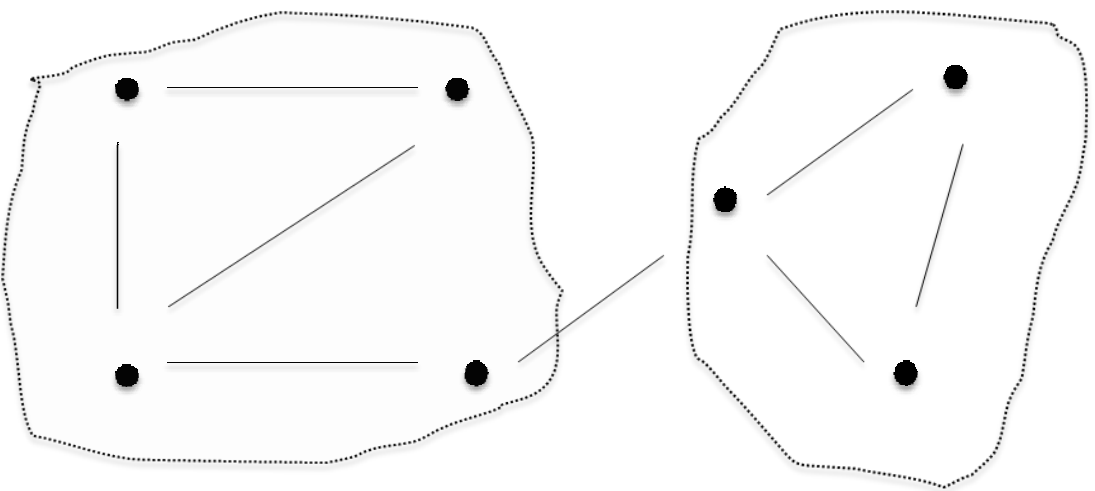
Resolution: 10^{-5} Morgans (Kbases)

Protein Interaction Network based model of Interactions

The path from genotype to phenotype could go through a network and this knowledge can be exploited



Groups of connected genes can be grouped in a supergene and disease dominance assumed: a mutation in any allele will cause the disease.



PIN based model of Interactions

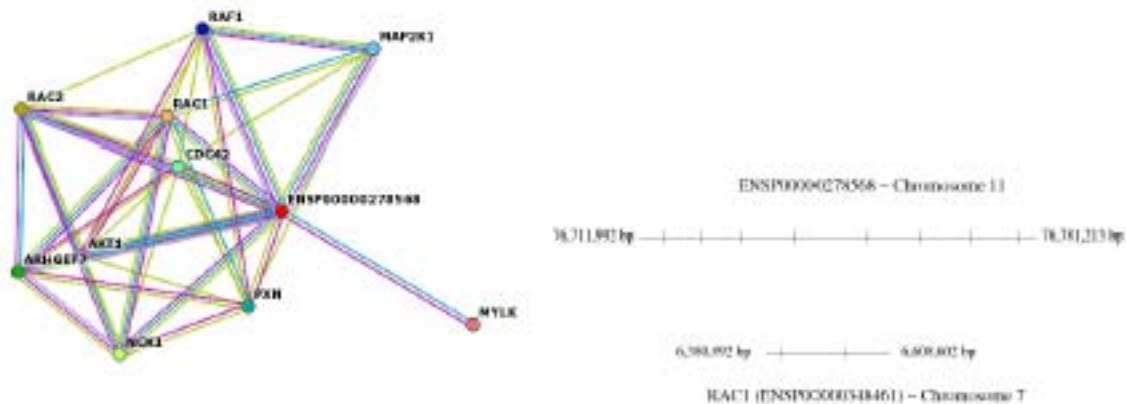
Emily et al, 2009

Single marker association

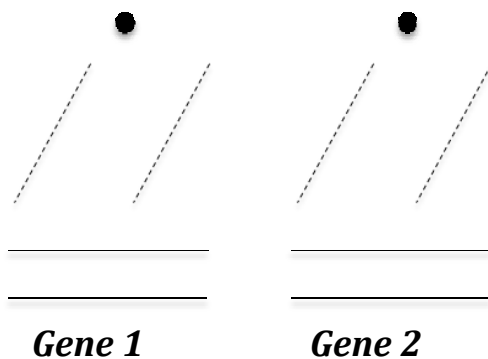


Single marker scan for T1 Diabetes in the WTCCC dataset

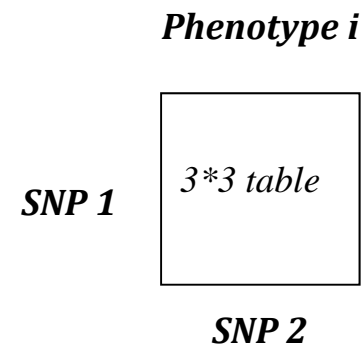
Protein Interaction Network



PIN gene pairs are allowed to interact



Interactions creates non-independence in combinations



Data, Integrative Genomics, Mapping and Functional Explanation

Data: *G, T, P, M, E & F*

Concepts: *Networks, Integrative Genomics, Systems Biology & Correlations*

Mappings: *Heritability, Causative Positions, QTLs, Interactions*

Functional Explanation: *Yet to come*

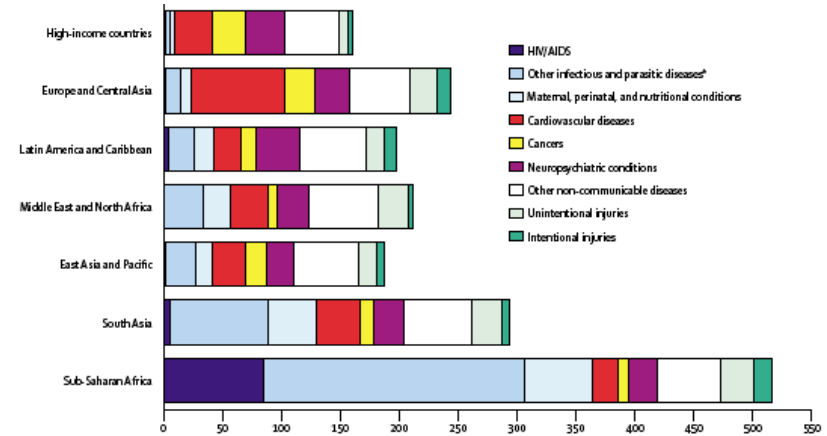
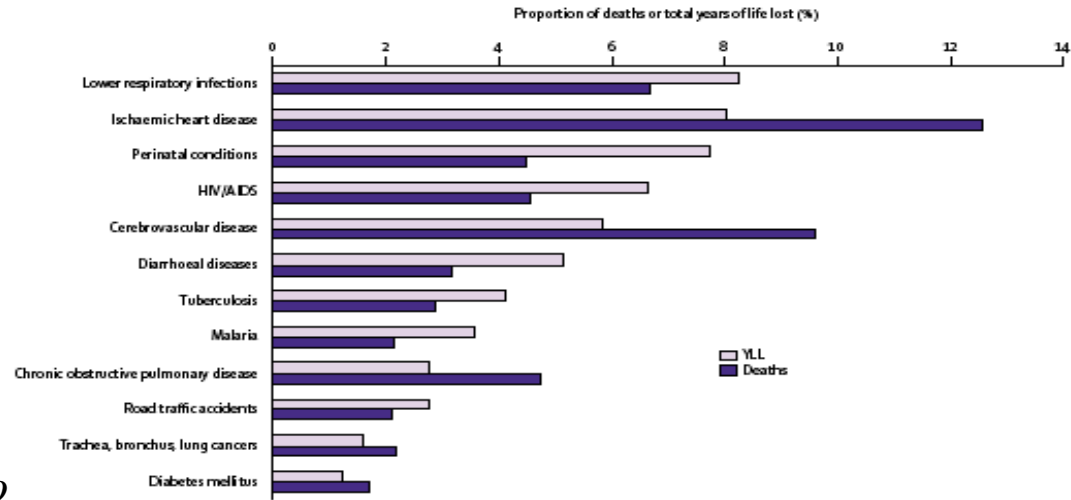
Cost of Disease

• *Most research in the bioscience is motivated by hope of disease intervention.*

• *Major WHO projects have tried to tabulate the costs of different diseases*

• *Genetic Diseases are diseases where there is genetic variation in the susceptibility.*

• *Even small improvements would save many billions*



Low-and-middle-income countries			High-income countries			
Cause	Deaths (millions)	% of total deaths	Cause	Deaths (millions)	% of total deaths	
1	Ischaemic heart disease	5.70	11.8%	Ischaemic heart disease	1.36	17.3%
2	Cerebrovascular disease	4.61	9.5%	Cerebrovascular disease	0.78	9.9%
3	Lower respiratory infections	3.41	7.0%	Trachea, bronchus, lung cancers	0.46	5.8%
4	HIV/AIDS	2.55	5.3%	Lower respiratory infections	0.34	4.4%
5	Perinatal conditions	2.49	5.1%	Chronic obstructive pulmonary disease	0.30	3.8%
6	Chronic obstructive pulmonary disease	2.38	4.9%	Colon and rectum cancers	0.26	3.3%
7	Diarrhoeal diseases	1.78	3.7%	Alzheimer's disease and other dementias	0.21	2.6%
8	Tuberculosis	1.59	3.3%	Diabetes mellitus	0.20	2.6%
9	Malaria	1.21	2.5%	Breast cancer	0.16	2.0%
10	Road traffic accidents	1.07	2.2%	Stomach cancer	0.15	1.9%

Table 1: Ten leading causes of death by income group, 2001

Computational Biology and Bioinformatics

<http://www.stats.ox.ac.uk/research/genome/projects>

11.10 Models of substitution I : Basic Models

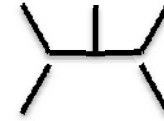
12.10 Models of substitution II : Complex Models

A
↓
T

18.10 Phylogenies I: Combinatorics

29.10 Phylogenies II: Distance, Parsimony & Likelihood

25.10 The Ancestral Recombination Graph & Pedigrees



26.10 Alignment Algorithms I Optimisation Alignment [AN]

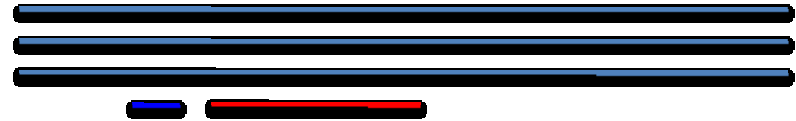
01.11 Alignment Algorithms II Statistical Alignment

ACT-T
-GTCT

02.11 Stochastic Grammars and their Biological Applications: Hidden Markov Models

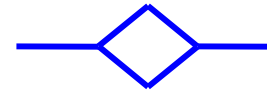
07.11 RNA structures

09.11 Challenges in Genome Annotation



14.11 Finding Signals in Sequences

16.11 Networks: Dynamics and Inference/Evolution



21.11 Grammars

23.11 Models of Evolution of Structures & Movements & Shapes & Grammars

29.11 Integrative Genomics: The Omics

30.11 Integrative Genomics: Mapping

