

Summary of this lecture

The cost of disease

Organism versus Model

The Central Dogma & Data

G - genetic variation

T - transcript levels

P - protein concentrations

M - metabolite concentrations

F – phenotype/phenome

G → F Mapping

General Function Enormous

Used for Disease Gene Finding

Can Include Biological Knowledge

Concepts

G → F Mapping

Models: Networks

Hidden Structures/ Processes

Knowledge

Evolution

Networks

Biological Networks

Physical-Chemical Networks

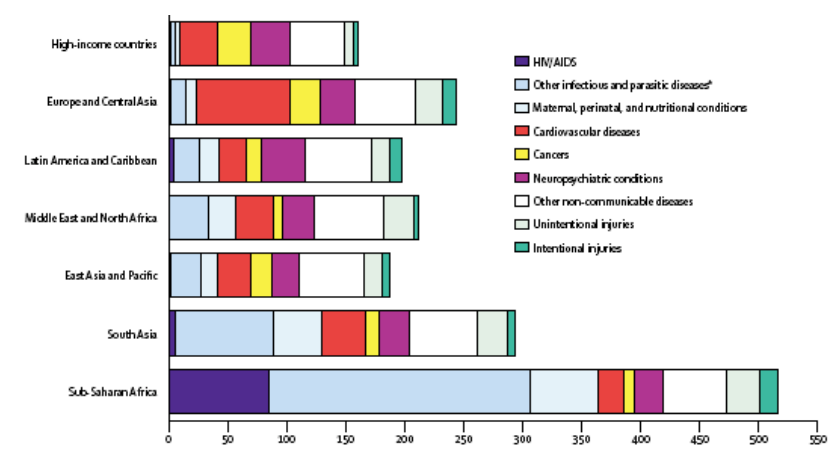
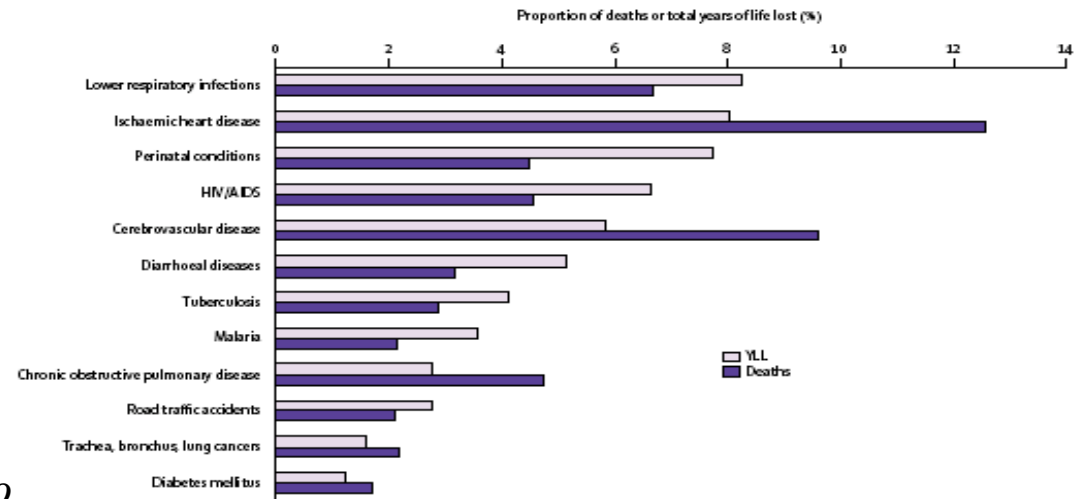
Statistical Networks

Comparative Biology and Model Organisms

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

What is a bacteria? A human being?

Central Dogma

DNA

RNA

Protein

Metabolism & Cell Structure

Organism

Prokaryote

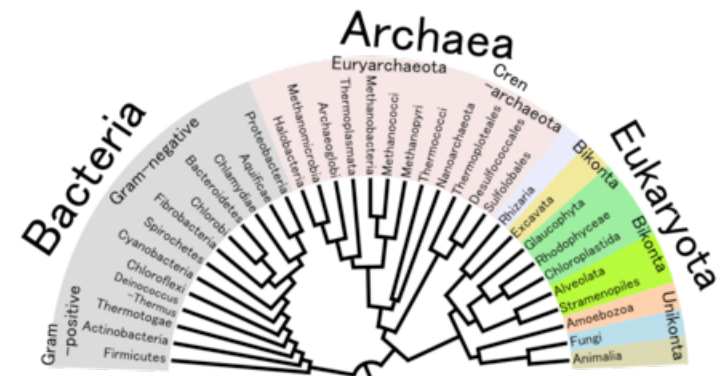
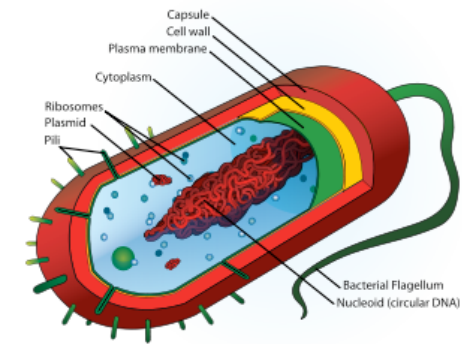
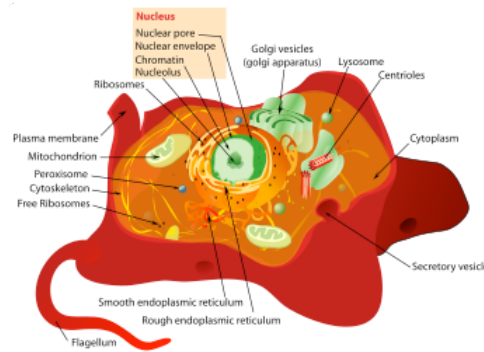
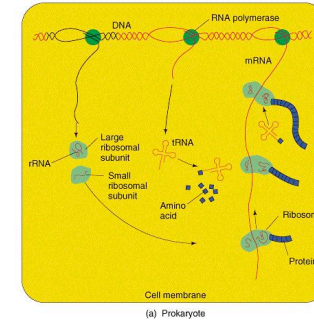
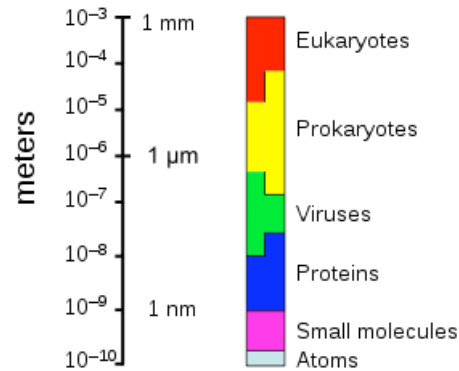
10^{10} atoms

Eukaryote

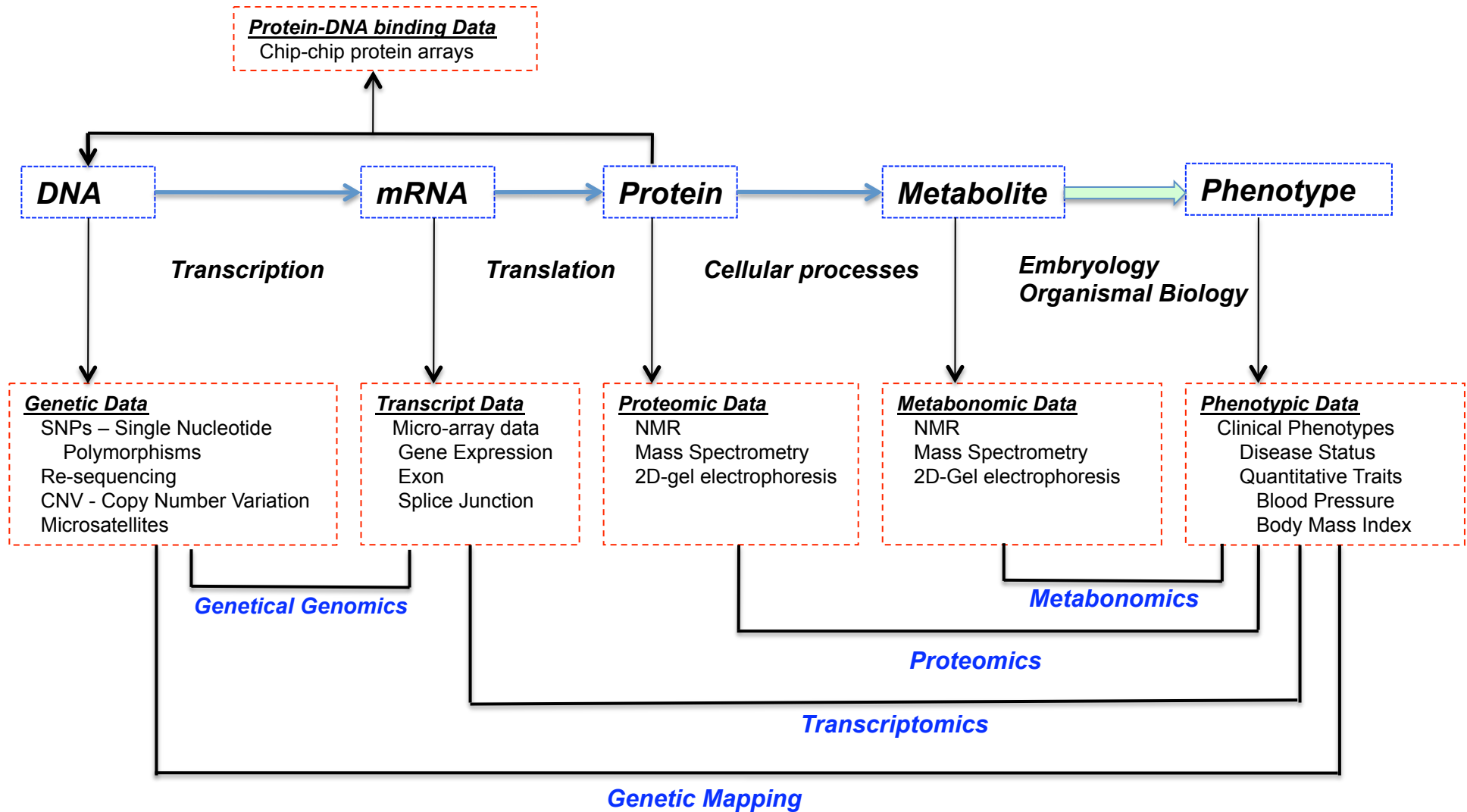
10^{13} atoms

Human

10^{14} cells



The Central Dogma & Data



Structure of Integrative Genomics

Classes DNA mRNA Protein Metabolite Phenotype

Parts

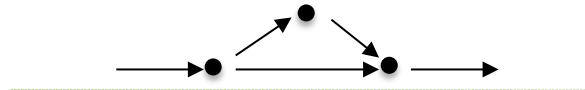
Concepts

***G* → *F* Mapping**

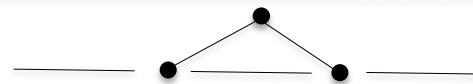


Models: Networks

Physical models:

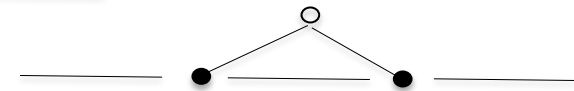


Phenomenological models:



Hidden Structures/ Processes

○ Unobserved/unobservable



Knowledge:

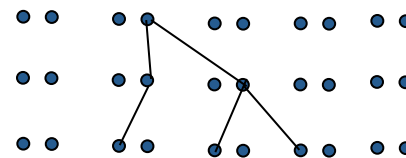
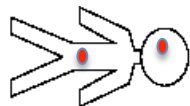
Externally Derived Constraints on which Models are acceptable

Evolution:

Cells in Ontogeny

Individuals/Sequences in a Population

Species



Analysis: **Data + Models + Inference** → **Model Selection**

Functional Explanation

G: Genomes

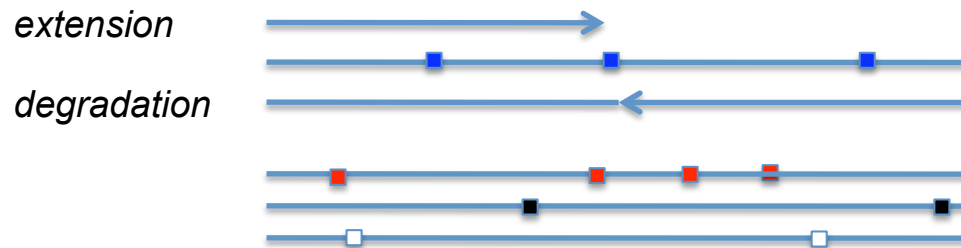
A diploid genome:



Key challenge: Making a single molecule observable!!

Classical Solution (70s): Many

De Novo Sequencing: Halted extensions or degradation



80s: From one to many: PCR – Polymerase Chain Reaction

00s: Re-sequencing: Hybridisation to complete genomes

Future Solution: One is enough!!

Observing the behavior of the polymerase

Passing DNA through millipores registering changes in current

G: Assembly and Hybridisation

Target genome

3×10^9 bp

(unobservable)

Reads

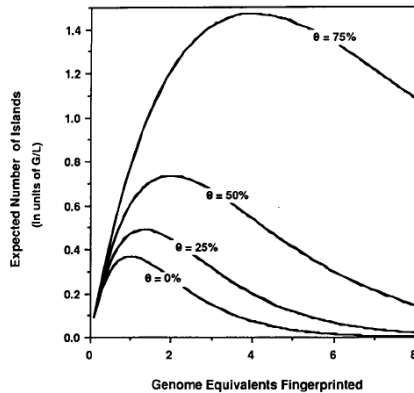
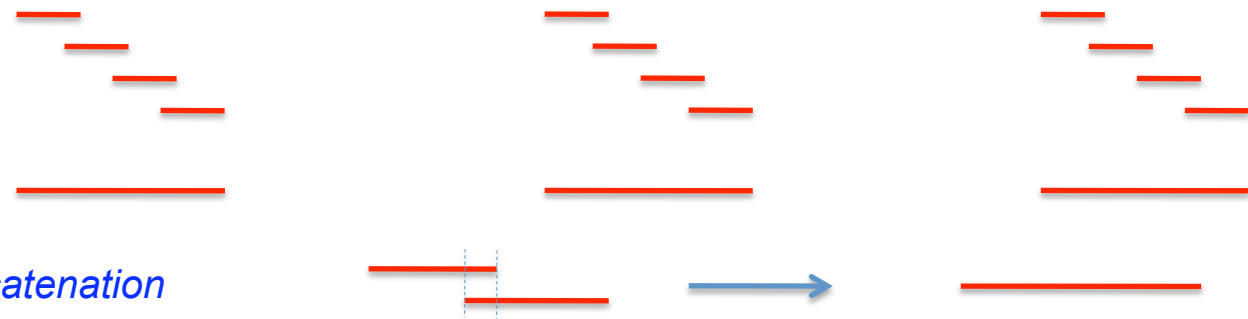
3-400 bp

(observable)

Contigs

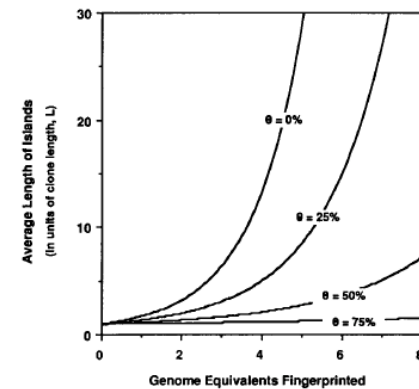
Sufficient overlap allows concatenation

Contigs and Contig Sizes as function of Genome Size (G), Read Size (L) and overlap (θ):

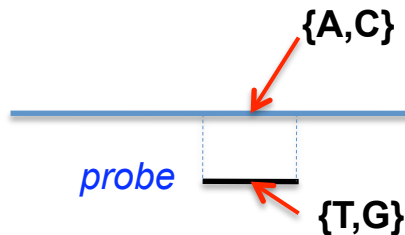


Approximate value of G/L

	Phage (15kb)	Cosmid (40kb)	Yeast (1Mb)
<i>E. coli</i>	267	100	4
<i>S. cerevisiae</i>	1333	500	20
<i>C. elegans</i>	5,667	2,125	85
Human	200,000	75,000	3,000



Complementary or almost complementary strings allow interrogation.



T - Transcriptomics

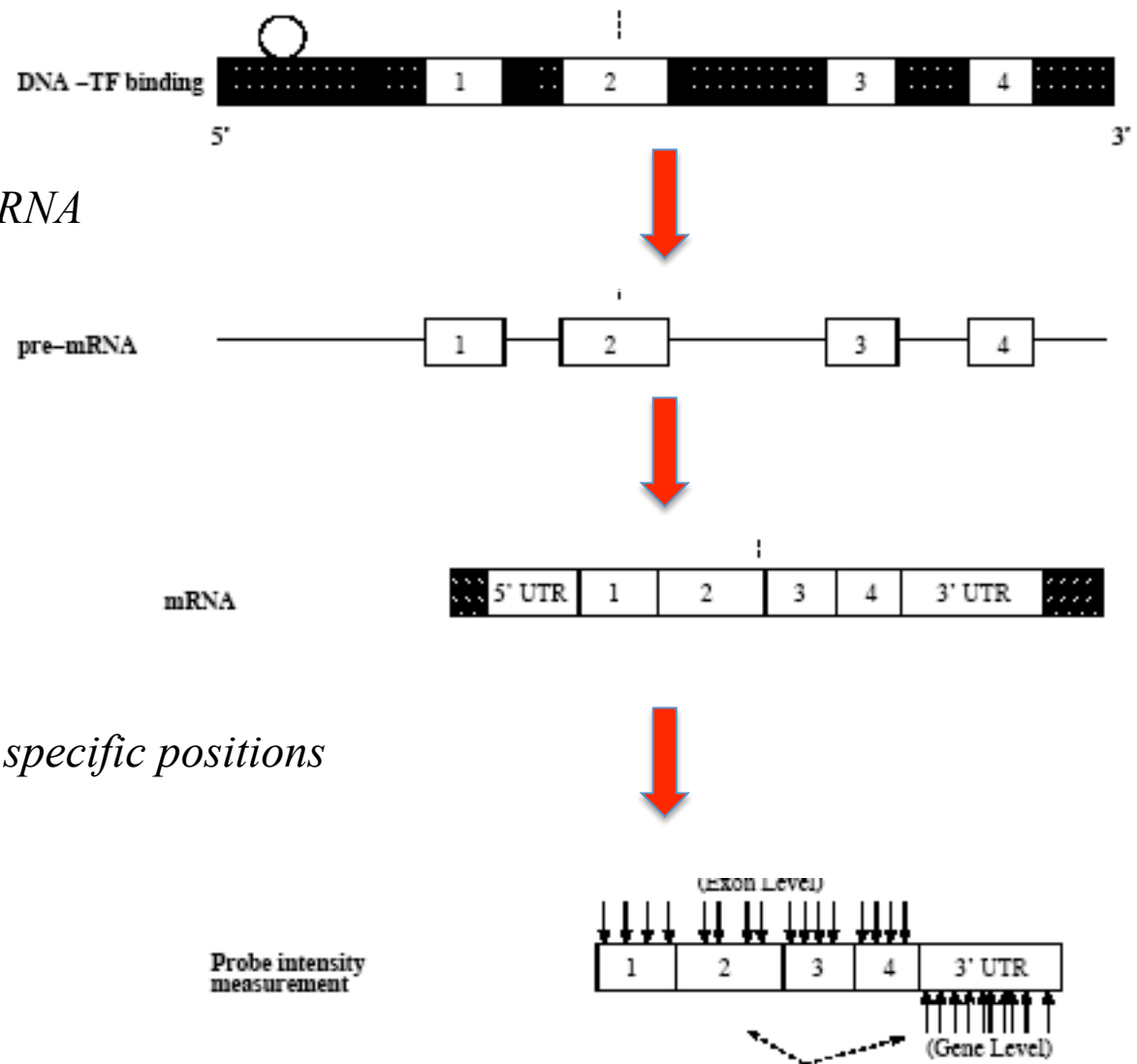
Classical Expression Experiment:

The Gene is transcribed into pre-mRNA

Pre-mRNA is processed into mRNA

Probes are designed hybridizing to specific positions

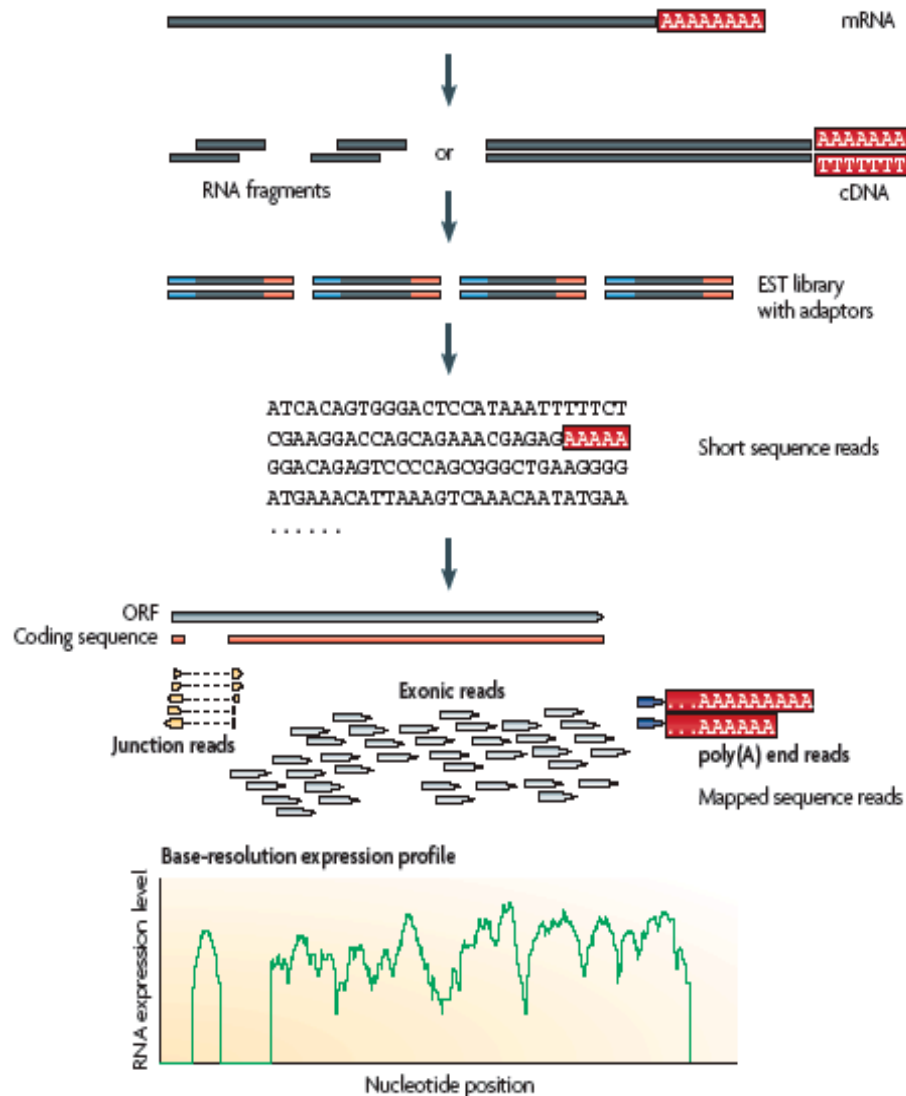
Measures transcript levels averaging of a set of cells.



T - Transcriptomics

RNA-Seq Expression Experiment:

Advantages - Discoveries



More quantitative in evaluating expression levels

More precise in positioning

Much more is transcribed than expected.

Transcription of genes very imprecise

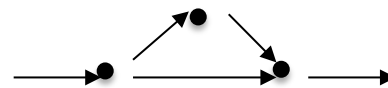
Concepts

$G \rightarrow F$ Mapping

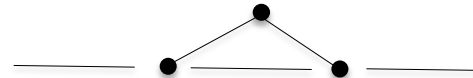


Models: Networks

Physical models:

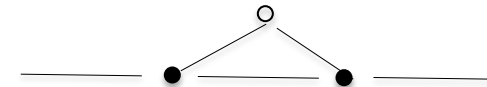


Phenomenological models:



Hidden Structures/ Processes

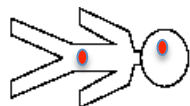
Unobserved/able ○



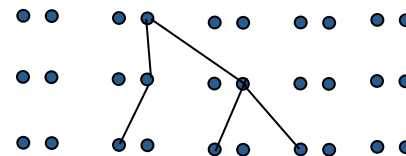
Knowledge: Externally Derived Constraints on which Models are acceptable

Evolution:

Cells in Ontogeny



Individuals/Sequences in a Population



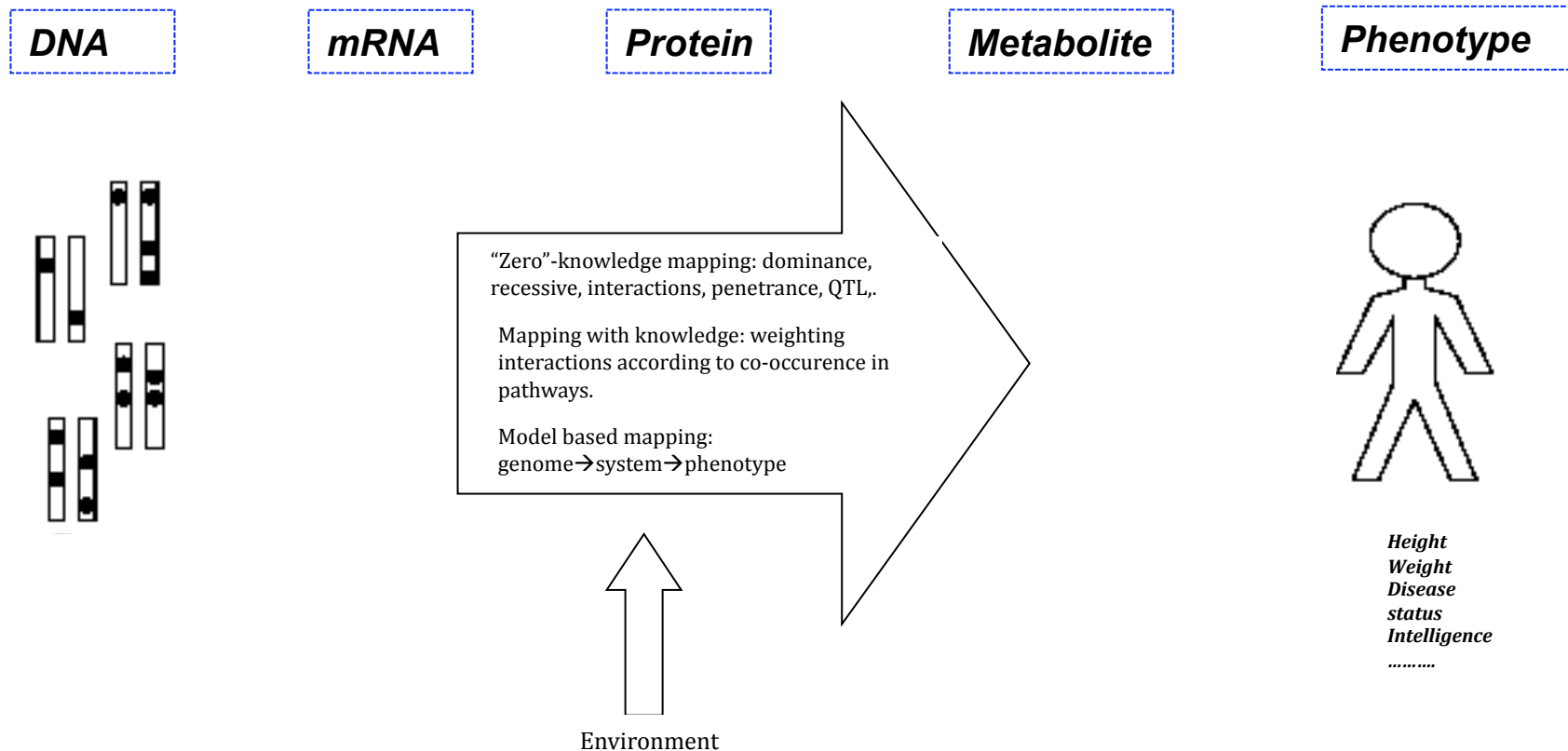
Species



$G \rightarrow F$

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

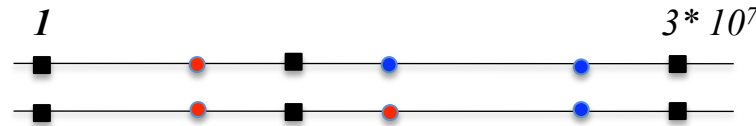
G is the most abundant data form - heritable and precise. F is of greatest interest.



The General Problem is Enormous

Set of Genotypes:

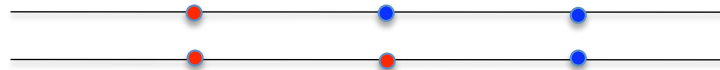
- Diploid Genome**



- 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**

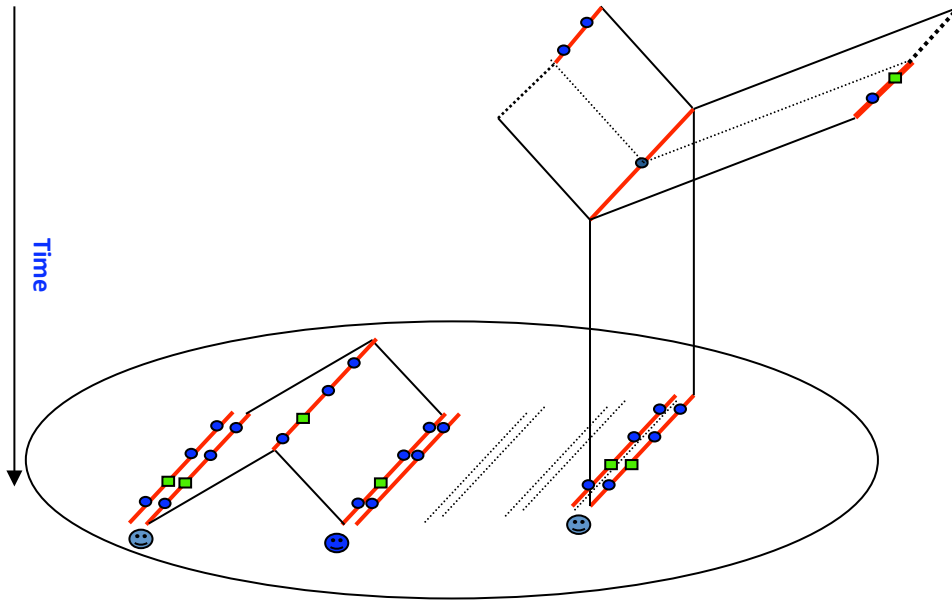
Epistasis: The effect of one locus depends on the state of another

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

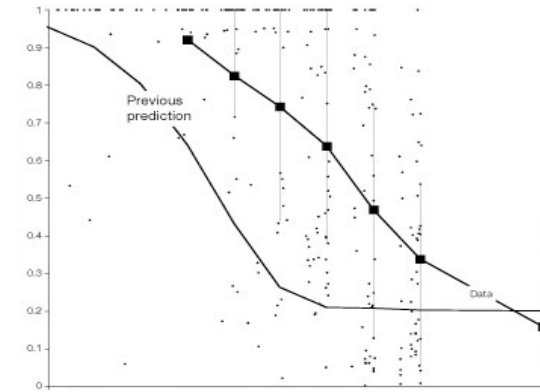
$$\sum_{i \text{ causative positions}} X_i(G_i) + \varepsilon$$

Genotype and Phenotype Co-variation: Gene Mapping

Sampling Genotypes and Phenotypes

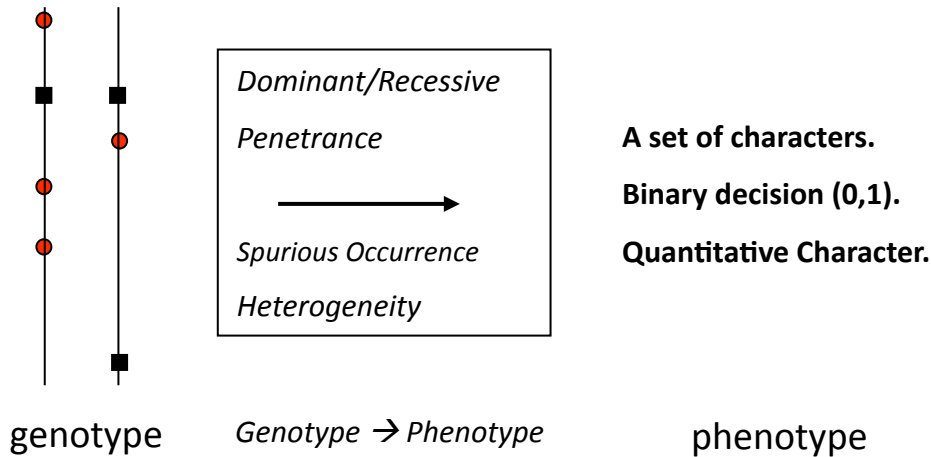


Decay of local dependency

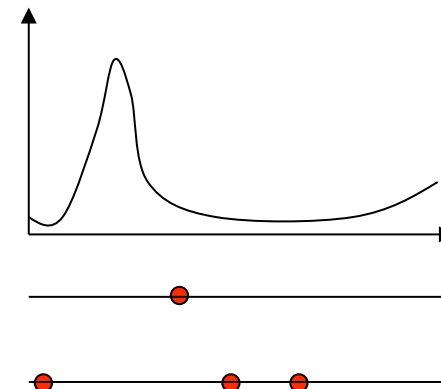


Reich et al. (2001)

Genotype -->Phenotype Function

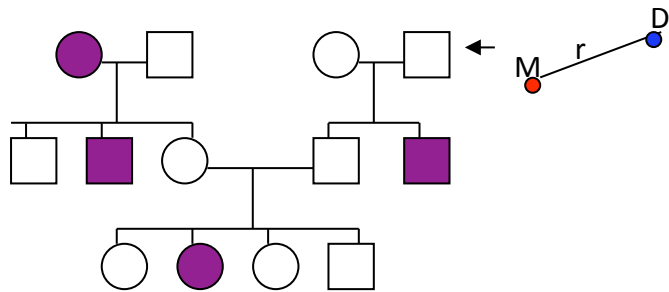


Result: The Mapping Function



Pedigree Analysis & Association Mapping

Pedigree Analysis:

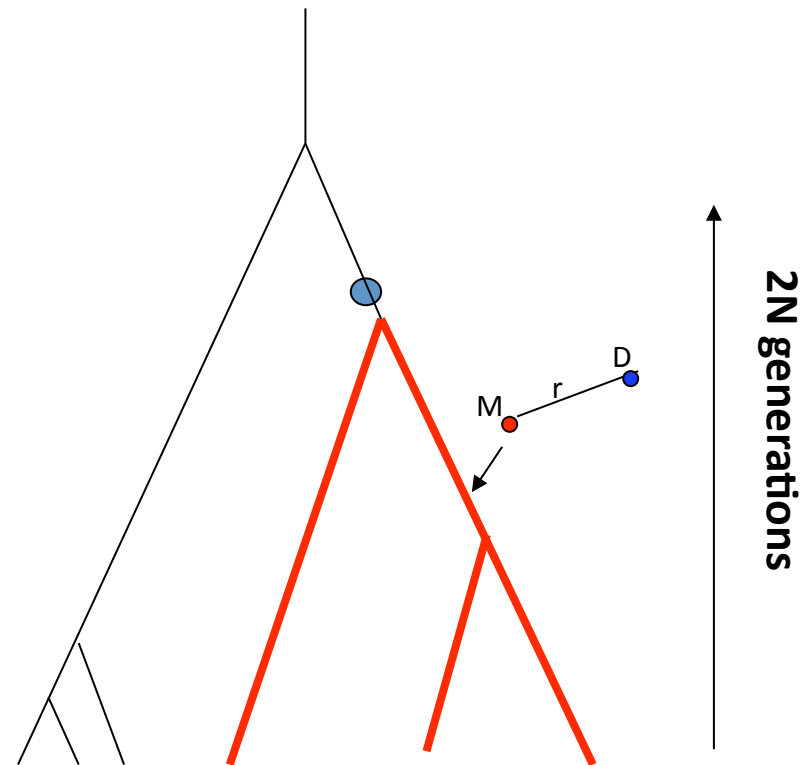


Pedigree known

Few meiosis (max 100s)

Resolution: cMorgans (Mbases)

Association Mapping:



Pedigree unknown

Many meiosis ($>10^4$)

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

Adapted from McVean and others

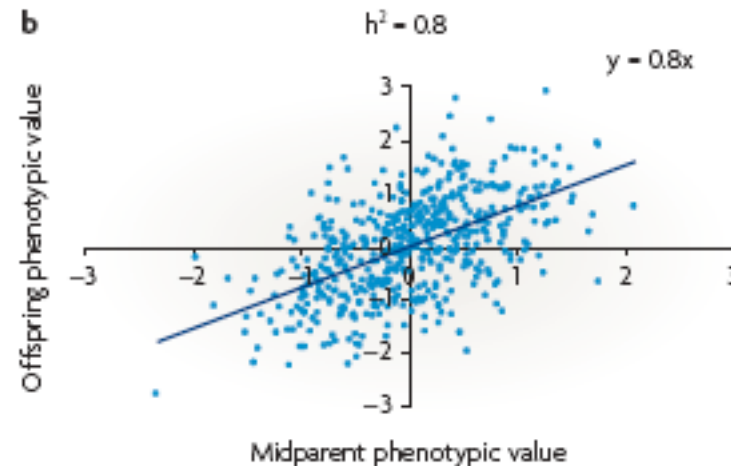
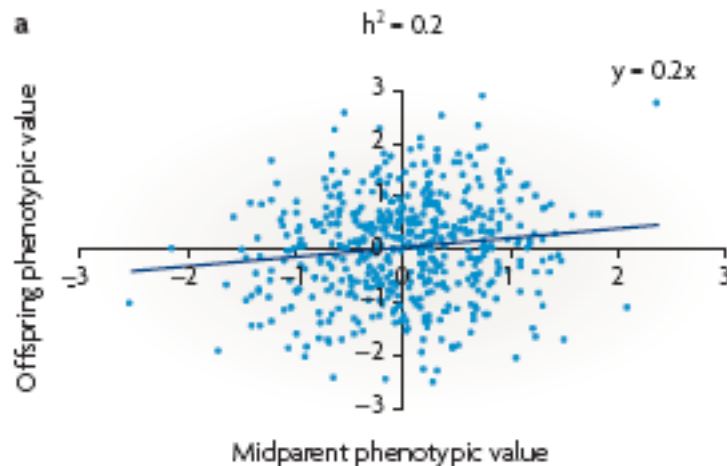
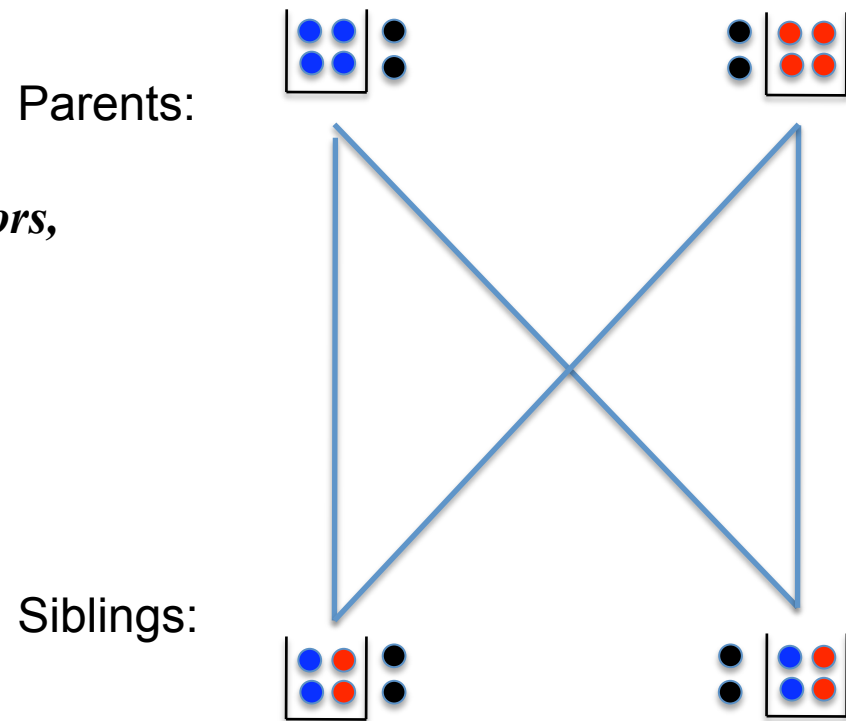
Heritability: Inheritance in bags, not strings.

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.

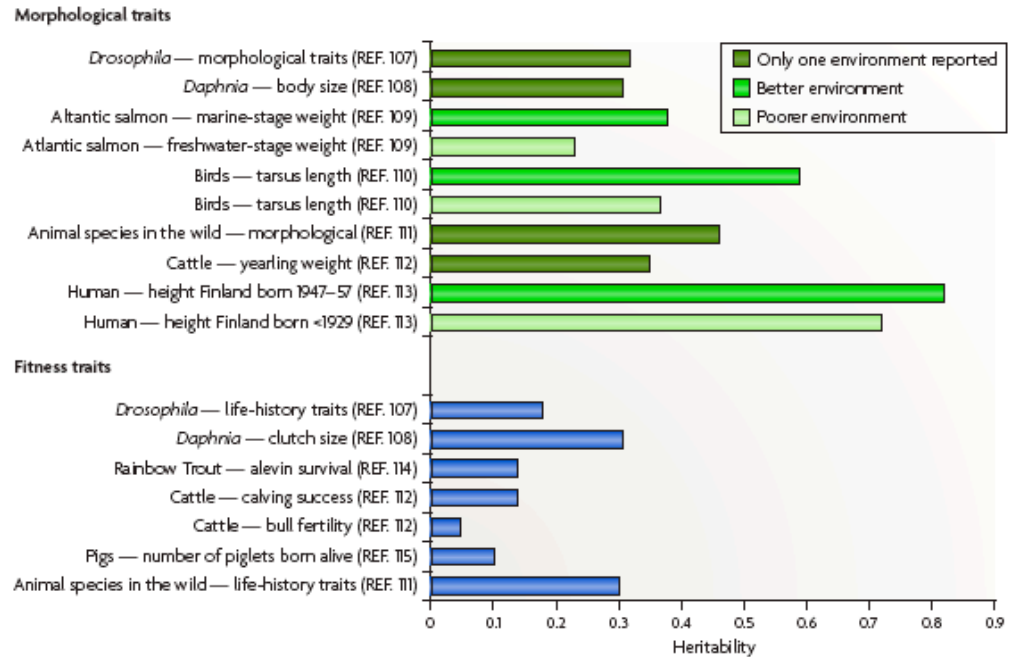
This allows calculation of relative effect of genetics and environment

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

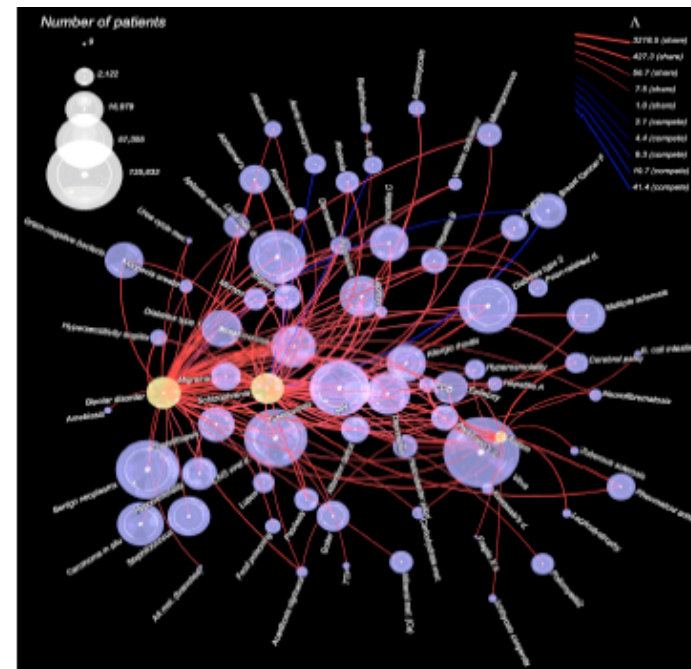
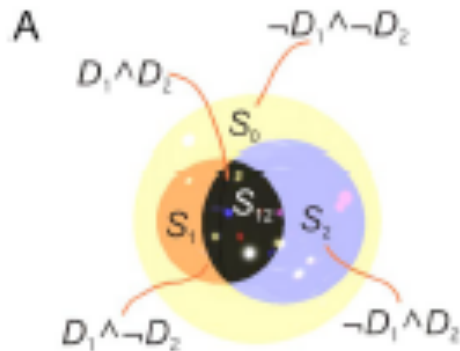


Heritability

Examples of heritability



Heritability of multiple characters:



Rzhetsky et al. (2006) Probing genetic overlap among complex human phenotypes PNAS vol. 104 no. 28 11694–11699

Visscher, Hill and Wray (2008) Heritability in the genomics era — concepts and misconceptions nATurE rEVIEWEWS | genetics volume 9.255-66

Networks in Cellular Biology

- *Dynamics*
- *Inference*
- *Evolution*

A. Metabolic Pathways

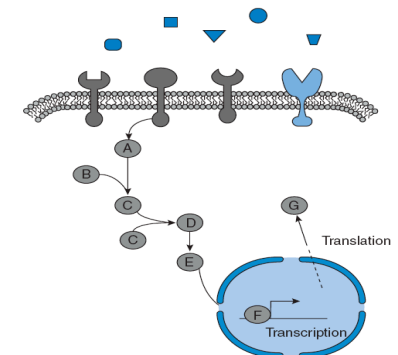
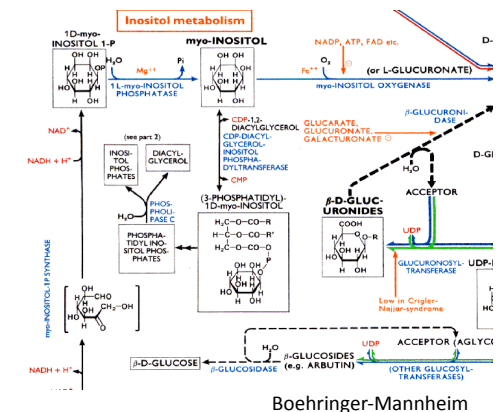
Enzyme catalyzed set of reactions controlling concentrations of metabolites

B. Regulatory Networks

Network of {Genes → RNA → Proteins}, that regulates each other transcription.

C. Signaling Pathways

Cascade of Protein reactions that sends signal from receptor on cell surface to regulation of genes.



Number of Networks

- *undirected graphs*

$$\alpha_n = 2^{\frac{n(n-1)}{2}}$$

- *Connected undirected graphs*

$$c_n = \alpha_n - \sum_{k=1}^{n-1} \binom{n-1}{k-1} c_k \alpha_{n-k}$$

- *Directed Acyclic Graphs - DAGs*

$$a_n = \sum_{k=1}^n (-1)^{k-1} \binom{n}{k} 2^{k(n-k)} a_{n-k}$$

- *Interesting Problems to consider:*

- *The size of neighborhood of a graph?*
- *Given a set of subgraphs, how many graphs have them as subgraphs?*

A repertoire of Dynamic Network Models

To get to networks:

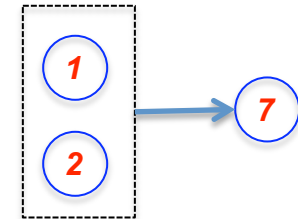
No space heterogeneity → molecules are represented by numbers/concentrations

Definition of Biochemical Network:

- *A set of k nodes (chemical species) labelled by kind and possibly concentrations, X_k*



- *A set of reactions/conservation laws (edges/hyperedges) is a set of nodes. Nodes can be labelled by numbers in reactions. If directed reactions, then an inset and an outset.*



- *Description of dynamics for each rule.*

ODEs – ordinary differential equations $\frac{dX_7}{dt} = f(X_1, X_2)$

Mass Action $\frac{dX_7}{dt} = cX_1X_2$

Time Delay $\frac{d\bar{X}(t)}{dt} = f(\bar{X}(t - \tau))$

Discrete Deterministic – the reactions are applied.

Boolean – only 0/1 values.

Stochastic

Discrete: the reaction fires after exponential with some intensity $I(X_1, X_2)$ updating the number of molecules

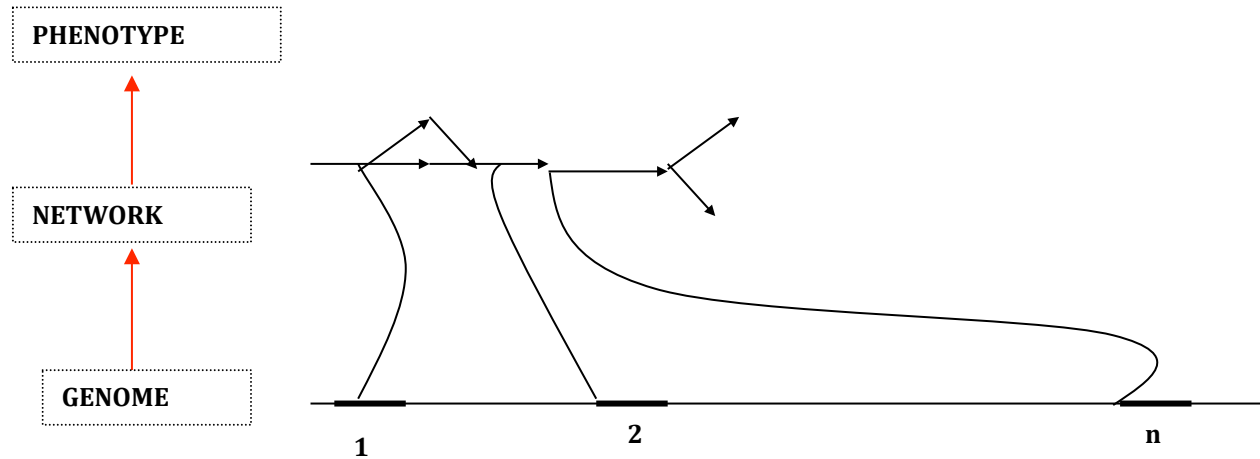
Continuous: the concentrations fluctuate according to a diffusion process.

Networks → A Cell → A Human

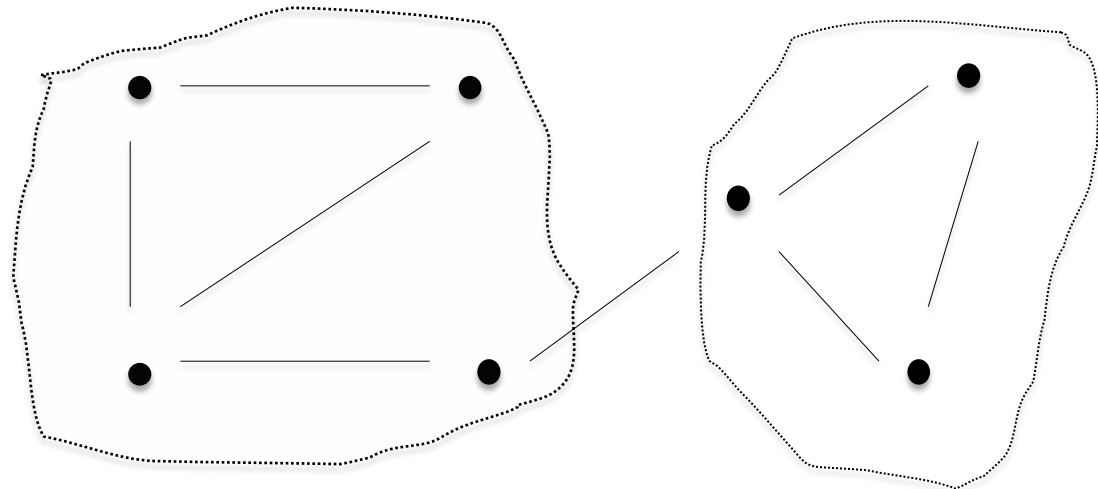
- *A cell has $\sim 10^{13}$ atoms.* 10^{13}
- *Describing atomic behavior needs $\sim 10^{15}$ time steps per second* 10^{28}
- *A human has $\sim 10^{13}$ cells.* 10^{41}
- *Large descriptive networks have 10^3 - 10^5 edges, nodes and labels* 10^5
- *What happened to the missing 36 orders of magnitude???*
- *Which approximations have been made?*
 - A Spatial homogeneity → 10^3 - 10^7 molecules can be represented by concentration* $\sim 10^4$
 - B One molecule (10^4), one action per second (10^{15})* $\sim 10^{19}$
 - C Little explicit description beyond the cell* $\sim 10^{13}$
- A Compartmentalisation can be added, some models (ie Turing) create spatial heterogeneity*
- B Hopefully valid, but hard to test*
- C Techniques (ie medical imaging) gather beyond cell data*

Protein Interaction Network based model of Interactions

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



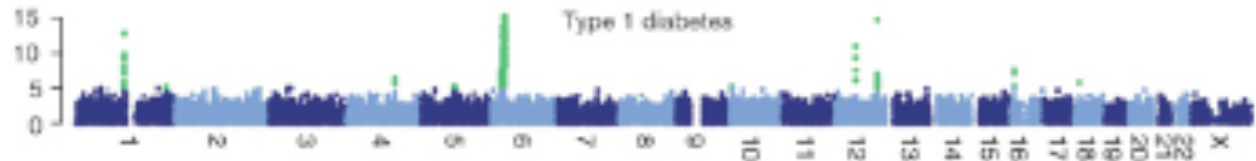
Groups of connected genes can be grouped in a supergene and disease dominance assumed: a mutation in any allele will cause the disease.



PIN based model of Interactions

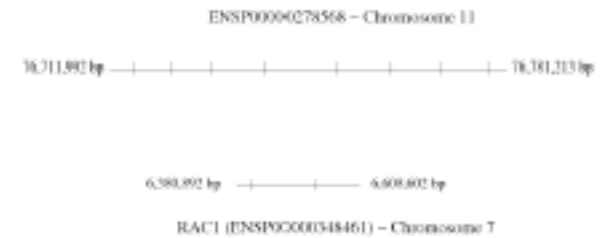
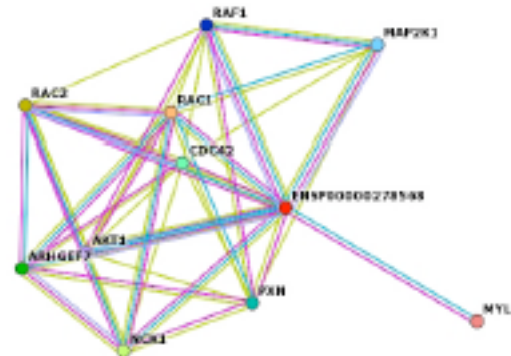
Emily et al, 2009

Single marker association



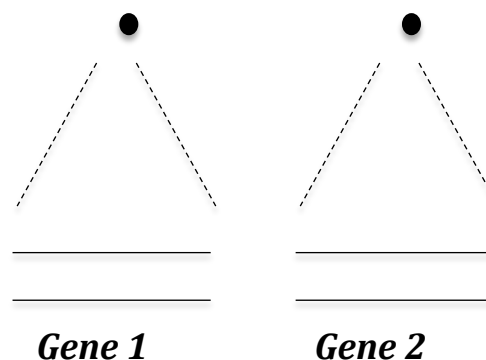
Single marker scan for T1 Diabetes in the WTCCC dataset

Protein Interaction Network



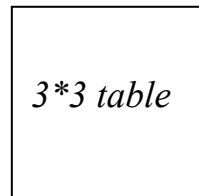
PIN gene pairs are allowed to interact

Interactions creates non-independence in combinations



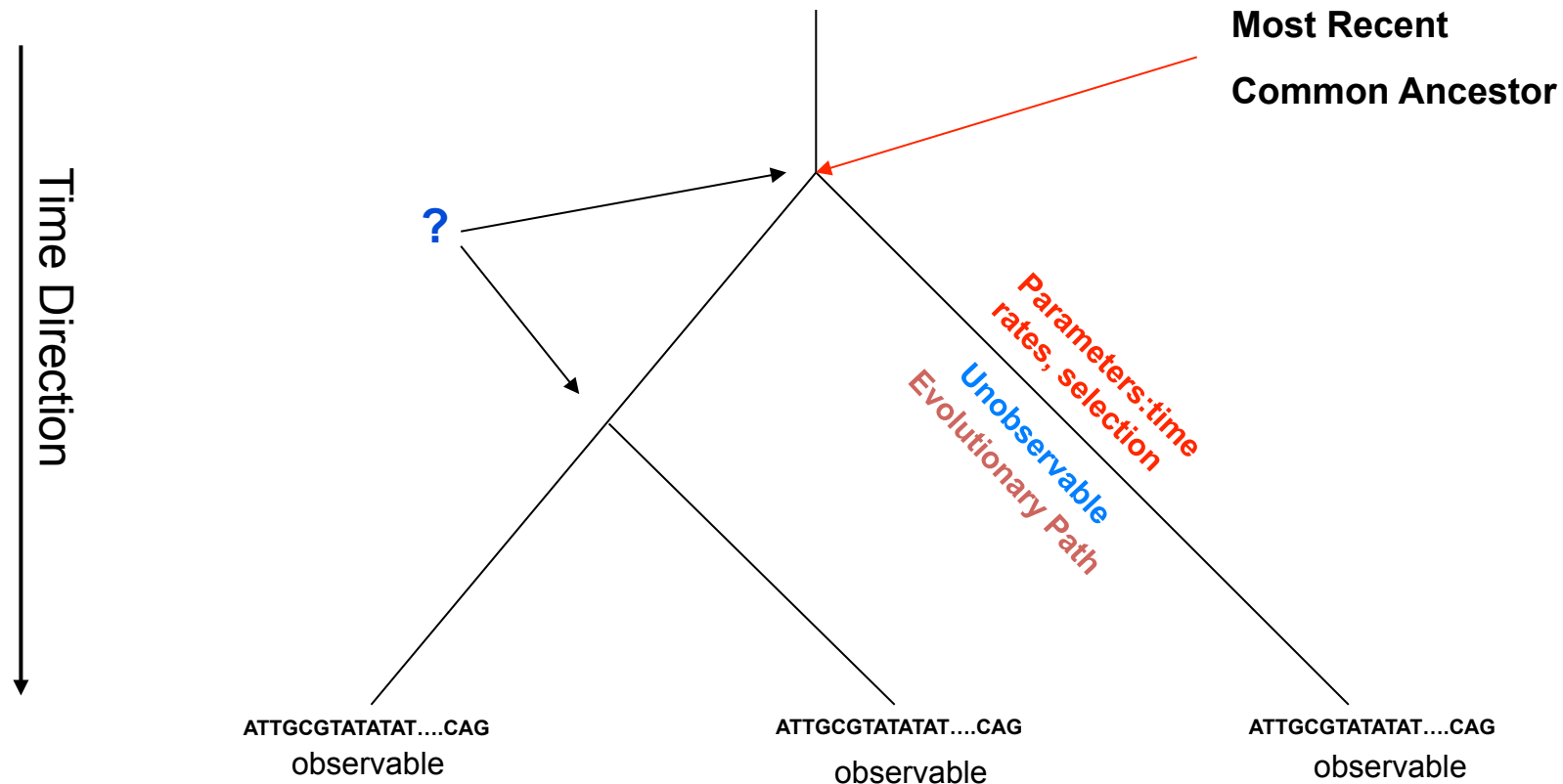
Phenotype i

SNP 1



SNP 2

Comparative Biology



Key Questions:

- Which phylogeny?
- Which ancestral states?
- Which process?

Key Generalisations:

- Homologous objects
- Co-modelling
- Genealogical Structures?

Comparative Biology: Evolutionary Models

<u>Object</u>	<u>Type</u>	<u>Reference</u>
Nucleotides/Amino Acids/codons	CTFS continuous time finite states	Jukes-Cantor 69 +500 others
Continuous Quantities	CTCS continuous time countable states	Felsenstein 68 + 50 others
Sequences	CTCS	Thorne, Kishino Felsenstein,91 + 40others
Gene Structure	Matching	DeGroot, 07
Genome Structure	CTCS MM	Miklos,
Structure		
RNA	SCFG-model like	Holmes, I. 06 + few others
Protein	non-evolutionary: extreme variety	Lesk, A; Taylor, W.
Networks	CTCS	Snijder, T (sociological networks)
Metabolic Pathways	?	
Protein Interaction	CTCS	Stumpf, Wiuf, Ideker
Regulatory Pathways	CTCS	Quayle and Bullock, 06
Signal Transduction	CTCS	Soyer et al.,06
Macromolecular Assemblies	?	
Motors	?	
Shape	- (non-evolutionary models)	Dryden and Mardia, 1998
Patterns	- (non-evolutionary models)	Turing, 52;
Tissue/Organs/Skeleton/....	- (non-evolutionary models)	Grenander,
Dynamics		
MD movements of proteins	-	
Locomotion	-	
Culture	analogues to genetic models	Cavalli-Sforza & Feldman, 83
Language		
Vocabulary	“Infinite Allele Model” (CTCS)	Swadesh,52, Sankoff,72, Gray & Aitkinson, 2003
Grammar		Dunn 05
Phonetics		Bouchard-Côté 2007
Semantics		Sankoff,70
Phenotype	Brownian Motion/Diffusion	
Dynamical Systems	-	

Summary of this lecture

The cost of disease

Organism versus Model

The Central Dogma & Data

G - genetic variation

T - transcript levels

P - protein concentrations

M - metabolite concentrations

F – phenotype/phenome

G → F Mapping

General Function Enormous

Used for Disease Gene Finding

Can Include Biological Knowledge

Concepts

G → F Mapping

Models: Networks

Hidden Structures/ Processes

Knowledge

Evolution

Networks

Biological Networks

Physical-Chemical Networks

Statistical Networks

Comparative Biology and Model Organisms