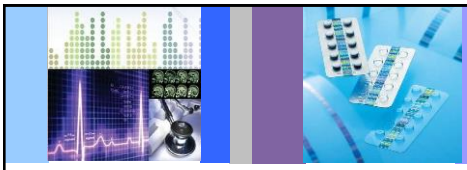
The present

One Size fits all



Pharmacogenomics

Yossi Levy

Outline

- What is pharmacogenomics
- Biological background
- Hardy-Weinberg Equilibrium and Linkage Disequilibrium
- Types of genetic studies
- GWA studies – design and analysis
- Candidate Gene studies – design and analysis

The Future

"My physician prescribed a customized formulation for me. Here's my DNA sequence."

Some applications of PGx

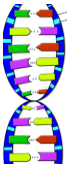
- Utilize genetic markers of efficacy for patient stratification
- Identify non-responders to investigational drug or SOC
- Identify markers of adverse drug reaction; modify dosing
- Improve benefit-risk ratio

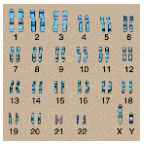
What is Pharmacogenomics?

- Pharmacogenomics (PGx) is the study of how genes affect the way our bodies respond to a medicine.
- An inter-disciplinary science involving
  - Biology and Genetics
  - Medicine
  - Pharmacology
  - Statistics

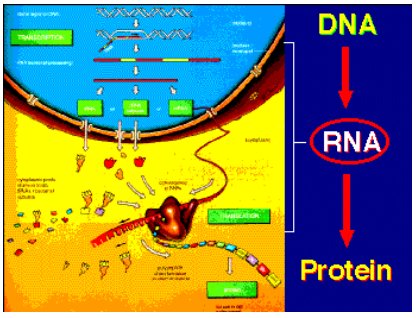
Genome and DNA

- **Genome** – contains all biological information
- Biological information is encoded in **DNA**
- DNA is divided to discrete units called **Genes**
- Each gene is represented by two copies, called **Alleles**
- Genes are packed into **Chromosomes**



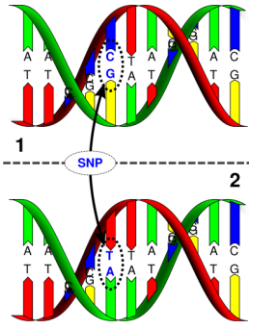


Central Dogma of Molecular Biology



SNPs

- **SNP** - Single Nucleotide Polymorphism
- Most common type of genetic variation
- Each SNP represent a difference in a single DNA base
- The SNP in the picture is CT or AC or AG or GT – they are all the same
- SNP can have 3 possible values: AA, Aa or aa

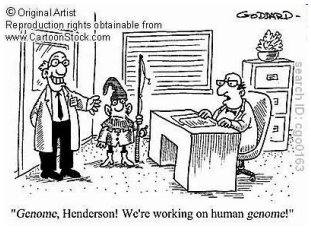


Polymorphism

- Some expressed traits are attributed to variation in DNA sequence
- When two individuals display different phenotypes in the same trait, they have two different alleles in the same gene.
- That gene is therefore said to be **polymorphic**.

The Human Genome

- 46 chromosomes – 23 pairs
- 2 meters of DNA
- 3.4B DNA bases
- 25000 genes
- 26M variable sites
- 10M SNPs



Alleles and expression

- **Genotype –** Combination of alleles
- **Homozygous** gene – both alleles are the same
- **Heterozygous** gene – alleles are different
- **Phenotype-** expression of genotype
- **A dominant allele** is almost always expressed
- **A recessive allele** is expressed only if there are two copies of that allele

HWE implications

- HWE implies constant alleles frequencies over generations
- HWE is violated:
  - in the presence of **population admixture** – a situation in which mating occurs between two populations for which the allele frequencies differ
  - in the presence of **population stratification** – combination of populations in which breeding occurs within but not between subpopulations
  - when mating occurs between relatives
  - In case of a genotyping error
- Population admixture or stratification can be explored using covariates, PCA and/or MDS

Hardy-Weinberg Equilibrium

- A theoretical description of the relationship between genotype and allele frequencies
- HWE denotes independence of the alleles at a single site between two homologous chromosomes
- Let p be the frequency of the dominant allele A and q and let be the frequency of the recessive allele a (p+q=1).
- The expected genotype frequencies are:

$$p_{AA} = p^2$$
$$p_{Aa} = 2pq = 2p(1-p)$$
$$p_{aa} = q^2 = (1-p)^2$$

Candidate polymorphism studies

- Consider polymorphism(s) within a gene
- There is an a priori hypothesis about functionality
- Primary hypothesis: the variable site under investigation is **functional**.
- That is, the given SNP (or set of SNPS) influence the disease trait directly

Types of genetics studies

Studies to investigate genotype-trait association within a population of **unrelated individuals**:

- Candidate polymorphism studies
- Candidate gene studies
- Fine mapping studies
- Gnome-wide association studies (GWAS)

Gnome Wide Association Studies

- Similar to candidate gene approach
- Aim to identify association between SNPs and trait
- Less hypothesis driven
- Involves the characterization of a much larger number of SNPs

Candidate gene studies

- Consider multiple SNPs within a gene
- SNPs are not assumed to be functional
- However, the selected SNPs may be associated to a functional SNP within the gene
- This association is called Linkage Disequilibrium

Linkage Disequilibrium

Expected allele distributions under independence

		Site 2		
		B	b	
Site 1	A	$n_{11} = Np_Ap_B$	$n_{12} = Np_Ap_b$	$n_{1\cdot} = Np_A$
	a	$n_{21} = Np_ap_B$	$n_{22} = Np_ap_b$	$n_{2\cdot} = Np_a$
		$n_{\cdot 1} = Np_B$	$n_{\cdot 2} = Np_b$	$N = 2n$

Observed allele distributions under LD

		Site 2		
		B	b	
Site 1	A	$n_{11} = N(p_Ap_B + D)$	$n_{12} = N(p_Ap_b - D)$	$n_{1\cdot}$
	a	$n_{21} = N(p_ap_B - D)$	$n_{22} = N(p_ap_b + D)$	$n_{2\cdot}$
		$n_{\cdot 1}$	$n_{\cdot 2}$	$N = 2n$

Linkage Disequilibrium

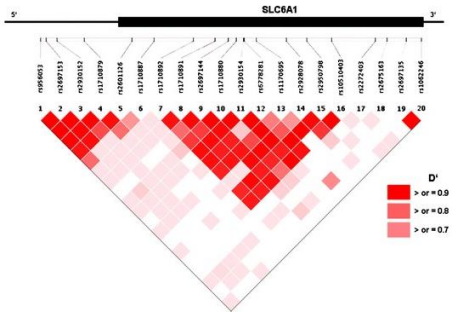
- Recall that in candidate gene studies and GWAS, studied SNPs may not be functional
- However, it is hoped that they are associated with the trait under consideration
- LD: an association in the alleles present at each of two sites present on a genome

Trait-genotype relationship

- Ultimate goal: identify SNP or set of SNPs that predict the phenotypic trait
- In pharmaceutical industry – the interesting trait is response to treatment

		SNP (G)		
Treatment (T)	Response (Y)	AA	Aa	aa
Active	Yes	n_{111}	n_{112}	n_{113}
	Undetermined	n_{121}	n_{122}	n_{123}
	No	n_{131}	n_{132}	n_{133}
Placebo	Yes	n_{211}	n_{212}	n_{213}
	Undetermined	n_{221}	n_{222}	n_{223}
	No	n_{231}	n_{232}	n_{233}

LD graphical presentation



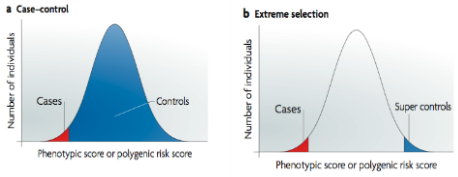
Typical GWAS study approach

- Data QC
  - Remove SNPs with >5% missing data and or nonrandom missingness
  - Remove SNPs with low Minor Allele Frequency
  - Remove SNPs that depart from HWE
  - Remove individuals with high percent of missing data
- Run logistic regression model for each of the SNPs
- Identify top SNPs with significant drug and SNP interaction
- Beware of multiple testing
- Try to model interactions between top SNPs (later)
- Identify SNPs for candidate gene study

Common approaches to SNP analysis

- Classical tests and measures of association (Chi-square, Fisher's exact test, Cochran-Armitage, etc.)
- Logistic regression
  - Look for significant T*G interaction
  - Allows for introduction of additional covariates
- Log linear model
  - Look for conditional independence of T and Y given G
- Bayesian testing: assume the parameter of the multinomial count data comes from a Dirichlet distribution
  - Works also when some cells has low/zero counts
  - Are you ready to pay the price?

Extreme phenotype approach



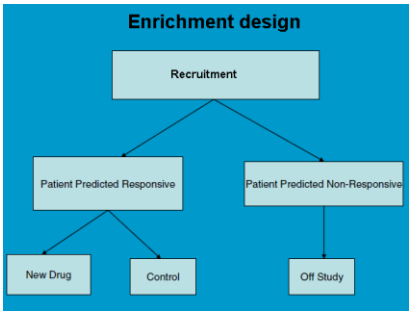
- Less variation and higher effect size lead to smaller sample size
- But, are test statistics still valid?

Sample size for GWA studies

- Nobody really knows how to calculate
- Considering that
  - Effect size for single SNP is usually low
  - Number hypotheses tests is very high
  - Not too liberal type I error level
  - Unknown LD structure

  Most publications recommend a sample size of thousands cases, and same number of controls
- Such sample size is unrealistic, even in the pharmaceutical industry
- Possible solution: enrichment of population

Candidate gene study design



What to do next?

- Check replicability
- Check predictability
- Build a predictive model combining several markers
- Look at GxG and GxE interactions
- Pathway analysis – does it all make biological sense?
- Look at additional clinical endpoints
- Design a candidate gene study

Candidate gene study design

Prospective-retrospective design

- Define the planned analysis prospectively
- Perform the analysis of historical clinical trial

Candidate gene study design