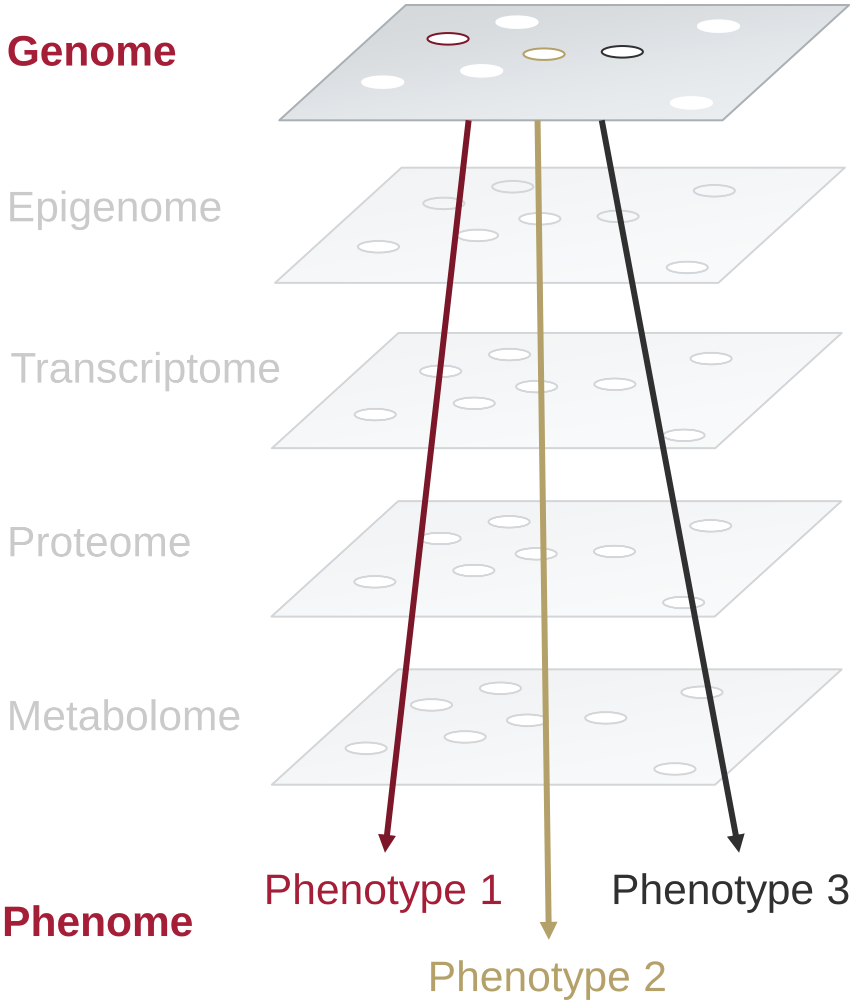
EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

FACULTY OF
SCIENCE

Linking genotype and phenotype: GWAS/PheWAS

Day 2

Single layer connection



Genome

Epigenome

Transcriptome

Proteome

Metabolome

Phenome

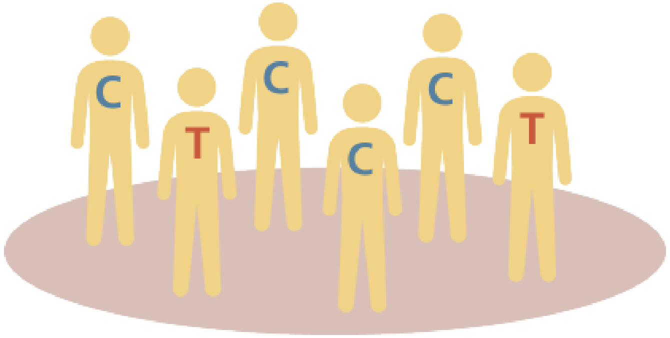
Phenotype 1

Phenotype 2

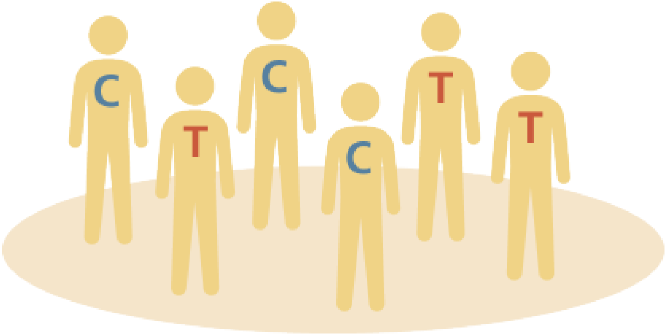
Phenotype 3

- Find association between the changes in the genome and particular characteristics in the phenome
  - Disease
  - Body size, weight
  - Eye/coat color
  - Gene expression

- Multiple approaches
  - Candidate gene approach
  - Genome wide approach
  - Phenotype centered approach



cases (n=1,000)
people with heart disease



controls (n=1,000)
people without heart disease

Motivation

- Observation:
  - Some characteristics of an individual seem to be inherited to their offspring, but differ between unrelated individuals
- Given:
  - Genetics for a set of individuals (e.g., SNPs, microsatellite markers, …)
  - Phenotype for same set of individuals (e.g. height, hair color, disease status, gene-expression, …)
- Goal:
  - Find genetic markers that explain the variance of the phenotype

Today’s focus

- GWAS
  - Genome-wide association study
  - Compare genome-wide set of genetic variants in many individuals to single trait
- PheWAS
  - Phenome-wide associations study
  - Compare many phenotypes in many individuals to single genetic variant (or other attribute) or single gene
  - Logical inverse to GWAS
- eQTL study
  - Association between the a risk SNP and the expression of a nearby gene (expression quantitative trait locus (eQTL))

Terminology

Trait: a distinguishing characteristic

Genotype: Genetic status of individual *i*

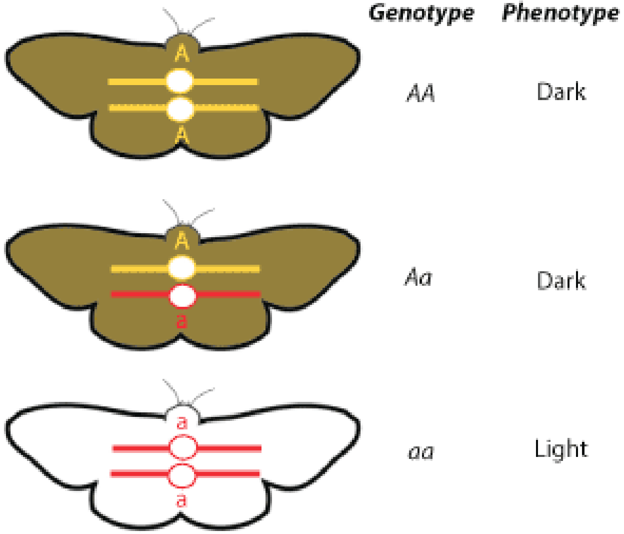
Phenotype: Status of a trait for individual *i*

Allele: Genetic state at a point in the genome

Locus: Position (or limited region) in genome

- Homozygous: maternal and paternal alleles are identical
- Heterozygous: maternal and paternal alleles differ

Haplotype: State of single set of chromosomes



Experimental design: GWAS

- Case-control
  - Compare 2 groups of individuals, one with trait/disease (“case”) and one without (“control)
  - Assumption: individuals in both groups provide unbiased allele frequency estimates from the underlying distribution
  - Examples: disease, hair color
- Quantitative traits
  - Collect measurements of trait for large group of individuals
  - Examples: height, biomarker concentration, gene expression

Statistical approach

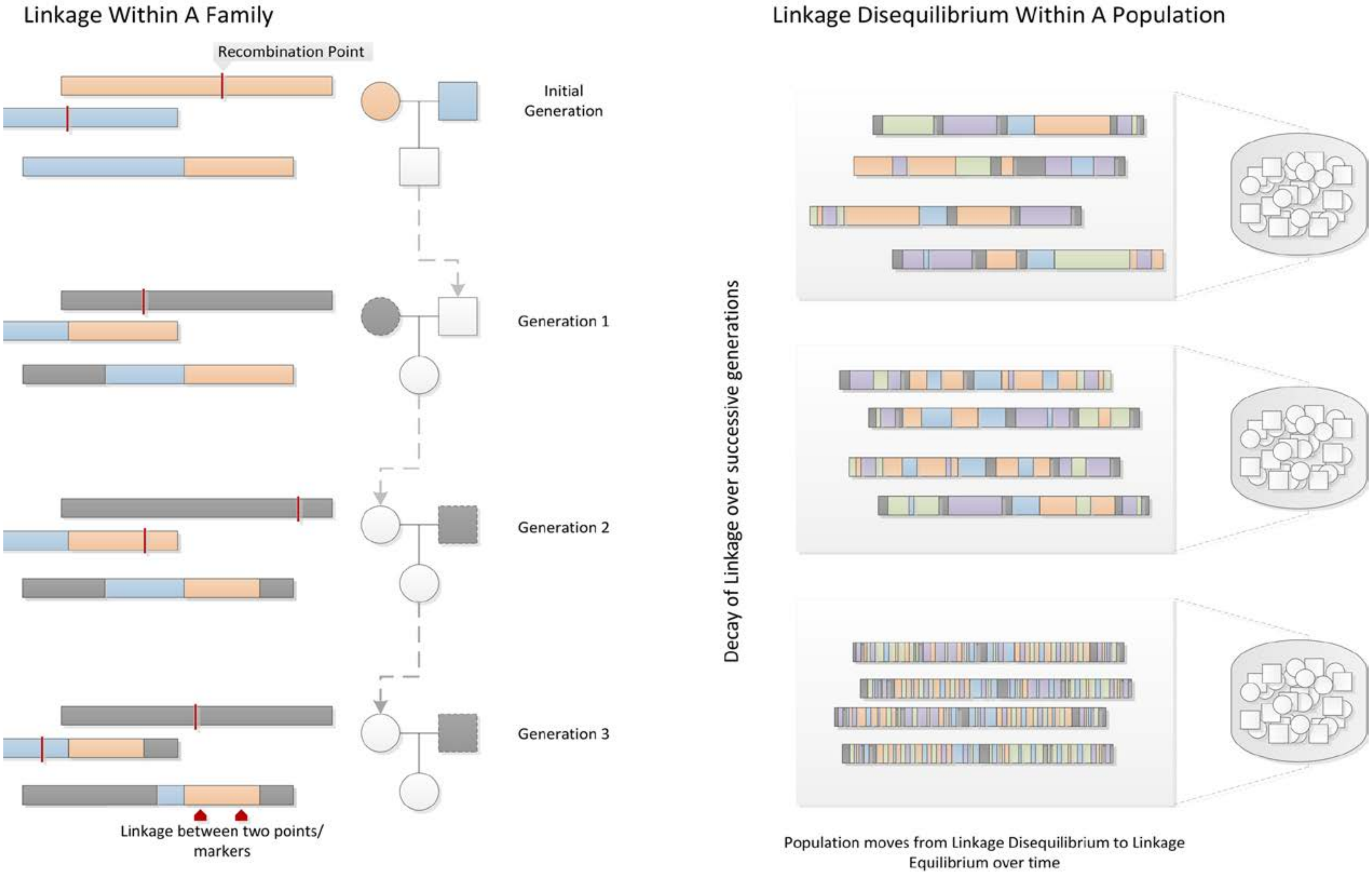
- **Objective:** Find genotypes/alleles significantly associated with phenotype

- Test for each SNP if allele frequency is different between case and control
  - Use **odds ratio** as measure of **effect size**: odds of disease with allele A/odds of disease with allele B
  - If allele frequency in case group higher than control group -> odds ratio > 1
  - **Significance** of odds ratio determined by **chi-squared test**

It’s not quite that easy

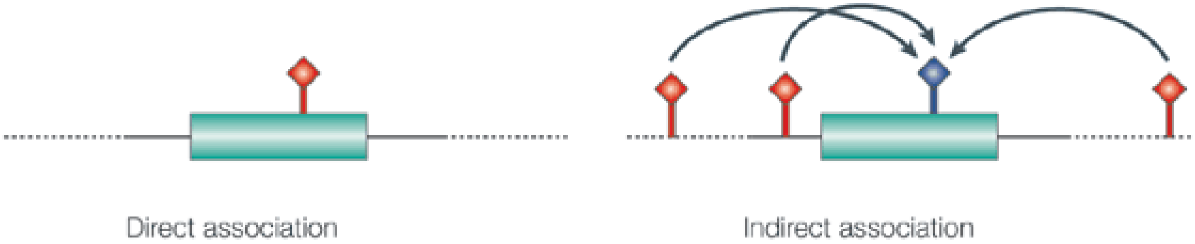
- Not SNP, but genes in **linkage disequilibrium** with SNP may be the once responsible for effect
- **Inheritance patterns:** Dominant and recessive inheritance of disease variants
- **Multiple testing correction:** chance of finding significant hits rise with number of statistical tests performed
- Confounding factors: geographic ancestry, sex, age, … require **population stratification**
- **Epistasis**: multiple loci (multiple SNPs) contribute to complex phenotype
- Not all SNPs are assayed (especially on SNP arrays): require **imputation**

Excursion: Linkage Disequilibrium


Linkage Within A Family
Recombination Point
Initial Generation
Generation 1
Generation 2
Generation 3
Linkage between two points/ markers
Linkage Disequilibrium Within A Population
Decay of Linkage over successive generations
Population moves from Linkage Disequilibrium to Linkage Equilibrium over time

Direct and indirect associations

- Direct association: identified SNP is directly involved in phenotype (e.g., disease causing)
- Indirect association: SNP is not causative, but in LD with true causative SNP



Association ≠ Causation

Disease penetrance

- Penetrance
  - Risk of developing disease at any point associated with a specific SNP
  - E.g. SNP rs6025 in Factor V Leiden associated with 6x risk increase for thrombosis, but most carriers clinically unaffected
  - Small genetic effect = low penetrance
- Heritability
  - Total effect of genetic variants at multiple alleles contributing to the overall disease risk
  - Can be estimated by twin studies
  - Example: heritability of 40$ = 40% of total variance in disease risk can be explained by genetic factors
- Common disease/common variant hypothesis
  - Common diseases are likely influenced by common low penetrance variants

Penetrance models

Multiple models of disease inheritance and penetrance possible for disease allele *A* possible

Common dominant: one or more copies of *A* increase risk of disease (i.e., *a/A* or *A/A*)

Common recessive: *A/A* required for disease

Additive: uniform linear increase in disease risk with each additional allele *A*

- If risk for disease is 3x for *A*, the risk for *A/A* is 2*3=6x higher
- Often used as default model in GWAS

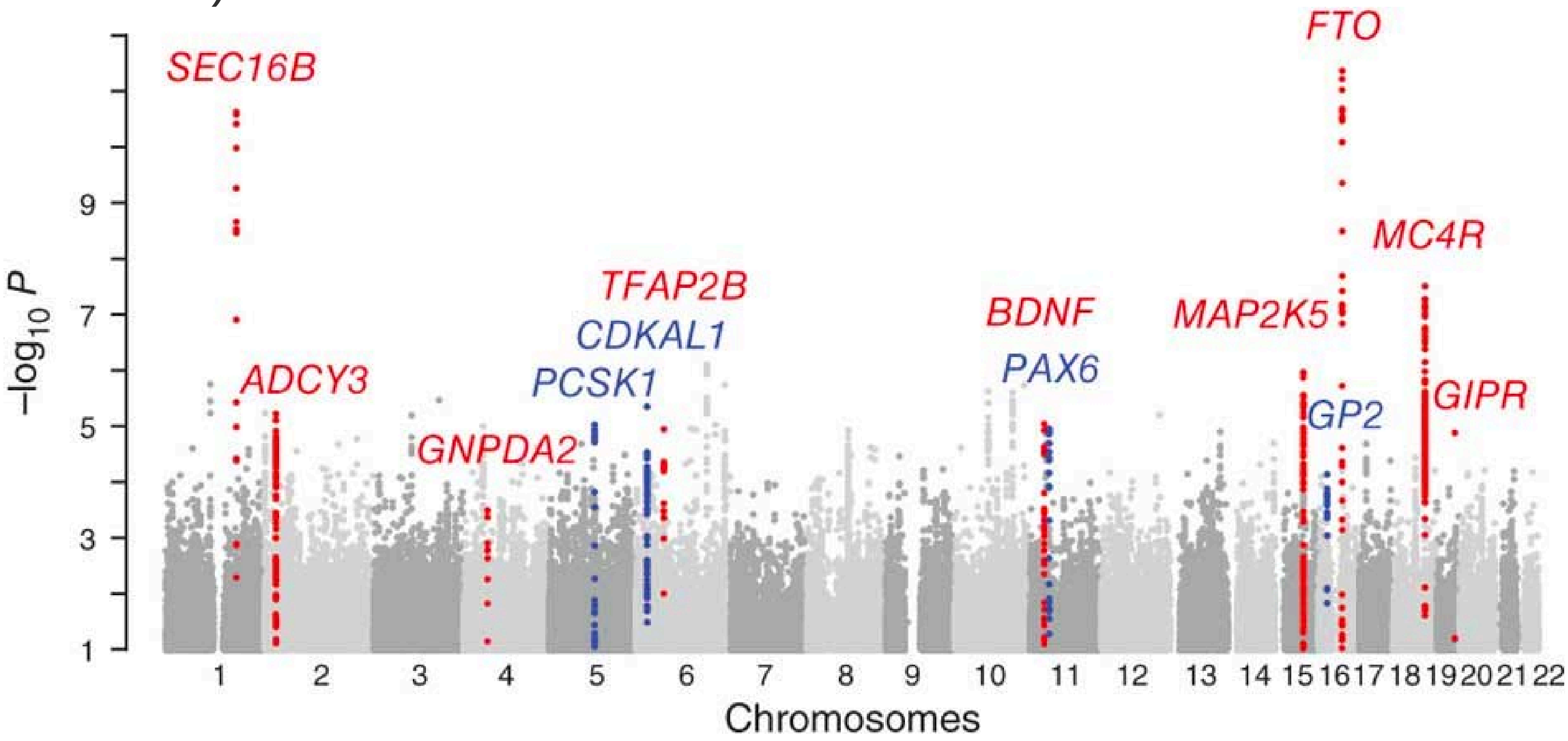
Multiplicative: risk of diseases increases by factor of *k* for each additional allele

- If risk for disease is 3x *A*, the risk for *A/A* is 3^2=9x higher)

Visualization

Manhattan Plot

- X axis: genome location
- Y axis: Negative logarithm of p-value
- Consider locations with significant p-value (here In red and blue)



1. Pick genotype encoding

- Allelic encoding
  - Test for association between one allele and a trait
  - Assumes Hardy-Weinberg, low penetrance and XXX
  - 2 cases for biallelelic locus:
    minor allele *a* and major allele *A*

- Genotypic encoding
  - Test for association between genotype and trait
  - For a biallelelic locus we have 3 unordered genotypes cases:
    a/a, a/A and *A/A*
  - Can be grouped to only contain two cases based on assumed inheritance model: e.g. for dominant A -> *a/a* vs (*a/A* or *A/A*)

- Power of statistical tests varies with encoding

2. Create contingency table

Contingency table for allelic encoding (allele counts)

	a	*A*	**Total**
Case	r_0	r_1	R
Control	s_0	s_1	S
Total	n_0	n_1	N

Contingency table for genotypic encoding (counts)

	a/a	*a/A*	*A/A*	**Total**
Case	r_0	r_1	r_2	R
Control	s_0	s_1	s_2	S
Total	n_0	n_1	n_2	N

Can be combined under dominant model

2. Calculate Effect size

- Calculation of relative risk of genotype: Only possible when exposure data of individuals is available over time

- **Odds ratio** used as alternative in Case-Control studies
  - Odds of event: P(event occurs)/P(event does not occur)

	a	A
Case	r_0	r_1
Control	s_0	s_1

  - Odds of allele *a* occurring in disease: r_0/s_0
  - Odds Ratio between risk of allele X and allele Y occurring in a disease patient:
    OR = odds of *a* in case/odds of *A* in case
    $= r_0/s_0 \,/\, r_1/s_1 = \; r_0\, s_1 \,/\, r_1\, s_0$

Interpretation of odds ratio

- OR = 1: no association
- OR > 1: allele a increases risk of disease
- OR < 1: allele A increases risk of disease

- For rare diseases OR ≈ genotype relative risk

3. Test for association

- **Null hypothesis**: risk of disease is identical between case and control groups
- Categorical data:
  - Chi-squared test
  - Cochran-Armitage trend test
  - Logistic regression models
- Quantitative data:
  - Linear regression models
- More complex models (include epistasis and confounding factors)
  - Linear mixed models
  - Bayesian approaches (incl. hierarchical models)

Chi-square test of independence

Observed

	a	*A*	Total
Case	r_0	r_1	R
Control	s_0	s_1	S
Total	n_0	n_1	N

Expected if independent (H_0)

	a	A	Total
Case	Rn_0/N	Rn_1/N	R
Control	Sn_0/N	Sn_1/N	S
Total	n_0	n_1	N

$$\chi^2 = \sum_{\text{all cells}} \frac{(\text{observed} - \text{expected})^2}{\text{expected}}$$

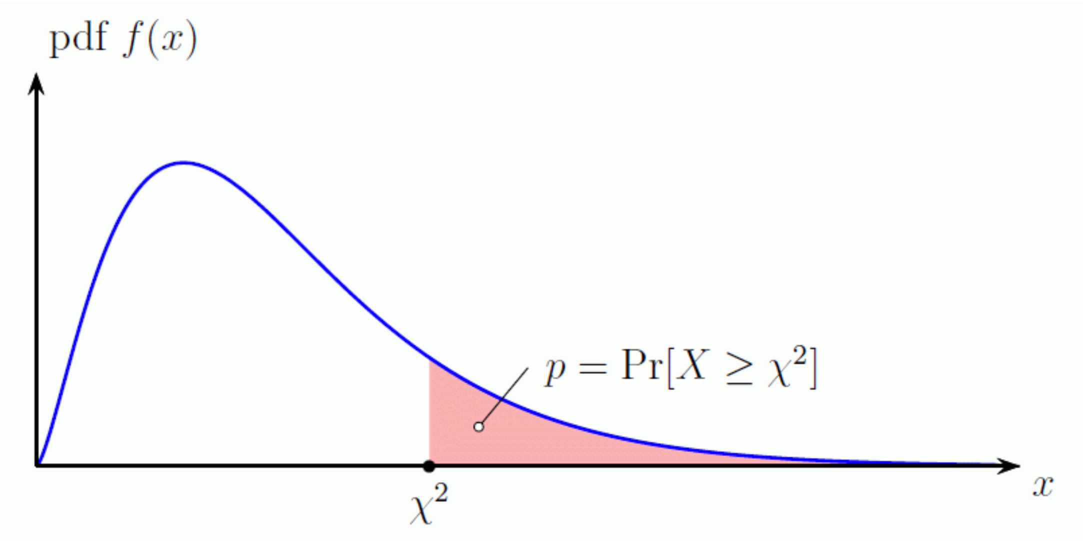
- Large χ^2 statistic corresponds to refusal of H_0
- In case of genotype encoding: 2 χ^2 test required

```
import numpy as np
from scipy.stats import chi2_contingency

obs = np.array([[10, 10, 20], [20, 20, 20]])
chi2, p, dof, ex = chi2_contingency(obs)
```

Obtain P-value for test statistic

p-value: probability of seeing the test statistic or something more extreme if H_0 was true



The p-value is thus the area under the χ^2 density to the right of the observed test statistic

```
from scipy.stats import chi2
p = 1 - chi2.cdf(<$\chi^2$ statistic>, <degrees of freedom>)
p = chi2.sf(<$\chi^2$ statistic>, <degrees of freedom>)
```

3. Control for multiple hypothesis testing

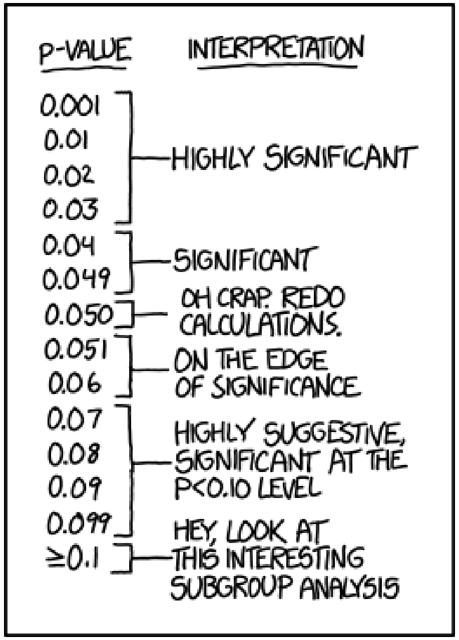
The more tests are being performed the higher the risk of making a type 1 error

➔ Correct for multiple tests before interpretation required

- In single test case:
  - Reject H_0 if p-value ≤ α (usually α ≤ 0.05)
- **Type 1 error** (false positive rate): probability of rejecting null hypothesis even though it’s true
- **Significance level**: proportion of FP that investigator is willing to tolerate (e.g., 5%)
- **Family-wise error rate** (FWER): probability of making one or more type 1 errors in set of tests

Standard correction methods

- Bonferroni correction
  - Adjust threshold p-value by number of tests
  - Reject H_0 if p-value $\alpha^* \leq \alpha/n$ (n=number of tests)
- Šidák correction
  - $\alpha^* \leq 1 - (1 - \alpha)^{1/n}$
- False Discovery Rate (FDR) approaches
  - E.g., as described by Benjamini and Hochberg
  - Control for expected number of false positives among predictions declared significant
- Permutation approaches

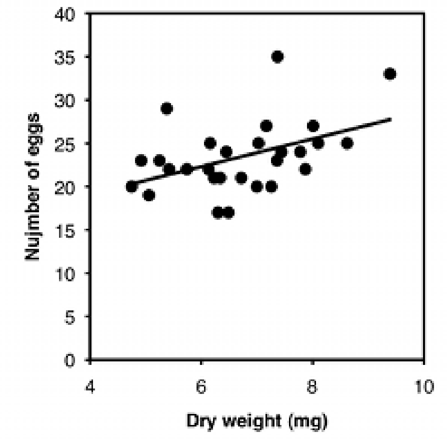

P-VALUE
INTERPRETATION
0.001
0.01
0.02
0.03
HIGHLY SIGNIFICANT
0.04
0.049
SIGNIFICANT
0.050
OH CRAP. REDO CALCULATIONS.
0.051
0.06
ON THE EDGE OF SIGNIFICANCE
0.07
0.08
0.09
0.099
HIGHLY SUGGESTIVE, SIGNIFICANT AT THE P<0.10 LEVEL
≥0.1
HEY, LOOK AT THIS INTERESTING SUBGROUP ANALYSIS

Source: xkcd

```
from statsmodels.sandbox.stats.multicomp import multipletests
rej, corrected_p, _, _ = multipletests(pvals, alpha=0.1,
                                        method='fdr_bh')
```

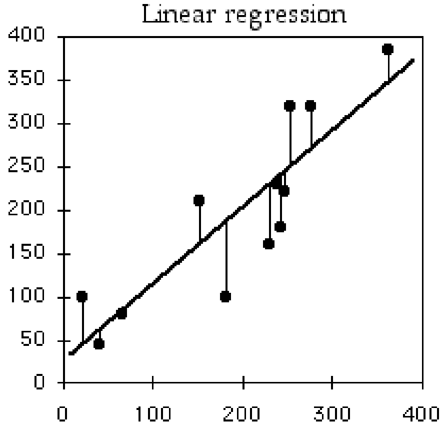
More complex: Regression models

- Terminology:
  - Dependent variable y: output/outcome whose variation is studied
  - Independent variables x: inputs/potential causes for variation



Goal
Find model to describe relationship between variables

$$f(x) = \beta x + \epsilon$$



- Categorical data: Logistic regression
- Quantitative data: Linear regression

Mathematical formulation
Linear Regression

Model

$$\overset{\text{response}}{y_i} = \underset{\text{intercept}}{\beta_0} + \underset{\text{slope}}{\beta_1} \overset{\text{feature}}{x_i}$$

Predict value of one variable through values of one or more other variables

H_0: independence of variables, i.e. $\beta_1 = 0$

Fit to data

Inputs:

Dependent variable Y: phenotype of individuals

y_i = continuous measurement of phenotype

Independent variable X: genotype of individuals at specific locus

x_i = 0 for phenotype *a/a*
x_i = 1 for phenotype *a/A*
x_i = 2 for phenotype *A/A*

Mathematical formulation
Logistic Regression

Logistic regression is similar to linear regression, but with binary outcomes

Outcome determined by an unobserved probability p_i

$$p_i = E[y_i | x_i]$$

equal to expected value of phenotype given genotype

Model

$$\underbrace{\ln\left(\frac{p_i}{1-p_i}\right)}_{\text{logit}(p_i)} = \beta_0 + \beta_1 x_i$$

Inputs:

Dependent variable y_i: phenotype of individuals

y_i = 0 for control (i.e. no disease)

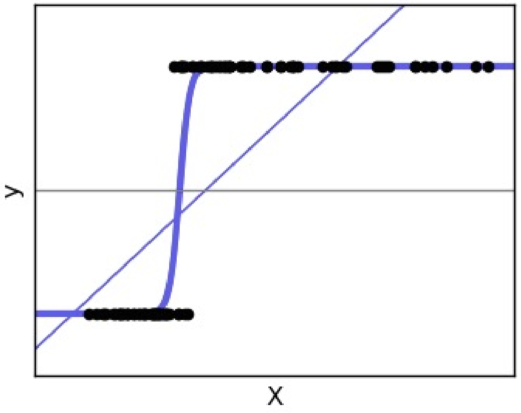
y_i = 1 for case (i.e. disease)

Independent variable x_i: genotype of individuals at locus

x_i = 0 for phenotype *a/a*

x_i = 1 for phenotype *a/A*

x_i = 2 for phenotype *A/A*



Interpreting model coefficients

- Test whether β_1 significantly differs from 0
  - Rejection of $H_{0,}$ i.e. assumption of independence
  - Roughly equivalent to Chi-square test
  - P-value is determined during model fitting process in python

- β_1
  - Linear model: linear association between a “unit” increase of x with a “unit” increase in outcome
  - Logistic model: e^{β_1} is estimate of odds ratio

Effect size: Strength of correlation

- Effect size in regression models:
  - variance of the experiment explained by the model
- Pearson's correlation coefficient *r*
  - On population level: Incorporates covariance of two variables and their independent standard deviations
  - On sample level: estimate covariance and standard deviation

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{1.1}$$

- with $\bar{y_i} = \frac{1}{n}\sum_{i=1}^{n} y_i$ being the mean of the observed data

- Coefficient of Determination R^2
  - Proportion of variance of dependent variable explained by independent variable
  - Total sum of squares: variability in the data

$$SS_{\text{tot}} = \sum_i (y_i - \bar{y})^2 \tag{2.1}$$

  - Residual sum of squares: average amount to which data differs from prediction

$$SS_{\text{res}} = \sum_i (y_i - \hat{y_i})^2 \tag{2.2}$$

  - Coefficient of Determination follows from these two

$$R^2 = \frac{SS_{\text{tot}} - SS_{\text{res}}}{SS_{\text{tot}}} = 1 - \frac{SS_{\text{res}}}{SS_{\text{tot}}} \tag{2.3}$$

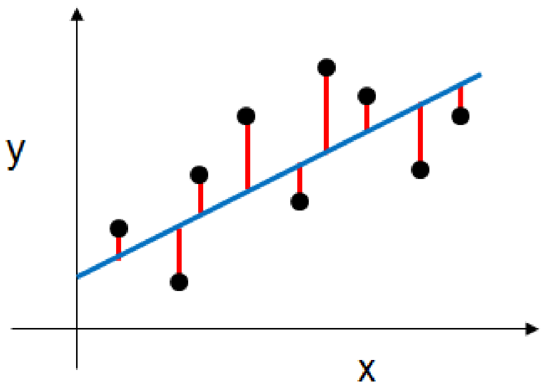
Proof for relationship between Coefficient of Determination R^2 and Squared Pearson Correlation Coefficient r can be found here: https://economictheoryblog.com/2014/11/05/proof/

- Adjusted R²
  - Normal R2 increases by increasing number of explanatory variables in model
  - Adjusts by degrees of freedom
  - For sample size n and d explanatory variables:

$$\bar{R}^2 = 1 - \frac{SS_{\text{res}}/(n-d-1)}{SS_{\text{tot}}/(n-1)} \tag{3.1}$$

Training

- Goal: Learn model coefficients $\beta_{0,}$ $\beta_{1,}$ … based on data
- Minimize the residuals for observed values for x and y
  - Residuals: remaining error between prediction and observed data, i.e. predicted - observed



Model Prediction

$$SS_{residuals} = \sum_{i=1}^{N} (\hat{y}_i - y_i)^2$$

Observed Result

```
import statsmodels.formula.api as smf
lm = smf.ols(formula=’<dep_col> ~ <indep_col>’,
             data=data).fit()
```

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Obtain relevant information in Python

```
import statsmodels.formula.api as smf

# Train model
lm = smf.ols(formula=’<dep_col> ~ <indep_col>’,
             data=data).fit()

# Print coefficients
lm.params

# Obtain confidence intervals for coefficients
lm.conf_int()

# Obtain p-values for coefficients
lm.pvalues

# Summary of model
lm.summary()
```

Logistic regression in Python

```
import statsmodels.formula.api as smf

# Train model
lm = smf.logit(formula=’<dep_col> ~ <indep_col>’,
             data=data).fit()

# Extract data sets for functions that do not support
# direct formula notation
import patsy
f = ’<dep_col> ~ <indep_col>’
y, X = patsy.dmatrices(f, df, return_type='dataframe’)

# Alternative to statsmodels: scikit-learn
from sklearn import linear_model
regr = linear_model.LinearRegression()
regr.fit(X, y)

logr = linear_model.LogisticRegression()
logr.fit(X, y)
```

More complex models

- Factor in other confounding variables in the model in multiple regression models

$$y = \beta_0 + \beta_1 \underset{\substack{\uparrow \\ \text{Feature 1}}}{x_1} + \beta_2 \underset{\substack{\uparrow \\ \text{Feature 2}}}{x_2} + \ldots + \beta_n \underset{\substack{\uparrow \\ \text{Feature n}}}{x_n}$$

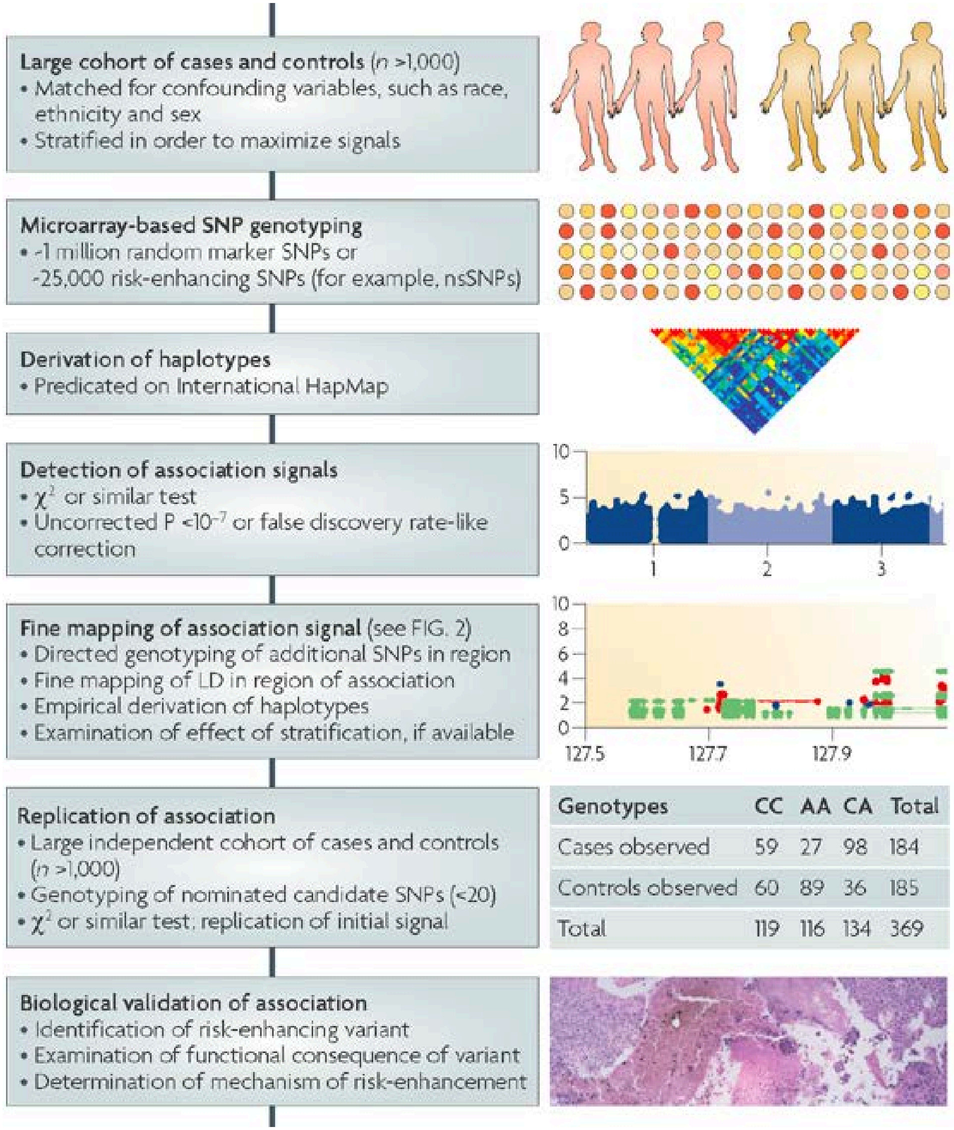
- Examples for features for confounding factors
  - Ethnicity of sample
  - Sex of sample
  - Sequencing batch

```
lm = smf.ols(formula=’<dep_col> ~ <indep_col_1> +
               <indep_col_2> + … +
                      <indep_col_n>’, data=data).fit()
```

Quality control

- Plot residuals vs predicted values
  - Any non-random effects?
- Replicate results in independent study
- Biological validation

Summary: GWAS workflow



Large cohort of cases and controls (*n* >1,000)
- Matched for confounding variables, such as race, ethnicity and sex
- Stratified in order to maximize signals

Microarray-based SNP genotyping
- ~1 million random marker SNPs or ~25,000 risk-enhancing SNPs (for example, nsSNPs)

Derivation of haplotypes
- Predicated on International HapMap

Detection of association signals
- χ^2 or similar test
- Uncorrected P <10^{-7} or false discovery rate-like correction

Fine mapping of association signal (see FIG. 2)
- Directed genotyping of additional SNPs in region
- Fine mapping of LD in region of association
- Empirical derivation of haplotypes
- Examination of effect of stratification, if available

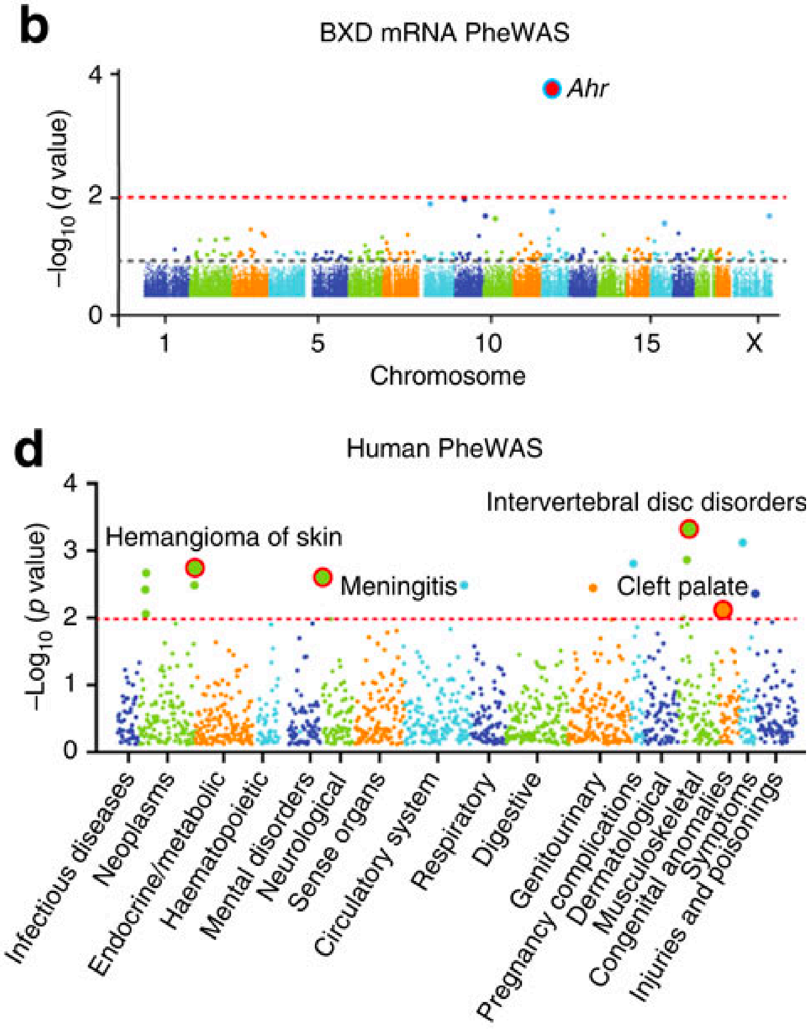
Replication of association
- Large independent cohort of cases and controls (*n* >1,000)
- Genotyping of nominated candidate SNPs (<20)
- χ^2 or similar test; replication of initial signal

Biological validation of association
- Identification of risk-enhancing variant
- Examination of functional consequence of variant
- Determination of mechanism of risk-enhancement

Genotypes	CC	AA	CA	Total
Cases observed	59	27	98	184
Controls observed	60	89	36	185
Total	119	116	134	369

PheWAS

- Compare single SNP against range of phenotypes
- Same statistical methods apply to PheWAS that are also used in GWAS
- Manhattan Plots
  - Phenotype Categories on x-axis
  - Negative logarithm of p-value on y-axis



Interesting Papers

- Papers
  - Bush and Moore, **Chapter 11: Genome-Wide Association Studies**, PLOS Comp Bio, 2012 (8)12
  - Clarke et al., **Basic statistical analysis in genetic case-control studies**, Nature Protocols, 2011 (6)2
  - Stephens and Balding, Bayesian statistical methods for genetic association studies, Nature Reviews Genetics, 2009 (10)
- Books
  - James, Witten, Hastie, Tibshirani, **An Introduction to Statistical Learning**, Springer Texts, 2013*
  - MacKay, Information Theory, Inference, and Learning Algorithms, Cambridge University Press, 2005**
- Blogs
  - Erik Minikel's blog: http://www.cureffi.org/2014/11/21/genetics-25/

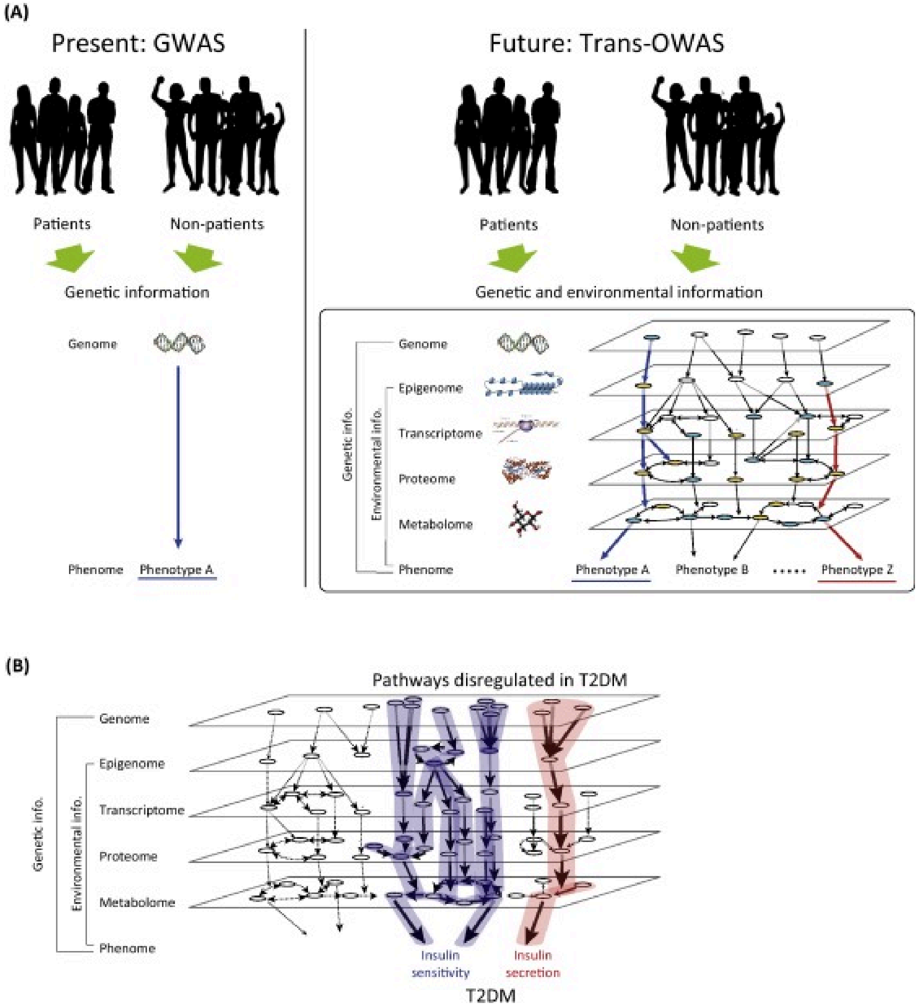
* Can be officially and freely downloaded from http://www-bcf.usc.edu/~gareth/ISL/

** Can be officially and freely downloaded from http://www.inference.phy.cam.ac.uk/itila/book.html

Some resources

- Databases
  - GWAS catalog (https://www.ebi.ac.uk/gwas/)
  - PheWAS catalog (https://phewas.mc.vanderbilt.edu/)
- Standalone tools
  - PLINK
  - Eigenstrat
- Python modules
  - pylmm
  - fastlmm
- R packages
  - PheWAS package
  - GenABEL

Outlook



Trends in Biotechnology