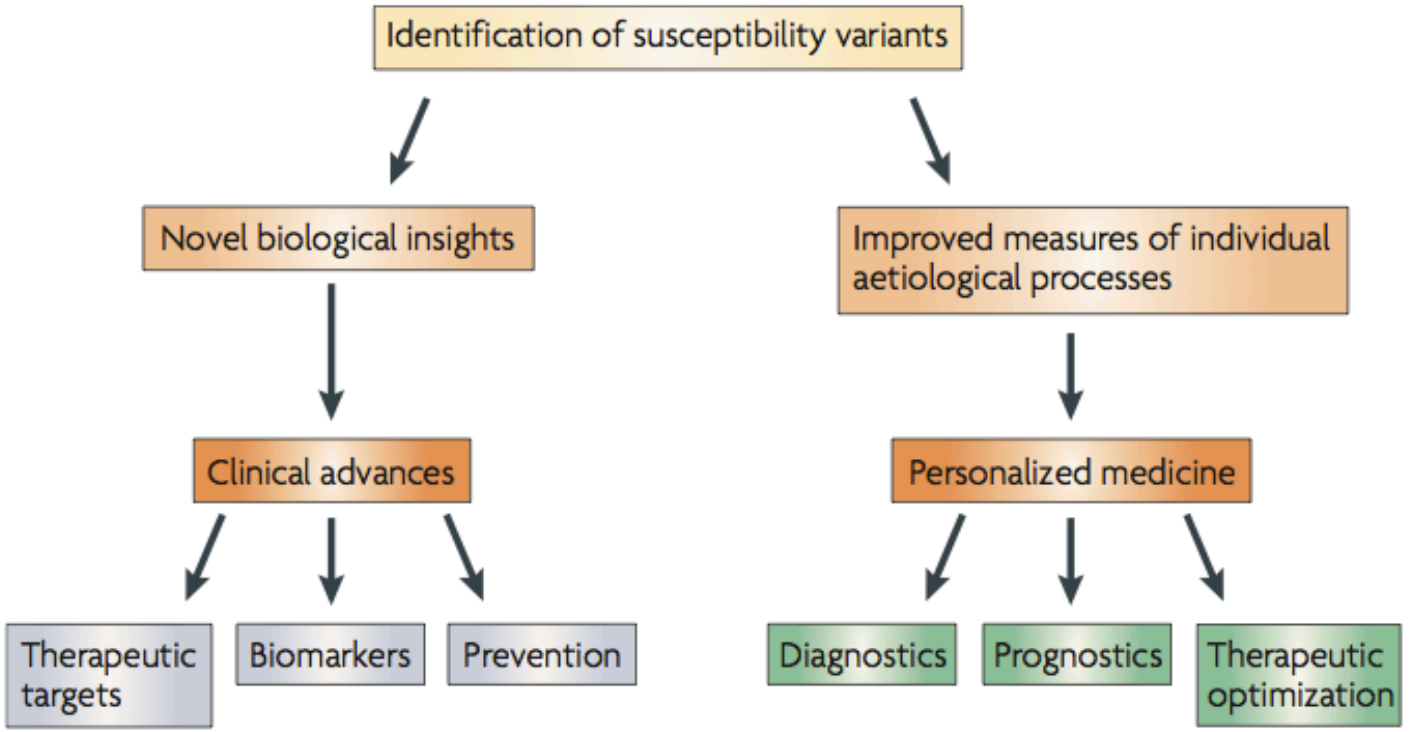
Genetic Association Analysis
--- impact of NGS



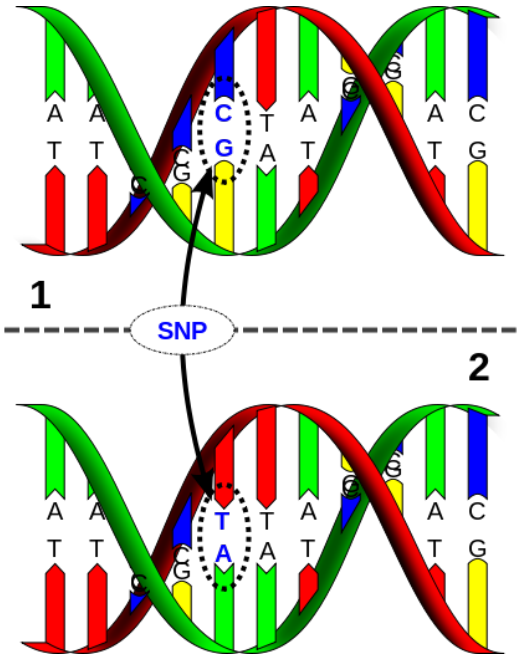
- One fundamental goal of genetics studies is to identify genetic variants causing phenotypic variations
- What does NGS have to offer?

Genome-wide association studies for complex traits: consensus, uncertainty and challenges. M I McCarthy, G R Abecasis, et al. Nature Review Genetics, 2008

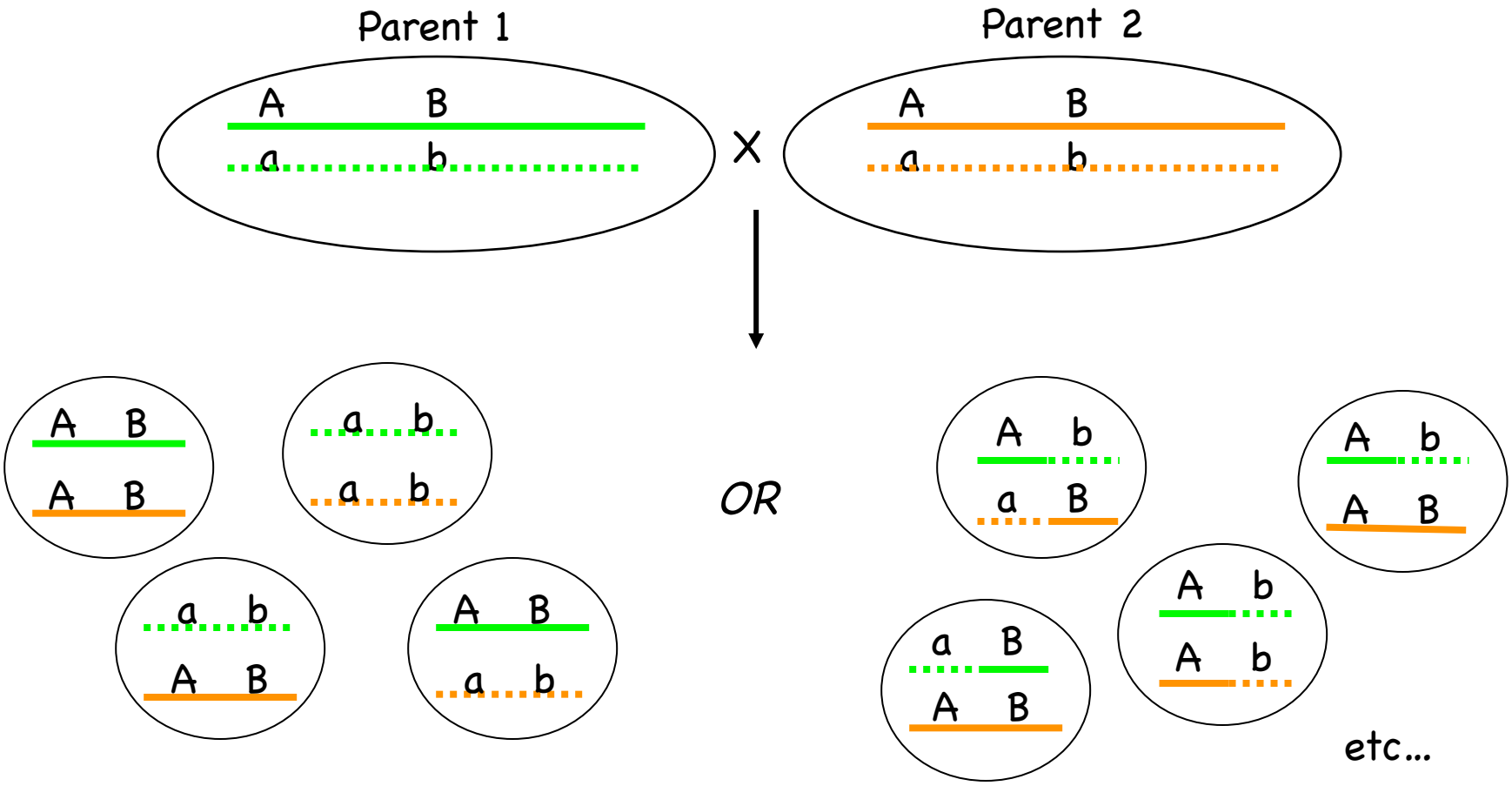
- Before NGS, what do people do?
  - Linkage analysis
  - Genome-wide association studies

- Genome-wide association studies (GWAS)
  - SNP (**s**ingle **n**ucleotide **p**olymorphism)
  - Technology: Microarray
  - Two major manufacturers:
  - Illumina and Affymetrix

http://en.wikipedia.org/wiki/Single-nucleotide_polymorphism



Linkage Disequilibrium



High LD -> No Recombination
(r^2 = 1) SNP1 "tags" SNP2

Low LD -> Recombination
Many possibilities

ASHG 2008 Hapmap Tutorial: http://hapmap.ncbi.nlm.nih.gov/tutorials.html.en

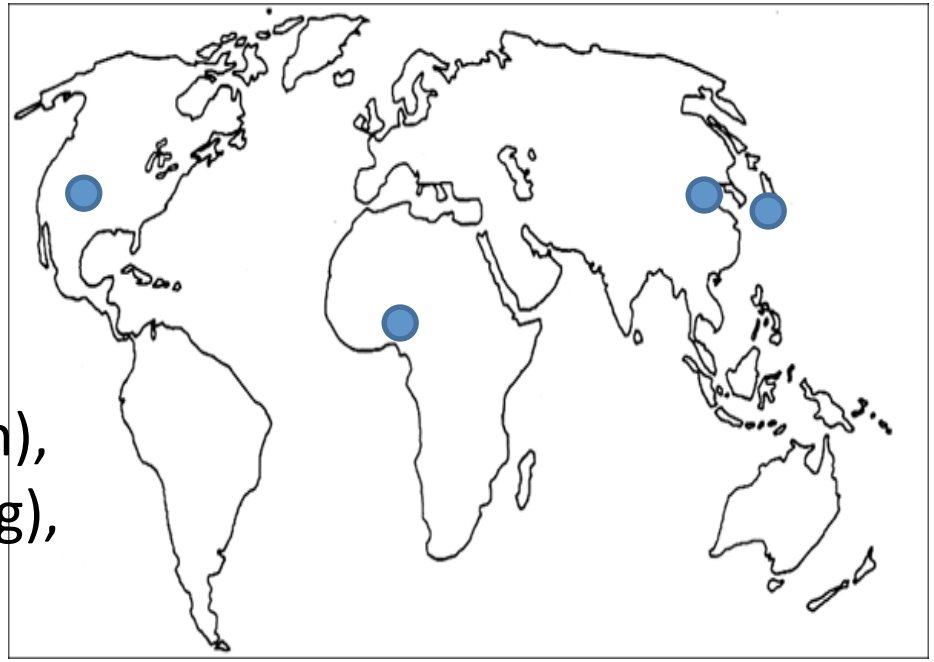
- SNPs on microarrays are "tagging" SNPs (reduce cost!!!)
- Selected based on linkage-disequilibrium structure
- How do we know the LD structure?

  The International HapMap Project

www.hapmap.org

The International HapMap Project

- Involved Illumina,
  Affymetrix,
  >20 institutions worldwide
- HapMap1 (2003) and
  Hapmap2 (2005)
  - 4 populations (270 indiv):
  CEU (NW European from Utah),
  CHB (Han Chinese from Beijing),
  JPT (Japanese from Tokyo),
  YRI (Yoruban from Nigeria)
- Hapmap3 (2010)
  - 11 populations (4+7, 1301 indiv)



www.hapmap.org

- In GWAS, only common SNPs (generally, with minor allele frequency > 5%) are considered
  - Only common SNPs can **"tag"** other common SNPs
  - The actual "causal" SNPs are usually not directly genotyped

- With NGS, we can:
  - Analyze rare variants
  - Get much better (highest possible) resolution

- But, are we there yet?
  - What are the challenges of analyzing rare variants?
  - What have we done?

- Challenge #1: Very limited statistical power
- A toy example:
- Suppose we wish to test the association between a gene (with alleles A and B) and human height. We collected 100 individuals from the population

	Scenario #1		Scenario #2	
	Allele A	Allele B	Allele A	Allele B
# of indiv	70	30	99	1
Avg height	6’	6’1’’	6’	6’1’’

Equal effect size for the variants in the two scenarios
Which scenario is more convincing about the association?

- Challenge #1: Very limited statistical power
- A toy example:
- Suppose we wish to test the association between a gene (with alleles A and B) and human height. We collected 100 individuals from the population

	Scenario #1		Scenario #2	
	Allele A	Allele B	Allele A	Allele B
# of indiv	70	30	99	1
Avg height	6’	6’1’’	6’	6’1’’

Equal effect size for the variants in the two scenarios

Which scenario is more convincing about the association?

- To maintain the same statistical power, a rare variant must have much larger effect size than a common variant.

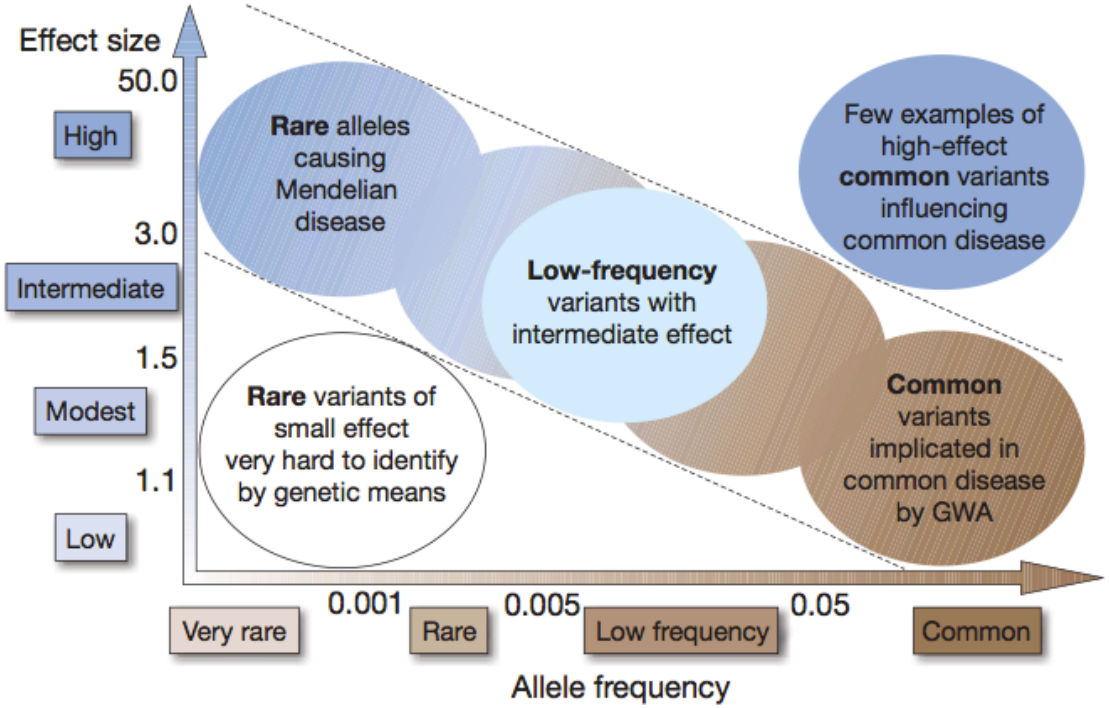
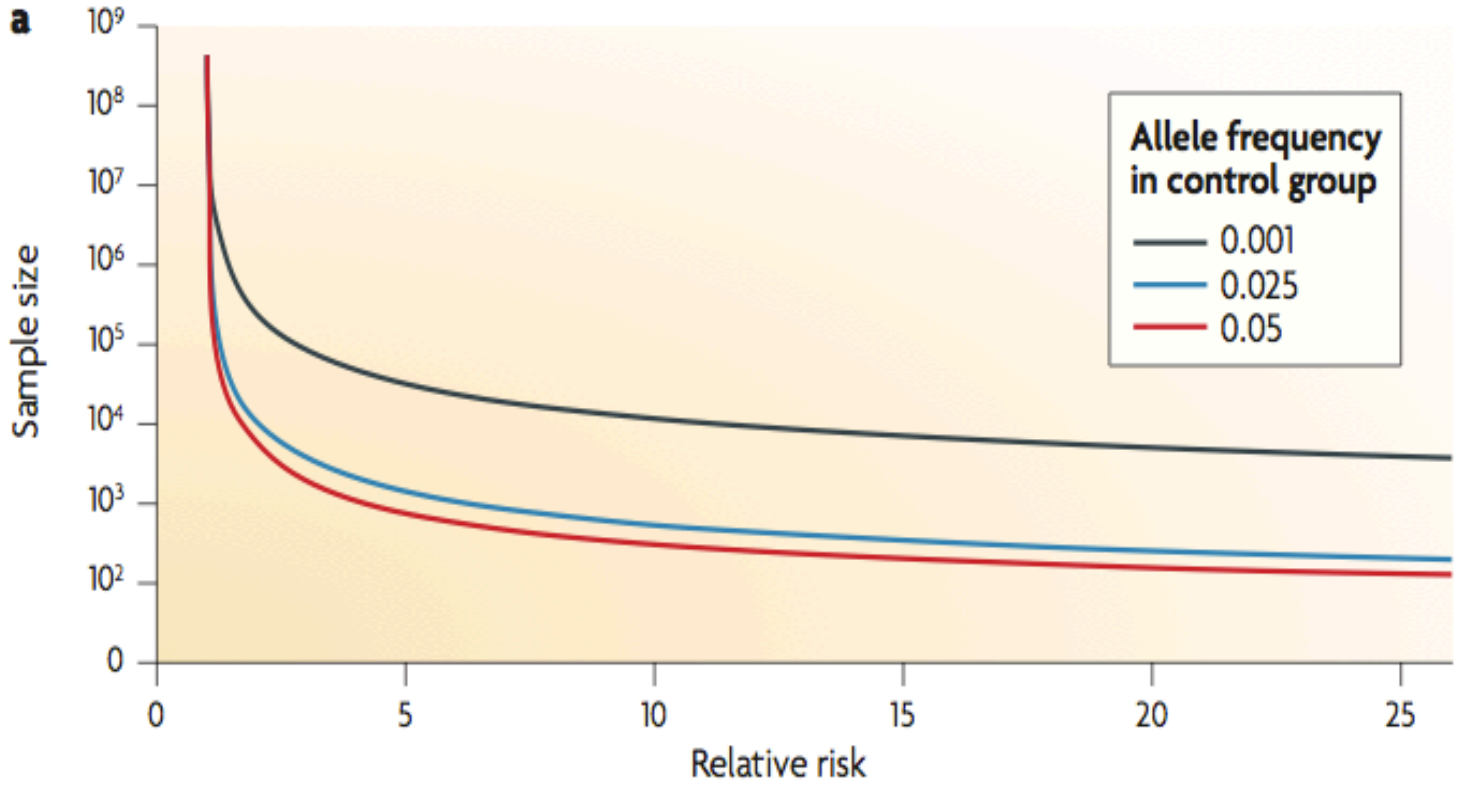


Figure 1 | Feasibility of identifying genetic variants by risk allele frequency and strength of genetic effect (odds ratio). Most emphasis and interest lies in identifying associations with characteristics shown within diagonal dotted lines. Adapted from ref. 42.

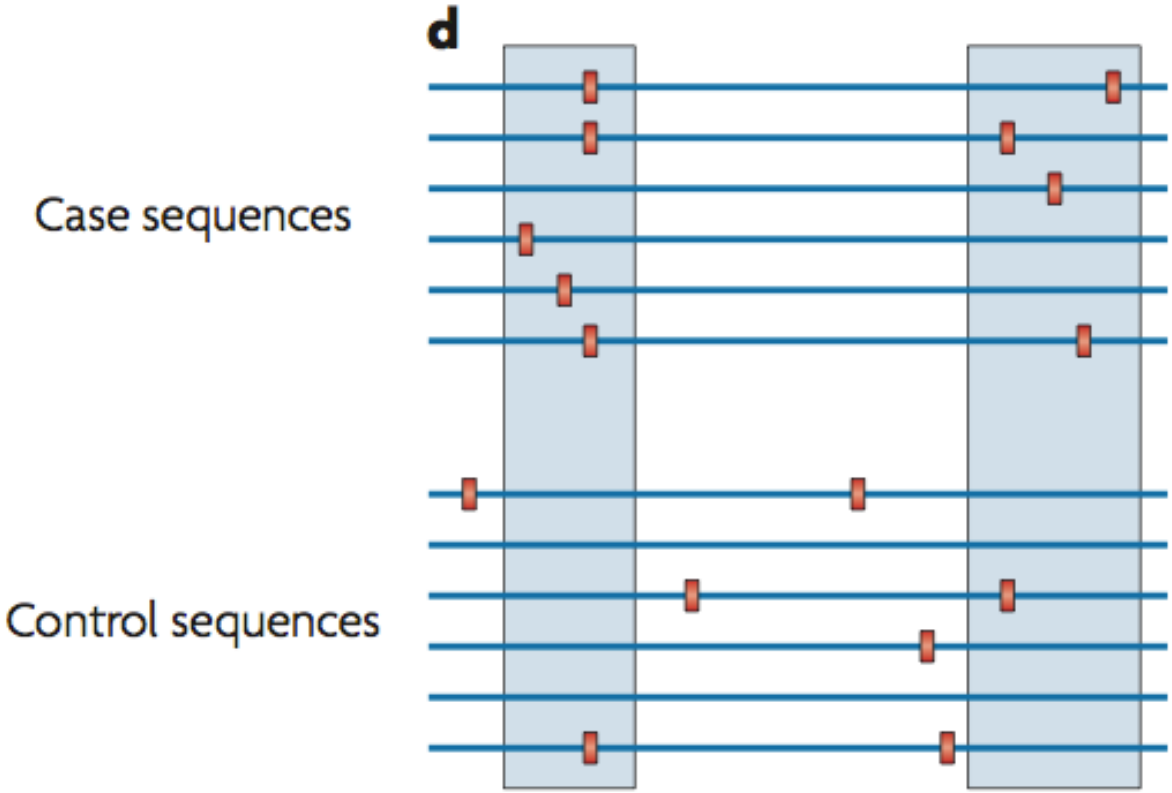
Finding the missing heritability of complex diseases. T A Manolio, F S Collions, N J Cox, et al. Nature Reviews. 2009

- With the same effect size, rare variants need much larger sample size to be detected than common variants



Statistical analysis strategies for association studies involving rare variants. V Bansal, O Libiger, A Torkamani and N J Schork. Nature Reviews Genetics. 2010.

- One strategy to deal with this problem is to create a "super-variant" by "collapsing" rare variants that belong to a functional unit (e.g. a gene)



Statistical analysis strategies for association studies involving rare variants. V Bansal, O Libiger, A Torkamani and N J Schork. Nature Reviews Genetics. 2010.

- Collapsing methods:
  - Burden tests
  - Kernel-based tests

- Sum tests
  - CAST (cohort allelic sums test)
    - Define a "super variant" X_C for each collapsing set C
    - X_C = 1 if the individual carries any of the rare variants in the collapsing set
  - CMC test (combined multivariate and collapsing test)
    - Extension of CAST
    - Including each common variant (without collapsing) and do multivariate test

A strategy to discover genes that carry multi-allelic or mono-allelic risk for common diseases: a cohort allelic sums test (CAST). Morgenthaler S, Thilly W G. Mut. Res. 2007

Methods for Detecting Associations with Rare Variants for Common Diseases: Application to Analysis of Sequence Data. Li B, Leal S M, 2008. Am J Hum Genet.

- In CAST and CMC tests, when a collapsing set is large enough, the "super-variant" for every individual will be 1

- A modification: Sum test
  - Define the super-variant X_C as the total number of rare variants within the collapsing set carried by an individual

Analysis of multiple SNPs in a candidate gene or region. Chapman J M, Whittaker. Genet Epidemiol. 2008.

- A further extension
  - weighted-sum test (w-Sum)
  - allows one to include variants of all allele frequency in a collapsing set
  - weight variants according to allele frequency so that rare variants are not overwhelmed by common variants

A Groupwise Association Test for Rare Mutations Using a Weighted Sum Statistic. Madsen B E, Browning S R. PLoS Genet. 2009.

Pooled association tests for rare variants in exon-resequencing studies. Price A L et al. Am J Hum Genetic, 2010.

- Pros and cons of burden tests
  - Pro: Degree of freedom is 1
  - Con: won't work when variants within a collapsing set affect the phenotype in different directions

- aSum (adaptive sum) test
  - Decide the sign of each variant by its marginal association with the trait
  - Account for possible opposite association direction
  - The cost is that degrees of freedom are consumed while estimating the signs from the data

A data-adaptive sum test for disease association with multiple common or rare variants. Han F, Pan W. Hum. Hered. 2010.

- Another class of tests that account for possible sign differences within a collapsing set are the **kernel-based tests**

- Kernel-based test
  - Two ways to understand it
  - A. If a set of variants contain some causal variants, then phenotype similarities should be correlated with the "genotype similarities" defined on these variants
  - B. Assuming the effects of a set of variants come from a distribution with zero mean and some variance, it tests whether the variance is zero or not
  - No assumptions about the direction of association

- Kernel-based test
  - Example: SKAT (Sequence Kernel Association Test)
  - A very popular R package
  - Use kernel methods to compute SNP-set level p-values efficiently
  - Allows adjusting for covariates
  - Flexible kernel choices (able to account for the interactions between variants)

Rare-variant association testing for sequencing data with the sequence kernel association test. Wu MC, Lee S, Cai T, Li Y, Boehnke M, Lin X. Am J Hum Genet. 2011

- Summary
  - Due to the low allele frequency, direct testing rare variants has very limited power
  - Assuming multiple causal variants fall in a pre-defined variant set, one can collapse the variants in the set and test on the set of variants
  - Burden tests work well when all variants in a collapsing set affect the phenotype in the same direction
  - Kernel-based test can deal with opposite association directions

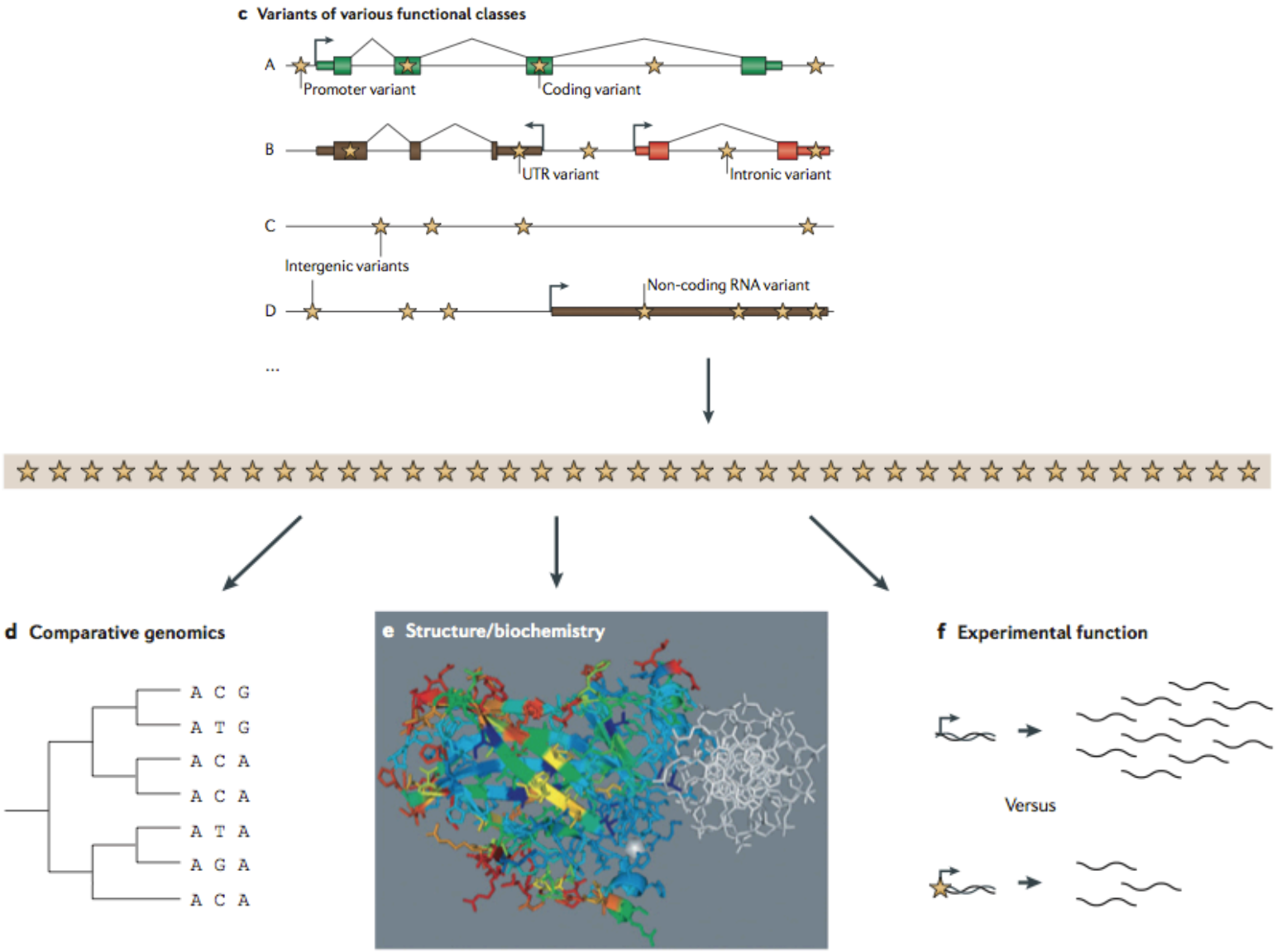
- Family-based study design – enriching rare variants
  - Rare variants may not longer be rare within a family
  - Traditional association tests that assume independence between samples are no longer valid
  - Relationships between family members need to be accounted for

- Testing rare variants in family-based design
  - Example: famSKAT (family-based SKAT)
  - Extension of the original SKAT method
  - Adding a variance component to the original SKAT model to account for familial relatedness between samples
  - Only available for quantitative trait yet

Sequence Kernel Association Test for Quantitative Traits in Family Samples.
H Chen, J B Meigs, J Dupuis. Genetic Epidemiology, 2013

- Challenge #2: Needles in haystack
  - A few causal variants in a huge number of variants
  - In statistical language: "multiple testing burden"
  - Need to reduce the total number of variants to be tested (and try to avoid missing true causal variants)

- Commonly used strategies
  - Targeted sequencing (e.g. Exome-Seq)
  - Filter variants by functional annotations (e.g. synonymous mutations)
  - More generally speaking, filter variants based on predicted "biological importance"
  - Rationale: a. reduce false positives; b. biologically unimportant variants usually have small effect sizes (hard to detect anyway)



Needles in stack of needles: finding disease-causal variants in a wealth of genomic data. G M Cooper, J Shendure. Nature Reviews Genetics. 2011.

Table 1 | **Tools for protein-sequence-based prediction of deleteriousness**

Name	Type	Information	URL	Refs
MAPP	Constraint-based predictor	Evolutionary and biochemical	http://mendel.stanford.edu/SidowLab/downloads/MAPP/index.html	27
SIFT	Constraint-based predictor	Evolutionary and biochemical (indirect)	http://sift.bii.a-star.edu.sg/	39
PANTHER	Constraint-based predictor	Evolutionary and biochemical (indirect)	http://www.pantherdb.org/	41
MutationTaster*	Trained classifier	Evolutionary, biochemical and structural	http://www.mutationtaster.org/	40
nsSNP Analyzer	Trained classifier	Evolutionary, biochemical and structural	http://snpanalyzer.uthsc.edu/	44
PMUT	Trained classifier	Evolutionary, biochemical and structural	http://mmb2.pcb.ub.es:8080/PMut/	38
polyPhen	Trained classifier	Evolutionary, biochemical and structural	http://genetics.bwh.harvard.edu/pph2/	35
SAPRED	Trained classifier	Evolutionary, biochemical and structural	http://sapred.cbi.pku.edu.cn/	42
SNAP	Trained classifier	Evolutionary, biochemical and structural	http://www.rostlab.org/services/SNAP/	36
SNPs3D	Trained classifier	Evolutionary, biochemical and structural	http://www.snps3d.org/	51
PhD-SNP	Trained classifier	Evolutionary and biochemical (indirect)	http://gpcr2.biocomp.unibo.it/~emidio/PhD-SNP/PhD-SNP_Help.html	37

*Also makes predictions for synonymous and non-coding variant effects: for example, splicing. MAPP, Multivariate Analysis of Protein Polymorphism; polyPhen, polymorphism phenotyping.

Needles in stack of needles: finding disease-causal variants in a wealth of genomic data. G M Cooper, J Shendure. Nature Reviews Genetics. 2011.

Table 2 | **Tools for nucleotide-sequence-based prediction of deleteriousness**

Name	Type	Information	URL	Refs
phastCons	Phylogenetic HMM	Evolutionary	http://compgen.bscb.cornell.edu/phast/	60
GERP	Single-site scoring	Evolutionary	http://mendel.stanford.edu/SidowLab/downloads/gerp/index.html	67
Gumby	Single-site scoring	Evolutionary	http://pga.jgi-psf.org/gumby/	21
phyloP	Single-site scoring	Evolutionary	http://compgen.bscb.cornell.edu/phast/	66
SCONE	Single-site scoring	Evolutionary	http://genetics.bwh.harvard.edu/scone/	68
binCons	Sliding-window scoring	Evolutionary	http://zoo.nhgri.nih.gov/binCons/index.cgi	69
Chai Cons	Sliding-window scoring	Evolutionary and structural	http://research.nhgri.nih.gov/software/chai	71
VISTA	Visualization tool (various scores)	Evolutionary	http://genome.lbl.gov/vista/index.shtml	70

GERP, Genomic Evolutionary Rate Profiling; HMM, hidden Markov model; SCONE, Sequence Conservation Evaluation.

Needles in stack of needles: finding disease-causal variants in a wealth of genomic data. G M Cooper, J Shendure. Nature Reviews Genetics. 2011.

- Despite so many efforts, not many rare variants were detected for common diseases
- Rare variant detection is much more successful for rare diseases

- A possible explanation: even with all the above efforts, the power may be still not enough?

- Or, rare variants may not contribute that much susceptibility for common disease?

LETTER

doi:10.1038/nature12170

Negligible impact of rare autoimmune-locus coding-region variants on missing heritability

Karen A. Hunt[1], Vanisha Mistry[1], Nicholas A. Bockett[1], Tariq Ahmad[2], Maria Ban[3], Jonathan N. Barker[4], Jeffrey C. Barrett[5], Hannah Blackburn[5], Oliver Brand[6], Oliver Burren[7], Francesca Capon[4], Alastair Compston[3], Stephen C. L. Gough[6], Luke Jostins[8], Yong Kong[9], James C. Lee[10], Monkol Lek[11], Daniel G. MacArthur[11], John C. Mansfield[12], Christopher G. Mathew[4], Charles A. Mein[13], Muddassar Mirza[4], Sarah Nutland[7], Suna Onengut-Gumuscu[14], Efterpi Papouli[4], Miles Parkes[10], Stephen S. Rich[14], Steven Sawcer[3], Jack Satsangi[15], Matthew J. Simmonds[6], Richard C. Trembath[16], Neil M. Walker[7], Eva Wozniak[13], John A. Todd[7], Michael A. Simpson[4], Vincent Plagnol[17] & David A. van Heel[1]

- 25 auto-immune risk genes' coding regions were sequenced on 40,000 individuals
- Rare variants in these genes have negligible contribution to auto-immune disease susceptibility

Negligible impact of rare autoimmune-locus coding-region variants on missing heritability. Hunt K A et al. Nature, 2013

- Summary
  - NGS technology offers an opportunity to discover disease susceptibility rare variants
  - Two major challenges in rare variant association studies:
    - Limited power due to low allele frequency
    - Too many rare variants (most are irrelevant)
  - Some strategies for rare variant association studies:
    - Collapsing
    - Family-based design
    - Variant filtering based on predicted deleteriousness