

# Next-Generation Sequencing Data Analyses

attending

**the 4<sup>th</sup> Short Course on Next-Generating  
Sequencing: Technology and Statistical Methods**

at Birmingham, AL, in December 2014

Robert Yu  
March 2015

From 12/15-18/2014, “**the 4<sup>th</sup> Short Course on NGS: Technology and Statistical Methods**”, a 4-day training course, was held in University of Alabama at Birmingham.

I’m **reintroducing some contents** that were talked and discussed from that 4-day Short Course.

The **intention** is to share the learning, to help us get attention to those ideas and approaches to analyzing sequencing data.

My quotes or directly referring to the content in the course-provided slides are for the convenience in our study and discussion. The **copyrights** belong to the original author(s) and their institutions



# What is this Short Course

- High demands on novel analysis strategies to deal with wealth of NGS data:
  - “**sound information extraction**”
  - “**sophisticated statistical methodologies and algorithms**”
- The National Human Genome Research Institute (NHGRI) sponsors this training course for “**exchanging of cutting-edge information and ideas, and fostering collaborations among methodologists, analysts, and biomedical investigators.**”
- University of Alabama at Birmingham (UAB)’s Biostats Dept has hosted this course for the 4<sup>th</sup> year.
- A 4-day course, running from 12/15 – 18/2014.
- ~4 talks/day; each talk lasted 1 hour plus 15-min discussion.



# The 4<sup>th</sup> Short Course at UAB





- 52 attended

	Female	Male	Total	
BS	4	1	5	9.6%
MS		4	4	7.7%
MD	2	5	7	13.5%
MD,PhD	1	1	2	69.2%
PharmD	2		2	
PhD	11	20	31	
PhD,MPH		1	1	
Total	20	32	52	

# Topics that this course has covered

1. Technical Overview of NGS (Platforms; chemistry; library construction) Shawn Levy, PhD – HudsonAlpha Institute of Biotechnology
2. Next Generation Sequencing in Methylation Studies Devin Absher, PhD - HudsonAlpha
3. Predicting Causal Variation Greg Cooper, PhD - HudsonAlpha  
StatGenLab, a virtual machine for genetic data analysis  
Jelai Wang, BSc – Wang Scientific Software Solutions, LLC
4. Software Demonstration & Hands-on Computing (R & Bioinformatics File Formats) Xiangqin Cui, PhD – UAB
5. Functional Genomics: Identifying and characterizing cis-regulatory sequences Dan Savic, PhD – HudsonAlpha
6. Variant Calling & Assembly NGS Data Degui Zhi, PhD - UAB
7. Rare Variants Analysis Michael Wu, PhD - Fred Hutchinson Cancer Research Center
8. Software Demonstration & Hands-on Computing (SKAT & NGS, Variant Calling) Michael Wu, PhD; Vinodh Srini, MS - UAB
9. Transcriptome Analysis Using NGS Xiangqin Cui, PhD - UAB
10. ChIP-Seq Data Analysis Hao Wu, PhD - Emory University
11. Analysis of whole-genome bisulfite sequencing data Karen Conneely, PhD - Emory University
12. Software Demonstration & Hands-on Computing (ChIP-Seq, BSmooth and DSS) Hao Wu, PhD; Karen Conneely, PhD
13. Computational Methods for Cancer Genome Interpretation Emidio Capriotti, PhD – UAB
14. Statistical Methods for NGS Data Nicholas Schork, PhD - J. Craig Venter Institute
15. 1000 Genomes & Beyond Fuli Yu, PhD - Baylor College of Medicine
16. Software Demonstration & Hands-on Computing (RNA-Seq) Degui Zhi, PhD

# Topic Listing of the Course

	<ol style="list-style-type: none"> <li>1. Technical Overview of NGS (Platforms; chemistry; library construction) <b>Shawn Levy, PhD – HudsonAlpha Institute of Biotechnology</b></li> <li>2. Next Generation Sequencing in Methylation Studies <b>Devin Absher, PhD - HudsonAlpha</b></li> <li>3. Predicting Causal Variation <b>Greg Cooper, PhD - HudsonAlpha</b>  <b>StatGenLab, a virtual machine for genetic data analysis</b>  <b>Jelai Wang, BSc – Wang Scientific Software Solutions, LLC</b></li> <li>4. Software Demonstration &amp; Hands-on Computing (R &amp; Bioinformatics File Formats) <b>Xiangqin Cui, PhD – UAB</b></li> </ol>	<b>Day 1</b>
	<ol style="list-style-type: none"> <li>5. Functional Genomics: Identifying and characterizing cis-regulatory sequences <b>Dan Savic, PhD – HudsonAlpha</b></li> <li>6. Variant Calling &amp; Assembly NGS Data <b>Degui Zhi, PhD - UAB</b></li> <li>7. Rare Variants Analysis <b>Michael Wu, PhD - Fred Hutchinson Cancer Research Center</b></li> <li>8. Software Demonstration &amp; Hands-on Computing (SKAT &amp; NGS, Variant Calling) <b>Michael Wu, PhD; Vinodh Srini, MS - UAB</b></li> </ol>	<b>Day 2</b>
	<ol style="list-style-type: none"> <li>9. Transcriptome Analysis Using NGS <b>Xiangqin Cui, PhD - UAB</b></li> <li>10. ChIP-Seq Data Analysis <b>Hao Wu, PhD - Emory University</b></li> <li>11. Analysis of whole-genome bisulfite sequencing data <b>Karen Conneely, PhD - Emory University</b></li> <li>12. Software Demonstration &amp; Hands-on Computing (ChIP-Seq, BSmooth and DSS) <b>Hao Wu, PhD; Karen Conneely, PhD</b></li> </ol>	<b>Day 3</b>
	<ol style="list-style-type: none"> <li>13. Computational Methods for Cancer Genome Interpretation <b>Emidio Capriotti, PhD – UAB</b></li> <li>14. Statistical Methods for NGS Data <b>Nicholas Schork, PhD - J. Craig Venter Institute</b></li> <li>15. 1000 Genomes &amp; Beyond <b>Fuli Yu, PhD - Baylor College of Medicine</b></li> <li>16. Software Demonstration &amp; Hands-on Computing (RNA-Seq) <b>Degui Zhi, PhD</b></li> </ol>	<b>Day 4</b>

# The topics that I'd like to reintroduce

1. Technical Overview of NGS (Platforms; chemistry; library construction) Shawn Levy, PhD – HudsonAlpha Institute of Biotechnology
2. Next Generation Sequencing in Methylation Studies Devin Absher, PhD - HudsonAlpha
3. Predicting Causal Variation Greg Cooper, PhD - HudsonAlpha  
StatGenLab, a virtual machine for genetic data analysis  
Jelai Wang, BSc – Wang Scientific Software Solutions, LLC
4. Software Demonstration & Hands-on Computing (R & Bioinformatics File Formats) Xiangqin Cui, PhD – UAB
5. Functional Genomics: Identifying and characterizing cis-regulatory sequences Dan Savic, PhD – HudsonAlpha
6. Variant Calling & Assembly NGS Data Degui Zhi, PhD - UAB
7. Rare Variants Analysis Michael Wu, PhD - Fred Hutchinson Cancer Research Center
8. Software Demonstration & Hands-on Computing (SKAT & NGS, Variant Calling) Michael Wu, PhD; Vinodh Srini, MS - UAB
9. Transcriptome Analysis Using NGS Xiangqin Cui, PhD - UAB
10. ChIP-Seq Data Analysis Hao Wu, PhD - Emory University
11. Analysis of whole-genome bisulfite sequencing data Karen Conneely, PhD - Emory University
12. Software Demonstration & Hands-on Computing (ChIP-Seq, BSmooth and DSS) Hao Wu, PhD; Karen Conneely, PhD
13. Computational Methods for Cancer Genome Interpretation Emidio Capriotti, PhD – UAB
14. Statistical Methods for NGS Data Nicholas Schork, PhD - J. Craig Venter Institute
15. 1000 Genomes & Beyond Fuli Yu, PhD - Baylor College of Medicine
16. Software Demonstration & Hands-on Computing (RNA-Seq) Degui Zhi, PhD

part 2



part 1



# First 2 topics.

Methodology overview

## Statistical Methods for Next Generation Sequencing Data

Nicholas J. Schork, Ph.D.  
J. Craig Venter Institute, La Jolla, CA &  
The University of California, San Diego, La Jolla, CA

1. Background: The limits of the contemporary GWAS
2. Analysis of rare variants in sequencing studies
3. Predicting the functional effect of variants
4. Population genetic analysis of rare variants
5. The human 'diplome' and the need to phase
6. 'Filtering' strategies for identifying causal variants

J. Craig Venter™  
I N S T I T U T E

## Rare Variant Analysis

Michael C. Wu

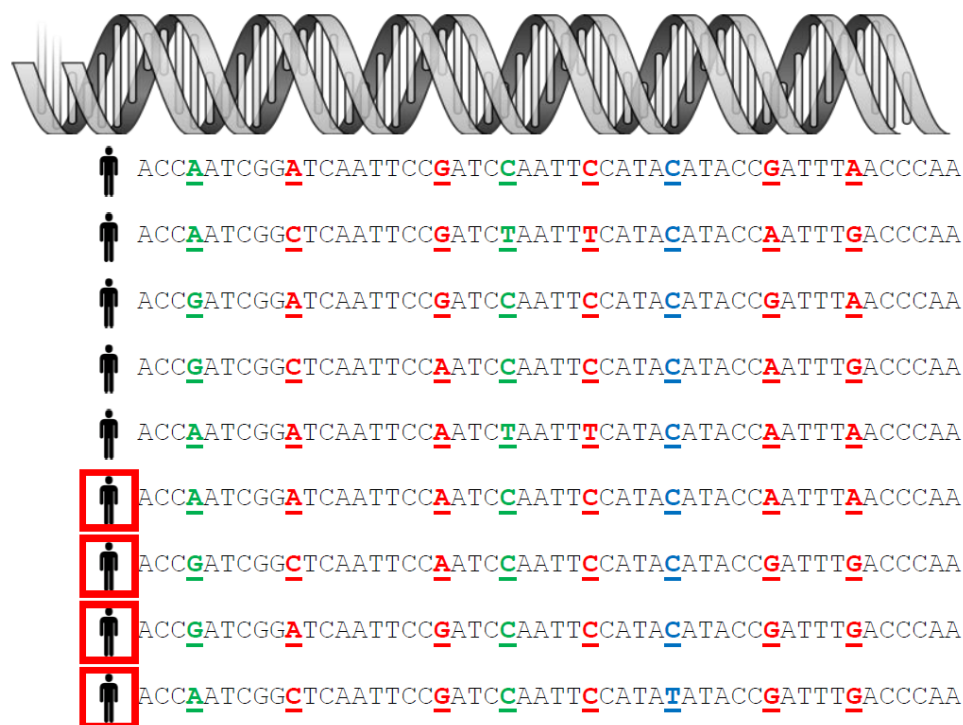
Division of Public Health Sciences  
Fred Hutchinson Cancer Research Center

<http://research.fhcrc.org/wu/>

Specific technique

# An Overview of the Big Picture

## DNA

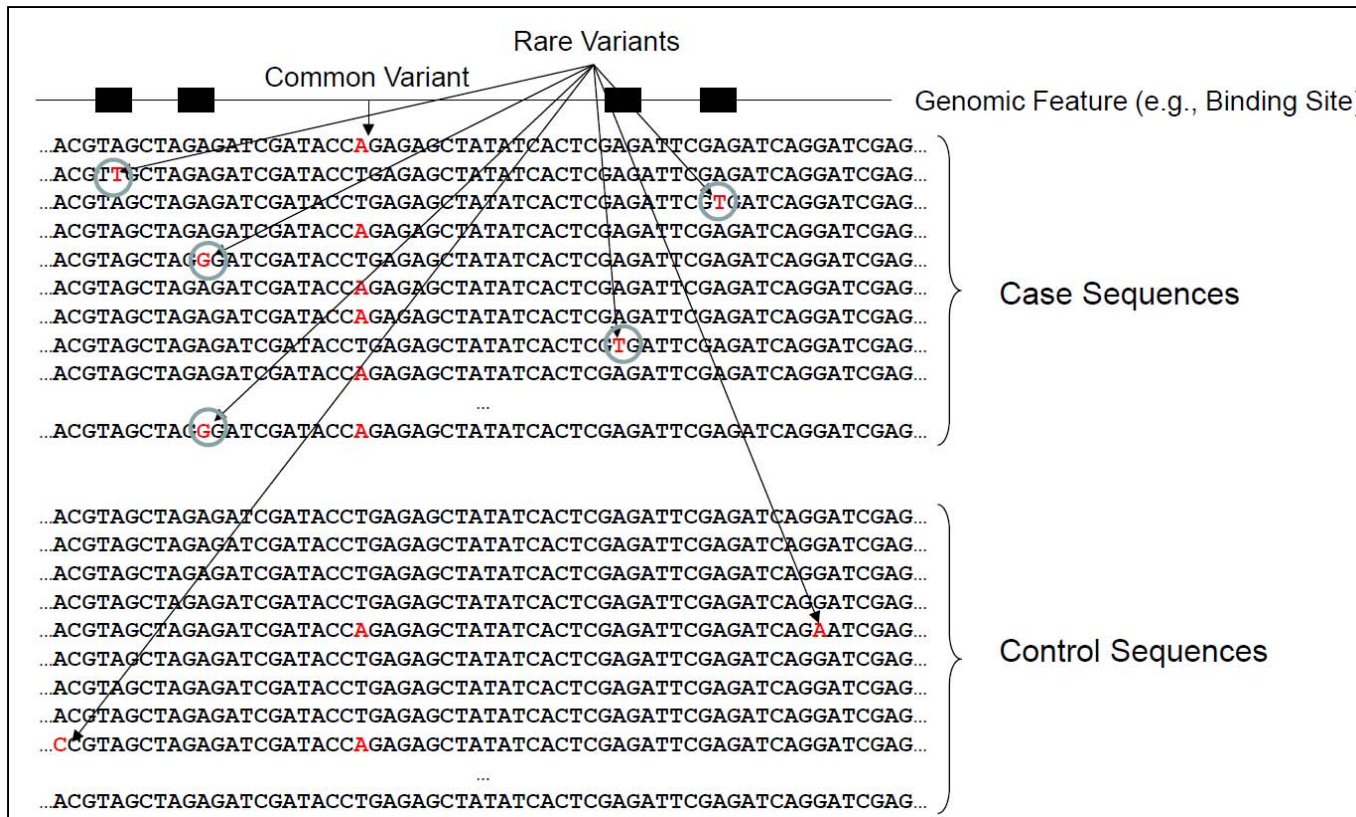


- Colored are SNP/variant
- Green: low density
- Red: high density+imputed
- Blue: picked by seq'd

$$freq(Allele_1) = \frac{count(Allele_1)}{count(Allele_1) + count(Allele_2)}$$



# MAF and rare variant



- Allelic frequency and MAF
- Common variants and rare variants
  - Detection of disease association with common variants (MAF  $\geq 5\%$ ), OR  $\sim 1.2 - 1.4$ , a small fraction (e.g. 4 – 10%) of diseases.
  - How to detect rare variants?
    - Statistical methods
    - Whole genome scan or region selection
    - Biological function and region identification
    - Causal variant identification
    - Etc.

# Referring to the original slides (part 1)

## **Statistical Methods for Next Generation Sequencing Data**

**Nicholas J. Schork, Ph.D.**

J. Craig Venter Institute, La Jolla, CA &  
The University of California, San Diego, La Jolla, CA

1. Background: The limits of the contemporary GWAS
2. Analysis of rare variants in sequencing studies
3. Predicting the functional effect of variants
4. Population genetic analysis of rare variants
5. The human 'diplome' and the need to phase
6. 'Filtering' strategies for identifying causal variants

**J. Craig Venter**  
I N S T I T U T E

Total 55 pages.

Copy of the slides made available to the  
pertinent audience.

In **the data**, we have  $N$  samples with  $p$  phenotypic variables for each, e.g. gene expression data.  $\rightarrow$  a  $\mathbf{Y}$  matrix

We have  $m$  genetic markers for each sample.  $\rightarrow$  an  $\mathbf{X}$  matrix

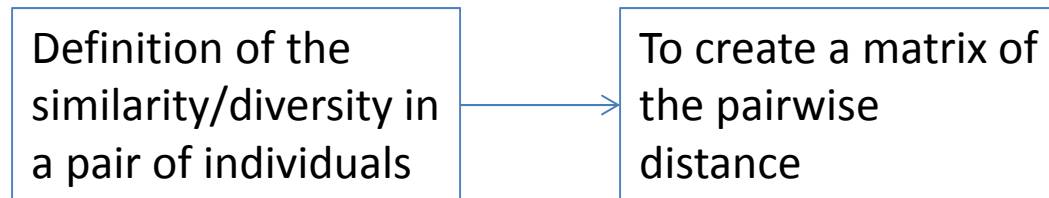
$$\mathbf{Y} = \begin{bmatrix} y_{00} & \cdots & y_{0p} \\ \vdots & \ddots & \vdots \\ y_{N0} & \cdots & y_{Np} \end{bmatrix}, \mathbf{X} = \begin{bmatrix} x_{00} & \cdots & x_{0m} \\ \vdots & \ddots & \vdots \\ x_{N0} & \cdots & x_{Nm} \end{bmatrix}$$

The  $y$ 's could be different phenotype data, and the  $x$ 's could be DNA data and/or covariates.

**The goal** is to 1) define the pairwise **similarity** or **diversity** within  $\mathbf{Y}$ , and the pairwise **similarity** or **diversity** within  $\mathbf{X}$ , and 2) seek correlation between  $\mathbf{Y}$  and  $\mathbf{X}$ , i.e.

$$\mathbf{Y} = \beta\mathbf{X} + \varepsilon$$

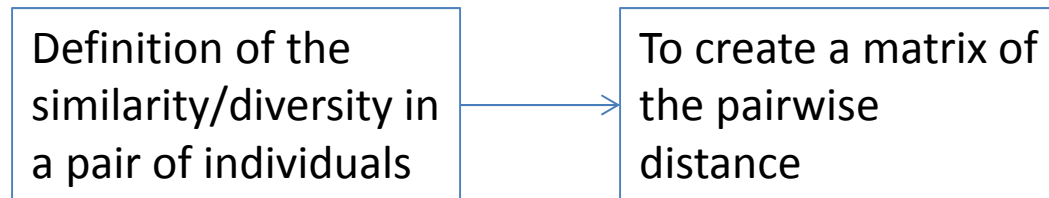
**The goal** is to 1) define the pairwise similarity or diversity within  $Y$ , and the pairwise similarity or diversity within  $X$ , and 2) seek correlation between  $Y$  and  $X$



**A set of methods** *measuring and comparing diversity, similarity or (genetic) distance between/among individuals*

- Set Method (by Hoh and Ott 2003)
- Diversity Method (by Jost 2007)
- Distance Dispersion (by Anderson 2006)
- AMOVA – analysis of molecular variance (by Excoffier, 1992)
- GAMOVA – a generalized AMOVA (by Schork, 2007)
  - MDMR (multivariate distance matrix regression)
  - *Genetic distance-based*
- etc.

**The goal** is to 1) define the pairwise similarity or diversity within  $Y$ , and the pairwise similarity or diversity within  $X$ , and 2) seek correlation between  $Y$  and  $X$



**A set of methods** *measuring and comparing diversity, similarity or (genetic) distance between/among individuals*

- Set Method (by Hoh and Ott 2003)
- ➔ • Diversity Method (by Jost 2007)
- ➔ • Distance Dispersion (by Anderson 2006)
- ➔ • AMOVA – analysis of molecular variance (by Excoffier, 1992)
- ➔ • GAMOVA – a generalized AMOVA (by Schork, 2007)
  - MDMR (multivariate distance matrix regression)
  - *Genetic distance-based*
- etc.

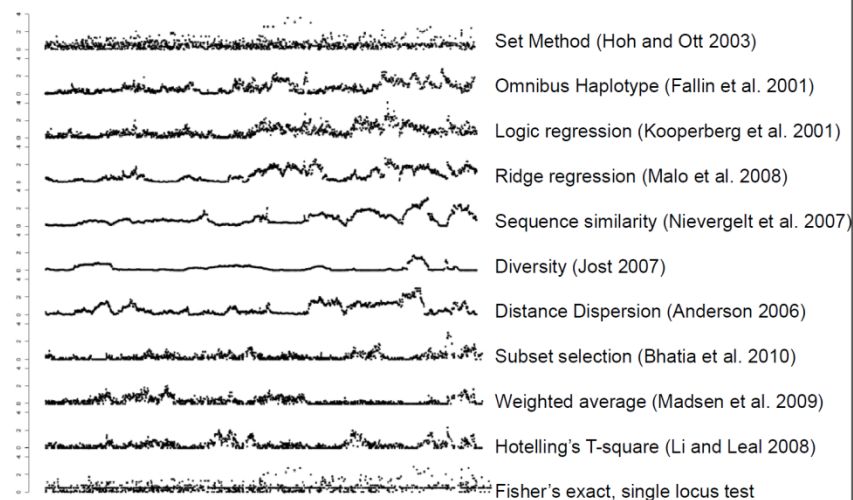
# N. Schork's slides: Example Application

## Sanofi/Scripps Study: Gene Sequence Variation and Obesity

- 298 Individuals (148 morbidly obese; 150 controls)
- Two endocannabinoid genes sequenced using Illumina GA (FAAH; MGLL)
- Standard assembly for SNP identification (60x coverage; 3 reads per variant)
- 242 variants identified in FAAH (many novel and rare): 31 kb of sequence
- 1232 variants identified in MGLL (many novel and rare): 157 kb of sequence
- FAAH: located on chromosome 1p33, known to hydrolyze anandamide (AEA), and other fatty acid amides
- MGLL: located on chromosome 3q21.3, a presynaptic enzyme that hydrolyzes 2-arachidonoylglycerol (2-AG), the most abundant endocannabinoid found in the brain

Harismendy et al. Genome Biol. 2010 Nov 30;11(11):R118. PMID: 21118518  
Bansal et al. Pac Symp Biocomput. 2011:76-87. PMID: 21121035

## Different Methods Applied to the MGLL Gene



Bansal et al. PSB 2011

## Genomic Features with Collapsed Variations

**Table 2.** P-values for association for each analysis method for specific sets of collapsed variations in the MGLL Gene

	NS	H3K27	FAAH TFBS	FOX2	Amidase
# of variants	5	29	4	14	5
Dispersion (Dis)	0.59	0.05	0.77	0.99	0.61
Diversity (Div)	0.43	0.42	0.81	0.33	0.46
MDMR Similarity (Sim)	0.19	0.21	0.05	0.14	0.41
Li & Leal (LL)	0.60	0.03	0.60	1.00	0.50
Subset Selection (SS)	1.00	0.01	0.60	0.75	0.60
Madsen & Browning (MB)	1.00	0.01	0.33	1.00	0.75
Logic Regression (LR)	0.23	0.18	0.39	0.22	0.48
Ridge Regression (RR)	0.35	0.09	0.06	0.33	0.54
PLINK Haplotype (Phap)	NA	0.92	NA	0.34	0.61
PLINK Set Analysis (Pset)	1.00	1.00	0.02	1.00	1.00
	NS	H3K27	MGLL TFBS	FOX2	Amidase
# of variants	9	100	11	3	0
Dispersion	0.28	0.99	0.02	0.72	NA
Diversity	0.77	0.65	0.73	0.64	NA
MDMR	0.81	0.07	0.67	0.29	NA
Li & Leal	1.00	1.00	1.00	0.75	NA
SubsetSelection	0.60	0.43	1.00	1.00	NA
Madsen & Browning	0.75	0.30	0.02	0.20	NA
Logic Regression	0.35	0.67	0.02	0.49	NA
Ridge Reg.	0.71	0.50	0.01	0.61	NA
PLINK Haplotype	NA	0.81	0.07	NA	NA
PLINK Set Analysis	1.00	0.43	0.05	1.00	NA

Different Procedures



# N. Schork's slides: Diversity Methods

## Diversity Methods: Summary Measures vs. Comparing Individual Sequences

Molecular Ecology (2008) 17, 4015–4026

doi: 10.1111/j.1365-294X.2008.03887.x

BIOMETRICS 62, 245–253  
March 2006

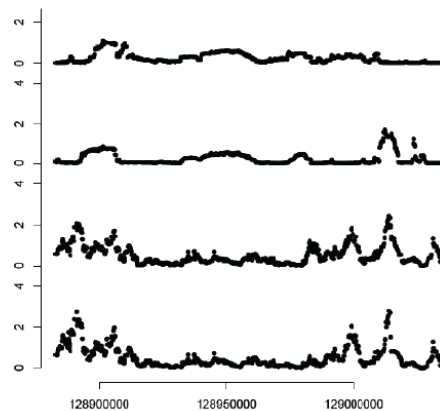
DOI: 10.1111/j.1541-0420.2005.00440.x

$G_{ST}$  and its relatives do not measure differentiation

LOU JOST  
Via Runtun, Baños, Tungurahua, Ecuador

$$\Delta = \left( \sum_{i=1}^k p_i^\lambda \right)^{(1/(1-\lambda))}$$

**Figure B.2.** Window-based association analysis for the MGLL gene assuming a diversity statistic with different exponents based on the work of Jost (2007). The  $\lambda$  values used to construct the graphs are, from the bottom panel to the top panel: 0.2, 0.5, 2.0, and 4.0.

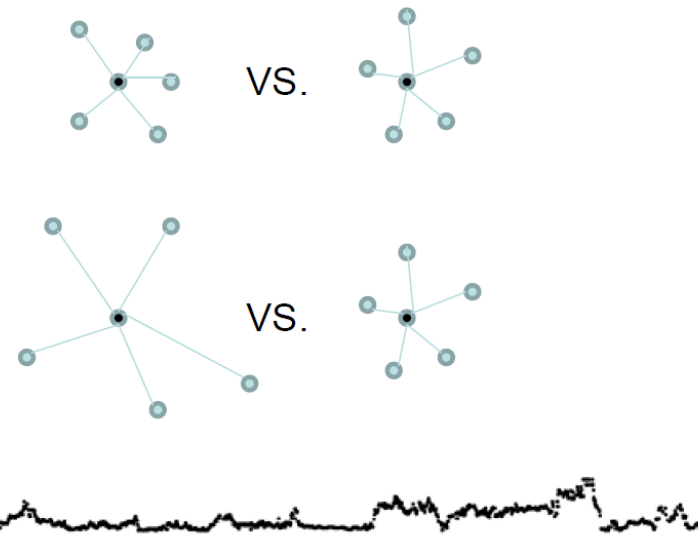


Summary Measure Approach

Distance-Based Tests for Homogeneity of Multivariate Dispersions

Marti J. Anderson

Department of Statistics, University of Auckland, Private Bag 92019, Auckland, New Zealand  
email: mja@stat.auckland.ac.nz



Sequence Diversity/Similarity Measure Approach

# N. Schork's slides: Distance-based Method

## Distance-Based Sequence Analysis for Associations: Simple Nucleotide-Level Identity-By-State Similarity Matrix

9

### DNA Sequence-Based Phenotypic Association Analysis

Nicholas J. Schork,<sup>\*,1,2,3,4,5</sup> Jennifer Wessel,<sup>\*,1,4,5</sup> and  
Nathalie Maito<sup>\*,4,5</sup>

*Advances in Genetics, Vol. 60*

#### 9. DNA Sequence Associations

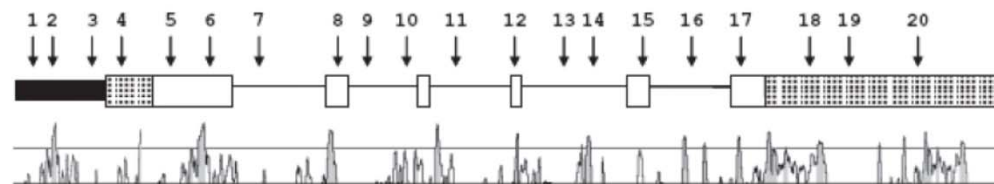
199

**Table 9.1.** Studies Suggesting That Multiple, Potential Interacting Variants Within a Gene or Specified Genomic Region Influence Phenotypic Expression

Gene	In vitro?	Phenotype	References
ADRB2	Yes	Bronchodilator response	Drysdale <i>et al.</i> (2000)
DRD4	No	Schizophrenia	Nakajima <i>et al.</i> (2007)
NRG1	No <sup>a</sup>	Schizophrenia and NRG1 mRNA levels	Law <i>et al.</i> (2006)
HTR2A	Yes	HTR2A gene expression	Myers <i>et al.</i> (2007)
ENT1	Yes	ENT1 gene expression	Myers <i>et al.</i> (2006)
CDA	Yes	CDA gene expression	Fitzgerald <i>et al.</i> (2006)
PCSK9	No	Lipoprotein levels	Kotowski <i>et al.</i> (2006)
NPC1L1	No	Lipoprotein levels	Cohen <i>et al.</i> (2006)
KRT1	Yes	KRT1 gene expression	Tao <i>et al.</i> (2006)
GH1	Yes	GH1 gene expression/ adult height	Horan <i>et al.</i> (2003)
DAT1 (SLC6A3)	Yes	DAT1 gene expression	Greenwood and Kelsoe (2003)
APOE	No	Lipid levels	Stengard <i>et al.</i> (2002)
SLC6A3	Yes	Parkinson's disease	Kelada <i>et al.</i> (2005)
CHGA	Yes	Catecholamine physiology	Wen <i>et al.</i> (2004)

<sup>a</sup>Note that the study of the NRG1 gene involved computational assessments of the functionality of gene variations rather than *in vitro* studies or just association studies.

### Sequence Diversity/Similarity Measure Approach



```

A . T . . C . . T . . G . . - . . C . . . . . T . T . . A . . - - - . . G . . G . . T . . G C . . T . . A . . - - - . . C . . G C T . . . . . C1
A . C . . C . . T . . G . . A . . C . . . . . T . G . . A . . A C T . . . . C . . G . . T . . G C . . C . . A . . - - - . . G . . G C T C G T . . . C2
A . T . . C . . T . . G . . - . . C . . . . . C . G . . A . . A C T . . . . G . . A . . T . . - C . . C . . G . . - - - . . C . . G C T . . . . . C3

G . C . . C . . G . . A . . - . . C . . . . . C . T . . G . . - - - . . C . . A . . A . . - - - . . T . . A . . - - - . . C . . G C T C G T C G T . . D1
A . C . . A . . G . . G . . A . . T . . . . . T . T . . G . . A C T . . . . G . . G . . A . . - C . . T . . G . . A A A . . C . . G C T C G T C G T . . D2
G . C . . A . . T . . G . . A . . C . . . . . C . T . . G . . - - - . . C . . G . . T . . - C . . T . . G . . A A A . . C . . G C T C G T . . . D3
    
```

Pan W. Relationship between genomic distance-based regression and kernel machine regression for multi-marker association testing. *Genet Epidemiol.* 2011 [Epub ahead of print]; PMID:21308765

## Relating Variation in Similarity to Outcomes: MDMR/GAMOVA

$$F = \frac{\text{tr}(\mathbf{HGH})/(M-1)}{\text{tr}[(\mathbf{I}-\mathbf{H})\mathbf{G}(\mathbf{I}-\mathbf{H})]/(N-M)}, \quad [4]$$

No *a priori* clustering or data reduction: test of predictors and variation in matrix

# Some review of population genetics (1)

- Fixation Index  $F_{st}$ , **Wright's F-statistics** – *a measure of population differentiation, usually using genetic polymorphism data, e.g. SNP, microsatellite markers. This method is good for 2-allele locus.*

If  $\bar{p}$  is the average frequency of an allele in the total population,  $\sigma_S^2$  is the variance in the frequency of the allele between different subpopulations, weighted by the sizes of the subpopulations, and  $\sigma_T^2$  is the variance of the allelic state in the total population,  $F_{ST}$  is defined as <sup>[1]</sup>

$$F_{ST} = \frac{\sigma_S^2}{\sigma_T^2} = \frac{\sigma_S^2}{\bar{p}(1 - \bar{p})}$$

## Estimation

---

In practice, none of the quantities used for the definitions can be easily measured. As a consequence, various estimators have been proposed. A particularly simple estimator applicable to DNA sequence data is:

$$F_{ST} = \frac{\pi_{\text{Between}} - \pi_{\text{Within}}}{\pi_{\text{Between}}}$$

where  $\pi_{\text{Between}}$  and  $\pi_{\text{Within}}$  represent the [average number of pairwise differences](#) between two individuals sampled from different sub-populations ( $\pi_{\text{Between}}$ ) or from the same sub-population ( $\pi_{\text{Within}}$ ). The average pairwise difference within a population can be calculated as the sum of the pairwise differences divided by the number of pairs. However, this estimator is biased when sample sizes are small or if they vary between populations. Therefore, more elaborate methods are used to compute  $F_{ST}$  in practice. Two of the most widely used procedures are the estimator by Weir & Cockerham (1984),<sup>[4]</sup> or performing an [Analysis of molecular variance](#). A list of implementations is available at the end of this article.



# Some review of population genetics (1)

- Fixation Index  $F_{st}$ , **Wright's F-statistics** – *a measure of population differentiation, usually using genetic polymorphism data, e.g. SNP, microsatellite markers. This method is good for 2-allele locus.*

If  $\bar{p}$  is the average frequency of an allele in the total population,  $\sigma_S^2$  is the variance in the frequency of the allele between different subpopulations, weighted by the sizes of the subpopulations, and  $\sigma_T^2$  is the variance of the allelic state in the total population,  $F_{ST}$  is defined as <sup>[1]</sup>

$$F_{ST} = \frac{\sigma_S^2}{\sigma_T^2} = \frac{\sigma_S^2}{\bar{p}(1 - \bar{p})}$$

## Estimation

In practice, none of the quantities used for the definitions can be easily measured. As a consequence, various estimators have been proposed. A particularly simple estimator applicable to DNA sequence data is:

$$F_{ST} = \frac{\pi_{\text{Between}} - \pi_{\text{Within}}}{\pi_{\text{Between}}}$$

where  $\pi_{\text{Between}}$  and  $\pi_{\text{Within}}$  represent the average number of pairwise differences between two individuals sampled from different sub-populations ( $\pi_{\text{Between}}$ ) or from the same sub-population ( $\pi_{\text{Within}}$ ). The average pairwise difference within a population can be calculated as the sum of the pairwise differences divided by the number of pairs. However, this estimator is biased when sample sizes are small or if they vary between populations. Therefore, more elaborate methods are used to compute  $F_{ST}$  in practice.

Two of the most widely used procedures are the estimator by Weir & Cockerham (1984),<sup>[4]</sup> or performing an Analysis of molecular variance. A list of implementations is available at the end of this article.

# Some review of population genetics (2)

## **Analysis of Gene Diversity in Subdivided Populations**

**(population structure/ genetic variability/heterozygosity/gene differentiation)**

MASATOSHI NEI

Center for Demographic and Population Genetics, University of Texas at Houston, Tex. 77025

*Communicated by Sewall Wright, August 6, 1973*

- **$G_{st}$ , Nei's G-statistics** – *a measure of gene diversity (or population differentiation), an extension of  $F_{st}$ , but using weighted average of  $F_{st}$  for all alleles in the case of multiallelic conditions.*

$$G_{ST} = D_{TS} / H_T = (H_T - H_S) / H_T = 1 - H_S / H_T.$$

where  $D_{TS}$  - between-subpopulation diversity,  $H_T$  – heterozygosities of total population, and  $H_S$  – heterozygosities of subpopulations.



# Diversity Method (1)

## $G_{ST}$ and its relatives do not measure differentiation

LOU JOST

*Via Runtun, Baños, Tungurahua, Ecuador*

### Abstract

$G_{ST}$  and its relatives are often interpreted as measures of differentiation between subpopulations, with values near zero supposedly indicating low differentiation. However,  $G_{ST}$  necessarily approaches zero when gene diversity is high, even if subpopulations are completely differentiated, and it is not monotonic with increasing differentiation. Likewise, when diversity is equated with heterozygosity, standard similarity measures formed by taking the ratio of mean within-subpopulation diversity to total diversity necessarily approach unity when diversity is high, even if the subpopulations are completely dissimilar (no shared alleles). None of these measures can be interpreted as measures of differentiation or similarity. The derivations of these measures contain two subtle misconceptions which cause their paradoxical behaviours. Conclusions about population differentiation, gene flow, relatedness, and conservation priority will often be wrong when based on these fixation indices or similarity measures. These are not statistical issues; the problems persist even when true population frequencies are used in the calculations. Recent advances in the mathematics of diversity identify the misconceptions, and yield mathematically consistent descriptive measures of population structure which eliminate the paradoxes produced by standard measures. These measures can be directly related to the migration and mutation rates of the finite-island model.

# Diversity Method (2)

## $G_{ST}$ and its relatives do not measure differentiation

LOU JOST

*Via Runtun, Baños, Tungurahua, Ecuador*

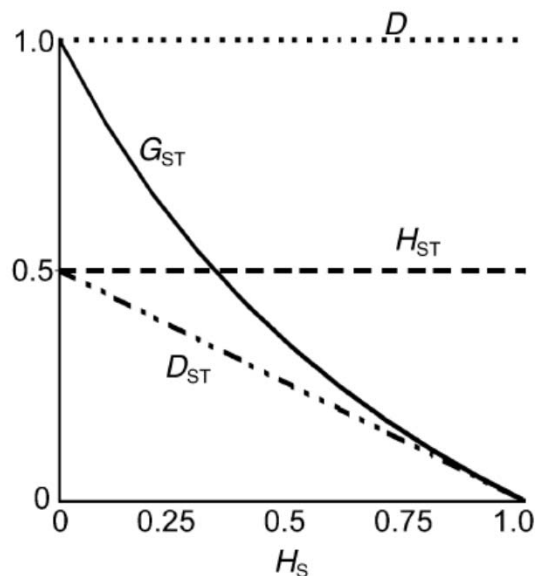


Fig. 1 Differentiation measures applied to two completely differentiated subpopulations (no shared alleles) for various values of mean within-subpopulation heterozygosity.  $G_{ST}$  and  $D_{ST}$  approach zero when mean heterozygosity is high, even though differentiation is 100% for these subpopulations. New measures  $D$  and  $H_{ST}$  correctly reflect differentiation regardless of the value of mean heterozygosity. All values are calculated using actual population frequencies, not sample frequencies; this is not a sampling issue.

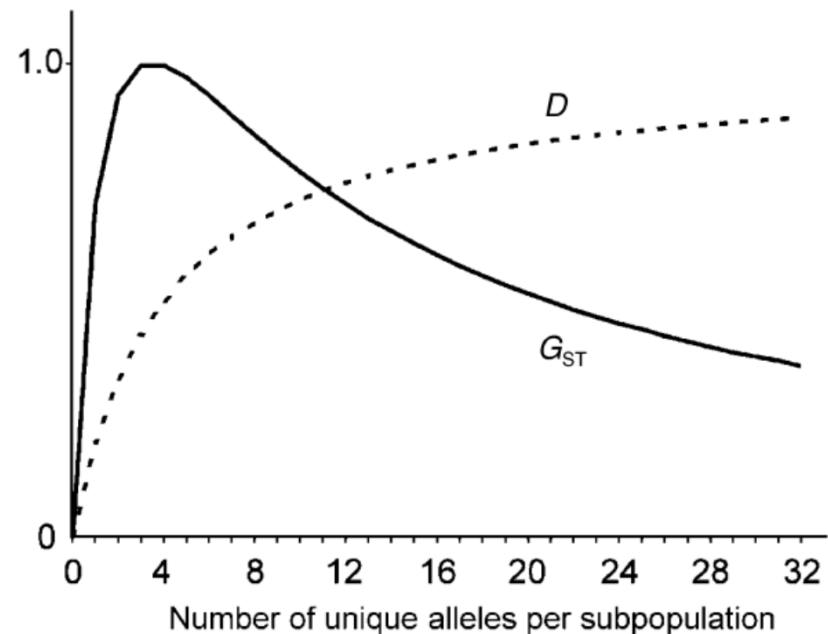


Fig. 2 Behaviour of  $G_{ST}$  and  $D$  as differentiation increases. We start with two identical subpopulations (four equally common alleles, 1000 individuals per allele per subpopulation). We then successively add unique alleles to each subpopulation (1000 individuals per allele) and graph  $G_{ST}$  and  $D$  (the measure of differentiation defined in the text).  $G_{ST}$  is normalized by dividing by its maximum value (0.0345). Even though differentiation increases steadily from left to right,  $G_{ST}$  reaches its maximum and then falls back to zero.  $G_{ST}$  is calculated from exact population allele frequencies, so this is not a sampling issue.

# Diversity Method (3)

## $G_{ST}$ and its relatives do not measure differentiation

LOU JOST

*Via Runtun, Baños, Tungurahua, Ecuador*

Why not derive new mathematically self-consistent descriptive measures of diversity and differentiation that really behave the way that geneticists thought their traditional measures behaved? The first step in such a program is to find a measure of genetic diversity that behaves correctly in common ratio comparisons and conservation genetics problems such as those just mentioned. We can then derive a formula to partition this diversity into truly independent within- and between-subpopulation components. The resulting pure between-subpopulation component can then be transformed into a meaningful, logically and mathematically consistent measure of relative differentiation to replace  $G_{ST}$ .

$$\text{Diversity } \Delta \equiv \left( \sum_{i=1}^k p_i^q \right)^{1/(1-q)} \quad (\text{eqn 4})$$

where  $p_i$  is the population frequency of the  $i$ -th allele and the exponent  $q$  determines the measure's sensitivity to allele frequencies. When  $q = 0$ , eqn 4 gives the allele number. When  $q$  approaches unity, eqn 4 gives (via calculus) the exponential of Shannon entropy

### *Partitioning true diversity*

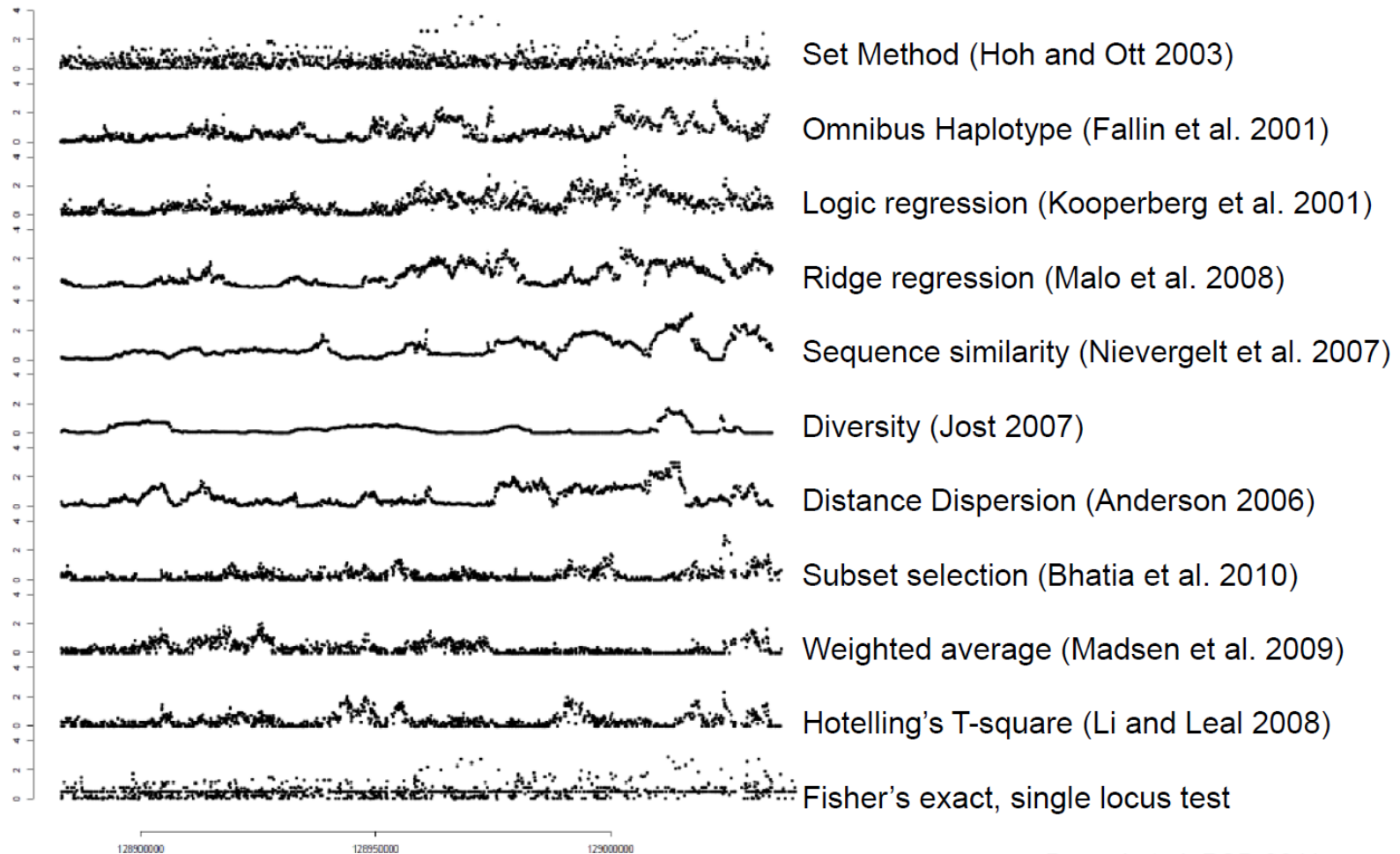
It can be proven (Jost 2007) that the decomposition of any true diversity  $\Delta_T$  into pure within- and between-subpopulation components ( $\Delta_S$  and  $\Delta_{ST}$ , respectively) must be multiplicative: the effective number of alleles in the pooled subpopulations ( $\Delta_T$ ) must equal the effective number of alleles per subpopulation times the effective number of distinct subpopulations:

$$\Delta_T = \Delta_S \cdot \Delta_{ST} \quad (\text{eqn 7})$$



# N. Schork's Slide of Results

## Different Methods Applied to the MGLL Gene



# MDMR/GAMOVA (1)

In **the data**, we have  $N$  samples with  $p$  phenotypic variables for each, e.g. gene expression data.  $\rightarrow$  a  $\mathbf{Y}$  matrix

We have  $m$  genetic markers for each sample.  $\rightarrow$  an  $\mathbf{X}$  matrix

$$\mathbf{Y} = \begin{bmatrix} y_{00} & \cdots & y_{0p} \\ \vdots & \ddots & \vdots \\ y_{N0} & \cdots & y_{Np} \end{bmatrix}, \mathbf{X} = \begin{bmatrix} x_{00} & \cdots & x_{0m} \\ \vdots & \ddots & \vdots \\ x_{N0} & \cdots & x_{Nm} \end{bmatrix}$$

The  $y$ 's could be different phenotype data, and the  $x$ 's could be DNA data and/or covariates.

**The goal** is to 1) define the pairwise similarity within  $\mathbf{Y}$ , and the pairwise similarity within  $\mathbf{X}$ , and 2) seek correlation between  $\mathbf{Y}$  and  $\mathbf{X}$ , i.e.

$$\mathbf{Y} = \beta\mathbf{X} + \varepsilon$$

**The goal** is to 1) define the pairwise **similarity** within **Y**, and the pairwise similarity within **X**, and 2) seek correlation between **Y** and **X**

Define the pairwise **similarity** within the matrix.

To create a matrix of **the pairwise distance**

$$S_{i,j}^{\text{IBS}} = \frac{\sum_{l=1}^L s_{i,j}^l (g_i^l, g_j^l)}{2L}$$

$$S_{i,j}^w = \frac{\sum_{l=1}^L w^l s_{i,j}^l (g_i^l, g_j^l)}{\sum_{l=1}^L w^l}$$

The distance could be

- Euclidean distance
- Transformation of  $r_{ij}$ , correlation coefficients.
- IBS allele sharing distance
- IBS allele sharing with weighting
- Weighting by nucleotide conservation across species
- Using ancestry information
- *etc.*



# MDMR/GAMOVA (2)

## Formation of a distance matrix

- Transforming a correlation coefficient,  $r$ , matrix to the distance,  $D$ , matrix

$$d_{ij} = \sqrt{2(1 - r_{ij})}.$$

## A Multivariate Multiple Regression Model

$$Y = X\beta + \varepsilon,$$

where  $Y$  is an  $N \times P$  matrix of phenotype, say expression of  $P$  genes from  $N$  individuals,  $X$  is an  $N \times M$  matrix of  $M$  genetic markers from  $N$  individuals, and  $\beta$  is an  $M \times P$  matrix of regression coefficients and  $\varepsilon$  is an error term.

The least-squares solution for  $\beta$  is  $\tilde{\beta} = (X'X)^{-1}X'Y$ , with the matrix of residual errors being

$$R = Y - \hat{Y} = Y - X\tilde{\beta} = (I - H)Y$$

where  $H = (X'X)^{-1}X'$  and is the traditional “hat” matrix.

# MDMR/GAMOVA (3)

Let  $\mathbf{G}$  be Gower's centered matrix

$$\mathbf{G} = \left( \mathbf{I} - \frac{1}{n} \mathbf{1}\mathbf{1}' \right) \mathbf{A} \left( \mathbf{I} - \frac{1}{n} \mathbf{1}\mathbf{1}' \right)$$

where  $\mathbf{1}$  is a  $N$ -dimensional column vector whose every element is 1 and  $\mathbf{I}$  is an  $N \times N$  identity matrix, and  $\mathbf{A} = (a_{ij}) = (-1/2 d_{ij}^2)$ .

An appropriate  $F$  statistic for assessing the relationship between the  $M$  predictor variables and variation in the dissimilarities among the  $N$  subjects with respect to the  $P$  variables is

$$F = \frac{\text{tr}(\mathbf{HGH}) / (M - 1)}{\text{tr}[(\mathbf{I} - \mathbf{H})\mathbf{G}(\mathbf{I} - \mathbf{H})] / (N - M)},$$

where “tr” stands for “trace”,  $\mathbf{H}$  is a hat matrix (projection matrix,  $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ ),  $\mathbf{G}$  is Gower's centered matrix, and  $F$  is the  $F$  statistics possessing the properties as in the  $F$  in ANOVA.

# MDMR/GAMOVA (4)

## Assessing Statistical Significance

- Distribution of the  $F$  statistic is determined by the particular distance matrix, i.e. different type of distance calculation could result in different  $F$  distribution.
- To generalize the  $F$  statistic, permutation tests to evaluate the probabilistic significance of the observed  $F$  is needed.
- Different predictor variables or subsets of variables can be tested for association with variation in a distance matrix.
- Then step-wise or variable selection procedures can be done as usually done in univariate standard multiple regression analysis.

## Example 1.

# Distance-based phenotypic association analysis of DNA sequence data

Doyoung Chung\*, Qunyuan Zhang, Aldi T Kraja, Ingrid B Borecki, Michael A Province

*From Genetic Analysis Workshop 17*  
Boston, MA, USA. 13-16 October 2010

The GAW17 provided a GWAS simulated data set containing

- 3 continuous phenotypes, Q1, Q2 and Q4
- 697 unrelated individuals, with information on their Age, Sex and Smoke
- 200 replications
- 3,205 autosomal genes with 24,487 SNPs; 3,132 of the SNPs having  $MAF \geq 0.05$

This study chose

- 13 risk genes that is associated with phenotype Q2
- All 697 unrelated individuals and 200 replications
- Rare variant (SNP) using  $MAF < 0.01$
- 508 noncausative genes for control analysis. In rare variant analysis, 125 of the 508 genes were omitted due to their MAF did not meet rare variant criteria.

## Example 1.

# Distance-based phenotypic association analysis of DNA sequence data

Doyoung Chung\*, Qunyuan Zhang, Aldi T Kraja, Ingrid B Borecki, Michael A Province

From Genetic Analysis Workshop 17

**MDMR as a gene-based association test** Boston, MA, USA. 13-16 October 2010

We calculated Euclidean distances using numerically coded genotypes of 13 Q2 risk genes for all possible pairs of the 697 unrelated individuals:

$$d(\mathbf{a}, \mathbf{b}) = \|\mathbf{a} - \mathbf{b}\| = [(\mathbf{a} - \mathbf{b}) \cdot (\mathbf{a} - \mathbf{b})]^{1/2}, \quad (4)$$

where the Euclidean distance is defined as the L2 norm between two individual genotype vectors  $\mathbf{a}$  and  $\mathbf{b}$ . Genotypes were coded as the number of minor alleles with no weighting of single-nucleotide polymorphisms (SNPs) was applied. For each gene and each Q2 simulation, we constructed a  $697 \times 697$  genotypic distance matrix  $\mathbf{D}$  and a  $697 \times 1$  phenotype matrix  $\mathbf{X}$ , which consists of the individual Q2 trait values, and used them to calculate a pseudo- $F$  statistic under the regression model that includes the Q2 trait as the sole independent variable. Each of the  $13 \times 200$  tests underwent 1,000 permutations in which the rows and columns of its raw genotype matrix (i.e., the individual-by-SNP matrix) were shuffled at random. The empirical  $p$ -value was determined as the frequency of observing more extreme pseudo- $F$  statistics in permutations than in the actual gene case. MDMRs were performed either using all variants within a gene or using only rare variants with minor allele frequency (MAF) less than 0.01. Similarly, we selected 508 noncausative (i.e., control) genes for Q2 and tested them using all 200 replications. We omitted a subset containing 125 genes from these 508 control genes for the rare-variant-only analyses because they contained no rare variants.

### Mantel test

The Mantel test measures the correlation between two distance matrices [8]. In our application, we calculated a phenotypic distance matrix and a genotypic distance matrix based on the Euclidean distance measure. The two distance matrices were then tested for correlation [9]. The genotypic distance matrix for the Mantel test was identical with that of the MDMR, whereas a  $697 \times 697$  distance matrix was calculated for each Q2 simulated replicate. Mantel tests were performed for the 13 Q2 risk genes using either all variants or only rare variants. Similarly, 508 control genes were tested for association using all variants, among which 383 genes continued to be tested using only rare variants.  $P$ -values were empirically determined using 1,000 permutations. We estimated the power and false-positive rates on the basis of the significance threshold value of 0.05 and compared them with the values from MDMR and collapsing analysis.

### Collapsing analysis

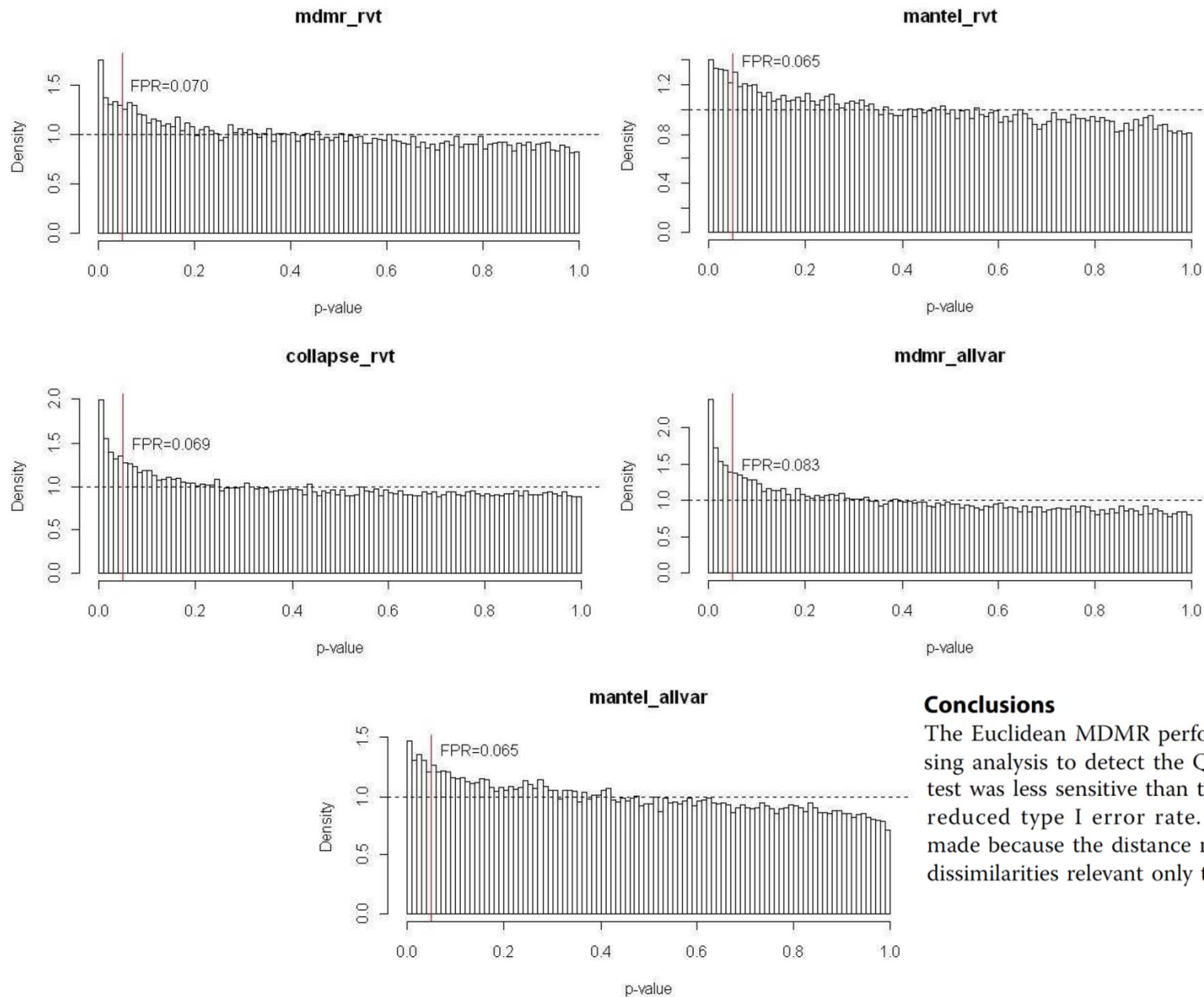
Collapsing analysis is a simple regression analysis that uses a collapsed variable [10] into which rare variants are collapsed in a binary manner based on the presence of any rare variant. Because our collapsing analysis excluded all “common” variants (defined by  $\text{MAF} > 0.01$ ), we also removed common variants in the other analyses to facilitate comparison. This allowed 12 Q2 risk genes to be compared, because one risk gene had no rare variants. Similarly, we tested 380 selected genes, simulated under the null hypothesis for Q2, for association with Q2 using all three methods. No correction for population structure or hidden relatedness was applied throughout this study.

**Table 1 True positive rates of five different strategies for the 13 Q2 risk genes**

Gene	Setting	MDMR using all variants	Mantel test using all variants	MDMR using only rare variants	Mantel test using only rare variants	Collapsing analysis using only rare variants
<i>BCHE</i>	1c + 28r (13s)	0.045	0.170	0.320	0.310	0.455
<i>GCKR</i>	1c (1s)	0.405	0.150	NA	NA	NA
<i>INSIG1</i>	1c + 4r (3s)	0.090	0.020	0.040	0.040	0.035
<i>LPL</i>	5c (1s) + 15r (2s)	0.045	0.135	0.060	0.125	0.040
<i>PDGFD</i>	5c + 6r (4s)	0.065	0.035	0.685	0.290	0.745
<i>PLAT</i>	4c + 25r (8s)	0.035	0.030	0.055	0.040	0.110
<i>RARB</i>	2c + 9r (2s)	0.105	0.145	0.410	0.115	0.155
<i>SIRT1</i>	1c + 23r (9s)	0.365	0.285	0.605	0.320	0.330
<i>SREBF1</i>	3c + 21r (10s)	0.030	0.110	0.380	0.205	0.690
<i>VLDLR</i>	4c + 23r (8s)	0.055	0.065	0.140	0.140	0.140
<i>VNN1</i>	1c (1s) + 6r (1s)	0.940	0.250	0.200	0.085	0.050
<i>VNN3</i>	6c (3s) + 9r (4s)	0.190	0.175	0.025	0.055	0.030
<i>VWF</i>	2c + 6r (2s)	0.180	0.080	0.285	0.080	0.190
Mean		0.196	0.127	0.267	0.150	0.248

The true positive rate was determined as the frequency of observing  $p$ -values less than 0.05 among 200 (replication)  $p$ -values for each gene. The "Setting" column shows the composition of SNPs within a gene: c, r, and s stand for common, rare, and signal SNPs, respectively. For example, *VNN1* has 1 common causal SNP and 6 rare SNPs, one of which is a signal. SNPs with MAF > 0.01 are defined as common.





## Conclusions

The Euclidean MDMR performed comparably to collapsing analysis to detect the Q2 causal genes. The Mantel test was less sensitive than these methods with a slightly reduced type I error rate. Potential progress can be made because the distance matrix appreciates genotypic dissimilarities relevant only to phenotypic dissimilarities.

**Figure 1 False positive rates of five strategies.** mdmr\_rvt, MDMR using only rare variants; mantel\_rvt, Mantel test using only rare variants; collapse\_rvt, collapsing analysis using only rare variants; mdmr\_allvar, MDMR using all variants; mantel\_allvar, Mantel test using all variants. The red vertical lines mark the significance threshold  $p$ -value of 0.05.

# N. Schork's slides: Methods Comparison

## Simulation-based Comparison of Methods

### Comparison of Statistical Tests for Disease Association with Rare Variants

SAONLI BASU, WEI PAN

<http://www.biostat.umn.edu/~weip/paper/RV2.pdf>


- Simulate a wide variety of settings: with LD, with opposite effect variants, with neutral variants, etc.
- Fit a number of different methods 
- The Kernel Machine Regression (KMR) which was shown to be equivalent to GAMOVA/MDMR similarity-based method was one of the most consistently best performers

Table 4: Empirical power for tests at nominal level  $\alpha$  based on 1000 replicates for a non-ideal case for 8 causal RVs with various association strengths  $OR = (3, 3, 2, 2, 2, 1/2, 1/2, 1/2)$  and a number of non-causal RVs. There is no LD among the RVs.

Test	$\alpha = 0.05$					$\alpha = 0.01$				
	# of neutral RVs					# of neutral RVs				
	0	4	8	16	32	0	4	8	16	32
UminP	.607	.532	.481	.417	.346	.318	.259	.227	.204	.142
Score	.869	.772	.721	.632	.483	.660	.532	.480	.356	.233
SSU	.895	<b>.835</b>	<b>.815</b>	<b>.774</b>	<b>.696</b>	.723	.662	.645	.583	.472
wSSU-P	.861	.776	.735	.685	.550	.606	.510	.460	.401	.258
SSUw	.867	.773	.732	.633	.501	.661	.550	.481	.355	.238
Sum	.682	.566	.465	.365	.258	.471	.348	.257	.172	.101
KMR(Linear)	<b>.897</b>	<b>.842</b>	<b>.824</b>	<b>.783</b>	<b>.707</b>	<b>.740</b>	<b>.678</b>	<b>.667</b>	<b>.619</b>	<b>.495</b>
KMR(Quad)	.893	.835	.815	.781	.698	.734	<b>.680</b>	<b>.663</b>	.608	<b>.484</b>
CMC(0.01)	.703	.669	.670	.670	.590	.511	.457	.470	.470	.383
CMC	.661	.544	.456	.336	.204	.461	.337	.235	.157	.086
wSum	.659	.548	.459	.335	.228	.460	.336	.236	.158	.093
aSum-P	.854	.745	.684	.574	.430	.670	.538	.430	.315	.207
Step-up	.839	.767	.724	.640	.527	.652	.564	.518	.413	.285
Seq-aSum	.892	.811	.757	.671	.528	<b>.752</b>	.620	.532	.438	.273
Seq-aSum-VS	.885	.807	.768	.686	.545	.729	.623	.567	.448	.293
KBAC	<b>.907</b>	.813	.763	.642	.436	<b>.737</b>	.607	.536	.399	.199
C-alpha-A	.892	.826	.802	.757	.655	.824	.732	.720	.653	.512
C-alpha-P	<b>.906</b>	<b>.844</b>	<b>.823</b>	<b>.775</b>	<b>.674</b>	.735	<b>.673</b>	<b>.661</b>	<b>.612</b>	<b>.496</b>
RBT	.810	.659	.603	.482	.301	.590	.429	.356	.250	.125

# Methods Comparison with Simulated Data

## Comparison of Statistical Tests for Disease Association with Rare Variants

SAONLI BASU, WEI PAN

*Division of Biostatistics, School of Public Health, University of Minnesota,  
Minneapolis, MN 55455*

November 30, 2010; Revised March 23, 2011

### Simulated data

We generated simulated data as in Wang and Elston (2008) and Pan (2009). Specifically, we simulated  $k$  SNVs with the sample size of 500 cases and 500 controls. Each RV had a mutation rate or MAF uniformly distributed between 0.001 and 0.01, while for a CV it was between 0.01 and 0.1. First, we generated a latent vector  $Z = (Z_1, \dots, Z_k)'$  from a multivariate normal distribution with a first-order auto-regressive (AR1) covariance structure: there was an correlation  $Corr(Z_i, Z_j) = \rho^{|i-j|}$  between

any two latent components. We used  $\rho = 0$  and  $\rho = 0.9$  to generate (neighboring) SNVs in linkage equilibrium and in linkage disequilibrium (LD) respectively. Second, the latent vector was dichotomized to yield a haplotype with MAFs each randomly selected. Third, we combined two independent haplotypes and obtained genotype data  $X_i = (X_{i1}, \dots, X_{ik})'$ . Fourth, the disease status  $Y_i$  of subject  $i$  was generated from the logistic regression model (1). For the null case, we used  $\beta = 0$ ; for non-null cases, we randomly selected 8 non-zero components of  $\beta$  while the remaining ones were all 0. Fifth, as in any case-control design we sampled 500 cases and 500 controls in each dataset.

We considered several simulation set-ups. Throughout the simulations, we fixed the test significance level at  $\alpha = 0.05$  (or  $\alpha = 0.01$  in a few cases), and used 500 permutations for each permutation-based method. The results were based on 1000 independent replicates for each set-up.

We used the R code of Wu et al (2010) implementing the KMR methods. We used the linear, IBS and quadratic kernels; since the first two performed similarly across all simulations, we present results for the linear and quadratic kernels. We used the R package `thgenetics` implementing the Step-up procedure, and a C++/R implementation of KBAC. We implemented all other tests in R. For the CMC test, we used the default cut-off of  $MAF \leq 0.05$  for RVs, though we explored using the cut-off  $\leq 0.01$  in a few cases.

# Comparison of Statistical Tests for Disease Association with Rare Variants

SAONLI BASU, WEI PAN

*Division of Biostatistics, School of Public Health, University of Minnesota,  
Minneapolis, MN 55455*

November 30, 2010; Revised March 23, 2011

Table 1: A summary of the properties of the tests to be compared: originally proposed to target CVs or RVs (or both), whether pooling over variants, whether sensitive to association directions (+/-), to a large number of non-causal RVs (nRVs) and to a few non-causal CVs (nCVs), requiring permutations for p-value calculations, capability to adjust for other covariates (Cov), applicability to other non-binary traits, whether can be formulated as testing on a variance component in a random-effects (R-E) model, and references for more details.

Test	Original target	Pool	Sens to +/-	Sens to nRVs	Sens to nCVs	Permut	Cov	Other traits	R-E	Refs
UminP	CV	No	No	No	No	No	Yes	Yes	No	3
Score	CV	No	No	No	No	No	Yes	Yes	Yes	1
SSU	CV	No	No	No	Yes	No	Yes	Yes	Yes	2
wSSU-P	Both	No	No	No	No	Yes	Yes	Yes	Yes	here
SSUw	CV	No	No	No	No	No	Yes	Yes	Yes	2
Sum	CV	No	Yes	Yes	Yes	No	Yes	Yes	No	2
KMR	CV	No	No	No	Yes	No	Yes	Yes	Yes	4, 5
CMC	RV	Yes	Yes	Yes	No	No	No	No	No	6
wSum	RV	Yes	Yes	Yes	Some	Some	No	No	No	7
aSum-P	Both	Yes	Some	Yes	Some	Yes	Yes	Yes	No	8
Step-up	RV	Yes	Some	Some	No	Yes	Yes	Yes	No	10
Seq-aSum	Both	Yes	Some	Some	Yes	Yes	Yes	Yes	No	here
Seq-aSum-VS	Both	Yes	Some	Some	No	Yes	Yes	Yes	No	here
KBAC	RV	No	Some	Some	Some	Yes	Some	No	No	11
C-alpha-A	RV	No	No	No	Yes	No	No	No	Yes	9
C-alpha-P	RV	No	No	No	Yes	Yes	No	No	Yes	9
RBT	RV	Yes	Some	Yes	No	Yes	No	No	No	12

Refs: 1. Clayton et al (2004); 2. Basu (2006); 3. Clayton & Bank (2007); 4. Kuo et al (2007)

Table 2: Type I error rates at nominal level  $\alpha$  based on 1000 replicates for 8 RVs plus a number of non-causal RVs. There is no LD among the RVs.

Test	$\alpha = 0.05$					$\alpha = 0.01$				
	# of neutral RVs					# of neutral RVs				
	0	4	8	16	32	0	4	8	16	32
UminP	.027	.027	.016	.011	.019	.003	.001	.004	.001	.002
Score	.043	.049	.040	.040	.040	.006	.009	.005	.005	.007
SSU	.044	.055	.045	.037	.043	.004	.013	.009	.005	.011
wSSU-P	.052	.051	.048	.048	.046	.008	.008	.014	.010	.008
SSUw	.041	.049	.039	.034	.040	.006	.011	.005	.005	.007
Sum	.047	.055	.041	.054	.038	.012	.007	.010	.010	.007
KMR(Linear)	.046	.056	.046	.042	.047	.007	.016	.011	.007	.012
KMR(Quad)	.046	.056	.047	.039	.046	.007	.016	.010	.006	.011
CMC(0.01)	.035	.053	.044	.055	.039	.008	.014	.010	.011	.009
CMC	.048	.053	.043	.056	.051	.010	.009	.011	.011	.007
wSum	.050	.057	.038	.059	.056	.010	.012	.011	.009	.006
aSum-P	.058	.064	.052	.063	.047	.012	.011	.010	.010	.011
Step-up	.046	.059	.056	.051	.051	.012	.011	.009	.009	.010
Seq-aSum	.044	.066	.056	.055	.059	.008	.013	.008	.008	.013
Seq-aSum-VS	.050	.058	.056	.051	.058	.011	.018	.011	.009	.013
KBAC	.058	.044	.053	.054	.046	.013	.007	.009	.012	.009
C-alpha-A	.045	.051	.042	.036	.043	.016	<b>.030</b>	<b>.022</b>	.010	.014
C-alpha-P	.050	.065	.058	.051	.055	.005	.016	.013	.006	.012
RBT	.045	.045	.050	.062	.044	.011	.010	.011	.011	.005



Table 3: Empirical power for tests at nominal level  $\alpha$  based on 1000 replicates for an ideal case for 8 causal RVs with a common association strength  $OR = 2$  and a number of non-causal RVs. There is no LD among the RVs.

Test	$\alpha = 0.05$						$\alpha = 0.01$					
	# of neutral RVs						# of neutral RVs					
	0	4	8	16	32	64	0	4	8	16	32	64
UminP	.441	.336	.296	.222	.175	.117	.142	.089	.094	.050	.043	.029
Score	.746	.632	.595	.471	.332	.245	.496	.391	.314	.221	.143	.073
SSU	.756	.702	.694	.626	.499	<b>.423</b>	.525	.479	.448	.379	.283	.205
wSSU-P	.821	.732	.714	.644	.514	.390	.573	.471	.407	.332	.222	.161
SSUw	.743	.638	.593	.477	.339	.268	.502	.389	.316	.218	.153	.082
Sum	<b>.951</b>	<b>.875</b>	<b>.808</b>	<b>.673</b>	.484	.313	<b>.859</b>	<b>.709</b>	<b>.605</b>	<b>.438</b>	.248	.116
KMR(Linear)	.762	.711	.699	.631	<b>.509</b>	<b>.438</b>	.548	.500	.473	.405	<b>.308</b>	<b>.234</b>
KMR(Quad)	.755	.707	.699	.629	.501	<b>.410</b>	.545	.497	.466	.403	.299	.215
CMC(0.01)	.853	.761	.702	.628	.484	.396	.672	.524	.452	.384	.268	<b>.218</b>
CMC	.938	.853	.777	.616	.399	.211	.831	.679	.570	.383	.196	.086
wSum	.940	.846	.782	.618	.424	.267	<b>.838</b>	<b>.687</b>	.568	.394	.216	.114
aSum-P	.933	<b>.858</b>	.780	.669	.499	.313	.781	.611	.534	.381	.257	.125
Step-up	.859	.801	.769	<b>.679</b>	<b>.521</b>	.335	.712	.608	.552	<b>.431</b>	.301	.135
Seq-aSum	.810	.705	.663	.547	.407	.312	.596	.470	.415	.320	.190	.128
Seq-aSum-VS	.798	.722	.692	.590	.420	.344	.598	.506	.452	.345	.216	.141
KBAC	<b>.960</b>	<b>.911</b>	<b>.867</b>	<b>.779</b>	<b>.600</b>	.388	<b>.858</b>	<b>.749</b>	<b>.680</b>	<b>.529</b>	<b>.317</b>	.160
C-alpha-A	.741	.687	.664	.597	.460	.364	.637	.580	.538	.446	.320	.234
C-alpha-P	.771	.712	.688	.627	.484	.378	.542	.492	.459	.402	<b>.305</b>	<b>.219</b>
RBT	<b>.941</b>	.849	<b>.784</b>	.664	.463	.321	.813	.667	<b>.587</b>	.424	.238	.121

Table 4: Empirical power for tests at nominal level  $\alpha$  based on 1000 replicates for a non-ideal case for 8 causal RVs with various association strengths  $OR = (3, 3, 2, 2, 2, 1/2, 1/2, 1/2)$  and a number of non-causal RVs. There is no LD among the RVs.

Test	$\alpha = 0.05$					$\alpha = 0.01$				
	# of neutral RVs					# of neutral RVs				
	0	4	8	16	32	0	4	8	16	32
UminP	.607	.532	.481	.417	.346	.318	.259	.227	.204	.142
Score	.869	.772	.721	.632	.483	.660	.532	.480	.356	.233
SSU	.895	<b>.835</b>	<b>.815</b>	<b>.774</b>	<b>.696</b>	.723	.662	.645	.583	.472
wSSU-P	.861	.776	.735	.685	.550	.606	.510	.460	.401	.258
SSUw	.867	.773	.732	.633	.501	.661	.550	.481	.355	.238
Sum	.682	.566	.465	.365	.258	.471	.348	.257	.172	.101
KMR(Linear)	<b>.897</b>	<b>.842</b>	<b>.824</b>	<b>.783</b>	<b>.707</b>	<b>.740</b>	<b>.678</b>	<b>.667</b>	<b>.619</b>	<b>.495</b>
KMR(Quad)	.893	.835	.815	.781	.698	.734	<b>.680</b>	<b>.663</b>	.608	<b>.484</b>
CMC(0.01)	.703	.669	.670	.670	.590	.511	.457	.470	.470	.383
CMC	.661	.544	.456	.336	.204	.461	.337	.235	.157	.086
wSum	.659	.548	.459	.335	.228	.460	.336	.236	.158	.093
aSum-P	.854	.745	.684	.574	.430	.670	.538	.430	.315	.207
Step-up	.839	.767	.724	.640	.527	.652	.564	.518	.413	.285
Seq-aSum	.892	.811	.757	.671	.528	<b>.752</b>	.620	.532	.438	.273
Seq-aSum-VS	.885	.807	.768	.686	.545	.729	.623	.567	.448	.293
KBAC	<b>.907</b>	.813	.763	.642	.436	<b>.737</b>	.607	.536	.399	.199
C-alpha-A	.892	.826	.802	.757	.655	.824	.732	.720	.653	.512
C-alpha-P	<b>.906</b>	<b>.844</b>	<b>.823</b>	<b>.775</b>	<b>.674</b>	.735	<b>.673</b>	<b>.661</b>	<b>.612</b>	<b>.496</b>
RBT	.810	.659	.603	.482	.301	.590	.429	.356	.250	.125

Table 5: Type I error (with  $OR = 1$ ) and power (with eight causal RVs with  $OR = (3, 1/3, 2, 2, 2, 1/2, 1/2, 1/2)$ ) for tests at nominal level  $\alpha = 0.05$  based on 1000 replicates for 8 RVs and a number of other non-causal RVs. There is LD among the RVs.

Test	$OR = 1$					$OR = (3, 1/3, 2, 2, 2, 1/2, 1/2, 1/2)$				
	# of neutral RVs					# of neutral RVs				
	0	4	8	16	32	0	4	8	16	32
UminP	.033	.027	.026	.016	.013	.489	.479	.452	.365	.318
Score	.034	.022	.025	.019	.023	.599	.538	.491	.380	.276
SSU	.040	.041	.052	.044	.036	.603	<b>.624</b>	<b>.635</b>	<b>.581</b>	<b>.574</b>
wSSU-P	.057	.043	.047	.062	.053	.566	.586	.609	.585	.491
SSUw	.035	.042	.049	.033	.034	.532	.561	.574	.506	.493
Sum	.049	.047	.059	.033	.049	.342	.312	.315	.258	.239
KMR(Linear)	.042	.045	.057	.046	.043	.611	<b>.630</b>	<b>.644</b>	<b>.597</b>	<b>.590</b>
KMR(Quad)	.038	.033	.041	.030	.025	.545	.563	.565	.493	.474
CMC	.045	.053	.056	.036	.060	.296	.283	.189	.182	.365
wSum	.045	.054	.056	.040	.063	.369	.297	.287	.191	.200
aSum-P	.050	.046	.061	.038	.053	.350	.323	.325	.258	.243
Step-up	.047	.060	.059	.042	.050	.524	.516	.532	.429	.409
Seq-aSum	.045	.062	.054	.056	.055	<b>.658</b>	.617	.596	.484	.416
Seq-aSum-VS	.043	.056	.058	.054	.049	<b>.658</b>	.606	.577	.472	.414
KBAC	.050	.054	.050	.053	.049	.497	.439	.426	.371	.275
C-alpha-A	.065	<b>.076</b>	<b>.092</b>	<b>.097</b>	<b>.110</b>	-	-	-	-	-
C-alpha-P	.050	.049	.062	.057	.048	<b>.629</b>	<b>.650</b>	<b>.668</b>	<b>.607</b>	<b>.598</b>
RBT	.047	.039	.036	.060	.056	.374	.343	.386	.357	.279

Table 6: Type I error (with  $OR = 1$ ) and power (with eight causal RVs with  $OR = (3, 1/3, 2, 2, 2, 1/2, 1/2, 1/2)$ ) for tests at nominal level  $\alpha = 0.05$  based on 1000 replicates for 8 RVs and a number of other non-causal RVs. There is LD among the 8 RVs and among other non-causal RVs, but no LD between the 8 RVs and non-causal RVs.

Test	$OR = 1$					$OR = (3, 1/3, 2, 2, 2, 1/2, 1/2, 1/2)$				
	# of neutral RVs					# of neutral RVs				
	0	8	16	32	64	0	8	16	32	64
UminP	.032	.018	.021	.014	.007	.506	.380	.324	.288	.208
Score	.029	.029	.028	.019	.021	.631	.480	.373	.241	.160
SSU	.049	.051	.035	.034	.034	.642	.553	<b>.475</b>	<b>.444</b>	<b>.334</b>
wSSU-P	.045	.060	.042	.050	.052	.606	.494	.424	.362	.269
SSUw	.045	.040	.027	.015	.036	.562	.450	.352	.272	.187
Sum	.046	.059	.046	.046	.046	.345	.229	.159	.110	.079
KMR(Linear)	.051	.056	.039	.040	.037	.649	<b>.568</b>	<b>.490</b>	<b>.459</b>	<b>.356</b>
KMR(Quad)	.046	.049	.022	.021	.017	.572	.487	.392	.331	.205
CMC	.046	.053	.040	.050	.047	.339	.235	.193	.124	.111
wSum	.048	.052	.041	.053	.048	.342	.237	.199	.133	.114
aSum-P	.052	.061	.049	.046	.052	.364	.239	.170	.113	.081
Step-up	.057	.055	.047	.048	.051	.554	.449	.378	.304	.213
Seq-aSum	.051	.053	.041	.046	.052	<b>.703</b>	<b>.584</b>	.453	.353	.249
Seq-aSum-VS	.053	.053	.048	.041	.054	<b>.701</b>	.572	.447	.351	.258
KBAC	.048	.058	.036	.053	.047	.527	.388	.321	.262	.180
C-alpha-A	<b>.076</b>	<b>.093</b>	<b>.084</b>	<b>.092</b>	<b>.118</b>	-	-	-	-	-
C-alpha-P	.055	.065	.043	.050	.047	<b>.669</b>	<b>.585</b>	<b>.504</b>	<b>.472</b>	<b>.340</b>
RBT	.057	.059	.049	.042	.054	.376	.285	.188	.141	.097



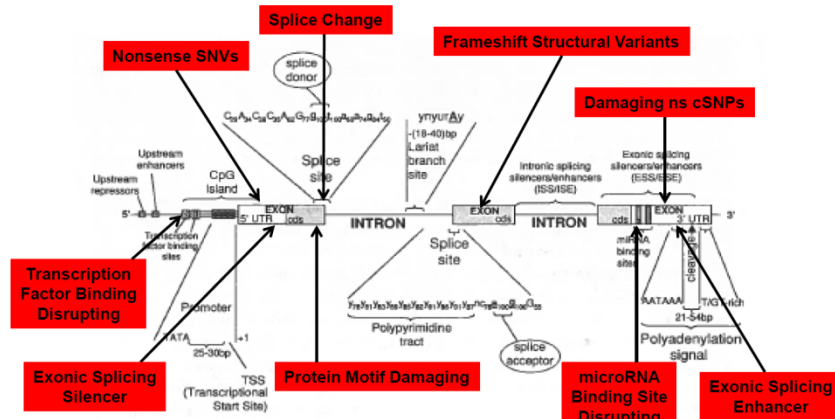
# Challenges in RV Analysis

## **Additional Issues with Rare Variant Analysis**

- Sequencing and Genotyping Errors
- Phasing and Diplotypic Effects
- Stratification
- The Use of *In Silico* Controls (e.g., 1000 Genomes Data)
- Moving Window vs. Annotation-Based Analyses
- Imputation
- Multiple Comparisons
- **Properties of Methods in Different Scenarios!**

# Functional Annotation, Predication

## Functional Annotations: *Bioinformatic* Predictions



**Figure 11.2** The anatomy of a gene. This figure illustrates some of the key regulatory regions that control the transcription, splicing and post-transcriptional processing of genes and transcripts. Polymorphisms in these regions should be investigated for functional effects

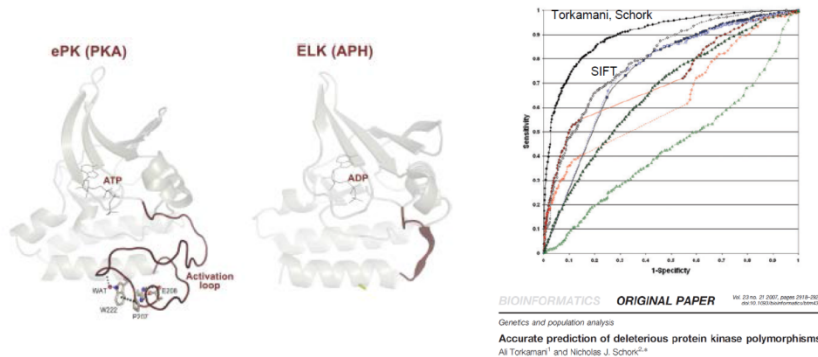
Plumpert and Barnes. "Predictive Functional Analysis of Polymorphisms: An Overview." in *Bioinformatics for Geneticists*. Wiley, 2007

We have developed methodology and tools for comprehensive bioinformatic WGS annot (Schork, Torkamani and colleagues: *Bioinformatics* 2008, 2009; *Cancer Research* (2009), *Nat Gen Rev* (2010), *Genomics* (2011))

## Functional Annotations: The Limits of Conservation

Torkamani, Kannan, Taylor, Schork. *PNAS* 105:9011-9016; 2008

Positions (residues/amino acids) of ~1000 disease causing variants in kinase proteins contrasted with the positions of ~1000 kinase variants not known to cause disease

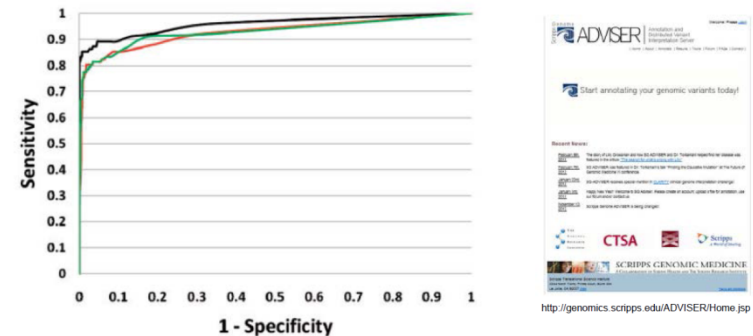


- **Review:** Lahiry, Torkamani, Schork, Hegele. *Nature Reviews Genetics* 11; 2010
- **Cancer Predictions:** Torkamani, Schork. *Cancer Research* 68; 2008

## Functional Annotations: Non-Coding Regions

Torkamani and Schork. *Bioinformatics* 24(16):1787-92; 2008

ENCODE features of the positions of 102 known disease-causing variants contrasted with the positions of 1049 non-disease-causing



Some features non-assay dependent; e.g., proximity to a TF start or end site

## Tools for *In Silico* Functional Prediction of Variants

- Model actual biophysical processes (e.g., protein structure, TF binding)
- Build classifiers using sequence information about the variants

Individual	Reference	Platform	Annotations
J. Venter	Venter (2007) [10]; Levy et al. (2007) [11]	Single sequencing	Disease, traits, ancestry
S. Quake	Quake et al. (2008) [12]	Microarray	Disease, traits, ancestry
Family with multiple members	Bruch et al. (2008) [13]	Complete genome	Specific disease mutations
NIH001	Moore et al. (2011) [14]	SNP	Specific disease mutations
NIH002	Moore et al. (2011) [14]	SNP	Disease, traits, ancestry
NIH003	Moore et al. (2011) [14]	SNP	Disease, traits, ancestry
NIH004	Moore et al. (2011) [14]	SNP	Disease, traits, ancestry
NIH005	Moore et al. (2011) [14]	SNP	Disease, traits, ancestry
NIH006	Moore et al. (2011) [14]	SNP	Disease, traits, ancestry
NIH007	Moore et al. (2011) [14]	SNP	Disease, traits, ancestry
NIH008	Moore et al. (2011) [14]	SNP	Disease, traits, ancestry
NIH009	Moore et al. (2011) [14]	SNP	Disease, traits, ancestry
NIH010	Moore et al. (2011) [14]	SNP	Disease, traits, ancestry

Tool	Website/reference	Purpose/theme
UCSC genome browser	<a href="http://www.genome.ucsc.edu/">http://www.genome.ucsc.edu/</a>	Position-specific functional organization of the genome
dbSNP	<a href="http://www.ncbi.nlm.nih.gov/projects/SNP/">http://www.ncbi.nlm.nih.gov/projects/SNP/</a>	Catalog variants with population-genetic annotations
OMIM	<a href="http://www.ncbi.nlm.nih.gov/omim">http://www.ncbi.nlm.nih.gov/omim</a>	Catalog known disease-causing mutations
HapMap	<a href="http://hapmap.ncbi.nlm.nih.gov/">http://hapmap.ncbi.nlm.nih.gov/</a>	Catalog variants with population-genetic annotations
COSMIC	<a href="http://www.sanger.ac.uk/perl/genetics/CGP/cosmic">http://www.sanger.ac.uk/perl/genetics/CGP/cosmic</a>	Catalog of somatic mutations from tumor sequencing
TAMAL	<a href="http://neoref.it/unc.edu/tamal/">http://neoref.it/unc.edu/tamal/</a>	Provides functional and population-genetic annotations
Variant analyzer	<a href="http://www.sanger.ac.uk/genetics/CGP/cosmic">http://www.sanger.ac.uk/genetics/CGP/cosmic</a>	Provides functional annotations
PharmGKB	<a href="http://www.pharmgkb.org/">http://www.pharmgkb.org/</a>	Pharmacogenetics variant annotations
HGDP selection browser	<a href="http://hgdp.uchicago.edu/cgi-bin/gbrowse/HGDP/">http://hgdp.uchicago.edu/cgi-bin/gbrowse/HGDP/</a>	Browser for assessing signs of selection in the human genome
Association database	<a href="http://www.genome.gov/gwasstudies">www.genome.gov/gwasstudies</a>	Results of genome wide association studies (GWAS)
SeattleSeq	<a href="http://gsos.washington.edu/SeattleSeqAnnotation/">http://gsos.washington.edu/SeattleSeqAnnotation/</a>	Variant annotation
Gene ontology	<a href="http://www.geneontology.org/">http://www.geneontology.org/</a>	Biological, molecular and cellular annotations
KEGG pathways	<a href="http://www.genome.jp/kegg/pathway.html">http://www.genome.jp/kegg/pathway.html</a>	Pathway analysis
DAVID	<a href="http://david.abcc.ncifcrf.gov/">http://david.abcc.ncifcrf.gov/</a>	Multiple annotations
UniProt	<a href="http://www.uniprot.org/">http://www.uniprot.org/</a>	Protein elements
Transfac	<a href="http://www.biobase-international.com">http://www.biobase-international.com</a>	Transcription factor databases
Genetoolkit website	<a href="http://www.genetoolkit.org">www.genetoolkit.org</a>	eQTL database

- Statistical RANKING algorithms are needed to prioritize variants in a study

# Population Issues in NGS Data Analysis

- Common variants – found in >5% of people in many populations
- Rare variants (<5% in the world population) *comprise the bulk of genetic variants and disproportionately important.*
- *RVs tend to be population specific.*

For example, in people with Native South American ancestry, a particular variant of a protein that transports cholesterol into cells is common and is strongly associated with low levels of high-density lipoprotein cholesterol, obesity and type 2 diabetes. European, Asian and African populations do not have this variant<sup>6</sup>.

Conversely, in dozens of studies in European populations, researchers have found 19 common single-nucleotide changes that are strongly associated with type 2 diabetes. In a further study of 6,000 people including European Americans, African Americans, Latinos, Japanese Americans and Native Hawaiians, 13 of these polymorphisms continue to be strongly associated with the disease<sup>7</sup>. Yet 5 of the 19 variants seem to have different effects in the different ethnic groups, and the role of one variant is unclear.

- Issues in using reference sequence

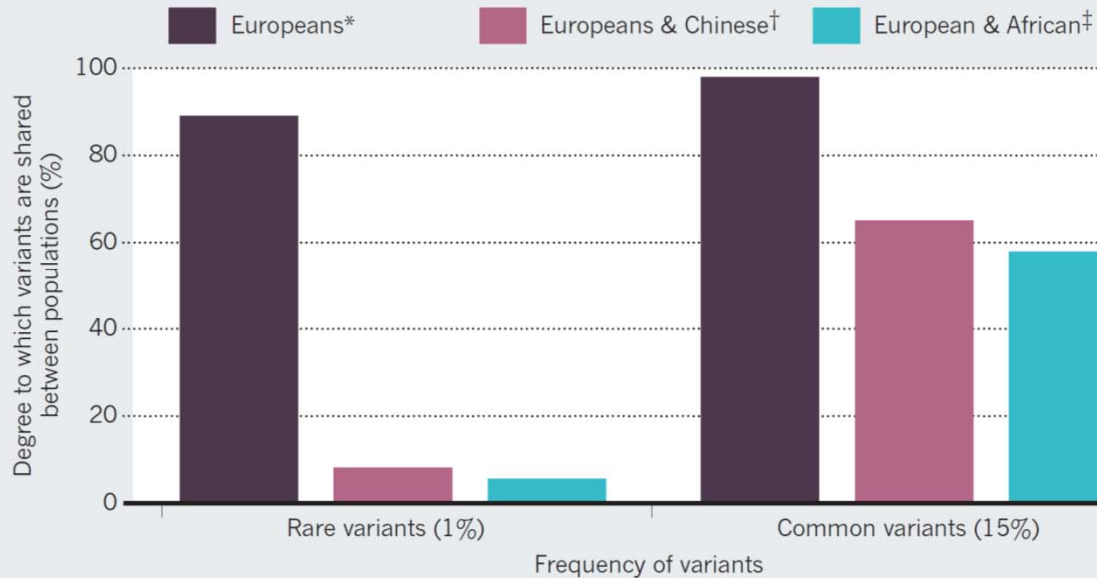
# Population Issues in NGS Data Analysis

SOURCE: REF. 5

## COMPARING THE UNCOMPARABLE

Ref: 14 JULY 2011 | VOL 475 | NATURE | 163

The rarer a genetic variant is within a population, the less likely it is to be found in all ethnic groups. One hundred people were sampled from each population.



\*Comparison of individuals of European descent in Utah and in Tuscany, Italy. † Han Chinese individuals from Beijing compared with Utah sample ‡ Yoruba individuals from Ibadan, Nigeria, compared with Utah sample.

## Example Issues:

- Determining individual ancestry or locus/allele-specific ancestry
- Unmatched (based on ancestry) cases and controls in a GWAS-seq = false positives
- Reference panel for determining the 'novelty' of a variant involves different ancestry

# Population Specific Alleles (Unique to Each Population)

Variant Type	Label	Populations			z-test p-values		
		AFR	EUR	ASN	AFR vs EUR	AFR vs ASN	EUR vs ASN
Total number of variants:		7614850	2024886	1294731			
Nonsense SNPs rate	1	0.500	0.840	0.842	6.931E-09	6.329E-07	4.910E-01
Frameshift Structural Variants rate	2	1.663	3.008	2.989	1.597E-34	6.239E-25	4.621E-01
Frameshift Insertion rate	3	0.657	1.274	1.383	6.368E-19	1.089E-18	2.006E-01
Frameshift Deletion rate	4	0.879	1.417	1.352	3.877E-12	1.584E-07	3.102E-01
Frameshift Rearrangement rate	5	0.127	0.316	0.255	2.614E-09	2.228E-04	1.572E-01
Splicing Change Variants rate	6	1.707	2.514	2.379	4.655E-14	7.112E-08	2.223E-01
Probably Damaging nscSNPs rate	7	10.103	15.472	15.602	1.136E-91	4.578E-69	3.853E-01
Possibly Damaging nscSNPs rate	8	5.991	7.744	8.233	7.313E-19	3.064E-21	6.111E-02
Protein motif damaging Variants rate	9	4.104	6.311	6.581	2.612E-39	3.043E-35	1.726E-01
TFBS Disrupting Variants rate	10	2.793	4.173	4.063	7.493E-69	2.764E-42	1.785E-01
miRNA-BS Disrupting Variants rate	11	0.948	1.170	1.081	2.405E-03	7.715E-02	2.286E-01
ESE-BS Disrupting Variants rate	12	5.835	7.260	7.283	1.696E-13	2.840E-10	4.689E-01
ESS-BS Disrupting Variants rate	13	2.460	3.013	2.865	6.435E-06	3.539E-03	2.232E-01
Total Likely Functional Variant rate	14	23.718	34.906	35.436	8.999E-170	1.234E-132	2.128E-01

Frequencies of funct pop spec variants: Greater in non-Africans      Highly significant AFR vs. non-AFR

- The rate of novel functional variants (not just homozygous) is significantly higher in non-Africans
- The rate is uniformly higher across ALL functional classes, not just ns cSNPs
- Selection has had less time to 'purify' the European and Asian population (i.e., replicated Lohmuller et al.)



# Phasing

## Diploidy and Compound Heterozygosity (CH)

Variants that cause dysfunction

Heterozygosity

...ATCGAGC**T**/CAGCGCGATAGC**G**/ACTAGCAT...

...ATCGAGC**T**AGCGCGATAGC**G**CTAGCAT... Maternal

Compensation

...ATCGAGC**C**AGCGCGATAGC**G**CTAGCAT... Paternal

or

Both gene homologs dysfunctional

...ATCGAGC**C**AGCGCGATAGC**G**CTAGCAT... Maternal

...ATCGAGC**T**AGCGCGATAGC**G**CTAGCAT... Paternal

## Phasing for Assessing 'Diploic' Phenomena

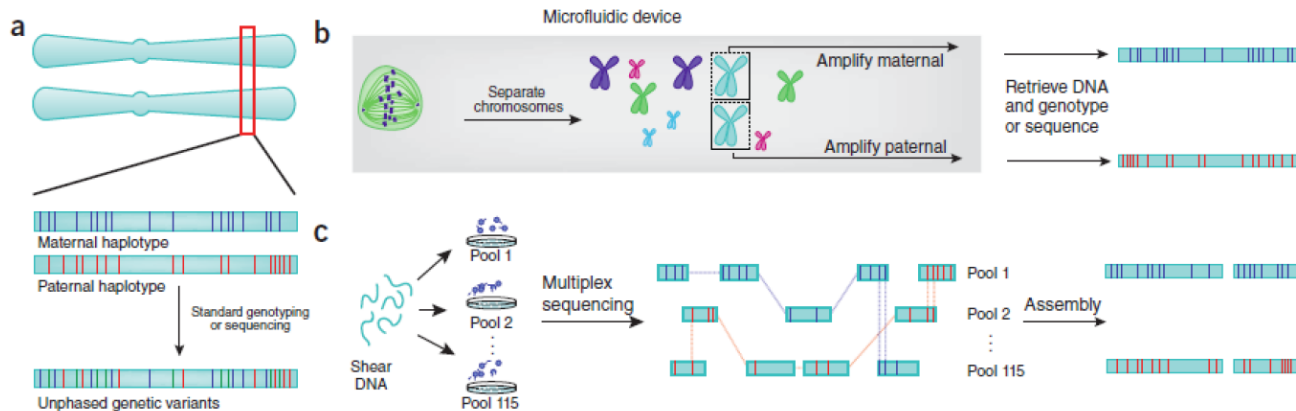
### Approaches to Resolving Phase

- Sequencing parents/relatives
- Population-based phasing (and imputation)
- Assembly of sequencing reads
- Separate chromosomes prior to sequencing

After phasing all variants:

1. Annotate positions of all variants (Human Genome hg18)
2. Predict likely functional effect of variants using bioinformatics pipeline
3. Assign disease risk alleles from association study databases
4. Explore regions of high heterozygosity/nucleotide content differences between homologous chromosomes

Torkamani et al. (in review)



### The next phase in human genetics

Vikas Bansal, Ryan Tewhey, Eric J. Topol & Nicholas J. Schork

Experimental haplotyping of whole genomes is now feasible, enabling new studies aimed at linking sequence variation to human phenotypes and disease susceptibility.

NATURE BIOTECHNOLOGY VOLUME 29 NUMBER 1 JANUARY 2011



# Network-based

## Genetic Networks and Network Analysis

NATURE | VOL 411 | 3 MAY 2001

**brief communications**

### Lethality and centrality in protein networks

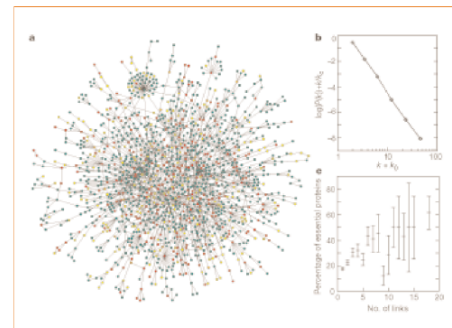
The most highly connected proteins in the cell are the most important for its survival.

H. Jeong\*, S. P. Mason†, A.-L. Barabási\*,  
Z. N. Oltvai†

Cell 144, March 18, 2011 ©2011

### Interactome Networks and Human Disease

Marc Vidal,<sup>1,2,\*</sup> Michael E. Cusick,<sup>1,2</sup> and Albert-László Barabási<sup>1,3,4,\*</sup>



**Figure 1** Characteristics of the yeast proteome. **a**, Map of protein-protein interactions. The largest cluster, which contains ~70% of all proteins, is shown. The colour of a node signifies the phenotypic effect of removing the corresponding protein (red, lethal; green, non-lethal; orange, slow growth; yellow, unknown). **b**, Connectivity distribution  $P(k)$  of interacting yeast proteins, giving the probability that a given protein is directly or indirectly connected to  $k$  other proteins. The horizontal axis indicates the number of proteins with more than 10 interactions. **c**, Percentage of essential proteins versus the number of links. The horizontal axis indicates the number of proteins with more than 10 interactions.

## Network Centrality Measures

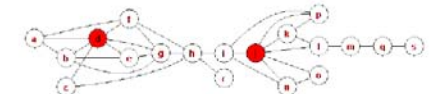
NATURE REVIEWS | **GENETICS** | VOLUME 12 | JANUARY 2011

### Network medicine: a network-based approach to human disease

Albert-László Barabási<sup>\*\*§</sup>, Natali Gulbahce<sup>\*\*||</sup> and Joseph Loscalzo<sup>§</sup>

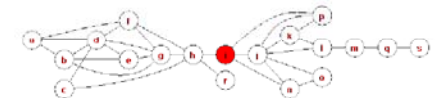
#### Degree Centrality

- Number of nodes connected to a given node
- How well a node is connected; direct influence



#### Closeness Centrality

- Sum of shortest distance (path) to all other nodes
- Inverse measure of centrality



#### Betweenness Centrality

- Frequency that *node*=shortest path between 2 nodes
- Control of communication between other nodes



Many other measures of node's importance in a network...

## Spectral Gap

# Referring to the original slides (part 2)

## Rare Variant Analysis

Michael C. Wu

Division of Public Health Sciences  
Fred Hutchinson Cancer Research Center

<http://research.fhcrc.org/wu/>

Total 100 pages for the topic slides.

Total 59 pages for software demo slides.

Copy of the slides made available to the  
pertinent audience.

# Dr. Wu's slides

## GWAS: Missing Heritability

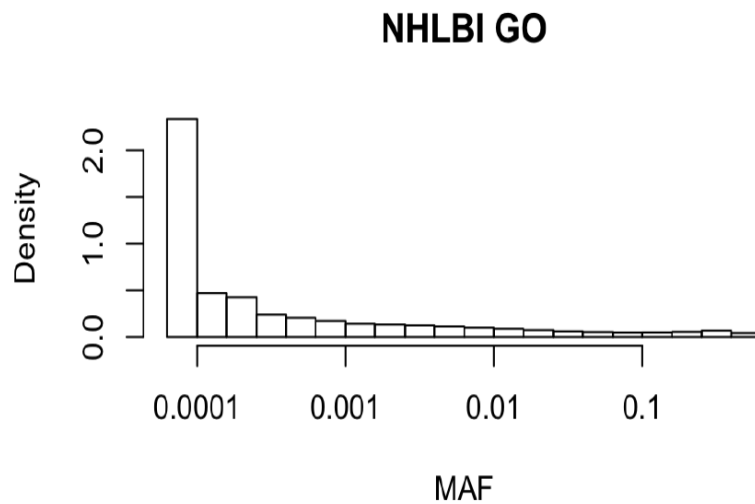
- GWAS focus on **common** variants ( $MAF \geq 5\%$ ) whose effects are small with  $RR \approx 1.2-1.5$ .
- **Missing heritability**: Significant GWAS SNPs explain a small proportion of disease heritability.
- Possible reasons:
  - ▶ GxG and GxE interactions?
  - ▶ Many common causal variants: Each with a small effect?
  - ▶ Epigenetics?
  - ▶ **Rare variants?**



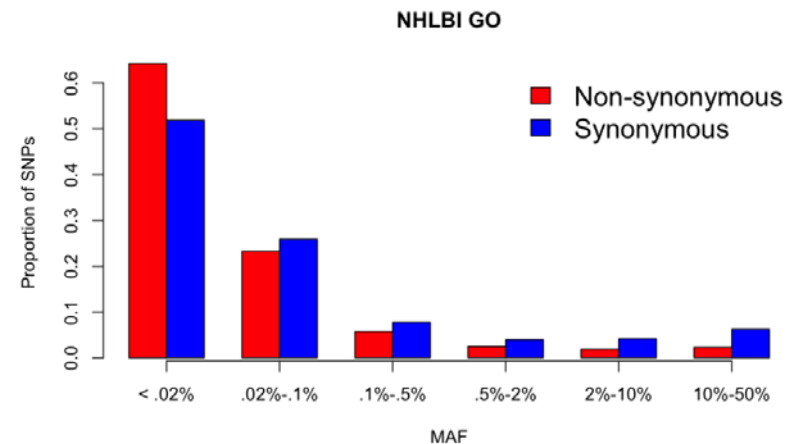
# Dr. Wu's slides

## Why rare variants?

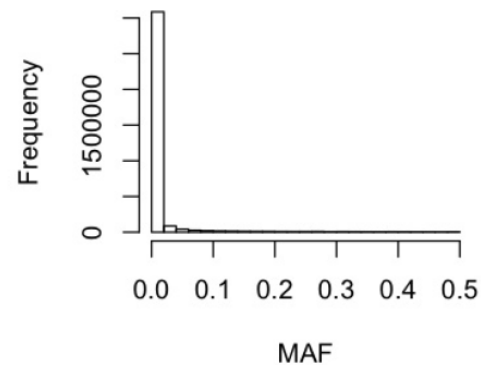
- Most of human variants are rare



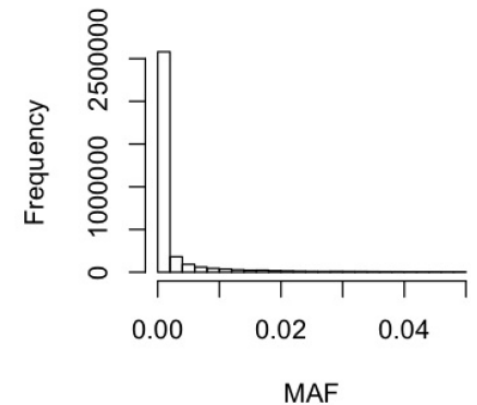
- Functional variants tend to be rare.



**A. MAF < 0.5**



**A. MAF < 0.05**



# Dr. Wu's slides

## Challenges: May Not “Observe” the Variant

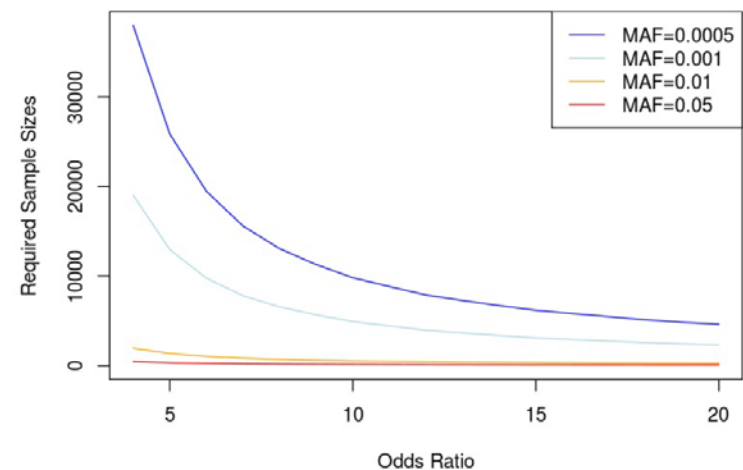
- Sample size necessary to observe a variant with MAF  $p$  with at least chance  $\theta$ :

$$n > \frac{\ln(1 - \theta)}{2\ln(1 - p)}$$

- For  $\theta = 99.9\%$ , the minimum required sample size is

MAF	0.1	0.01	0.001	0.0001
Required $n$	33	344	3453	345337

## Challenges: Power Depends on Allele Frequency





# Dr. Wu's slides

## Approach and Notation

### General Approach

- GWAS Analysis Unit: Individual SNP
- Sequencing Study Unit: region (e.g. gene, moving window, exons, etc)
- Operationally:
  - 1 Test effect of variants within single region
  - 2 Correct for multiple testing

Then for a single region:

- $y_i$  = trait value for person  $i$
- $\mathbf{G}_i$  = vector of genotypes for the particular group of variants
- $\mathbf{Z}_i$  = vector of any additional covariates (e.g. demographics/environment)

**Goal:** Test for association between  $y$  and  $G$  while adjusting for  $Z$ .

# Dr. Wu's slides

## Collapsing Tests

- Aggregate rare variant information in a region into a summary dose variable
  - CAST
  - CMC
  - MZ (GRANVIL)
  - Weighted Sum Test
- Most powerful if all rare variants are causal variants with the same effect sizes (and association directions).

### Burden Tests

- Collapse rare variants

Y	G <sub>1</sub>	G <sub>2</sub>	G <sub>3</sub>	G <sub>4</sub>	C
1	1	0	0	0	1
1	0	1	0	0	1
1	0	0	1	1	2
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
0	0	0	0	0	0
0	0	0	0	0	0
0	0	0	0	0	0



### What is Burden test?

Burden test is a type of tests that assess the cumulative effects of multiple variants in a genomic region.

Burden tests are based on collapsing or summarizing the rare variants within a region by a single value, which is then tested for association with the trait of interest.

### Ref:

- Li, B., and Leal, S.M. (2008). Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. *Am. J. Hum. Genet.* 83, 311–321.
- Madsen, B.E., and Browning, S.R. (2009). A groupwise association test for rare mutations using a weighted sum statistic. *PLoS Genet.* 5, e1000384.
- Morgenthaler, S., and Thilly, W.G. (2007). A strategy to discover genes that carry multi-allelic or mono-allelic risk for common diseases: a cohort allelic sums test (CAST). *Mutat. Res.* 615, 28–56.
- Li, B., and Leal, S.M. (2009). Discovery of rare variants via sequencing: implications for the design of complex trait association studies. *PLoS Genet.* 5, e1000481.
- Price, A.L., Kryukov, G.V., de Bakker, P.I., Purcell, S.M., Staples, J., Wei, L.J., and Sunyaev, S.R. (2010). Pooled association tests for rare variants in exon-resequencing studies. *Am. J. Hum. Genet.* 86, 832–838.
- Han, F., and Pan, W. (2010). A data-adaptive sum test for disease association with multiple common or rare variants. *Hum. Hered.* 70, 42–54.
- Morris, A.P., and Zeggini, E. (2010). An evaluation of statistical approaches to rare variant analysis in genetic association studies. *Genet. Epidemiol.* 34, 188–193.

# Dr. Wu's slides

## Collapsing Tests:

### CAST: Binary Collapsing

Cohort Allele Sum Test

$$C_i = I(\sum_{j=1}^p G_{ij} > 0) = \begin{cases} 1 & \text{rare variants in the region for subject } i \\ 0 & \text{otherwise} \end{cases}$$

### Count Collapsing: (MZ, GRANVIL, ANRV, other names)

$$C_i = \sum_{j=1}^p G_{ij} = \# \text{ of rare variants observed in the region for subject } i$$

Use the model:

$$y_i = \beta_0 + \beta C_i + \alpha' Z_i$$

and we can test  $H_0 : \beta = 0$ .

CMC test: extends these by collapsing (sub)groups of variants in a region and testing the subgroups between cases and controls.

## Collapsing Tests: Early Weighted Methods

### Weighted Collapsing: (Madsen & Browning)

- For variant  $j$ , set  $\hat{w}_j = 1 / \sqrt{q_j(1 - q_j)}$  where  $q_j$  is the MAF in controls

$$C_i = \sum_{j=1}^p \hat{w}_j * G_{ij}$$

- Construct a wilcoxon statistic comparing between cases/controls and use Permutation for significance
- Idea: we want to up-weight rarer variants

### Weighted Collapsing: (Unsupervised)

Set  $\hat{w}_j = 1 / \sqrt{q_j(1 - q_j)}$  where  $q_j$  is the MAF in all samples

$$C_i = \sum_{j=1}^p \hat{w}_j * G_{ij}$$

Test using regression model again.

## Supervised Collapsing Methods

Supervised: the outcome (phenotype or case/control status) is used to estimate weights

- Similar to the weighted collapsing, we can introduce weights that adjust to the magnitude and direction of effect by introducing new weights and then computing

$$C_i = \tilde{w}_1 G_{i1} + \tilde{w}_2 G_{i2} + \dots + \tilde{w}_p G_{ip}.$$

- Theoretical optimal weights:  $\tilde{w}_j = \beta_j$  the true LOR
- EREC Test:  $\tilde{w}_j = \hat{\beta}_j + \delta$  where  $\hat{\beta}_j$  is an initial (hopefully good) estimate for  $\beta_j$  and  $\delta$  is a constant (usually 1 or 2)
- Han and Pan:  $\tilde{w}_j = 1$  if  $\hat{\beta}_j > 0$  and  $\tilde{w}_j = 1$  if  $\hat{\beta}_j < 0$
- Others...

Permutation or bootstrapping MUST be used for significance.

# Dr. Wu's slides

## Understanding Collapsing Methods

- Count collapsing restrict to rare variants (with MAF < threshold) and collapse rare variants (dose= $C_i$ ).
- If all  $\beta$ 's in the regression model are the same, the model becomes

$$y_i = \alpha_0 + \alpha' \mathbf{Z}_i + \beta C_i + \varepsilon_i$$

where  $C_i = G_{i1} + G_{i2} + \dots + G_{ip}$  = number of rare variants in the region.

### When Collapsing Methods are Optimal:

- 1 **ALL** rare variants in a region are causal
  - 2 All rare variant effects ( $\beta$ 's) have **SAME** direction and magnitude
- Similar ideas for other collapsing methods.
  - Is it realistic to anticipate all variants have same direction and (after weighting) magnitude? Perhaps not.

## Burden Test: Mixed effect directions

- Lose power if variants have positive and negative effects.

Y	G <sub>1</sub>	G <sub>2</sub>	G <sub>3</sub>	G <sub>4</sub>	C
1	1	0	0	0	1
1	0	1	0	0	1
1	0	0	0	0	0
·	·	·	·	·	·
·	·	·	·	·	·
·	·	·	·	·	·
0	0	0	0	0	0
0	0	0	1	0	1
0	0	0	0	1	1



⇒ No Power !

## Sequence Kernel Association Test (SKAT)

- Intuition: Compare pair-wise similarity in phenotype between subjects to pair-wise similarity in genotypes at the rare variants
- Similarity in genotype is measured by way of a **Kernel**
  - ▶ Kernel:  $K(\mathbf{G}_i, \mathbf{G}_{i'})$  measures similarity between subjects  $i$  and  $i'$
  - ▶ The Kernel determines the form of the underlying trait model
- SKAT uses a score test to generate a  $p$ -value:

$$Q = \frac{(\mathbf{y} - \hat{\mathbf{y}})' \mathbf{K} (\mathbf{y} - \hat{\mathbf{y}})}{\hat{\sigma}^2}$$

where  $\mathbf{K}$  as  $n \times n$  matrix with  $i, i'^{th}$  term  $K(\mathbf{G}_i, \mathbf{G}_{i'})$

- ▶ Asymptotics are used to get  $p$ -value analytically
- Default version of SKAT is considerably simpler (new few slides)

Default version:

- Earlier model:

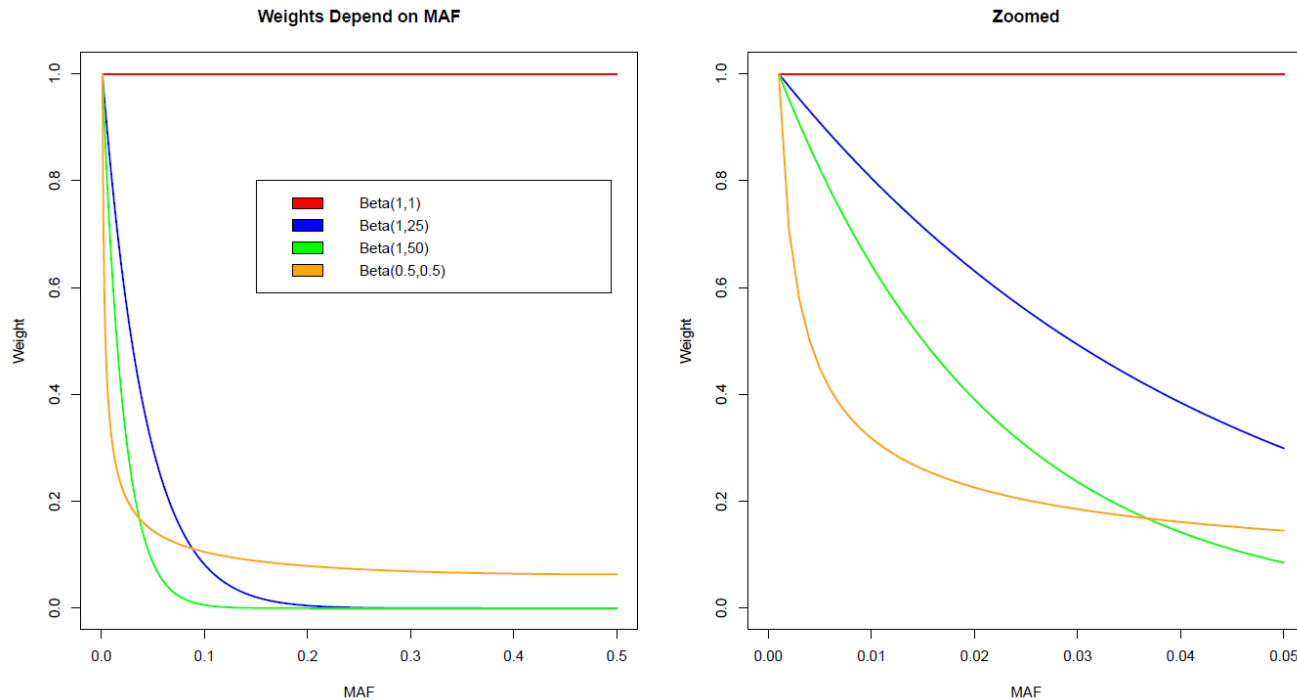
$$y_i = \alpha_0 + \mathbf{Z}_i' \boldsymbol{\alpha} + \beta_1 \mathbf{G}_{i1} + \beta_2 \mathbf{G}_{i2} + \cdots + \beta_p \mathbf{G}_{ip} + \varepsilon_i$$

- Goal: Test  $H_0 : \beta_1 = \beta_2 = \dots = \beta_p = 0$
- Standard LR test requires  $p$ -df test (low power).
- Instead: assume each  $\beta_j \sim$  distribution  $F(0, w_j \tau)$  where  $w_j$  is a weight for variant  $j$ .
- Then  $H_0 : \beta_1 = \beta_2 = \dots = \beta_p = 0 \Leftrightarrow H_0 : \tau = 0$
- Can use a variance component score test via a connection with mixed models.

## Weights in the Sequence Kernel Association Test (SKAT)

- Upweight rarer variants
- We assume weights  $w_j = \text{decreasing function of MAF } \pi_j$ .
- Example:  $w_j = \text{Beta}(\pi_j; a_1, a_2)$ , where  $\text{Beta}(\cdot)$  is *Beta* function.
- Optimal  $w_j$  is an indicator for whether the  $j^{\text{th}}$  variant is causal — never known a priori!
- Estimation of weights (e.g. EREC) would require permutation or bootstrapping for significance.

### Example of Weights





## SKAT Statistic

- SKAT = weight sum of individual score statistics:

$$Q_{SKAT} = \sum_{j=1}^p w_j U_j^2$$

where  $U_j = \mathbf{G}'_j(\mathbf{y} - \hat{\mathbf{y}})$  is the score statistic for  $\beta_j$  in the model:

$$y_i = \alpha_0 + \boldsymbol{\alpha}'\mathbf{Z}_i + \beta_j G_{ij} + \varepsilon_i$$

- Calculations of  $Q$  only require fitting the NULL model — which is the same for any region of the genome:

$$y_i = \alpha_0 + \boldsymbol{\alpha}'\mathbf{Z}_i + \varepsilon_i$$

so we only need to fit the model once!

## SKAT: p-value Calculation

- The SKAT statistic asymptotically can be written as

$$Q_{SKAT} = \sum_{j=1}^p \lambda_j Z_j^2$$

where the  $Z_j$  are  $N(0, 1)$  and the  $\lambda_j$  are mixture weights which can be computed analytically.

- $p$ -value can then be calculated as a mixture of  $\chi^2$  using many different methods: we choose the Davies approach.
- Allows for fast computation

### Computational Speed

Assuming  $n = 1000$  subjects and regions of 30kb:

Sequence Length	300 Kb	3 MB	3Gb (Whole Genome)
Time	2.5 sec	25 sec	7 hrs

on a 2.33 GHz Laptop with 6Gb of memory.

## C-alpha Test (Another Similarity-based Test):

- C-alpha test:
  - ▶ Based on Neyman's (1966)  $c(\alpha)$  test for mixture of "biased" coins (overdispersion)
  - ▶ Requires permutation for significance
  - ▶ Case-control traits only
  - ▶ Like SKAT, C-alpha is better than collapsing methods when variants are not unidirectional in effect
- SKAT is a generalization of C-alpha test:
  - ▶ C-alpha and SKAT (with a dichotomous trait and no covariates) are the same if we set

$$K(\mathbf{G}_i, \mathbf{G}_{i'}) = \mathbf{G}_i' \mathbf{G}_{i'}$$

(i.e. setting the weights to 1) and then use permutation to assess significance.

## SKAT vs. Collapsing

- Collapsing tests are more powerful when a large % of variants are causal and effects are in the same direction.
- SKAT is more powerful when a small % of variants are causal, or the effects have mixed directions.
- Both scenarios can happen when scanning the genome.
- Best test to use depends on the underlying biology.
  - Difficult to choose which test to use in practice.
- Questions:
  - Which group of variants test? I.e. what is the threshold for "rare"?
  - Which type of test should I use? Variance component or burden?
- Truth is unknown: depends on the situation
- Omnibus tests: work well across situation

## SKAT-O: An Optimal Unified Strategy

- We can also construct a score statistic for Collapsing analysis:  
 $Q_{collapse}$
- Unified test statistic:

$$Q_{opt}(\rho) = \rho Q_{collapse} + (1 - \rho) Q_{SKAT}, \quad 0 \leq \rho \leq 1$$

- Note: SKAT ( $\rho = 0$ ) and collapsing ( $\rho = 1$ ) are special cases!
- **Idea: Use data to adaptively estimate  $\rho$  to maximize power, i.e. minimize p-value, and account for having  $\rho$  estimated.**

### SKAT-O: Key Features

- Optimal: good for scenarios where SKAT works **AND** scenarios where collapsing works well
- Still permits analytical p-value computation.

## Simulations:

- Generate sequence data using a coalescent population genetics model.
- Most variants are rare and a good number have LD: for example, in a 30kb region:

# Variants	MAF
626 true	
159 (25%)	$\leq 10^{-4}$
441 (71%)	$\leq 10^{-3}$
511 (88%)	$\leq 10^{-2}$

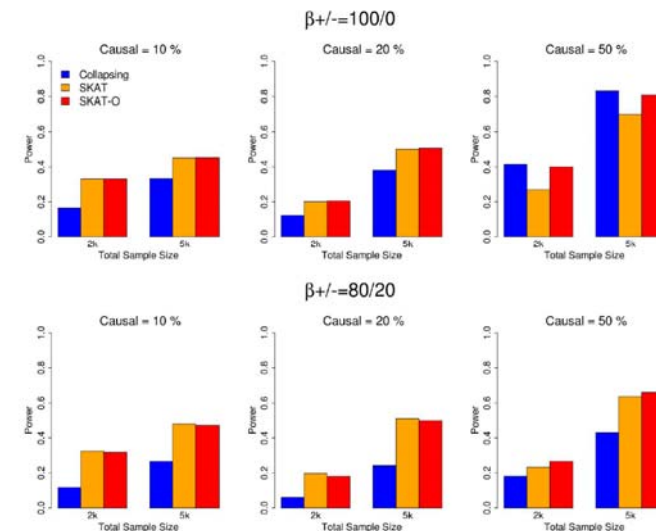
## Simulations: Type I Error Rate (SKAT)

Type I error at  $\alpha = 10^{-6}$  level

Total Sample Size	Continuous Traits	Binary Trait
500	$5.9 \times 10^{-7}$	$1.0 \times 10^{-8}$
1000	$8.0 \times 10^{-7}$	$2.3 \times 10^{-7}$
2500	$8.4 \times 10^{-7}$	$5.6 \times 10^{-7}$
5000	$8.8 \times 10^{-7}$	$7.0 \times 10^{-7}$

SKAT is known to be conservative for binary traits and small  $n$ . —  
Note: **Newer version can correct for this!**

## Simulations: Power Results



## Rare variant Meta-analysis

- Meta-analysis is an effective approach to combine data from multiple studies.
- Rare variant meta-analysis: **desirable properties**
  - ▶ Use summary statistics
  - ▶ Same power as mega-analysis (joint analysis)
  - ▶ Account for varying levels of heterogeneity of genetic effects across studies.

### Single study $k$

- Score statistic of variant  $j$

$$S_{kj} = \sum_{i=1}^n g_{ijk} (y_{ij} - \hat{\mu}_{ij}) / \hat{\phi}_k$$

- SKAT and Burden test statistics:

$$Q_{SKAT} = \sum_{j=1}^p (w_{jk} S_{kj})^2, \quad Q_{Burden} = \left( \sum_{j=1}^p w_{jk} S_{kj} \right)^2$$

- SKAT-O (combined approach):

$$T = \min_{0 \leq \rho \leq 1} P_\rho$$

where  $P_\rho$  is the p-value of

$$Q_\rho = (1 - \rho) Q_{SKAT} + \rho Q_{Burden}$$

## Multi-Study Model

- For the  $k^{th}$  study ( $k = 1, \dots, K$ ),
  - ▶ Genotype  $\mathbf{G}_{ki} = (g_{ki1}, \dots, g_{kip})'$
  - ▶ Covariates  $\mathbf{X}_{ki} = (x_{ki1}, \dots, x_{kiqu_k})'$
  - ▶ Model:

$$g(\mu_{ik}) = \mathbf{X}_{ki} \boldsymbol{\alpha}_k + \mathbf{G}_{ki} \boldsymbol{\beta}_k$$

- ▶ Test  $H_0: \boldsymbol{\beta}_k = 0$  ( $k = 1, \dots, K$ )

## Input Summary Statistics for meta-analysis

- Input summary statistics from each study

- ▶ MAF
- ▶  $S_{kj}$ : score statistic of each marker
- ▶ Between-variant relationship matrix ( $p \times p$ )

$$\boldsymbol{\Phi}_k = \mathbf{G}'_k \mathbf{P}_k \mathbf{G}_k,$$

$$\text{where } \mathbf{P}_k = \mathbf{V}_k^{-1} - \mathbf{V}_k^{-1} \mathbf{X}_k (\mathbf{X}'_k \mathbf{V}_k^{-1} \mathbf{X}_k)^{-1} \mathbf{X}'_k \mathbf{V}_k^{-1}$$

## Meta-SKAT: Homogeneous genetic effects

- Meta-SKAT assuming homogeneous genetic effects:

$$Q_{hom\_meta\_SKAT} = \sum_{j=1}^p \left( \sum_{k=1}^K w_{kj} S_{kj} \right)^2$$

- Meta-Burden:

$$Q_{meta\_Burden} = \left( \sum_{j=1}^p \sum_{k=1}^K w_{kj} S_{kj} \right)^2$$

- Meta-SKAT-O:

$$Q_{hom\_meta}(\rho) = (1 - \rho)Q_{hom\_meta\_SKAT} + \rho Q_{meta\_Burden}$$

- Test statistics are essentially **identical** to those of the **mega analysis SKAT and burden test**  
 $\Rightarrow$  **As powerful as mega-analysis**
- P-values can be computed using the Davies method.  
 $\Rightarrow$  **Fast computation**
- SKAT-O can be conducted with adaptively selecting  $\rho$ .

## Meta-SKAT: Heterogeneous genetic effects

- Assume genetic effects vary between studies
  - ▶  $\beta_1, \dots, \beta_K$  are iid
  - ▶  $E(\beta_{kj}) = 0$ ,  $\text{var}(\beta_j) = w_{kj}\tau$  and  $\text{cor}(\beta_{kj}, \beta_{kj'}) = \rho$ .
- Multivariate score-based analog of the univariate random effect model meta-analysis.
- P-values can be calculated analytically
- Useful for meta analysis of studies of the same ethnicity or different ethnicities.
- Meta-SKAT assuming heterogeneous genetic effects:

$$Q_{het\_meta\_SKAT} = \sum_{j=1}^p \sum_{k=1}^K (w_{kj} S_{kj})^2$$

- Meta-SKAT-O:

$$Q_{hom\_meta}(\rho) = (1 - \rho)Q_{het\_meta\_SKAT} + \rho Q_{meta\_Burden}$$

## Meta-SKAT for multi-ethnicities:

- Multi-ethnic studies:
  - ▶ within-group homogeneity and between-group heterogeneity
  - ▶  $\beta_k = \beta_l$  for the same group and  $\beta_k \perp \beta_l$  for the different groups
- Meta-SKAT with  $B$  ancestry groups

$$Q_{het\_meta\_SKAT} = \sum_{j=1}^p \sum_{b=1}^B \left( \sum_{k=k_{b-1}+1}^{k_b} w_{kj} S_{kj} \right)^2$$

- Meta-SKAT-O:

$$Q_{hom\_meta}(\rho) = (1 - \rho)Q_{het\_meta\_SKAT} + \rho Q_{meta\_Burden}$$

## Comments

- Quality control:
    - ▶ Are the observed variants really variants?
    - ▶ Batch effects
    - ▶ Some standard pipelines now in place
  - Population stratification:
    - ▶ Common strategy: just use same PCs from common variant analysis to correct for PS
    - ▶ Some evidence that rare variants require special accommodation (much larger number of PCs)
  - Accommodating common variants:
    - ▶ What do you do with common variants?
    - ▶ (a) Assess joint effect with rare variants
    - ▶ (b) Adjust for effect of common variants
  - Prediction
    - ▶ In a new population (sample), we're unlikely to see the same variants and we're likely to see a lot of variants not previously observed
  - Prioritization of individual variants
    - ▶ How to choose individual causal variants?
    - ▶ Some work on variable selection methods, but no ability to control type I error.
    - ▶ Bioinformatics and functionality tools may be useful
  - Incorporation of functional information and other genomic data
- Design Choices
    - ▶ Want to enrich for variants (extreme phenotypes)
    - ▶ Some of these designs require specialized methods
    - ▶ Stuck with the design chosen
  - Is rare variant analysis worthwhile?
    - ▶ Huge sample sizes required to even observe the variant
    - ▶ Despite hypotheses, \*relatively\* few associations have been discovered
    - ▶ Perfect confounding with environment?
    - ▶ What's the real public health impact?
    - ▶ Perhaps too early to tell.



## SKAT Package

SKAT package has functions to:

1. test an association between SNP sets and continuous/binary phenotypes and
2. compute power/sample size for future sequence association studies.

### Getting R and SKAT Package

#### ► Downloading R:

<http://cran.r-project.org/>

#### ► Obtaining SKAT:

```
install.packages("SKAT")  
library(SKAT)
```

## Example Dataset

SKAT package provides an example dataset (SKAT.example)

```
library(SKAT)  
data(SKAT.example)  
names(SKAT.example)
```

```
attach(SKAT.example)
```

```
hist(apply(Z,2,mean)/2, xlab = "MAF",  
      main = "MAFs of Variants")
```

## SKAT: Simple Usage

To test an association, you first need to run SKAT Null Model function to get parameters and residuals from the null model of no association, and then to run SKAT to compute a p-value.

```
# continuous trait
obj<-SKAT_Null_Model(y.c ~ X, out_type="C")
SKAT(Z, obj)$p.value

# dichotomous trait
obj<-SKAT_Null_Model(y.b ~ X, out_type="D")
SKAT(Z, obj)$p.value
```

## SKAT-O: Omnibus (Combined) Test of collapsing and SKAT

The test statistic of the combined test is

$$Q_{\rho} = (1 - \rho)Q_S + \rho Q_B,$$

where  $Q_S$  is a test statistic of SKAT, and  $Q_B$  is a score test statistic of weighted burden test. Thus,  $\rho = 0$  results in the original weighted linear kernel SKAT, and  $\rho = 1$  results in the weighted burden test. You can specify  $\rho$  value using the `r.corr` parameter (default: , `r.corr=0`).

```
SKAT(Z, obj, r.corr=0)$p.value
SKAT(Z, obj, r.corr=0.9)$p.value
SKAT(Z, obj, r.corr=1,
  weights = rep(1, ncol(Z)))$p.value
summary(glm(y.b ~ apply(Z,1,sum)+X,
  family = "binomial"))
```

## SKAT-O: Optimal Adjustment

If `method='optimal.adj'`,  $\rho$  is selected from a grid of eight points  $\rho = (0, 0.1^2, 0.2^2, 0.3^2, 0.4^2, 0.5^2, 0.5, 1)$  to maximize the power. If you want to use the original implementation of SKAT-O, use `method='optimal'`. We recommend to use `'optimal.adj'`, since it has a better type I error control in the tail area.

```
SKAT(Z, obj, method="optimal")$p.value
```

```
SKAT(Z, obj, method="optimal.adj")$p.value
```

## Combined test of rare and common variants

If you want to test the combined effects of common and rare variants, you can use `SKAT_CommonRare` function.

```
# Combined sum test (SKAT-C and Burden-C)
```

```
SKAT_CommonRare(Z, obj)$p.value
SKAT_CommonRare(Z, obj, r.corr.rare=1,
  r.corr.common=1 )$p.value
```

```
# Adaptive test (SKAT-A and Burden-A)
```

```
SKAT_CommonRare(Z, obj, method="A")$p.value
SKAT_CommonRare(Z, obj, r.corr.rare=1, r.corr.common=1,
  method="A" )$p.value
```

## Accommodating PLINK Formats

```
# note that "PlinkExample/" is a directory we download
# the data
#
# Create the MW File
File.Bed<-"./PlinkExample/Example1.bed"
File.Bim<-"./PlinkExample/Example1.bim"
File.Fam<-"./PlinkExample/Example1.fam"
File.SetID<-"./PlinkExample/Example1.SetID"
File.SSD<-"./PlinkExample/Example1.SSD"
File.Info<-"./PlinkExample/Example1.SSD.info"

# To use binary ped files, you have to generate SSD file first.
# If you already have a SSD file, you do not need to call this
Generate_SSD_SetID(File.Bed, File.Bim, File.Fam, File.SetID,
  File.SSD, File.Info)
```

## Running SKAT with PLINK

Now you can open SSD and Info file and run SKAT.

```
FAM<-Read_Plink_FAM(File.Fam, Is.binary=FALSE)
y<-FAM$Phenotype
```

```
# To use a SSD file, please open it first.
# After finishing using it, you must close it.
```

```
SSD.INFO<-Open_SSD(File.SSD, File.Info)
```

```
# Number of samples
SSD.INFO$nSample
```

```
# Number of Sets
SSD.INFO$nSets
```

```
obj<-SKAT_Null_Model(y ~ 1, out_type="C")
out<-SKAT.SSD.All(SSD.INFO, obj)
out
```

## Running SKAT with PLINK: Covariates

If you have a plink covariate file, you can use Read\_Plink\_FAM\_Cov file to read both FAM and covariate files.

```
File.Cov<-"./PlinkExample/Example1.Cov"
FAM_Cov<-Read_Plink_FAM_Cov(File.Fam, File.Cov,
  Is.binary=FALSE)
```

```
# First 5 rows
FAM_Cov[1:5,]
```

```
# Run with covariates
X1 = FAM_Cov$X1
X2 = FAM_Cov$X2
y<-FAM_Cov$Phenotype
```

```
obj<-SKAT_Null_Model(y ~ X1 + X2, out_type="C")
out<-SKAT.SSD.All(SSD.INFO, obj)
out
```