Genome-wide Association (GWA) Studies

Data Quality and Methods of Analysis

Nancy J. Cox, Ph.D.

The University of Chicago

Overview

- Practical current issues raised by some of the presentations in the meeting
- Implications for future studies

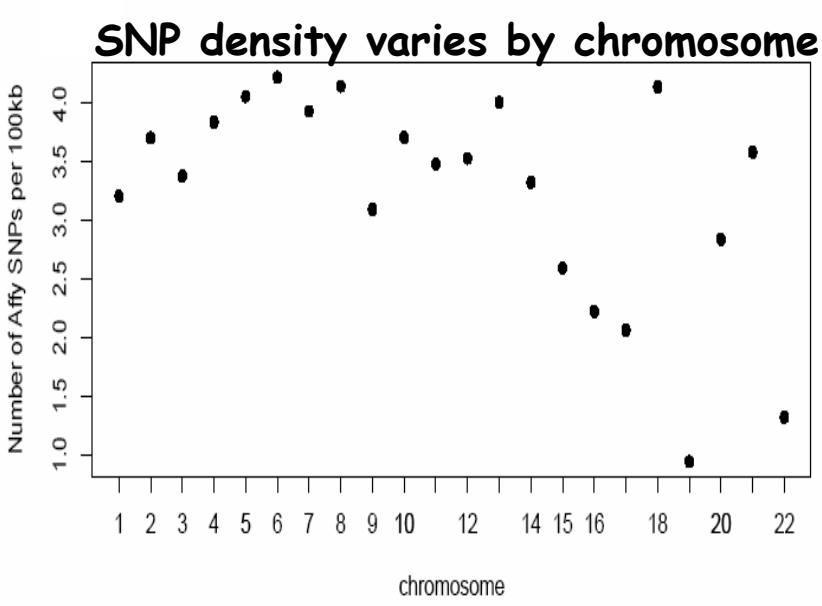
To .cel or not to .cel

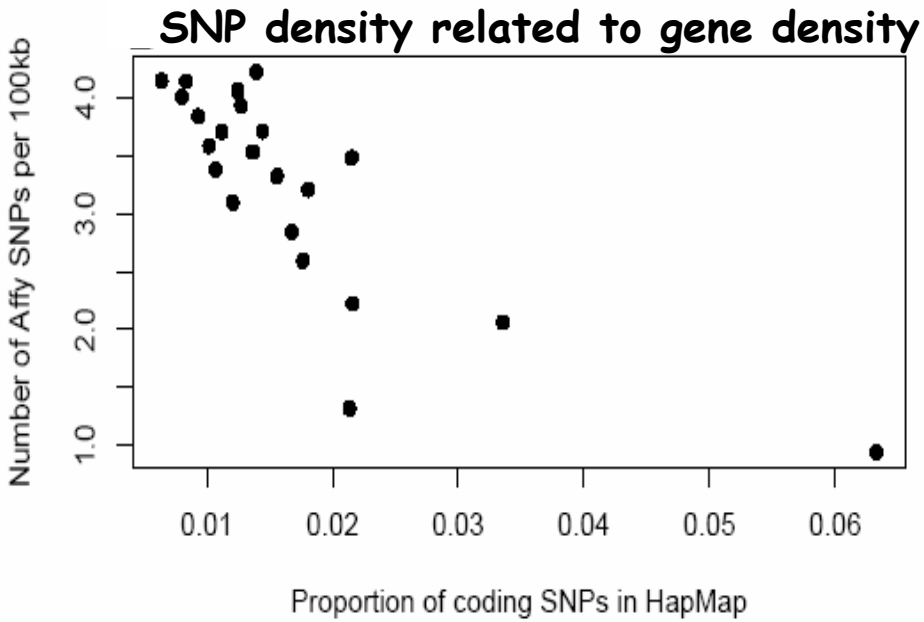
- Allele calling algorithms may stabilize relatively quickly
- Algorithms for identifying structural variation are still being developed and will continue to evolve for some time. You will need intensity data in order to utilize the latest approaches for assaying structural variation.

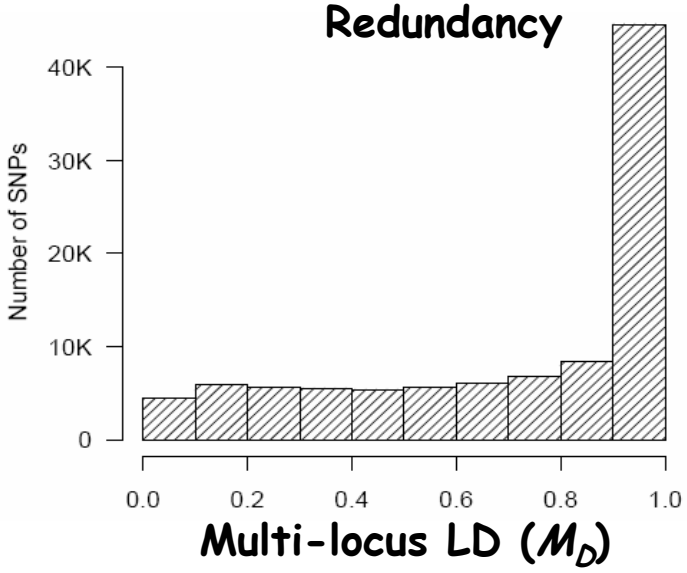
Biases in Coverage and Characteristics of High Throughput Platforms

- Must usually measure that with respect to the HapMap – which is subject to its own biases
- ENCODE is invaluable, and there will be an increasing wealth of resequencing information available for genes that can be used in this context

Affymetrix GeneChip® Human Mapping 100K Set



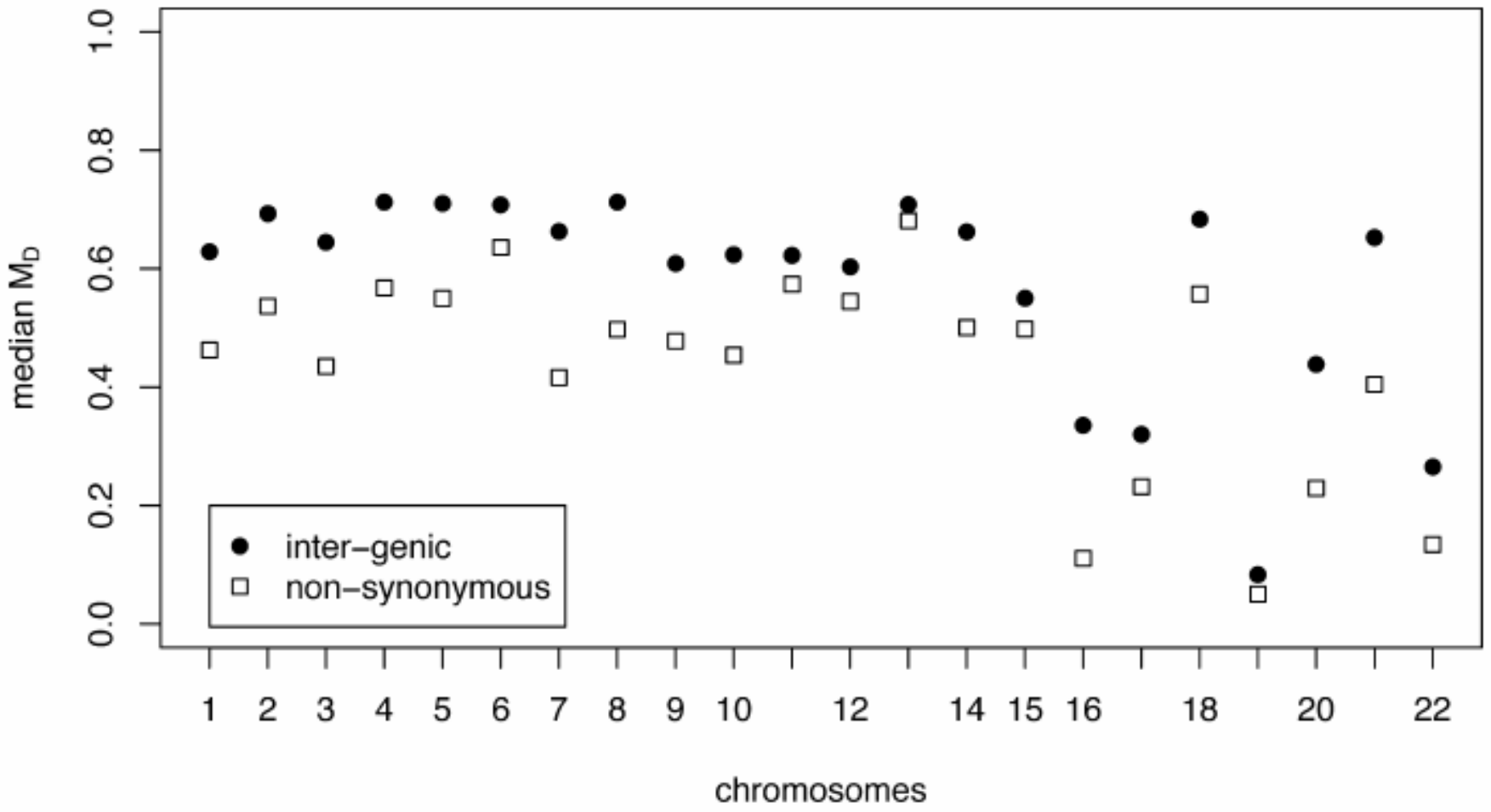




Nicolae et al. (2006) *PLoS Genetics* 2:e67

Genome Coverage:

- 116,204 SNPs
- Mean intermarker distance: 8.5 kb
- Median intermarker distance: 23.6 kb
- Average Heterozygosity: 0.30
- Genome within 100kb of a SNP: 92%
- Genome within 10kb of a SNP: 40%



Median of multi-locus measure of LD for inter-genic (>2 kb from gene) and non-synonymous SNPs

Key Points

- High throughput genotyping requires use of SNPs that can be reliably genotyped
  - Genes with recent duplications, gene families with high sequence homology are often poorly interrogated in high throughput platforms

Functional Patterns

- genes associated at $p \leq 10^{-3}$

Biological Processes (N=18,484)	T2D GWA	100K platform	P-value
Cell adhesion	3.7%	1.7%	.0010
Neuronal Activities	4.1%	2.0%	.0014
Developmental Processes	10.7%	7.7%	.0177
Immunity and defense	1.2%	4.8%	.0002

Molecular Functions (N=12,454)	T2D GWA	100K platform	P-value
Cell adhesion molecule	4.1%	1.6%	.001
Nucleic acid binding	11.7%	8.3%	.033
Proteases	2.5%	1.2%	.039
Defense/immunity protein	0.3%	1.8%	.041

Pathways (N=3,730)	T2D GWA	100K platform	P-value
VEGF signaling	9.3%	1.3%	1.7×10^{-15}
Endothelin signaling	5.6%	1.7%	4.2×10^{-4}
p53 pathway feedback loops 2	3.7%	1.6%	0.04

Intriguing Challenge

- Pathway/annotation information comes at the gene level
- May want to weight by how well the gene is interrogated by the platform
- Can only determine interogation relative to HapMap or resequencing (if available)

Genetic Epidemiology 30: 718–727 (2006)

Testing Untyped Alleles (TUNA)—Applications to Genome-Wide Association Studies

Dan L. Nicolae*

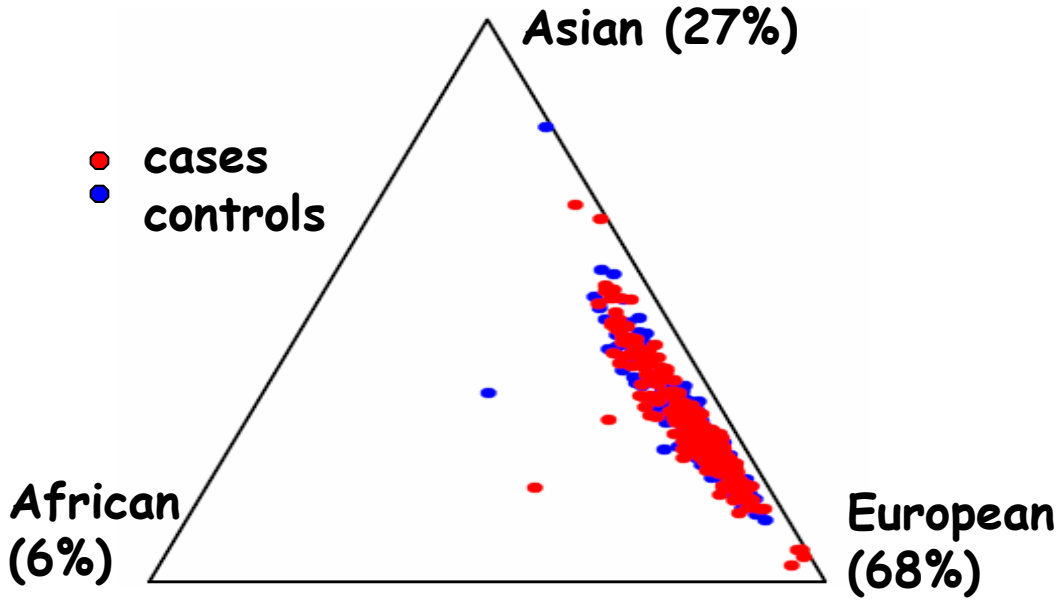
Departments of Medicine and Statistics, The University of Chicago, Chicago, Illinois

Haplotype	A_1 - T - A_2 - A_3 - A_4	Frequency
H_1	1 - 0 - 0 - 0 - 0	0.058
H_2	0 - 1 - 0 - 1 - 0	0.300
H_3	1 - 1 - 0 - 1 - 0	0.050
H_4	1 - 1 - 1 - 0 - 1	0.558
H_5	0 - 1 - 1 - 0 - 1	0.017
H_6	1 - 1 - 0 - 0 - 1	0.017

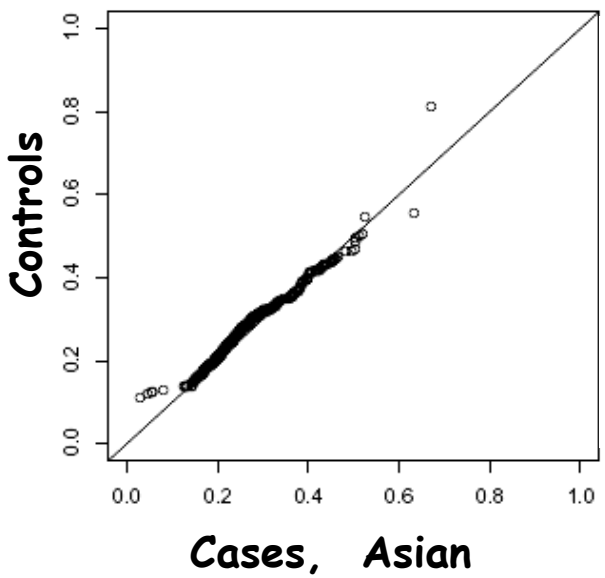
TUNA

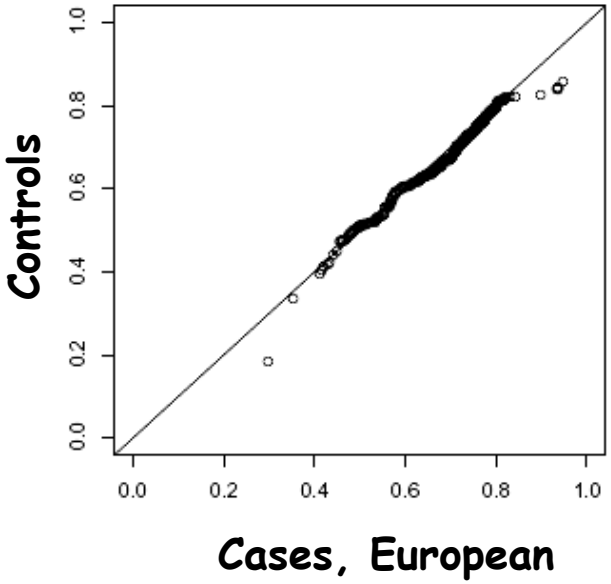
- For high-density screens, can be used for in silico follow-up
  - Set low threshold for "in silico" follow-up of primary screen and TUNA "type" every SNP in the vicinity of a signal to decide which to actually genotype
- Can convert lower density screens to higher density
- Can be used to combine data across platforms (not computationally intensive)

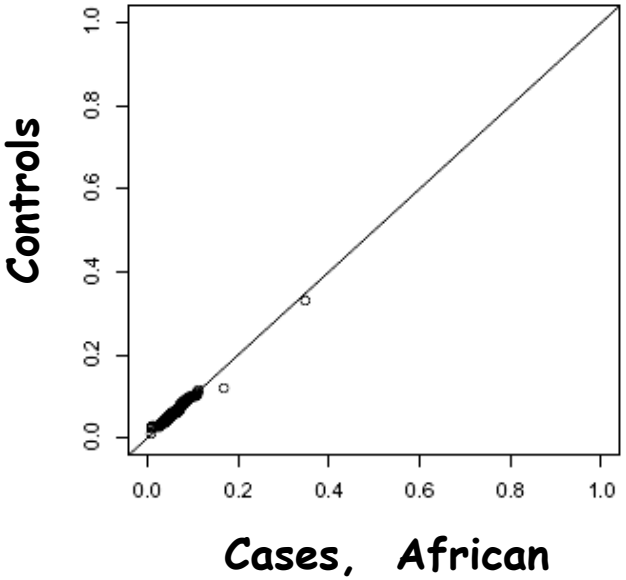
Admixture Proportions, Cases vs. Controls



- no significant difference between cases and controls
- spurious associations are not likely







Linking Platform Genotypes to HapMap Genotypes

- Does require strand orientation
- Information not only is not easy to obtain, but is inconsistent
- About 6700 of the SNPs on the 100K set were ambiguous and could be sorted out only using the Affymetrix web site SNP information on HapMap SNPs

REMINDERS

- Stage 1 screens are often focusing on cases ascertained from families used in previous linkage screens
- Increases power for the stage 1 screen
- HAS MAJOR IMPLICATIONS FOR REPLICATION AND EXTENSION STUDIES

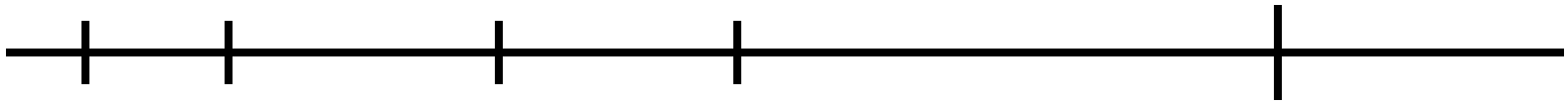
q increasing →

con ran case fam+

Effects on Association

q increasing →

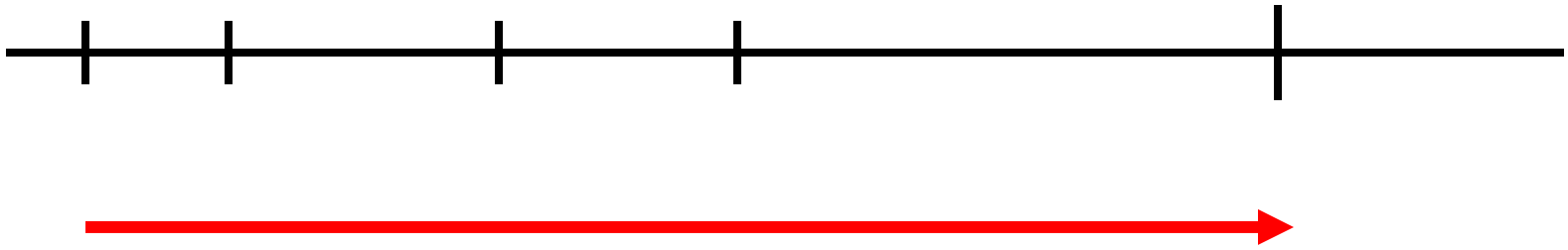
con ran case fam+ link++



Effects on Association

q increasing →

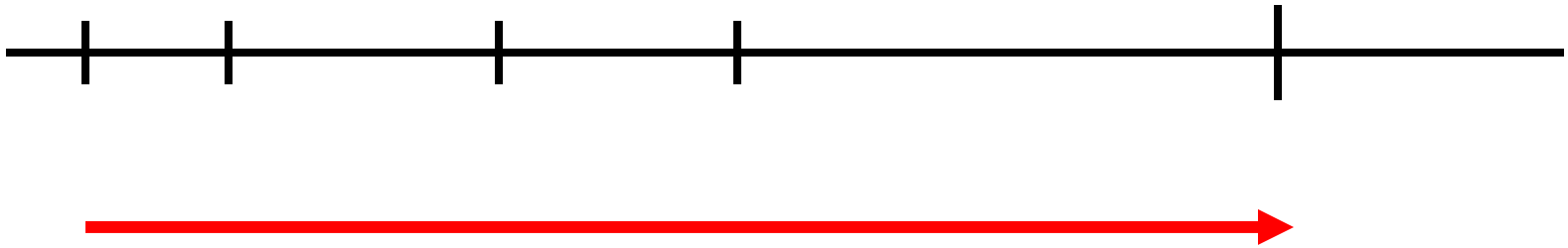
con ran case fam+ link++



Effects on Association

q increasing →

con ran case fam+ link++




Kelly Slater
Photo: Divine

GWA – Short Half-Life Studies?

- Is there merit in doing 10's of GWA studies per phenotype?
- How many are "enough" (and enough for what)?
- Do we measure success in GENES discovered or in larger scale level understanding of the phenotypes?

Colleagues and Collaborators

University of Chicago

Nancy Cox Lab – Geoffrey Hayes, Maggie Ng, Anna Pluzhnikov, Cheri Roe, Jaqui Wittke-Thompson, William Wen, Ying Sun

Dept. of Biochemistry and Molecular Biology – Graeme Bell, Takafumi Tsuchiya, Kazuaki Mayami

Dept. of Human Genetics – Mark Abney, Anna Di Rienzo, Carole Ober, Jonathan Pritchard

Dept. of Statistics – Dan Nicolae, Mary Sara McPeek