National Cancer Institute

The Integration of ENCODE into the Study of the Complexity of Cancer Susceptibility

Stephen J. Chanock, M.D.

Director, Division of Cancer Epidemiology & Genetics

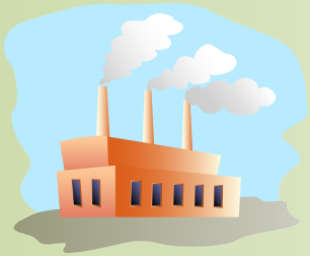
July 1, 2015

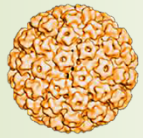
U.S. DEPARTMENT OF HEALTH AND HUMAN SERVICES

National Institutes of Health

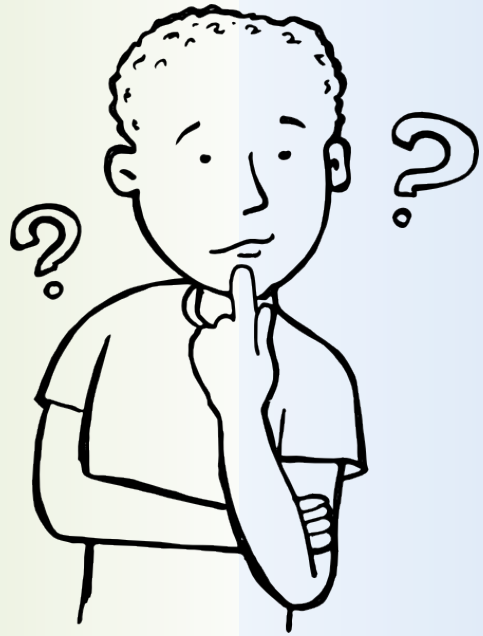
Etiology of Cancer

Environment





Lifestyle



Genetic Susceptibility

Mutations

SNPs

STR's

Chromosomes

(Epigenetics)

100%

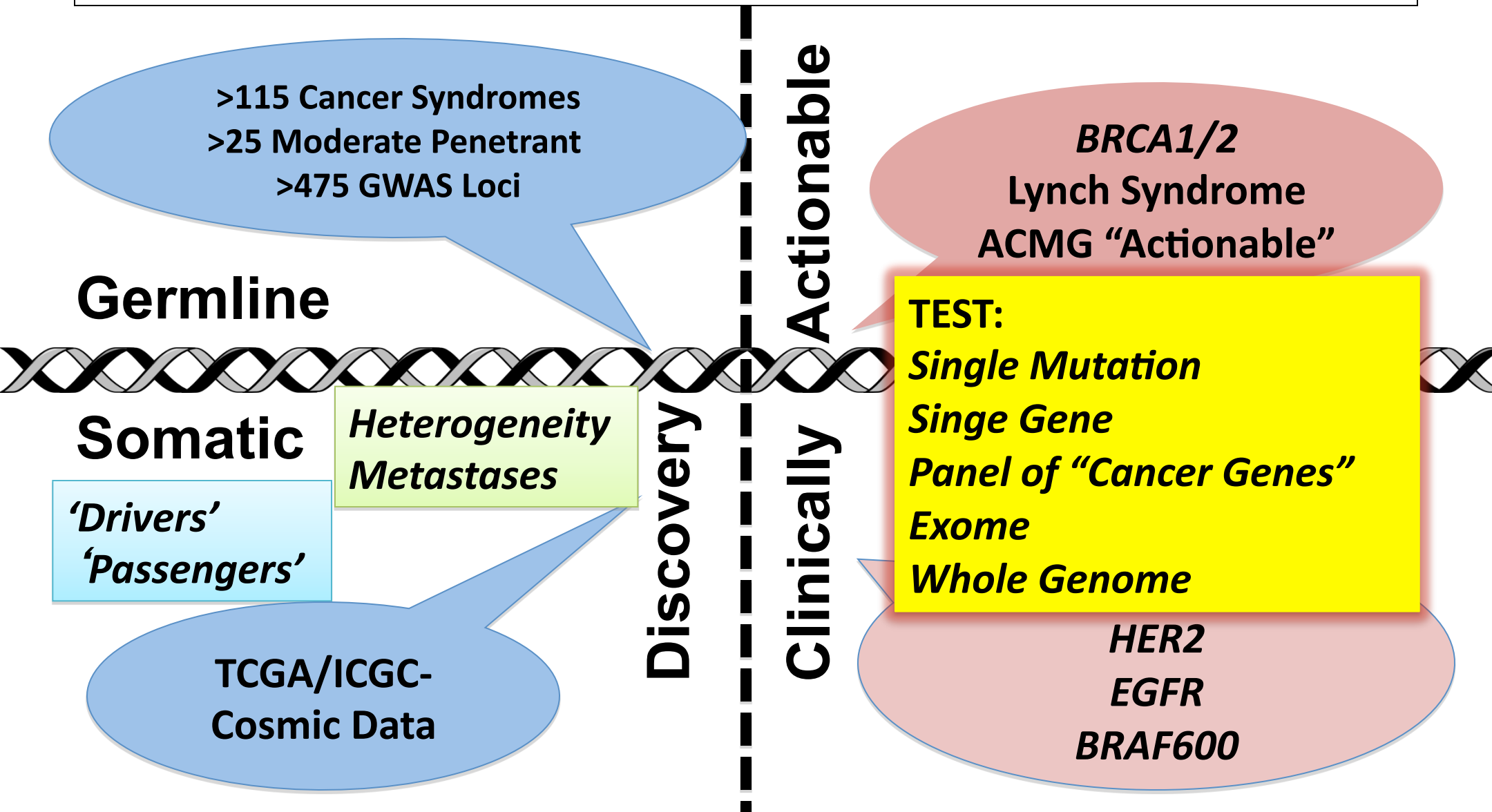
"Triggers"

"Chance"

100%

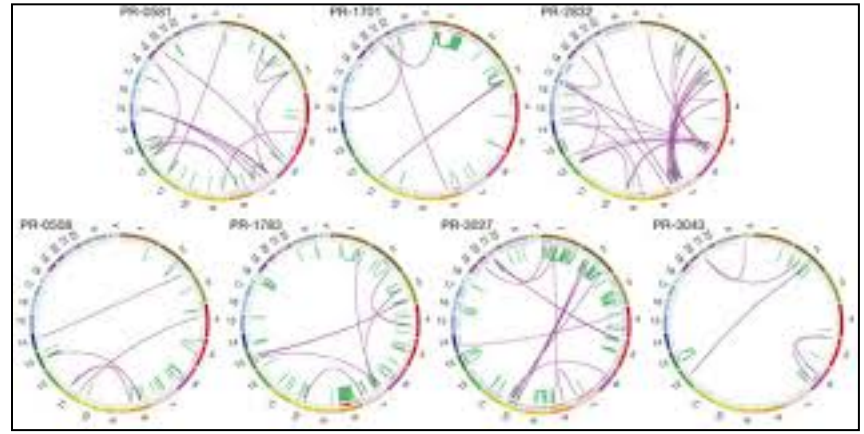
"Set Point"

Cancer Genomics: 4 Spaces



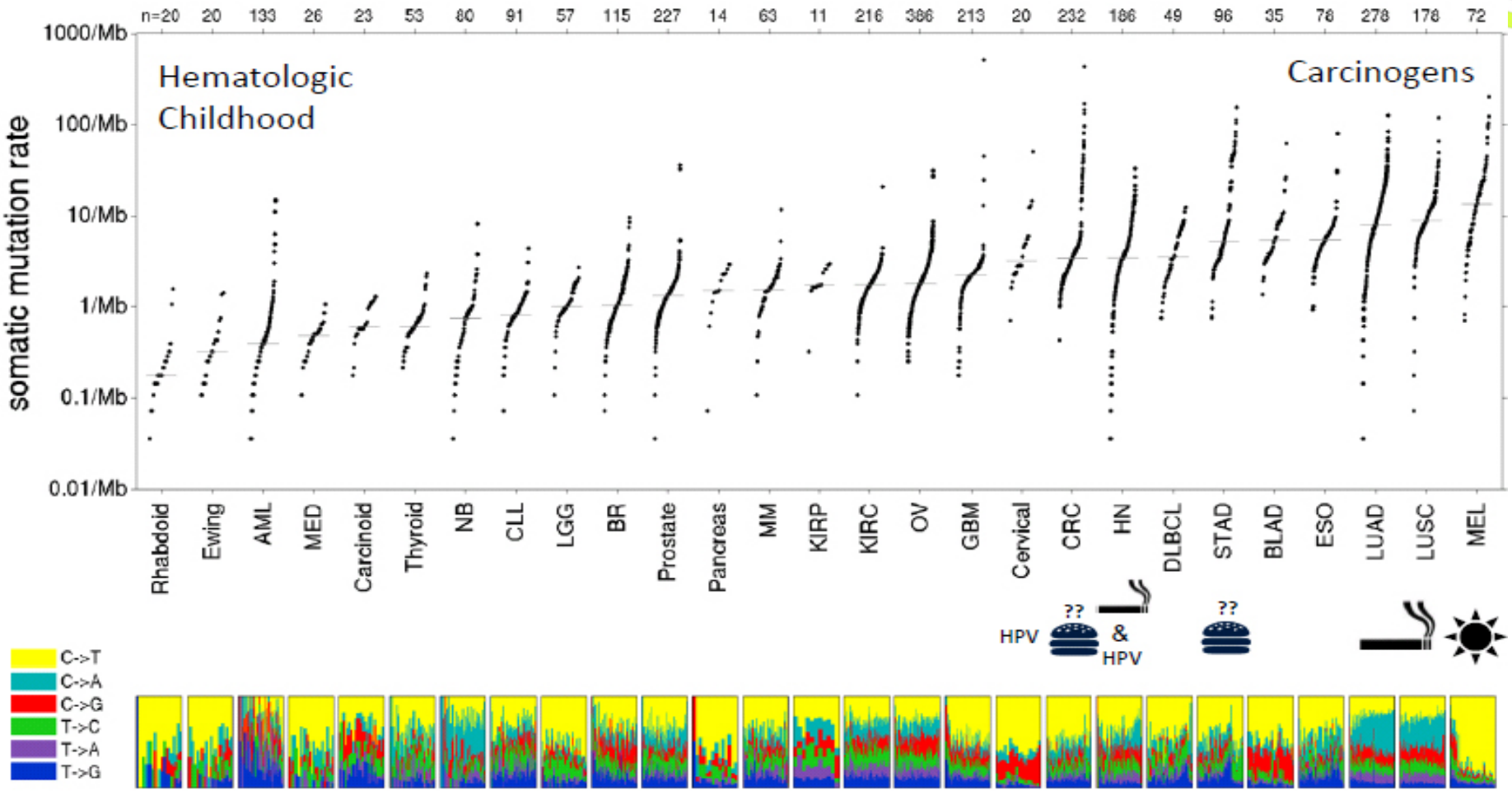
What Happens When There is More than One Genome?

Challenge of Cancer Genomics



TCGA: Driven by the Numbers

But not yet validated in the laboratory….



Value of Frequency in Generating Hypotheses
But, *further* laboratory work is needed…

The Cancer Genome Atlas

Evidence for Heritability of Cancer

1866 Broca observed heritability based on familial breast cancer

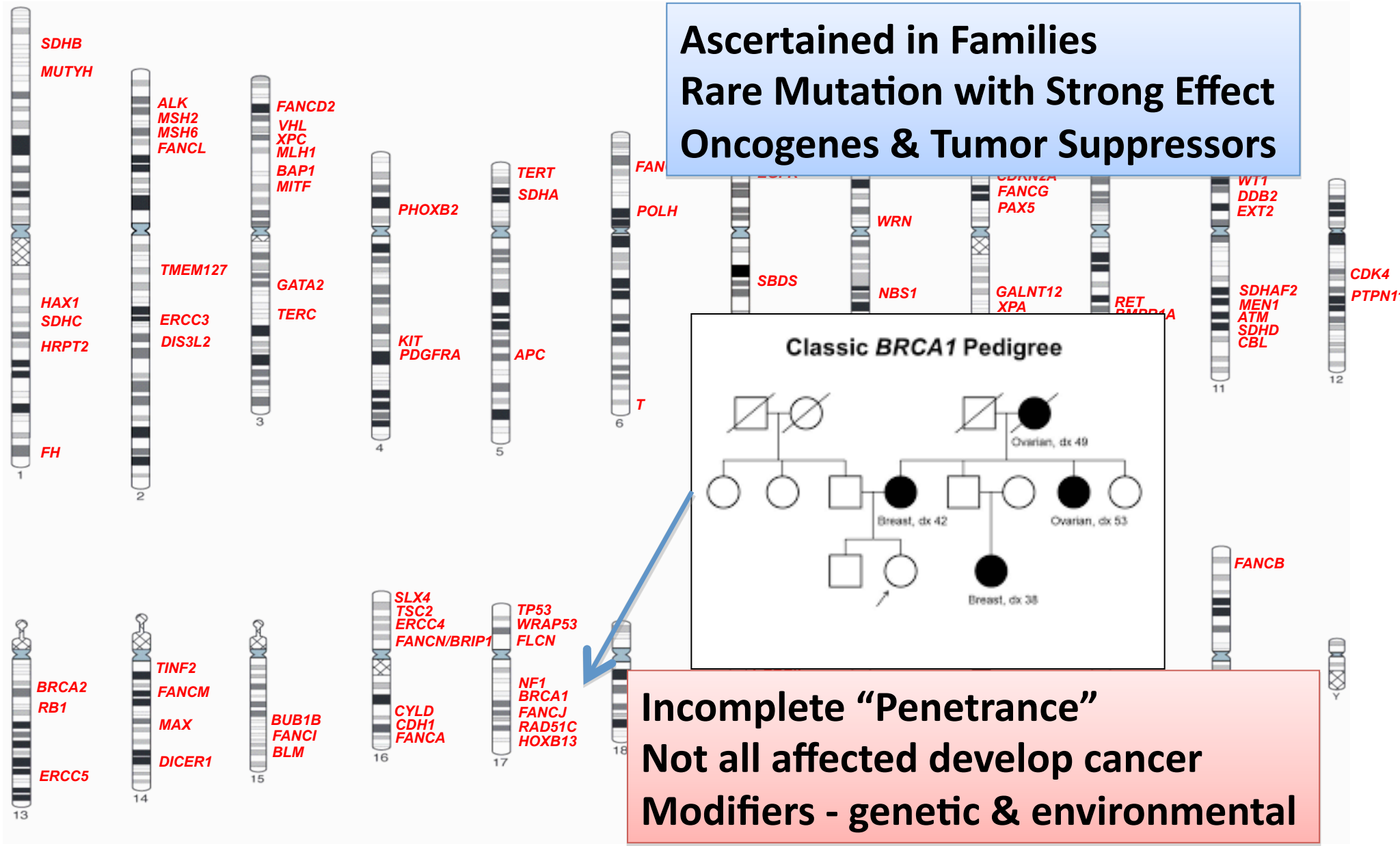
Interim Twin/Family/Sibling studies…

1969 Li-Fraumeni observed familial clustering (*TP53*)

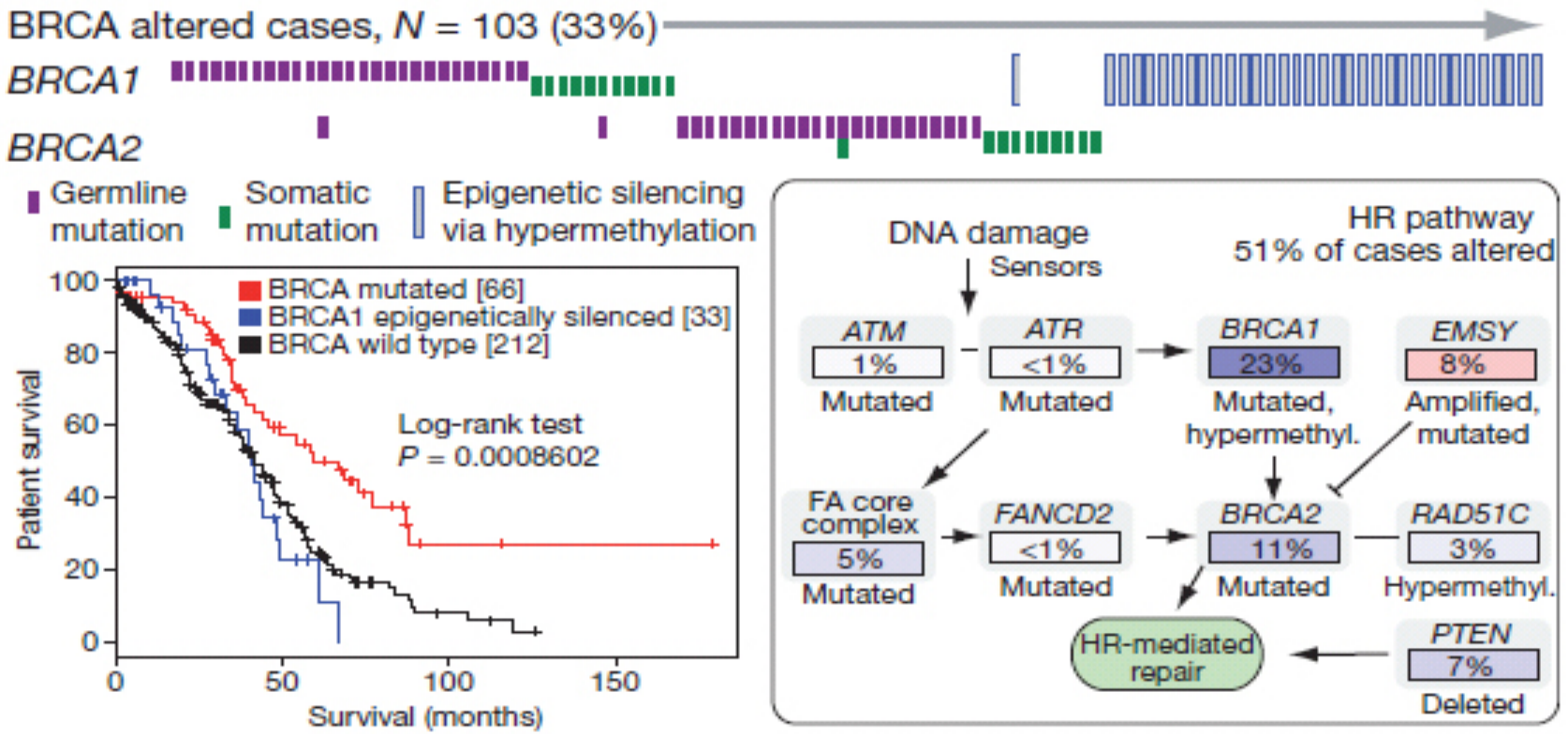
1971 Knudson postulated “two-hit” hypothesis for retinoblastoma

1991 Positional cloning of a familial breast cancer gene (*BRCA1)*

>115 Genes Mutated in Cancer Susceptibility Syndromes



TCGA: Lessons Learned from the Data
Survival Analyses
Impact of germline or somatic mutations



High Penetrance Mutations & Somatic Alterations

Nearly 50%
"High Frequency"
Somatic Mutations

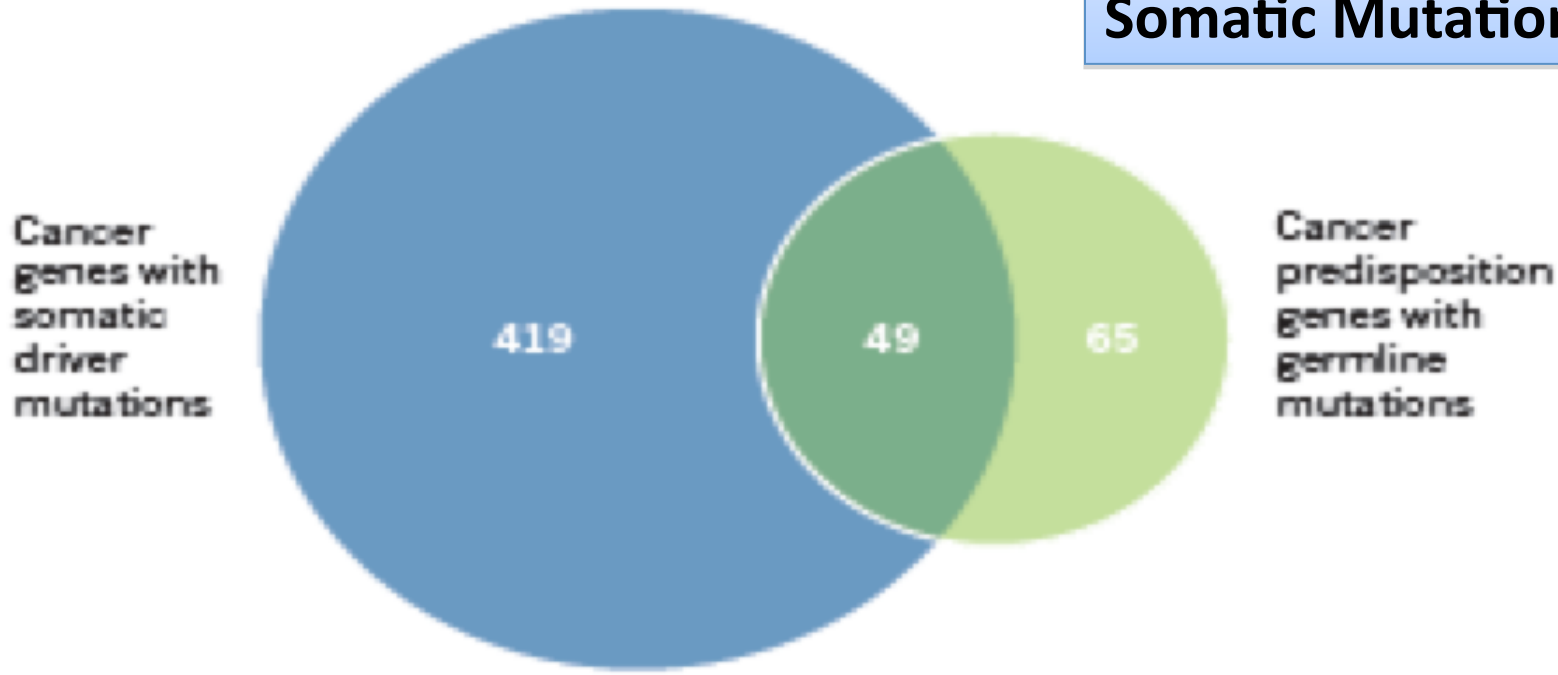


Figure 3 | Overlap between somatically mutated cancer genes and cancer predisposition genes (CPGs). 468 genes with somatic driver mutations in cancers are recorded in the COSMIC database of which 49 are also included within the 114 CPGs.

Rahman *Nature* 2014

Search for Common Variants in Complex Diseases

Reproducible Technology

SNP Microarray Chip

>5 M genotyped SNPs across genome

>30 M imputed SNPs across genome

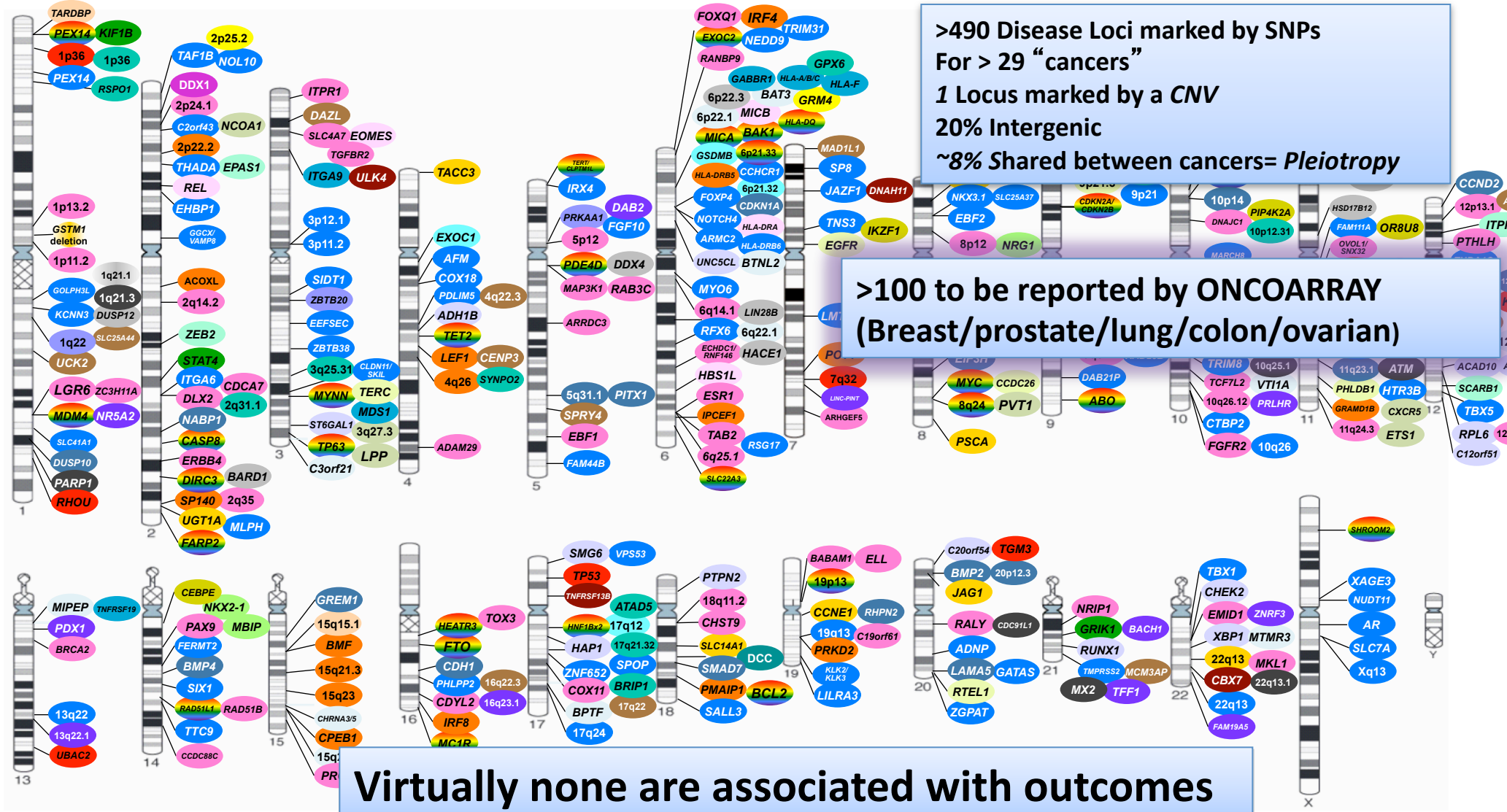
High Concordance > 99.5%/assay

'Markers' across the genome

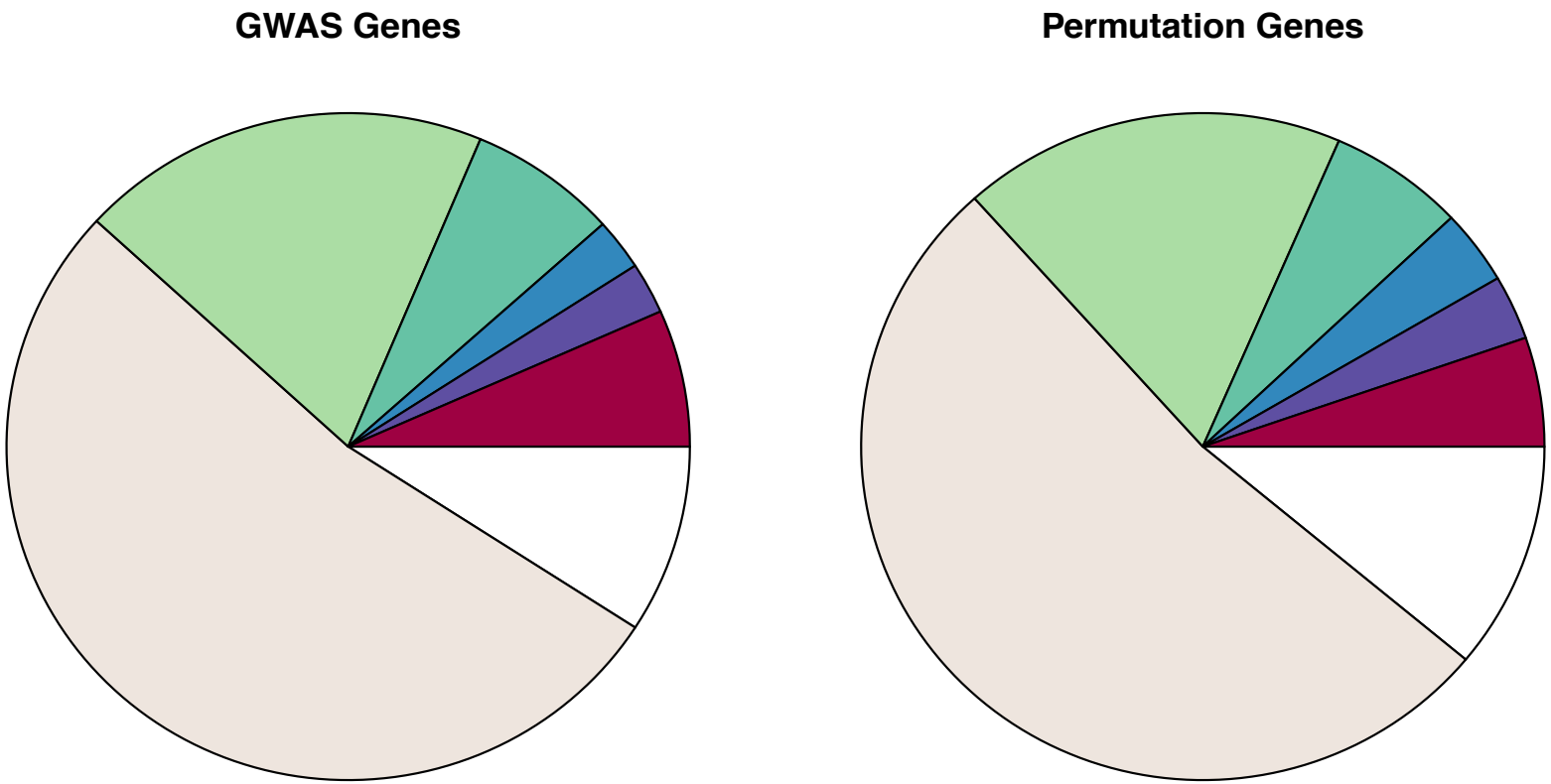
Commitment to Mapping

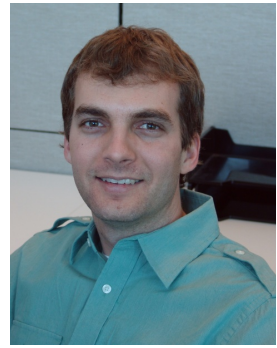
Creates a multiple testing problem

Published Cancer GWAS Etiology Hits: July 2015



GWAS Signals & Somatic Mutations: No Strong Correlation Redundant Pathways- 'NOT Drivers'





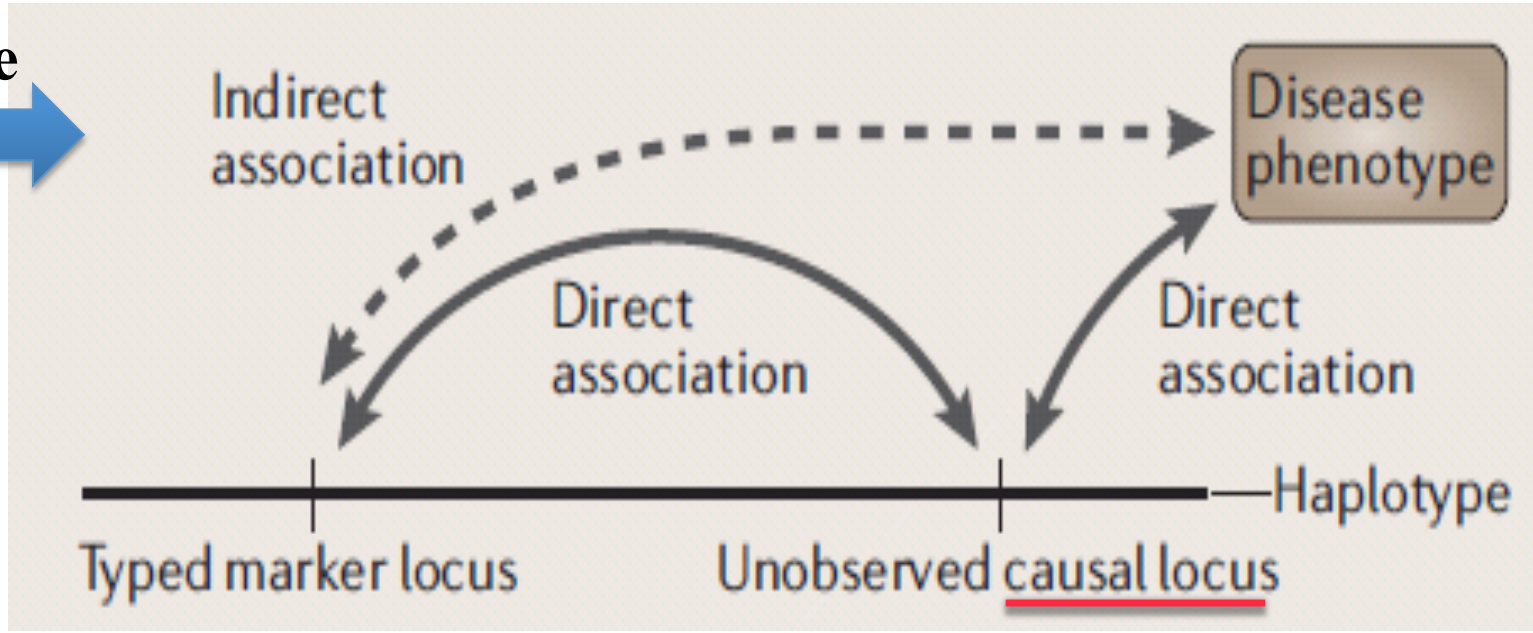
M Machiela in Revision

Interpretation:
Correlation does not imply causation

GWAS designs provide no mechanism to distinguish statistical association from causation

Linkage

GWAS



NGS

Balding, Nature Genetics Review 2006

Can we use ENCODE to prioritize SNPs for follow-up?

OPEN ACCESS Freely available online

PLOS | GENETICS

GPA: A Statistical Approach to Prioritizing GWAS Results by Integrating Pleiotropy and Annotation

Dongjun Chung[1,2☯], Can Yang[1,3,4☯], Cong Li[5], Joel Gelernter[3,6,7,8], Hongyu Zhao[1,5,7,9]*

1 Department of Biostatistics, Yale School of Public Health, New Haven, Connecticut, United States of America, **2** Department of Public Health Sciences, Medical University of South Carolina, Charleston, South Carolina, United States of America, **3** Department of Psychiatry, Yale School of Medicine, New Haven, Connecticut, United States of America, **4** Department of Mathematics, Hong Kong Baptist University, Hong Kong, China, **5** Program in Computational Biology and Bioinformatics, Yale University, New Haven, Connecticut, United States of America, **6** VA CT Healthcare Center, West Haven, Connecticut, United States of America, **7** Department of Genetics, Yale School of Medicine, West Haven, Connecticut, United States of America, **8** Department of Neurobiology, Yale School of Medicine, New Haven, Connecticut, United States of America, **9** VA Cooperative Studies Program Coordinating Center, West Haven, Connecticut, United States of America

Abstract

Results from Genome-Wide Association Studies (GWAS) have shown that complex diseases are often affected by many genetic variants with small or moderate effects. Identifications of these risk variants remain a very challenging problem. There is a need to develop more powerful statistical methods to leverage available information to improve upon traditional approaches that focus on a single GWAS dataset without incorporating additional data. In this paper, we propose a novel statistical approach, GPA (Genetic analysis incorporating Pleiotropy and Annotation), to increase statistical power to identify risk variants through joint analysis of multiple GWAS data sets and annotation information because: (1) accumulating evidence suggests that different complex diseases share common risk bases, i.e., pleiotropy; and (2) functionally annotated variants have been consistently demonstrated to be enriched among GWAS hits. GPA can integrate multiple GWAS datasets and functional annotations to seek association signals, and it can also perform hypothesis testing to test the presence of pleiotropy and enrichment of functional annotation. Statistical inference of the model parameters and SNP ranking is achieved through an EM algorithm that can handle genome-wide markers efficiently. When we applied GPA to jointly analyze five psychiatric disorders with annotation information, not only did GPA identify many weak signals missed by the traditional single phenotype analysis, but it also revealed relationships in the genetic architecture of these disorders. Using our hypothesis testing framework, statistically significant pleiotropic effects were detected among these psychiatric disorders, and the markers annotated in the central nervous system genes and eQTLs from the Genotype-Tissue Expression (GTEx) database were significantly enriched. We also applied GPA to a bladder cancer GWAS data set with the ENCODE DNase-seq data from 125 cell lines. GPA was able to detect cell lines that are biologically more relevant to bladder cancer. The R implementation of GPA is currently available at http://dongjunchung.github.io/GPA/.

Citation: Chung D, Yang C, Li C, Gelernter J, Zhao H (2014) GPA: A Statistical Approach to Prioritizing GWAS Results by Integrating Pleiotropy and Annotation. PLoS Genet 10(11): e1004787. doi:10.1371/journal.pgen.1004787

Editor: Hua Tang, Stanford University, United States of America

Received February 17, 2014; **Accepted** September 29, 2014; **Published** November 13, 2014

Funding: This study was supported by National Institutes of Health grants RC2 DA028909, R01 DA12690, R01 DA12849, R01 DA18432, R01 AA11330, R01 AA017535, P50 AA12870, R01 GM59507, R01 DA030976, UL1 RR024139, and the VA Cooperative Studies Program of the Department of Veterans Affairs, Office of Research and Development. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* Email: hongyu.zhao@yale.edu

☯ These authors contributed equally to this work.

Introduction

Hundreds of genome-wide association studies (GWAS) have been conducted to study the genetic bases of complex human traits. As of January, 2014, more than 12,000 single-nucleotide polymorphisms (SNPs) have been reported to be significantly associated with at least one complex trait (see the web resource of GWAS catalog [1] http://www.genome.gov/gwastudies/). Despite of these successes, these significantly associated SNPs can only explain a small portion of genetic contributions to complex traits/diseases [2]. For example, human height is a highly heritable trait whose heritability is estimated to be around 80%, i.e., 80% of variation in height within the same population can be attributed to genetic effects [3]. Based on large-scale GWAS, about 180 SNPs have been reported to be significantly associated with human height [4]. However, these loci together only explain about 5-10% of variation in height [2,4,5]. This phenomenon is referred to as the "missing heritability" [2,6,7].

Identifying the source of this missing heritability has drawn much attention from researchers, and progress has been made towards explaining the apparent discrepancy. The role of a much greater-than-expected set of common variants (minor allele frequency (MAF) ≥ 0.01) has been shown to be critical in explaining the phenotypic variance [8]. Instead of only using genome-wide significant SNPs, Yang et al. [9] reported that, by using all genotyped common SNPs, 45% of the variance for human height can be explained. This result suggests that a large proportion of the heritability is not actually missing: given the limited sample size, many individual effects of genetic markers are too weak to pass the genome-wide significance, and thus those variants remain undiscovered. So far, people have found similar genetic architectures for many other complex traits [10], such as

YES

&

NO

OPEN ACCESS Freely available online

PLOS | GENETICS

Comprehensive Functional Annotation of 77 Prostate Cancer Risk Loci

Dennis J. Hazelett[1]*, Suhn Kyong Rhie[1], Malaina Gaddis[2], Chunli Yan[1], Daniel L. Lakeland[3], Simon G. Coetzee[4], Ellipse/GAME-ON consortium[5¶], Practical consortium[6¶], Brian E. Henderson[5], Houtan Noushmehr[4], Wendy Cozen[7], Zsofia Kote-Jarai[6], Rosalind A. Eeles[6,8], Douglas F. Easton[9], Christopher A. Haiman[5], Wange Lu[10], Peggy J. Farnham[2], Gerhard A. Coetzee[1]*

1 Departments of Urology and Preventive Medicine, Norris Cancer Center, University of Southern California Keck School of Medicine, Los Angeles, California, United States of America, **2** Department of Biochemistry and Molecular Biology, Keck School of Medicine, University of Southern California, Los Angeles, California, United States of America, **3** Sonny Astani Department of Civil and Environmental Engineering, University of Southern California, Los Angeles, California, United States of America, **4** Department of Genetics, University of São Paulo, Ribeirão Preto, Brazil, **5** Department of Preventive Medicine, Norris Cancer Center, University of Southern California Keck School of Medicine, Los Angeles, California, United States of America, **6** The Institute of Cancer Research, Sutton, United Kingdom, **7** USC Keck School of Medicine, Norris Comprehensive Cancer Center, University of Southern California, Los Angeles, California, United States of America, **8** Royal Marsden National Health Service (NHS) Foundation Trust, London and Sutton, United Kingdom, **9** Centre for Cancer Genetic Epidemiology, Department of Oncology, University of Cambridge, Cambridge, United Kingdom, **10** Eli and Edythe Broad Center for Regenerative Medicine and Stem Cell Research, Department of Biochemistry and Molecular Biology, Keck School of Medicine, University of Southern California, Los Angeles, California, United States of America

Abstract

Genome-wide association studies (GWAS) have revolutionized the field of cancer genetics, but the causal links between increased genetic risk and onset/progression of disease processes remain to be identified. Here we report the first step in such an endeavor for prostate cancer. We provide a comprehensive annotation of the 77 known risk loci, based upon highly correlated variants in biologically relevant chromatin annotations— we identified 727 such potentially functional SNPs. We also provide a detailed account of possible protein disruption, microRNA target sequence disruption and regulatory response element disruption of all correlated SNPs at $r^2 \geq 0.5$. 88% of the 727 SNPs fall within putative enhancers, and many alter critical residues in the response elements of transcription factors known to be involved in prostate biology. We define as risk enhancers those regions with enhancer chromatin biofeatures in prostate-derived cell lines with prostate-cancer correlated SNPs. To aid the identification of these enhancers, we performed genomewide ChIP-seq for H3K27-acetylation, a mark of actively engaged enhancers, as well as the transcription factor TCF7L2. We analyzed in depth three variants in risk enhancers, two of which show significantly altered androgen sensitivity in LNCaP cells. This includes rs4907792, that is in linkage disequilibrium ($r^2 = 0.91$) with an eQTL for NUDT11 (on the X chromosome) in prostate tissue, and rs10486567, the index SNP in intron 3 of the JAZF1 gene on chromosome 7. Rs4907792 is within a critical residue of a strong consensus androgen response element that is interrupted in the protective allele, resulting in a 56% decrease in its androgen sensitivity, whereas rs10486567 affects both NKX3-1 and FOXA-AR motifs where the risk allele results in a 39% increase in basal activity and a 28% fold-increase in androgen stimulated enhancer activity. Identification of such enhancer variants and their potential target genes represents a preliminary step in connecting risk to disease process.

Citation: Hazelett DJ, Rhie SK, Gaddis M, Yan C, Lakeland DL, et al. (2014) Comprehensive Functional Annotation of 77 Prostate Cancer Risk Loci. PLoS Genet 10(1): e1004102. doi:10.1371/journal.pgen.1004102

Editor: Vivian G. Cheung, University of Michigan, United States of America

Received October 1, 2013; **Accepted** November 14, 2013; **Published** January 30, 2014

Funding: The work reported here was funded by the National Institutes of Health (NIH) [CA109147, U19CA148537 and U19CA148107 to GAC; 5T32CA009320-27 to HN and NIDH/NHGRI U54HG006996 to PJF] and David Mazzone Awards Program (GAC) and 5T32GM067587 for MG. The scientific development and funding of this project were in part supported by the Genetic Associations and Mechanisms in Oncology (GAME-ON): a NCI Cancer Post-GWAS Initiative. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: hazelett@usc.edu (DJH); coetzee@usc.edu (GAC)

¶ Membership of the Ellipse/GAME-ON consortium and the Practical consortium is provided in the Acknowledgments.

Introduction

The basic goal of research into human genetics is to connect variation at the genetic level with variation in organismal and cellular phenotype. Until recently, inferences about such connections have been limited to the kind associated with heritable disorders and developmental syndromes. Such variations often turn out to be the result of disruptions to protein coding sequences of critical enzymes for an affected pathway. Recent advances in genomics and medicine have begun to illuminate a sea of variation of a more subtle variety, not always the result of mutation of protein coding sequences. In particular, genome-wide association studies (GWAS) have identified thousands of variants associated with hundreds of disease traits [1]. These variants, typically encoded by single nucleotide polymorphisms (SNPs), are given landmark status and called 'index-SNPs' (they are also frequently referred to in the literature as 'tag-SNPs') as the reference for disease or phenotype association in that region. The vast majority

GWAS Signals

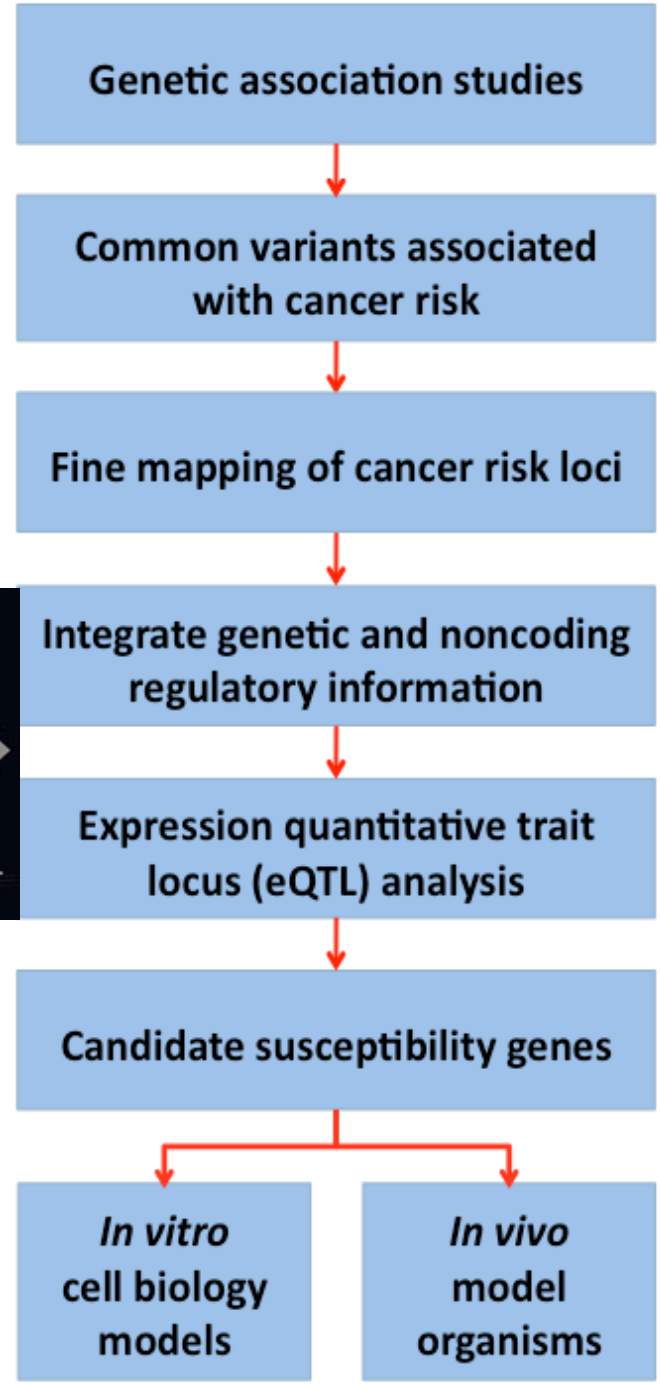
“One by One” Investigation

Insights into Biology

Perturbations of Redundant Pathways/Processes

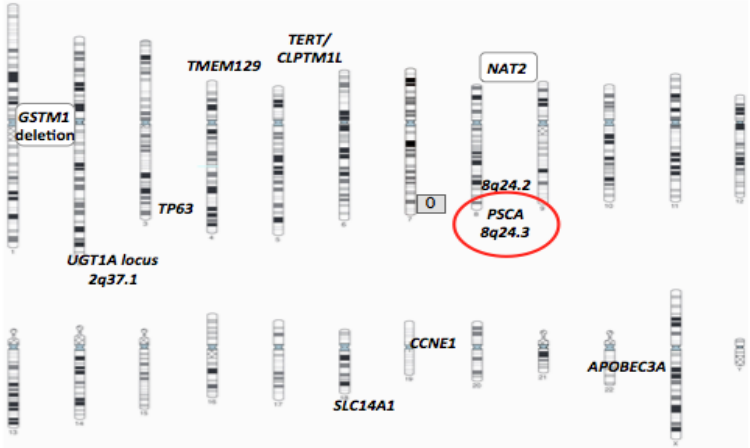
Not Causal

Instead…..
Functional Contribution

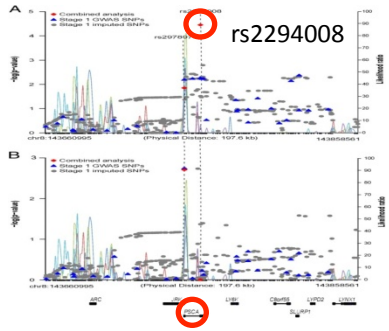


Bladder Cancer GWAS Discovery ➡ Clinical Trial
Target Prostate Stem Cell Antigen (PSCA)
PI: M Prokunina-Olsson

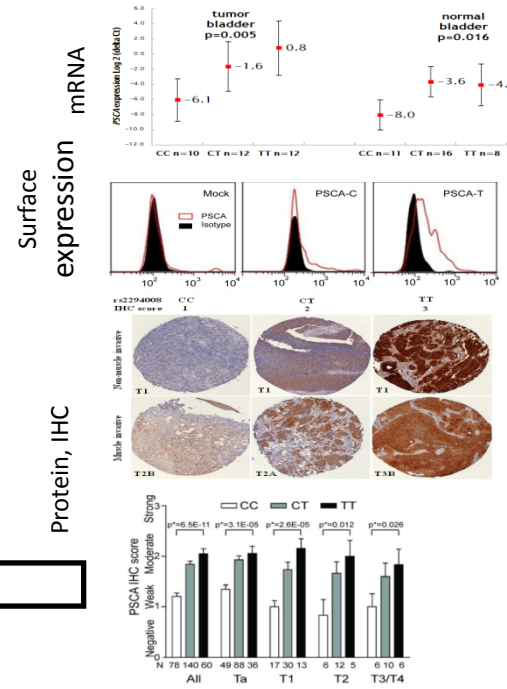
PSCA, 8q24.3
Discovered 2009



Fine Mapping
Genotyping & Imputation



Functional Studies
Risk allele T
⇧mRNA expression

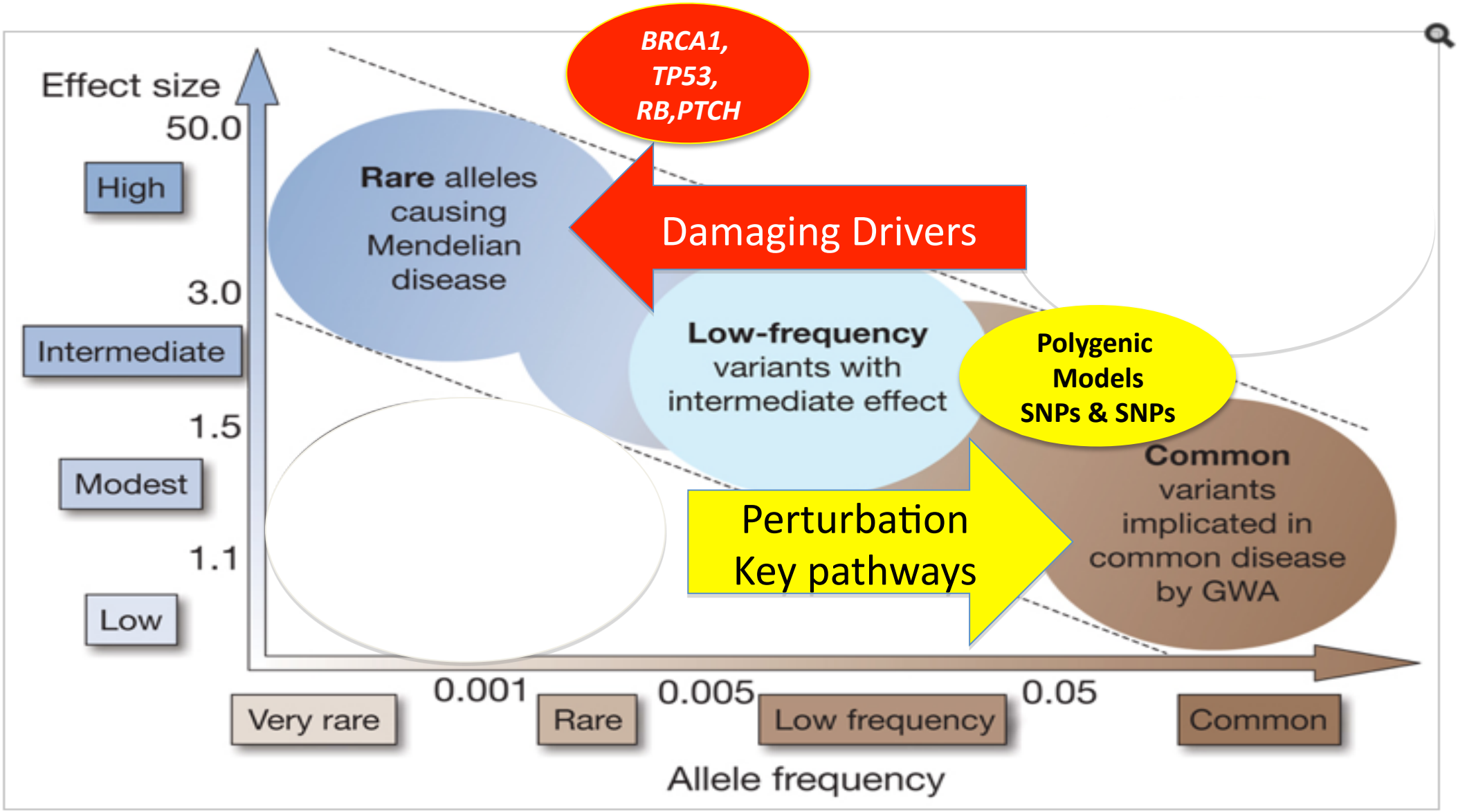


Translational application
rs2294008 predicts PSCA expression in tumors
Anti-PSCA therapy?

Possible Clinical Trial
Therapeutic humanized anti-PSCA antibody for bladder cancer

Mila Prokunina-Olsson

Architecture of Genetic Susceptibility of Cancer Defining Distinct Spaces



What Fraction of the Polygenic Component Contributes to Each Cancer?

- 13 cancer GWAS
  - 49,492 cases
  - 34,131 controls (often used in > 1 study)
- Use genotyped SNPs
- **Explains 10-50% of variability on the liability scale**

Across Cancer Types

Table 2 Estimates of first-degree familial relative risk from familial registries and GWAS

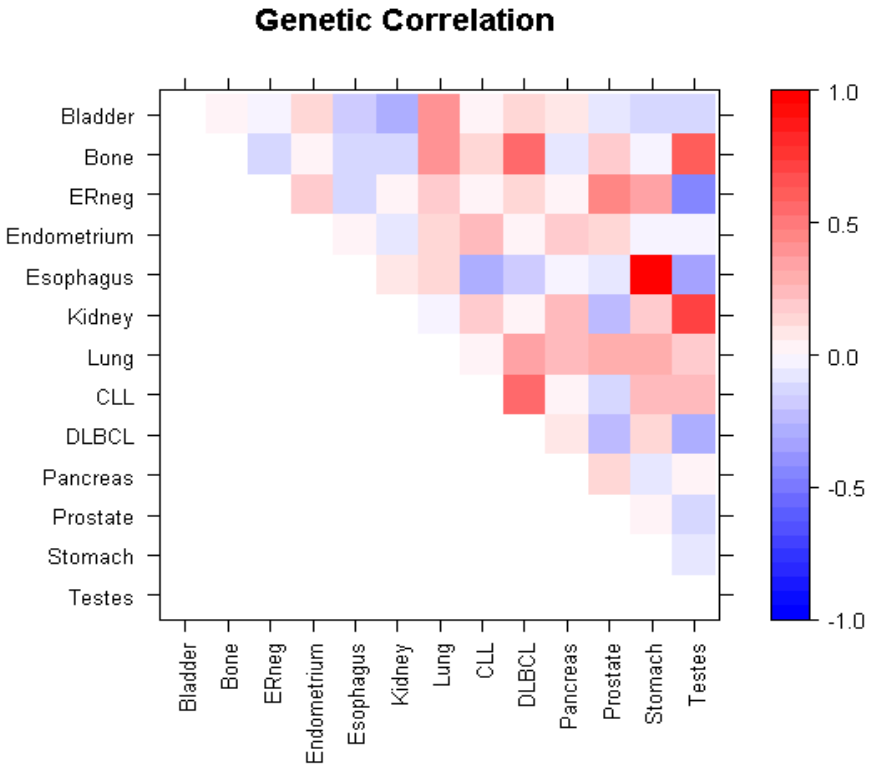
Cancer	Sweden			Iceland[A]	Utah[A]	GWAS
	All 1st Degree Relationships	Parent/Child	Sibling			
Bladder	1.69 (1.33-2.14)	1.53 (1.16–1.99)	3.30 (1.70–5.78)	1.68 (1.39-2.05)	1.8 (1.4–2.3)	1.37 (1.25-1.50)
Breast (ER-)						1.28 (0.98-1.63)
Endometrium	3.02 (2.33-3.92)	2.85 (2.08–3.82)	3.97 (1.97–7.13)	1.86 (1.31-2.62)	1.4 (1.1–1.8)	1.56 (1.25-1.92)
Esophagus		2.14 (0.77-4.70)		2.09 (1.30-3.31)	1.3 (0.2–10.0)	1.63[B] (1.27-2.05)
Glioma	1.67 (1.43-1.94)		3.31 (2.08-5.02)	1.41 (0.74-2.40)	2.3 (0.99-4.5)	1.19 (0.91-1.54)
Kidney	1.78 (1.33-2.39)	1.52 (1.06–2.11)	4.52 (2.15–8.35)	2.30 (1.89-2.80)	2.1 (1.3–3.5)	1.54 (1.07-2.13)
Lung						
European	1.70 (1.42-2.05)	1.64 (1.34–2.00)	2.61 (1.29–4.68)	2.00 (1.83-2.16)	2.4 (1.9–3.0)	1.42 (1.28-1.57)
Asian						1.31[B] (1.16-1.46)
Lymphoma						
CLL	8.5 (6.1-11.7)				6.1 (4.75-7.65)	2.28 (1.86-2.77)
DLBCL	9.8 (3.1-31.0)					1.40 (1.15-1.68)
Osteosarcoma						12.7 (8.27-19.1)
Pancreas		1.68 (1.16–2.35)		2.33 (1.83-2.96)	2.1 (1.3–3.2)	1.35 (1.12-1.62)
Prostate	2.75 (2.32-3.25)	2.71 (2.26–3.22)	4.91 (1.28–12.7)	1.89 (1.75-2.01)	2.1 (1.9–2.2)	1.51 (1.32-1.72)
Stomach	1.99 (1.47-2.71)	1.72 (1.19–2.40)	8.82 (3.50–18.3)	1.90 (1.74-2.05)	2.0 (1.1–3.7)	1.94[B] (0.95-3.49)
Testes	7.07 (5.34-9.37)	4.31 (2.05–7.95)	8.50 (6.01–11.7)	3.52 (1.18-7.37)	1.8 (0.4–8.6)	3.09 (1.41-6.05)

Sampson - under review

Shared Heritability from GWAS

13 Distinct Cancers

(49,492 cases and 34,131 shared controls)



Shared factors:

Some expected

- Testes & Kidney
- CLL & DLBCL
- Bladder & Lung (smoking)

Others not…

- DLBCL & Osteosarcoma

Josh Sampson
+ 280 co-authors

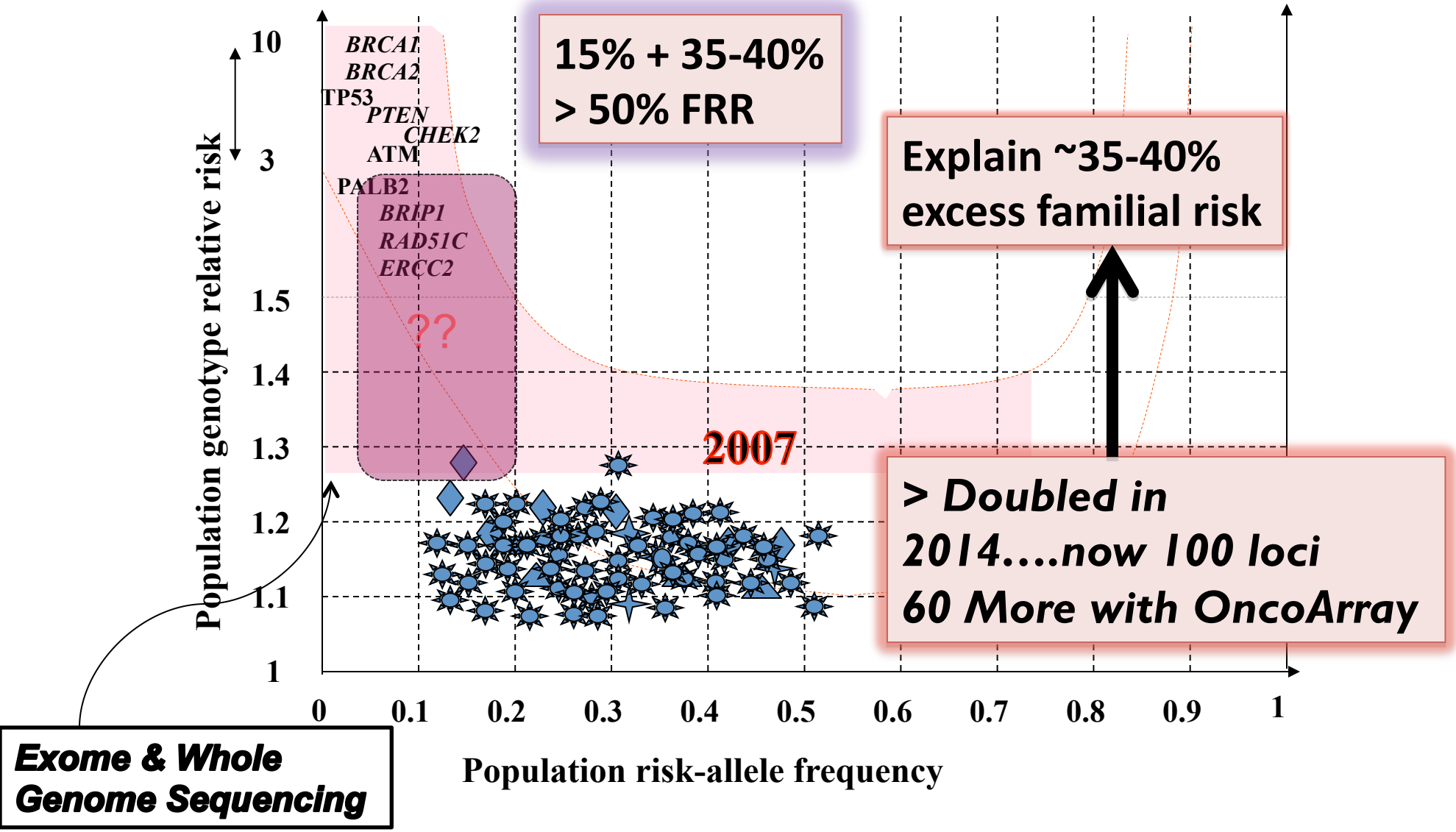
Prediction is difficult,
Especially about the future.

Yogi Berra

Dan Quayle

Niels Bohr

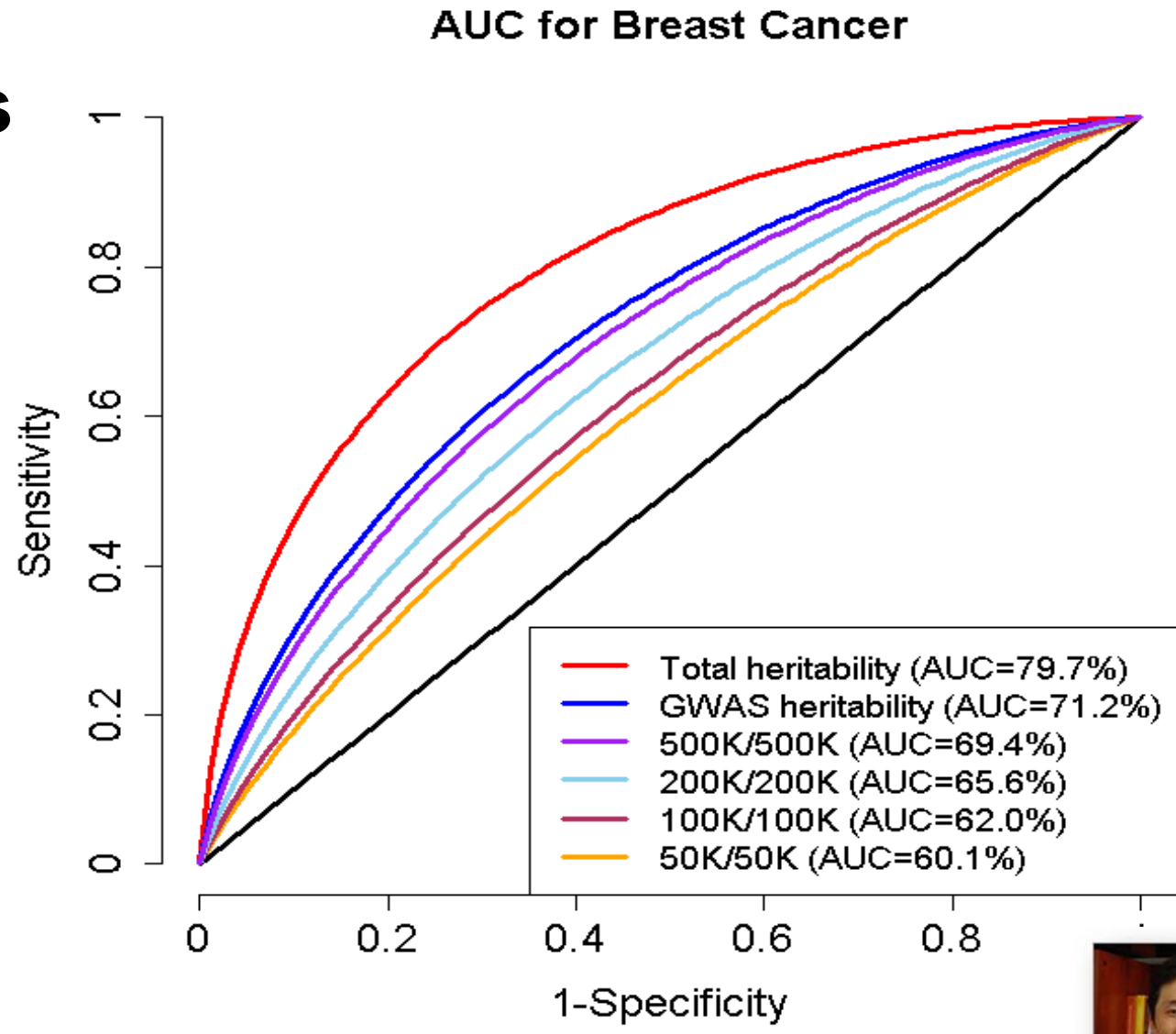
Genetic Predisposition to Breast Cancer
European Population



2015 Projections For General Population

So far, we explain ~35-40% familial risk

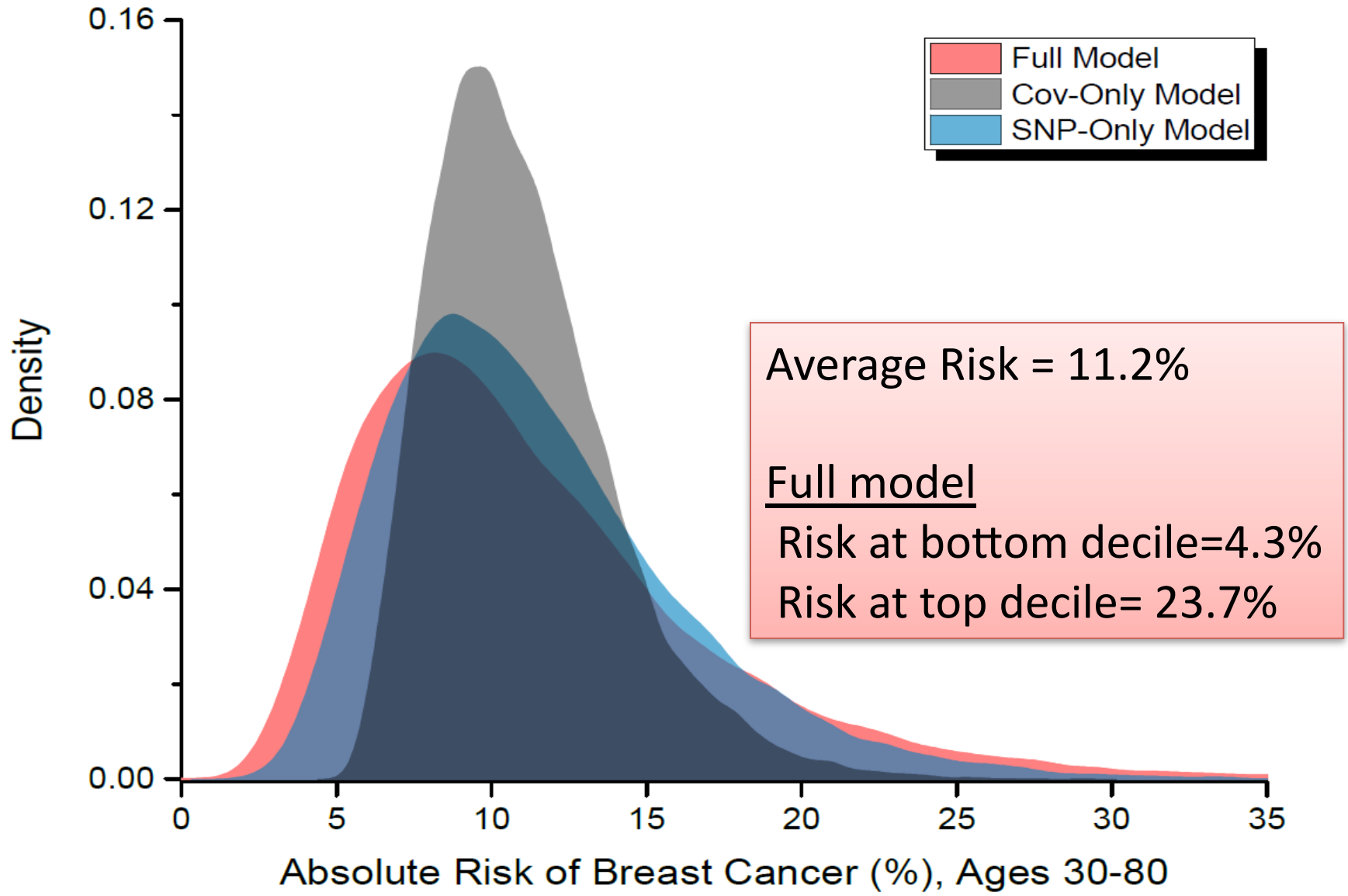
Change in Absolute Risk For Screening



- Total heritability corresponds to 2-fold sibling relative risk.
- GWAS Heritability: ~3000 SNPs explain 1.4 fold sibling relative risk

JuHyun Park

Projected Distribution of Absolute Lifetime Risk (Age 30-80) of Breast Cancer for US Caucasian Women



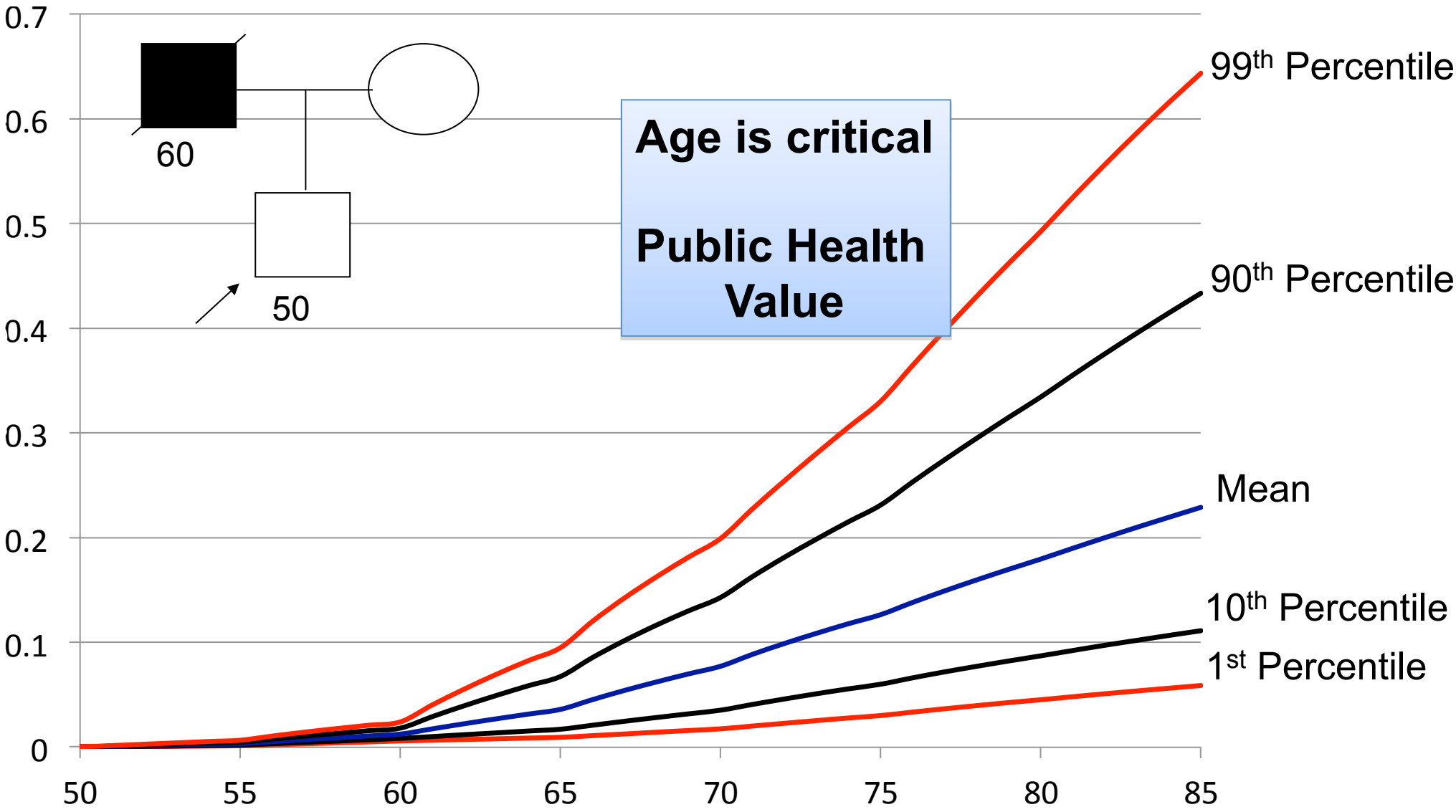
Average Risk = 11.2%

Full model

Risk at bottom decile=4.3%

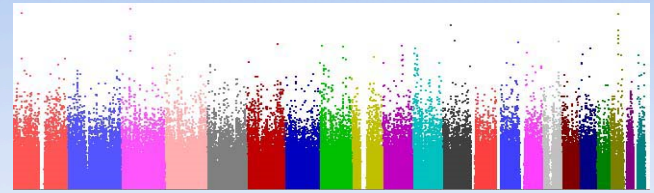
Risk at top decile= 23.7%

Predicted Prostate Cancer Risk by SNP Profile Distribution (76 SNPs)



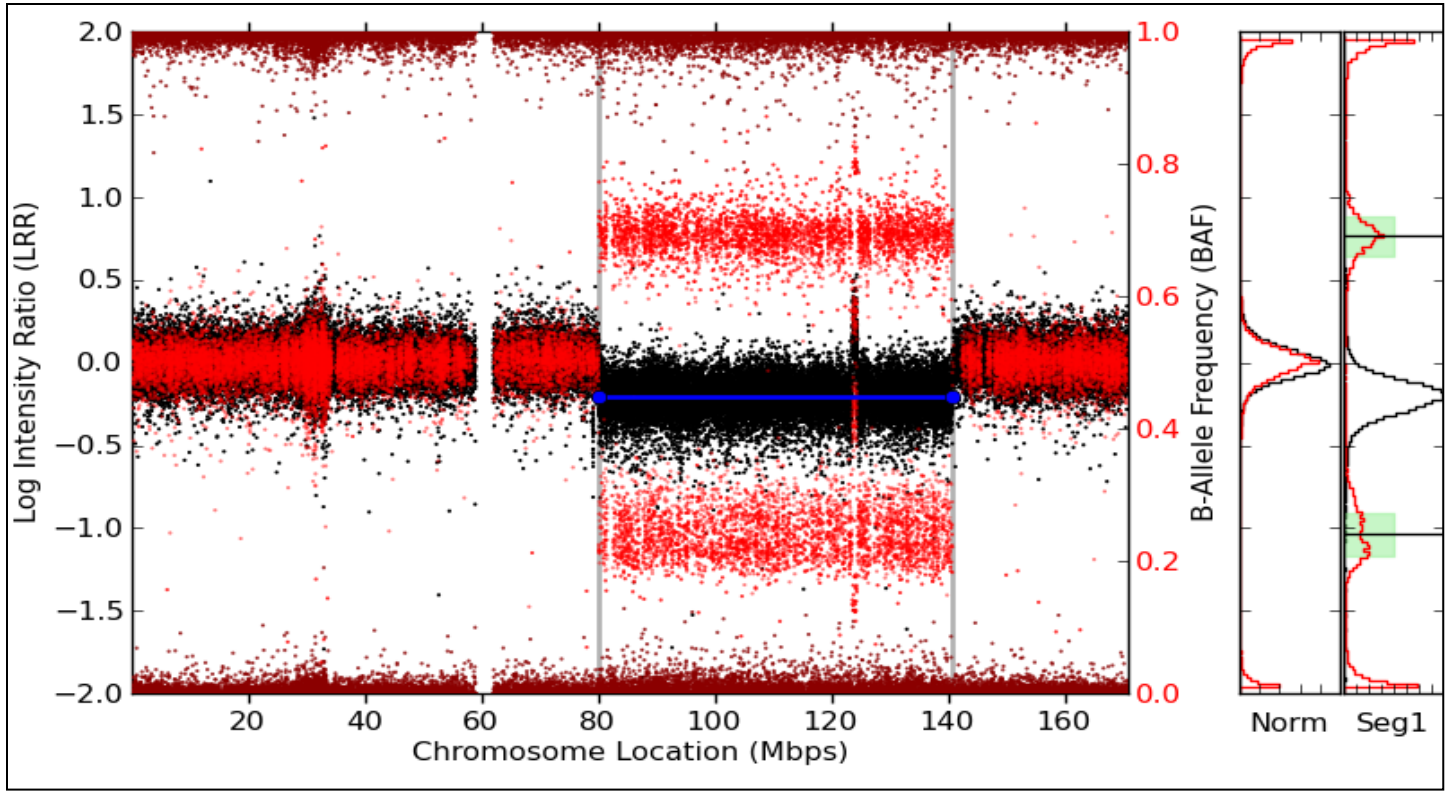
Antoniou *pers commun*

Unexpected Findings



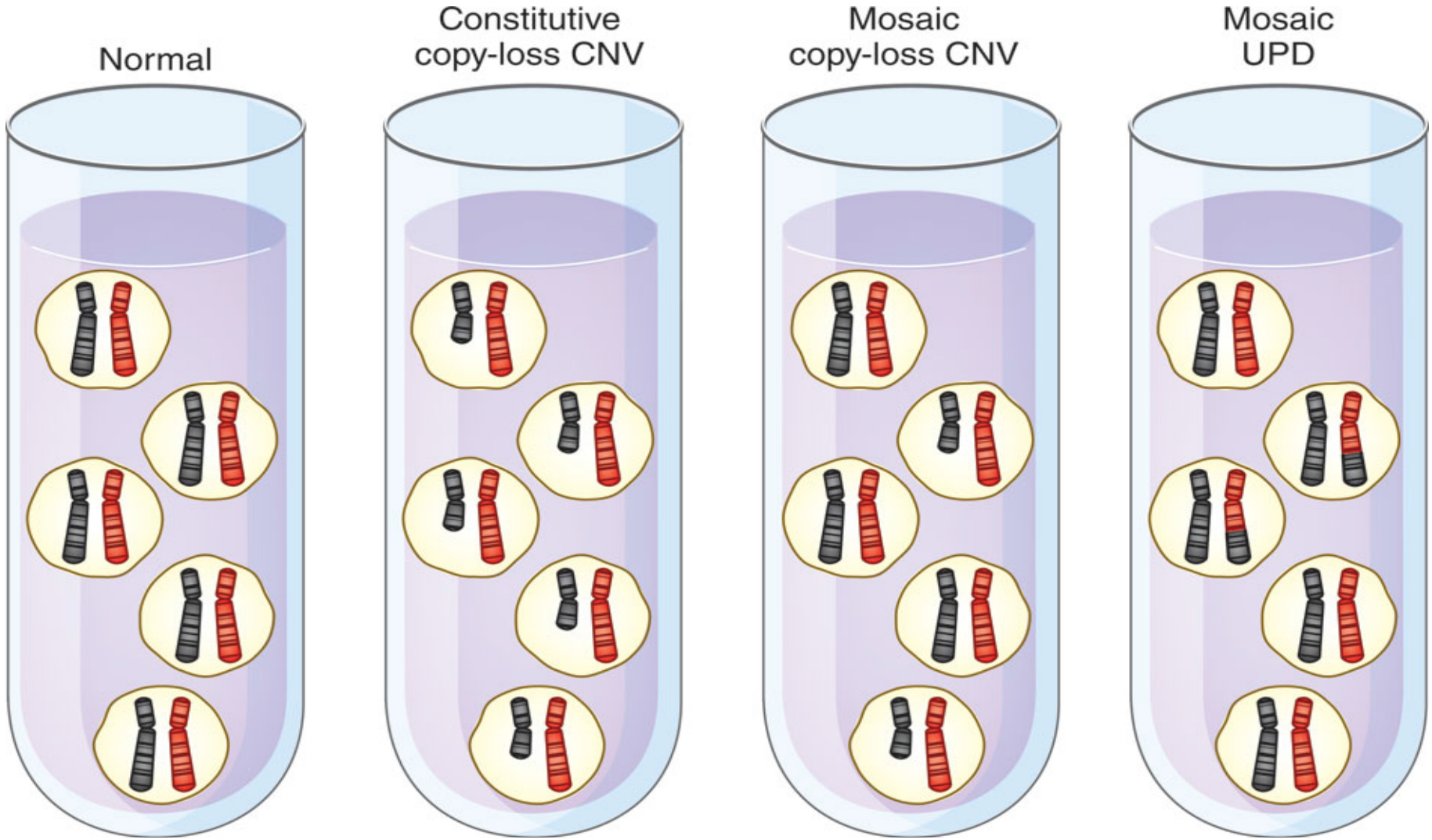
Genome-wide association studies

Large chromosomal abnormalities, structural variation, aneuploidy in germline DNA



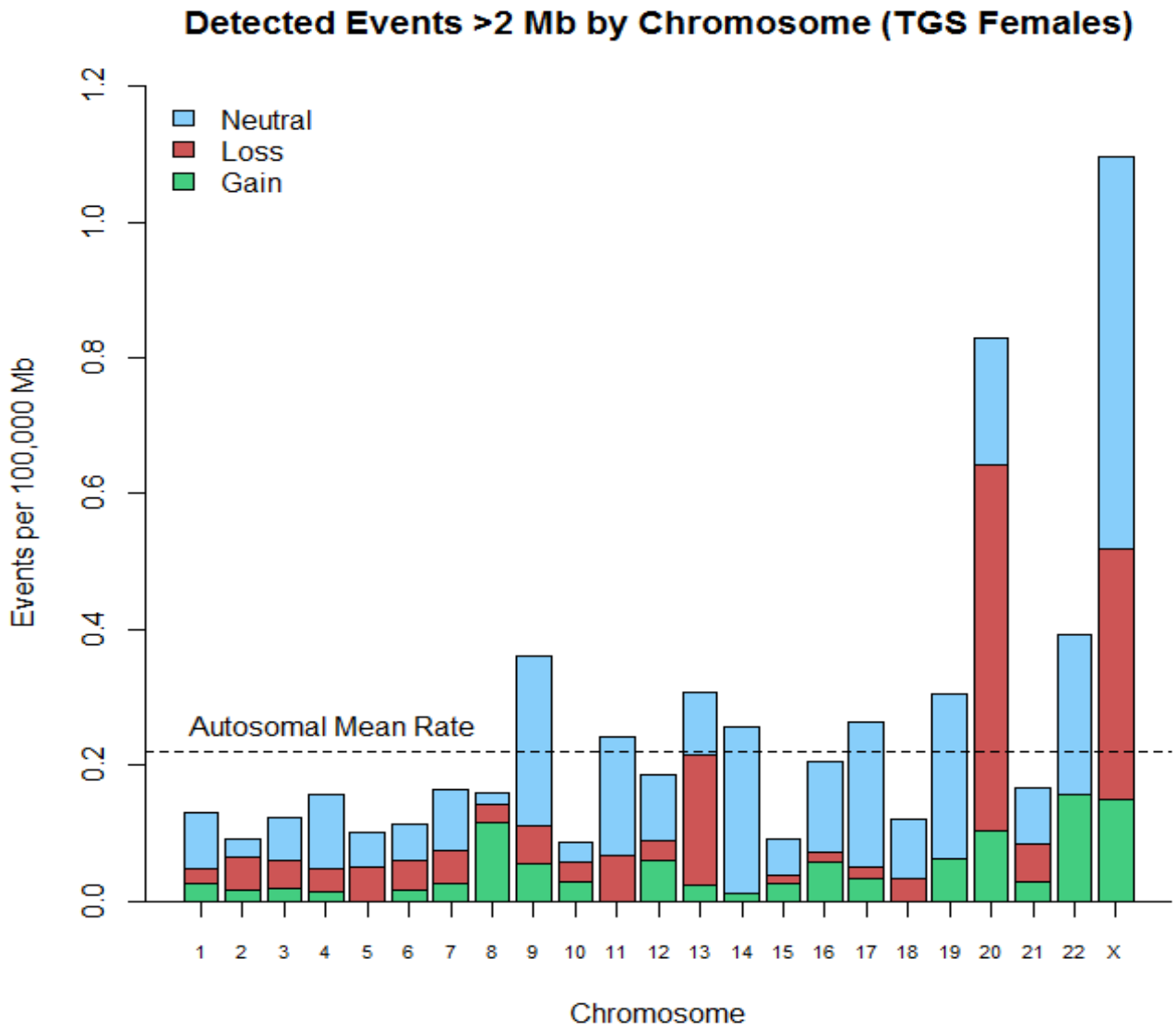
Rodriguez-Santiago *AJHG* 2010
Jacobs et al *Nature Genetics* 2012
Laurie et al *Nature Genetics* 2012

Somatic Mosaicism- the Dynamic Genome



Rate of Mosaicism by Chromosome:

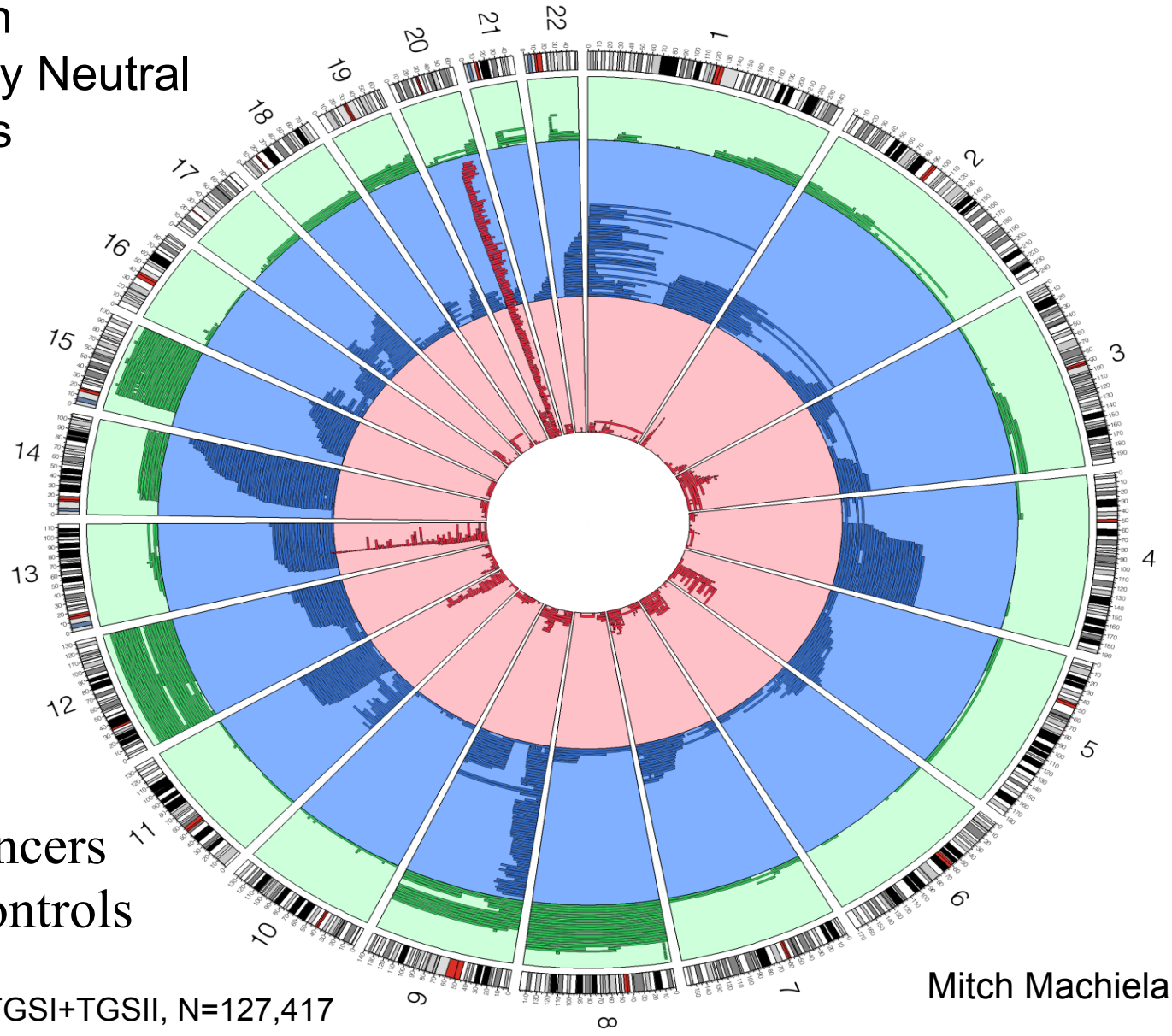
Adjusted for Chromosomal Size



Combined Sample Detected Events >2Mb

(N=1,330 events in 127,417 individuals)

Mosaic Gain
Mosaic Copy Neutral
Mosaic Loss



Non-Heme Cancers
Cancer-free Controls

* Combined GENEVA+TGSI+TGSII, N=127,417

AJHG 2015

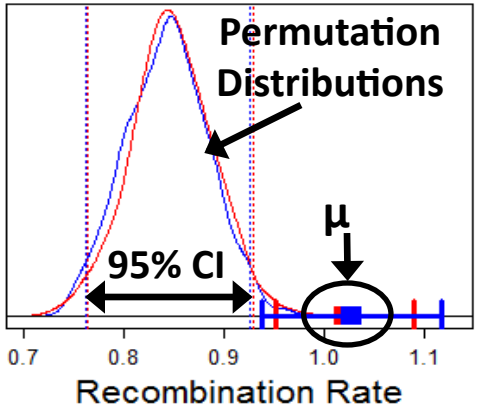
Mitch Machiela

Breakpoint Analysis of Large Mosaic Regions

- 688 Interstitial Events
- 543 Telomeric Copy Neutral Events

- Examined
  - 200kb Windows
  - 500 Permutation
- Enrichment of ENCODE elements?

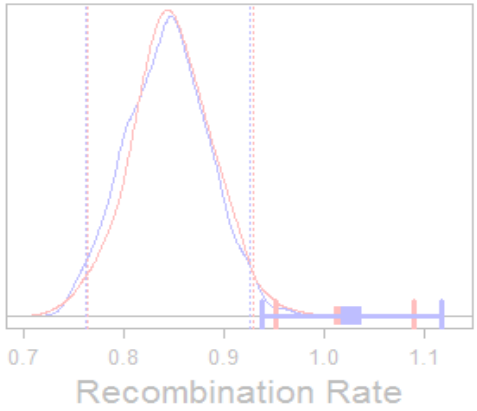
ENCODE Features around Breakpoint Regions

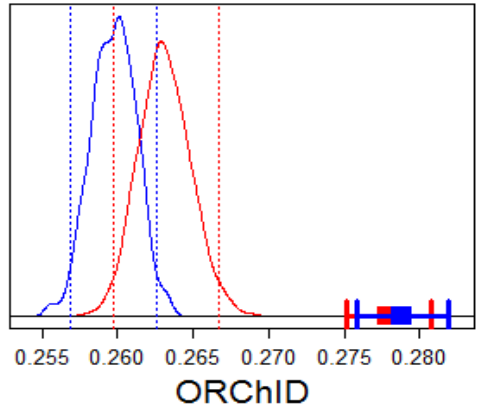


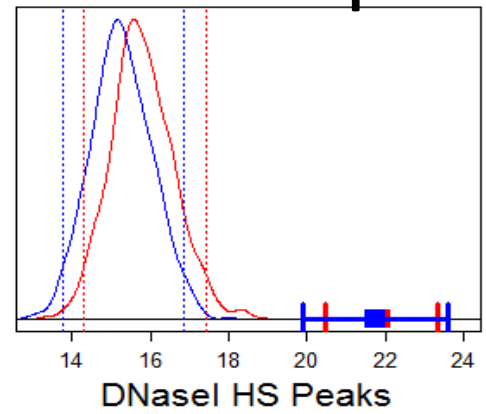
549 Telomeric Copy Neu
688 Interstitial Loss

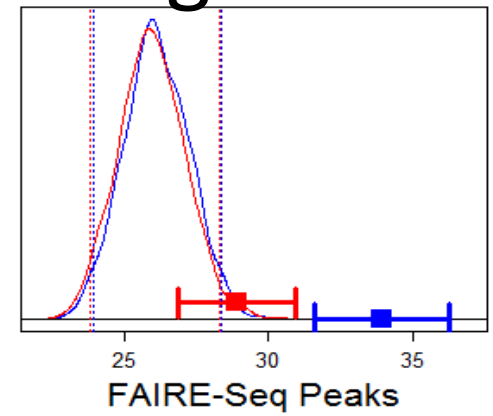
*GENEVA+TGSI+TGSII events >2Mb

ENCODE Features around Breakpoint Regions







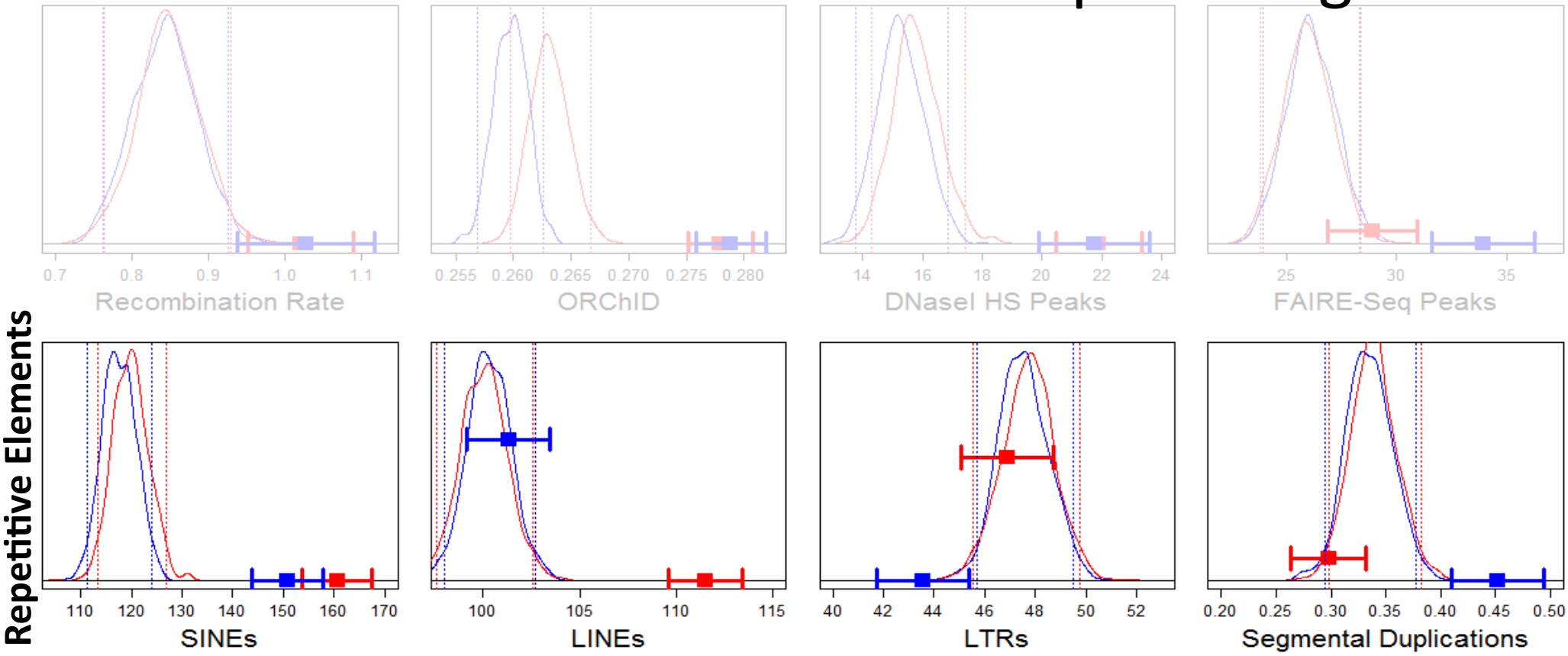


Open Chromatin

Telomeric Copy Neutral
Interstitial Loss

*GENEVA+TGSI+TGSII events >2Mb

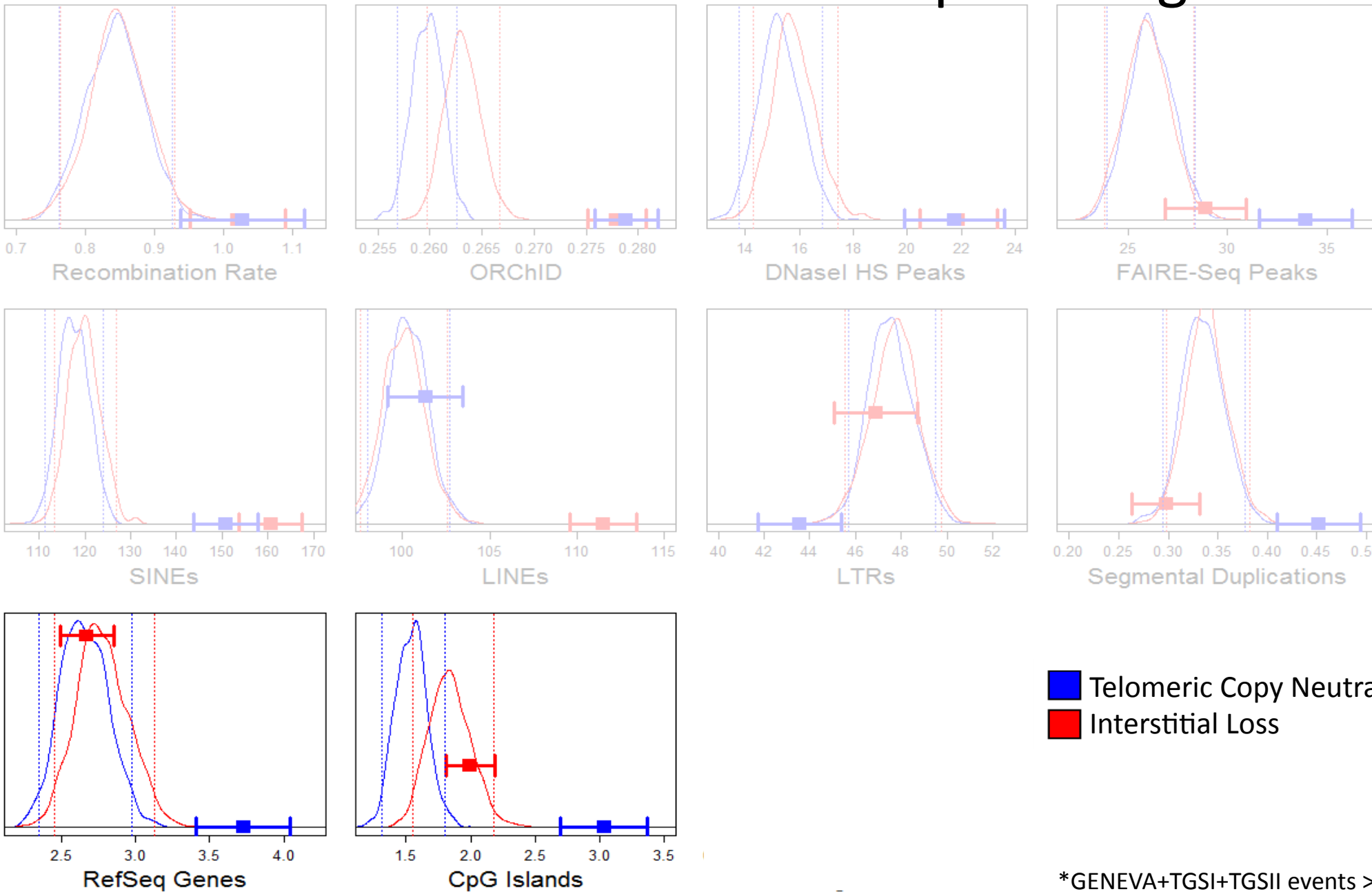
ENCODE Features around Breakpoint Regions



Telomeric Copy Neutral
Interstitial Loss

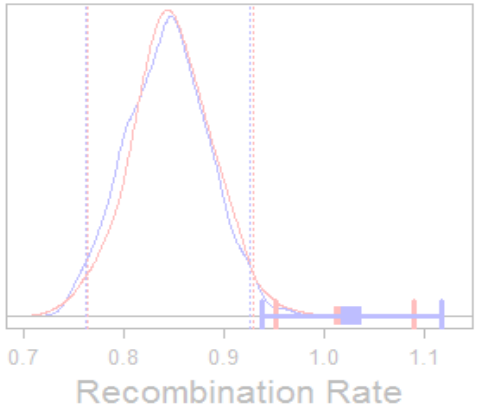
*GENEVA+TGSI+TGSII events >2Mb

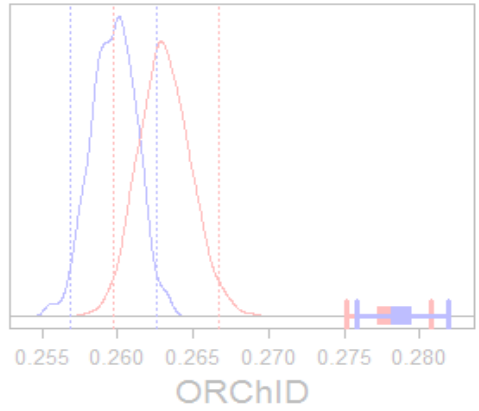
ENCODE Features around Breakpoint Regions

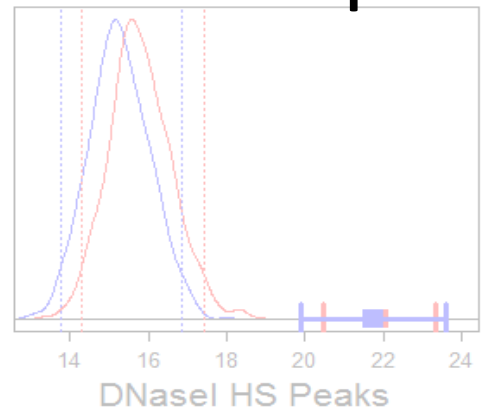


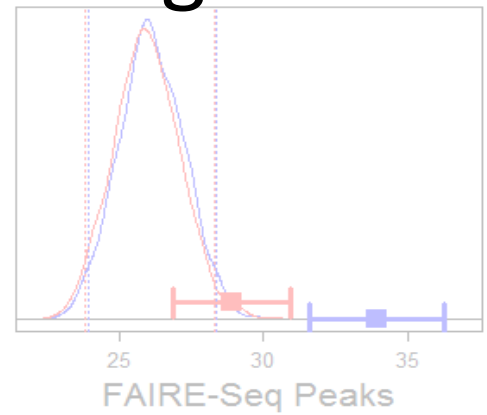
*GENEVA+TGSI+TGSII events >2Mb

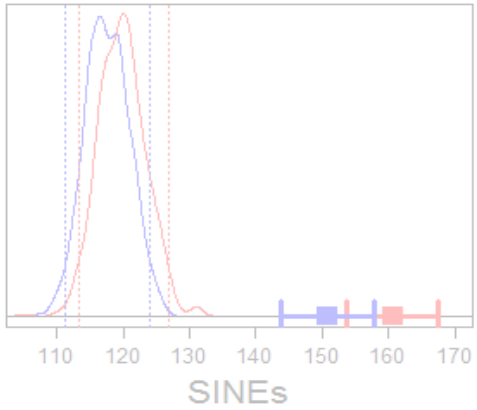
ENCODE Features around Breakpoint Regions

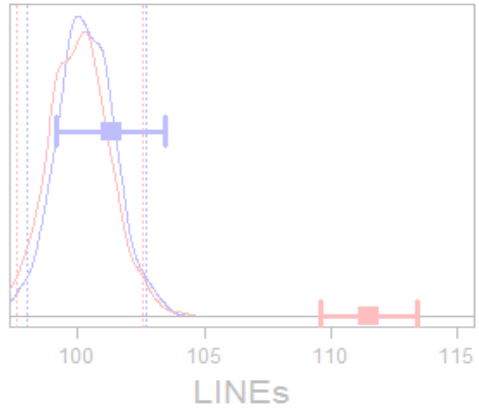


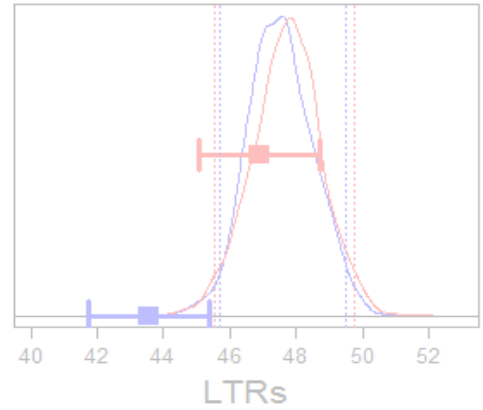


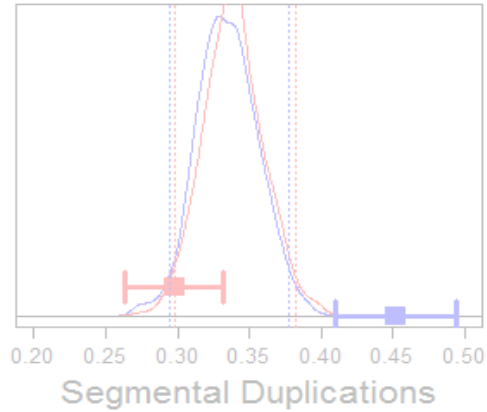


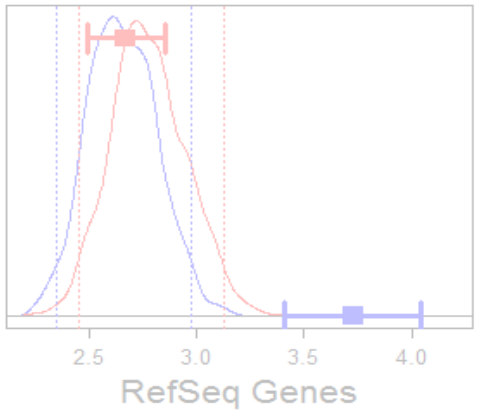


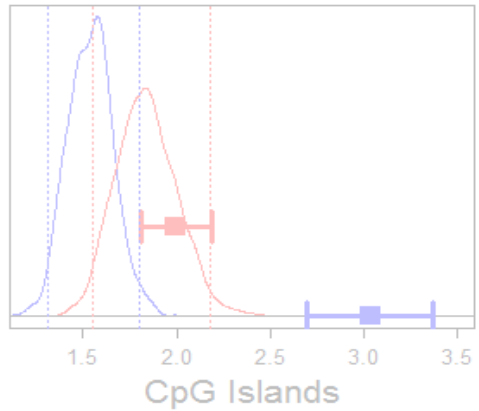


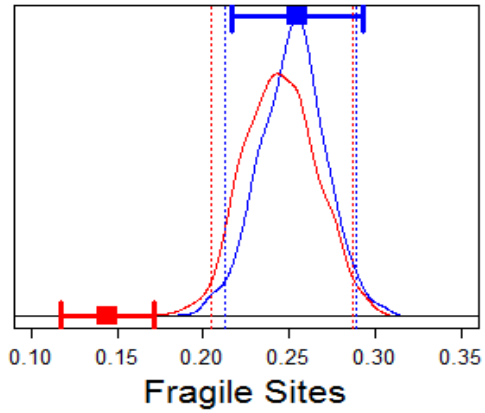








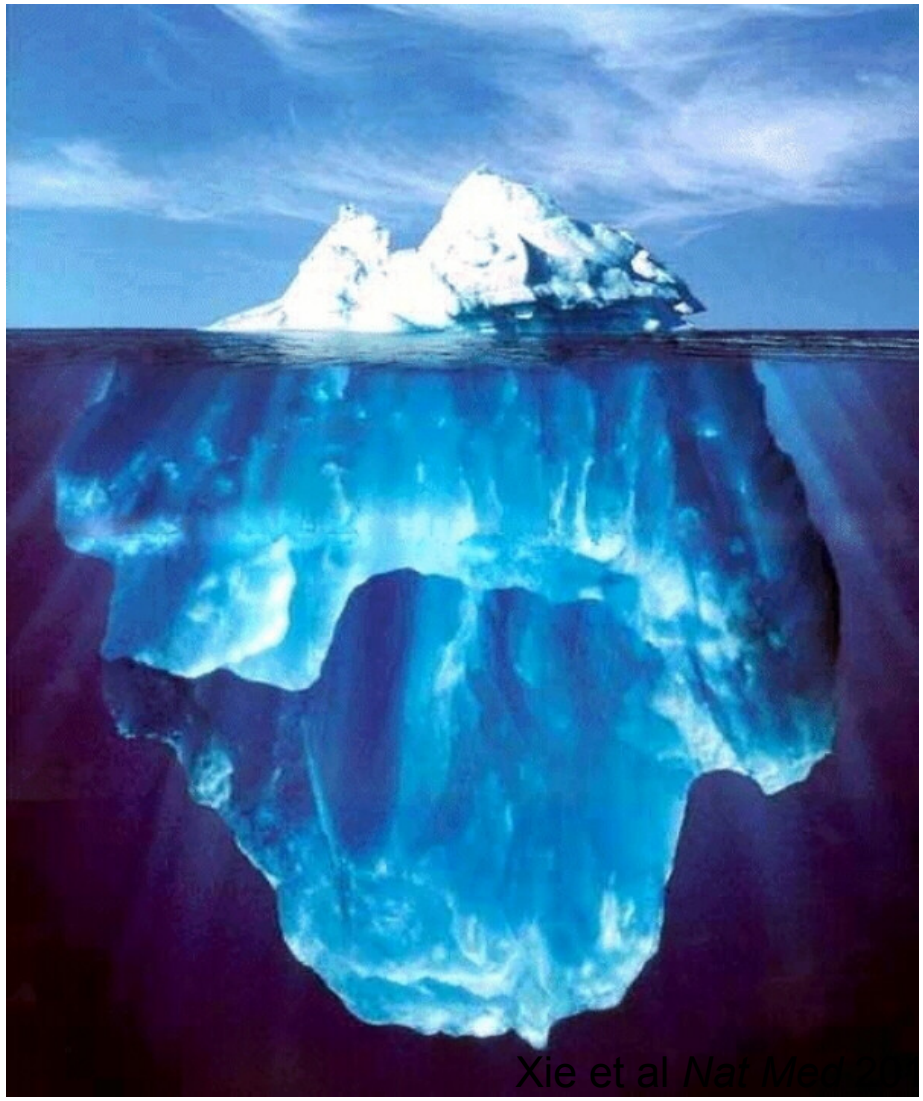




Telomeric Copy Neutral
Interstitial Loss

*GENEVA+TGSI+TGSII events >2Mb

Detectable Mosaicism: Tip of the Iceberg?



Large events detected by SNP arrays/aCGH

Detect smaller events with new algorithms for NGS (NEJM 2014)

"U" shape curve
Seen in very young & aging population

Significance for aging diseases

Xie et al *Nat Med* 2014

Current Challenges of Explaining Susceptibility

- Tissue Specificity
  - Tissue of origin
  - Adjacent cells
  - Immunological Modulation
    - Example: Selective Success of Immune Blockade (PD-1)
- Timing of Effect
- Interaction with environmental stimuli

Immense Value of ENCODE

Scientific

- Spectacular Resource for Understanding the Functional Basis of Susceptibility
  - Prioritization of variants
- Opportunity to Explore Novel Elements
  - Individual
  - Interactions

Cultural

- **Team Science**
  - **Short Term**
  - **Long Term**
- **Establish Thresholds & Standards**
  - **Driven by Questions at hand**

Acknowledgements

NCI-DCEG

Joseph Fraumeni
Robert Hoover
Peggy Tucker
Meredith Yeager
Mitch Machiela
Nilanjan Chatterjee
Juhyn Park
Zhaoming Wang
Weiyin Zhou
Nathaniel Rothman
Mila Prokunina-Olsson
Joshua Sampson
Victoria Fisher

NCI-Special Experts

David Hunter-HSPH
Peter Kraft-HSPH
Montse Garcia-Closas-ICR

More than 400 Collaborators

acknowledgements

The OncoArray Consortium

The GAME-ON Network

HARVARD SCHOOL OF PUBLIC HEALTH

Dartmouth-Hitchcock NORRIS COTTON CANCER CENTER

Keck Medicine of USC

NIH NATIONAL CANCER INSTITUTE

ICR The Institute of Cancer Research

UNIVERSITÉ LAVAL

Queensland Institute of Medical Research

CANCER RESEARCH UK

MOFFITT CANCER CENTER

GenomeCanada

CIHR IRSC Canadian Institutes of Health Research / Instituts de recherche en santé du Canada

GenomeQuébec

NIH National Institute of Environmental Health Sciences
Your Environment. Your Health.

CIDR Center for Inherited Disease Research Johns Hopkins University

UNIVERSITY OF CAMBRIDGE