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
Etiology of Cancer

Environment
- Factory
- Radiation

Lifestyle
- Scale
- Pipe
- Bottle
- Glass

Genetic Susceptibility
- Mutations
- SNPs
- STR’s
- Chromosomes
  (Epigenetics)

100% "Triggers"

"Chance"

100% "Set Point"
Cancer Genomics: 4 Spaces

Germline
- >115 Cancer Syndromes
- >25 Moderate Penetrant
- >475 GWAS Loci

Somatic
- ‘Drivers’ ‘Passengers’
- TCGA/ICGC-Cosmic Data
- Heterogeneity Metastases

Clinically Actionable
- BRCA1/2 Lynch Syndrome
- ACMG “Actionable”
- TEST:
  - Single Mutation
  - Singe Gene
  - Panel of “Cancer Genes”
  - Exome
  - Whole Genome
- HER2
- EGFR
- BRAF600
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…
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

Ascertained in Families
Rare Mutation with Strong Effect
Onogenes & Tumor Suppressors

Incomplete “Penetrance”
Not all affected develop cancer
Modifiers - genetic & environmental
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

>490 Disease Loci marked by SNPs
For > 29 “cancers”
1 Locus marked by a CNV
20% Intergenic
~8% Shared between cancers= Pleiotropy

>100 to be reported by ONCOARRAY
(Breast/prostate/lung/colon/ovarian)

Virtually none are associated with outcomes
GWAS Signals & Somatic Mutations: No Strong Correlation
Redundant Pathways- ‘NOT Drivers’

GWAS Genes

Permutation Genes

Number of Mutations (per 100 individuals)

0 0
0−1
1−2
2−3
3−4
4−5
5<
Interpretation:
Correlation does not imply causation

GWAS designs provide no mechanism to distinguish statistical association from causation

Balding, Nature Genetics Review 2006
GPA: A Statistical Approach to Prioritizing GWAS Results by Integrating Pleiotropy and Annotation

Dongjun Chung1,2,3, Cen Yang3,4,5, Cong Liu2,6, Joel Gelenter2,6,7, Hongyu Zhao1,5,7,9

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 remains a very challenging problem. There is a need to develop more powerful statistical methods to leveraging 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 genetic variants with small or moderate effects. Identifications of these risk variants through joint analysis of multiple GWAS datasets and functional annotations to seek association signals, and it can also perform hypothesis testing to test the presence of pleiotropy, enrichment and 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 analyses, 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 gene 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 ChIP-seq data set, and highly correlated SNPs. To aid the identification of these enhancers, we performed genome-wide ChIP-seq for H3K27-acetylation, a histone marker of active enhancers, in two of which show significantly altered androgen sensitivity in LNCaP cells. This includes rs4907792, that is in linkage disequilibrium (r2) with marker A, and rs4951013, that is in linkage disequilibrium (r2) with marker B. These markers were used to infer the presence of a risk allele, resulting in a 56% decrease in its androgen response element that is interrupted in the protective allele, resulting in a 56% decrease in its androgen response element in the prostate. These findings are consistent with recent studies that have shown that androgen receptor activity is associated with prostate cancer risk. In summary, we propose GPA as a powerful tool to improve GWAS power and identify risk variants for complex diseases.
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

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

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

Mila Prokunina-Olsson
Architecture of Genetic Susceptibility of Cancer Defining Distinct Spaces

- **BRCA1, TP53, RB, PTCH**
- **Rare alleles causing Mendelian disease**
- **Low-frequency variants with intermediate effect**
- **Common variants implicated in common disease by GWA**
- **Polygenic Models SNPs & SNPs**
- **Perturbation Key pathways**

Effect size
- High
- Intermediate
- Modest
- Low

Allele frequency
- Very rare
- Rare
- Low frequency
- Common
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

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

\(^{A}\) Added in 2011
\(^{B}\) Added in 2013
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

1994

Population genotype relative risk

Population risk-allele frequency

15% + 35-40% > 50% FRR

Explain ~35-40% excess familial risk

> Doubled in 2014….now 100 loci
60 More with OncoArray

Exome & Whole Genome Sequencing
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.
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)

Age is critical
Public Health Value

Antoniou pers commun
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

Detected Events >2 Mb by Chromosome (TGS Females)

Events per 100,000 Mb

Chromosome
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+TGS1+TGSII, N=127,417
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

Permutation Distributions

95% CI

μ

Recombination Rate

549 Telomeric Copy Neu
688 Interstitial Loss

*GENEVA+TGSI+TGSII events >2Mb
ENCODEx Features around Breakpoint Regions

- Recombination Rate
- ORChID
- DNasel HS Peaks
- FAIRE-Seq Peaks

Telomeric Copy Neutral
Interstitial Loss

*GENEVA+TGSI+TGSII events >2Mb
ENCODE Features around Breakpoint Regions

Recombinant Rate
ORChID
DNasel HS Peaks
FAIRE-Seq Peaks

SINEs
LINEs
LTRs
Segmental Duplications

Telomeric Copy Neutral
Interstitial Loss

*GENEVA+TGI+TGSII events >2Mb
ENCODE Features around Breakpoint Regions

- Telomeric Copy Neutral
- Interstitial Loss

*GENEVA+TGS1+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