Next-generation sequencing Course

February 23, 2015
Marylyn D. Ritchie and Sarah A. Pendergrass
Outline

Learning objectives are:

• Technologies for genome/exome sequencing
• Study designs for sequencing experiments
• Learn about new methodological approaches to look for both rare variation, as well as the combination of common and rare variation associated with traits
• Discuss what we can learn from sequence data as well as ethical concerns
DNA sequence variation
Lecture 1
I want to do a genetic study on my phenotype of interest. What do I do next?
If only it were that easy.....

• What’s your research question?
• What type of genetic variation are you interested in?
  – Rare or common? Both? Either?
• How many samples do you have access to?
  – What type of samples and when/how were they extracted?
  – How much DNA do you have?
• How many variants do you want to (or can you) analyze?
  – Sample all variation in the genome?
  – Tag variants?
• What’s your data analysis plan?
• What’s your timeline?
• What’s your budget?
• What do you wish to accomplish?
Options for Genotyping SNPs

THROUGHPUT

- Low
  - Taqman
  - Sequenom
- High
  - OpenArray
  - Illumina VeraCode
  - Affymetrix Axiom
  - Affymetrix Gene Chip Array
  - Illumina Infinium

SNP Plex:
- 1
- 10-30
- 16-256
- 48-384
- ~750K
- >906K (6.0)
- 300K-5M

# Samples Per “Run”
- 384
- 384
- 144-12
- 96
- 96
- 1-5*
- 96-4*
Hardware for Genotyping
Genome-wide Common Variant Panels

- 10,000-5 million SNPs
- Affymetrix, Illumina
- Random SNPs – spaced across the genome
- Selected haplotype tag SNPs
- Copy Number Probes
- Considerations for genotyping...
  - Mix up Cases/Controls/ethnic groups
  - Run duplicates
  - Run trios
  - Run HapMap Controls (Positive Control)
  - Blanks/no blanks? (Negative Control)
Published Genome-Wide Associations through 12/2013
Published GWA at $p \leq 5 \times 10^{-8}$ for 17 trait categories

As of 02/20/15, the catalog includes 2,111 publications and 15,396 SNPs.
The case of the missing heritability

When scientists opened up the human genome, they expected to find the genetic components of common traits and diseases. But they were nowhere to be seen. Brendan Maher shines a light on six places where the missing loot could be stashed away.
The case of the missing heritability

When scientists opened up the human genome, they expected to find the genetic components of common traits and diseases. But they were nowhere to be seen. Brendan Maher shines a light on six places where the missing lost could be stashed away.

Missing Heritability

• Under our nose
• Out of sight
• In the architecture
• Underground networks
• Lost in diagnosis
• The great beyond

Biology is complex
Molecular biology is complex
Enabling discoveries with Next-Generation GWAS

- Rare Variants: Small Effect Size
- Rare/Intermediate Variants: Intermediate Effect Size
- Very Rare Variants: Large Effect Size

SEQ or ARRAYS

Next-Gen Arrays

1st Gen Arrays

MAF

Rare

Common

SEQ

Large

Small
Timeline

1. Release of complete human genome
2. HapMap Consortium, collection of common variation
3. Common Variant-Common Disease did not explain much heritability
4. Began to investigate rare variants
5. 1000 Genomes Project
The International HapMap

- Project started in 2003
- Designed to determine frequencies and patterns of association between common SNPs

<table>
<thead>
<tr>
<th>HapMap details</th>
<th># of SNPs Genotyped</th>
<th>Targeted SNPs</th>
<th>Populations Studied</th>
</tr>
</thead>
<tbody>
<tr>
<td>Phase I</td>
<td>1 million</td>
<td>Prioritized coding SNPs to attain 1 SNP for each 5-kb region</td>
<td>CEU,YRI,CHB,JPT</td>
</tr>
<tr>
<td>Phase II</td>
<td>3 million</td>
<td>Prioritized non-synonymous SNPs in coding regions</td>
<td>CEU,YRI,CHB,JPT</td>
</tr>
<tr>
<td>Phase III</td>
<td>1.4 million</td>
<td>Prioritized rare variants</td>
<td>CEU,YRI,CHB,JPT, ASW,CHB,GIH,LWK, MXL,MKK,TSI</td>
</tr>
</tbody>
</table>
The ‘Common Disease-Common Variant’ Hypothesis and Familial Risks

Kari Hemminki¹,²*, Asta Försti¹,², Justo Lorenzo Bermejo¹

¹ Division of Molecular Genetic Epidemiology, German Cancer Research Center (DKFZ), Heidelberg, Germany, ² Center for Family and Community Medicine, Karolinska Institute, Huddinge, Sweden

Abstract

The recent large genotyping studies have identified a new repertoire of disease susceptibility loci of unknown function, characterized by high allele frequencies and low relative risks, lending support to the common disease-common variant (CDCV) hypothesis. The variants explain a much larger proportion of the disease etiology, measured by the population attributable fraction, than of the familial risk. We show here that if the identified polymorphisms were markers of rarer functional alleles they would explain a much larger proportion of the familial risk. For example, in a plausible scenario where the marker is 10 times more common than the causative allele, the excess familial risk of the causative allele is over 10 times higher than that of the marker allele. However, the population attributable fractions of the two alleles are equal. The penetrance mode of the causative locus may be very difficult to deduce from the apparent penetrance mode of the marker locus.


Editor: A. Cecile J. W. Janssens, Erasmus University Medical Center, Netherlands

Received January 30, 2008; Accepted May 16, 2008; Published June 18, 2008

Copyright: © 2008 Hemminki et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits

Penn State
Want to learn more about GWAS?

Bioinformatics challenges in genome-wide association studies (GWAS).
De R¹, Bush WS, Moore JH.

Abstract
Genome-wide association studies (GWAS) are a powerful tool for investigators to examine the human genome to detect genetic risk factors, reveal the genetic architecture of diseases and open up new opportunities for treatment and prevention. However, despite its successes, GWAS have not been able to identify genetic loci that are effective classifiers of disease, limiting their value for genetic testing. This chapter highlights the challenges that lie ahead for GWAS in better identifying disease risk predictors, and how we may address them. In this regard, we review basic concepts regarding GWAS, the technologies used for capturing genetic variation, the missing heritability problem, the need for efficient study design especially for replication efforts, reducing the bias introduced into a dataset, and how to utilize new resources available, such as electronic medical records. We also look to what lies ahead for the field, and the approaches that can be taken to realize the full potential of GWAS.

PMID: 24870131 [PubMed - indexed for MEDLINE]
The 1000 Genomes Project

“Provide deep characterization of human genome sequence”

• Expand the investigation of causal variants to include rare variants
• Project started in 2008
• Genotype 95% of 1+% variation

Details for 1000 Genomes three pilot projects

<table>
<thead>
<tr>
<th>Pilot Data sets</th>
<th>Populations</th>
<th>Samples</th>
<th>Coverage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Trio</td>
<td>2</td>
<td>6</td>
<td>20-40x</td>
</tr>
<tr>
<td>Low coverage</td>
<td>4</td>
<td>179</td>
<td>2-4x</td>
</tr>
<tr>
<td>Exon (8,140 exons ~5% of exome)</td>
<td>7</td>
<td>697</td>
<td>20-50x</td>
</tr>
</tbody>
</table>
The 1,000 Genomes Project

Sequence 1,000 genomes to complete the picture of genetic variation

Achieve a nearly complete catalog of common human genetic variants with frequency 1% or higher.

Project Goals

1. Accelerate fine-mapping efforts in gene regions identified through genome-wide association studies or candidate gene studies

2. Improve the power of future genetic association studies by enabling design of next-generation genotyping microarrays that more fully represent human genetic variation

3. Enhance the analysis of ongoing and already completed association studies by improving our ability to “impute” or “predict” untyped genetic variants
Properties of Variants Found

Known = present in dbSNP129
91% SNPs found in coding regions were already present in dbSNP

The 1000 Genomes Project Consortium. “A map of human genome variation from population-scale sequencing.” Nature 2010
An integrated map of genetic variation from 1,092 human genomes

The 1000 Genomes Project Consortium*

By characterizing the geographic and functional spectrum of human genetic variation, the 1000 Genomes Project aims to build a resource to help to understand the genetic contribution to disease. Here we describe the genomes of 1,092 individuals from 14 populations, constructed using a combination of low-coverage whole-genome and exome sequencing. By developing methods to integrate information across several algorithms and diverse data sources, we provide a validated haplotype map of 38 million single nucleotide polymorphisms, 1.4 million short insertions and deletions, and more than 14,000 larger deletions. We show that individuals from different populations carry different profiles of rare and common variants, and that low-frequency variants show substantial geographic differentiation, which is further increased by the action of purifying selection. We show that evolutionary conservation and coding consequence are key determinants of the strength of purifying selection, that rare-variant load varies substantially across biological pathways, and that each individual contains hundreds of rare non-coding variants at conserved sites, such as motif-disrupting changes in transcription-factor-binding sites. This resource, which captures up to 98% of accessible single nucleotide polymorphisms at a frequency of 1% in related populations, enables analysis of common and low-frequency variants in individuals from diverse, including admixed, populations.
Better Reference Catalog?

Which database should be used as a reference?
  – Which database was most recently released?
  – Which contains the most variants?
  – Which ethnicities were included?
  – What is the allele frequency distribution captured?
New Content for Next-gen GWAS Arrays

Rich content to explore new hypotheses and enable new discoveries

Sequence to discover SNPs >1% MAF (1000-Genomes project)

Leverage the power of LD to select tagSNPs and remove redundancy

Include progressively more SNPs at lower allele frequencies (5%, 2.5%, 1%)

<table>
<thead>
<tr>
<th>Project</th>
<th>Year</th>
<th>Cumulative SNPs found</th>
<th>Tag SNPs needed for max coverage</th>
<th>Lower limit of allele frequency targeted</th>
<th>% variation tagged ($r^2 &gt; 0.8$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>HapMap</td>
<td>2003-2007</td>
<td>3M</td>
<td>0.6M</td>
<td>5%</td>
<td>&gt;90%</td>
</tr>
<tr>
<td>1kG Pilot Project</td>
<td>2008-2009</td>
<td>17M</td>
<td>2.5M</td>
<td>2.5%</td>
<td>~80%</td>
</tr>
<tr>
<td>1kG Full Project</td>
<td>2010</td>
<td>38M</td>
<td>5.0M</td>
<td>1%</td>
<td>&gt;90%</td>
</tr>
</tbody>
</table>
The spectrum of known variation is expanding at an unprecedented rate...

HapMap – 3.5 Million SNPs

1kGP Pilot Phase – 17 Million SNPs

1kGP Main Project – 38 Million SNPs
...and resetting the reference point for GWAS.
So maybe it’s in the rare(r) variants

Common Disease, Multiple Rare (and Distant) Variants

Richard Robinson*

Freelance Science Writer, Sherborn, Massachusetts, United States of America

Genome-wide association (GWA) studies have emerged as a potentially powerful tool for discovery of new genes for common diseases, such as Alzheimer’s disease and stroke. But the common interpretation of GWA findings might be incorrect in many cases, according to a new study by Samuel Dickson, David Goldstein, and colleagues in this issue of PLoS Biology. Their results suggest that the signals in these studies may not always be pointing to a few common gene variants, as assumed by most researchers, but instead to many rare variants, each of which causes relatively few cases, and each of which may be relatively far away from the site identified in the GWA study.

A GWA study compares DNA sequence variation in cases to that of controls. In the case of Alzheimer’s disease, the “common disease, many rare variants” hypothesis, and the difficulty in finding culprit genes was that these modest effects make the genes very difficult to recognize.

But an alternative explanation is also possible, that the disease is caused by multiple strong-effect variants, each of which is found in only a few people (the “common disease, many rare variants” hypothesis). Instead of the common signpost pointing to a common weak-effect variant, it might be pointing to many strong-effect variants. To distinguish this scenario from the common interpretation, the authors refer to associations between rare higher-impact variants and common markers as “synthetic associations”.

In the world of synthetic associations, the
Other reference panels

Go-NL
GENOME of the NETHERLANDS

Ultra-sharp genetic group portrait of the Dutch

What genetic variation is to be found in the Dutch indigenous population? Detailed knowledge about this is not only interesting in itself, it also helps to extract useful biomedical information from Dutch biobanks. The Dutch biobank collaboration BBMRI-NL has initiated the extensive Rainbow Project "Genome of the Netherlands" (GoNL) because it offers unique opportunities for science and for the development of new treatments and diagnostic techniques. A close-up look at the DNA of 750 Dutch people-250 trio’s of two parents and an adult child-plus a global genetic profile of large numbers of Dutch will disclose a wealth of new information, new insights, and possible applications.

News
- GoNL in the Dutch press
- GoNL in Nature Genetics
- Prizes at Genetics Retreat 2014
- Search GoNL snps online
- Position paper online

Archive
- July 2014
- March 2014
- July 2013
- June 2013

http://www.nlgenome.nl/
Other reference panels

Explore genetic variation interactively. Compare entire cohorts in seconds with SQL-like queries. Compute transition/transversion ratios, genome-wide association, allelic frequency and more.

Process big genomic data easily. Run batch analyses like principal component analysis and Hardy-Weinberg equilibrium on as many samples as you like, in minutes or hours, with just a little code.

Use Google’s infrastructure and big data expertise. Store one genome or a million using Google Genomics and take advantage of the same infrastructure that powers Search, Maps, YouTube, Gmail and Drive.
Other reference panels

**UK10K**

*Rare Genetic Variants in Health and Disease (2010-2013)*

---

**What is UK10K?**

The UK10K project will enable researchers in the UK and beyond to better understand the link between low-frequency and rare genetic changes, and human disease caused by harmful changes to the proteins the body makes.

Although many hundreds of genes that are involved in causing disease have already been identified, it is believed that many more remain to be discovered. The UK10K project aims to help uncover them by studying the genetic code of 10,000 people in much finer detail than ever before.

---

**Project Design**

Not all genetic changes are harmful or lead to disease, so the project is taking a two-pronged approach to identify rare variants and their effects:

- by studying and comparing the DNA of 4,000 people whose physical characteristics are well documented, the project aims to identify those changes that have no discernible effect and those that may be linked to a particular disease;

---

[http://www.uk10k.org/](http://www.uk10k.org/)
I’m interested in other variation besides SNPs...

Perhaps we should look at CNVs
CNVs are Associated with Various Phenotypes

Copy number polymorphism in Fcgr3 predisposes to glomerulonephritis in rats and humans

The Influence of CCL3L1 Gene-Containing Segmental Duplications on HIV-1/AIDS Susceptibility

A Chromosome 8 Gene-Cluster Polymorphism with Low Human Beta-Defensin 2 Gene Copy Number Predisposes to Crohn Disease of the Colon

Strong Association of De Novo Copy Number Mutations with Autism

Rare Structural Variants Disrupt Multiple Genes in Neurodevelopmental Pathways in Schizophrenia
ARTICLES

Origins and functional impact of copy number variation in the human genome

Donald F. Conrad¹ *, Dalila Pinto² *, Richard Redon¹,³, Lars Feuk²,⁴, Omer Gokcumen⁵, Yujun Zhang¹, Jan Aerts¹, T. Daniel Andrews¹, Chris Barnes¹, Peter Campbell¹, Tomas Fitzgerald¹, Min Hu¹, Chun Hwa Ihm⁵, Kati Kristiansson¹, Daniel G. MacArthur¹, Jeffrey R. MacDonald², Ifeijelo Onyiah¹, Andy Wing Chun Pang²§, Sam Robson¹, Kathy Stirrups¹, Armand Valsesia¹, Klaudia Walter¹, John Wei², Wellcome Trust Case Control Consortium*, Chris Tyler-Smith¹, Nigel P. Carter¹, Charles Lee⁶, Stephen W. Scherer²,⁷ & Matthew E. Hurles¹

ARTICLES

Genome-wide association study of CNVs in 16,000 cases of eight common diseases and 3,000 shared controls

The Wellcome Trust Case Control Consortium*
“Take-Home” Quotes from CNV Paper

• “We have demonstrated that high-confidence CNV calls can be assigned in large, real-world case-control samples for a substantial proportion of the common CNVs estimated to be present in the human genome.

• We have identified directly several CNV loci that are associated with common disease. Such loci could contribute to disease pathogenesis.

• However, the loci identified are well tagged by SNPs and, hence, the associations can be, and were, detected indirectly via SNP association studies.

• Among the CNVs that we could type well, those not well tagged by SNPs have the same overall association properties as those which are well tagged.”
GWAS is fine, and CNVs are cool, but I want to detect ALL variation in my samples!

• SEQUENCE is the SOLUTION
Questions???