Emerging Resources for Genomic Assays That Power Discoveries in Diverse Populations

Janina M. Jeff, PhD, MS
Global Bioinformatics Specialist
The Central Dogma

DNA → Transcription → RNA → Translation → Protein

Replication
Central Dogma Career Path for Scientist

PhD → Post-Doc → Assistant Pro.
Central Dogma Career Path for Scientist

PhD

Interdisciplinary Graduate Program
~80-100 PhD students
~stipend, no teaching
~3-4 rotations

Crawford Lab
Genetic Epidemiology and Statistical Genetics
PhD Human Genetics
MS Applied Statistics

Pushed to my deepest limits
1. Imposter syndrome → building confidence (publications and oral presentations)
2. Fear of failing → understanding success cannot exist without failure (graduating and getting post doc offers)
Central Dogma Career Path for Scientist

Original post-doc advisor
New York, NY
most diverse biobank

Erwin Bottinger, MD

Communal post-doc
Adopted post-doc advisors

Pushed to my deepest limits
1. Competitive climate → Defining your own success (forming new career opportunities: mentorship, teaching)
2. Different mentorship styles → Lead to independence (independently submitting abstracts and grants)

Eimear Kenny, PhD
Ruth Loos, PhD
Central Dogma Career Path for Scientist

PhD → Post-Doc → Assistant Pro.

Big Pharma
Direct to Consumer
Government
Biotechnology

Science
Lifestyle
Teaching and Mentorship
Research
Central Dogma Career Path for Scientist Does Not Exists

PhD → Post-Doc → Bioinformatician

Biotechnology

Science

Lifestyle

Teaching and Mentorship

Research

illuminar®
Emerging Resources for Genomic Assays That Power Discoveries in Diverse Populations

Janina M. Jeff, PhD, MS
Global Bioinformatics Specialist
An Integrated View of Complex Disease

A combination of genetic and environmental factors contributes to disease

Arrays

Genomic Variants

Complex Disease

Environment

Clinical Database

Sequencing
The Evolution of Large-Scale Genomics
10 years of Genome-Wide Association Studies (GWAS)


- Common variants
  - HumanHap-300
  - 550,660,610, 1M

- Common and rare variants
  - Omni family
  - Core family

- Multi-ethnic / imputation
  - Common and rare variants

- Population based studies
  - Common and rare variants

- Multi-Ethnic family
- Screening Array Program (GSA, ASA, MEA etc)

For Research Use Only. Not for use in diagnostic procedures.
Global Screening Array Consortium Members

- Global participation
- Large population cohorts, Biobanks, DTC, Researchers, Pharmaceutical companies, Health care providers, Service providers
- Millions of samples available within the next 2-3 years
- Total of ~200 consortia members from 55 institutions
Global Commitment to Population Studies
5 Million Samples Globally Sold

- European: 10,000
- USA: 2,085,000
- Hispanic: 659,568
- Asian: 1,245,000
- India: 1,000,000
New Applications on a Global Scale

Increasing Application Breadth

- DTC
- Health Care Providers
- Research
- Cancer
- Pharma

- 2,225,000
- 1,600,000
- 884,568
- 110,000
- 450,000
Multi-Ethnic-GWAS backbone

Two reasons to embrace multi-ethnic genomics studies

- Seamless replication
- Improved Understanding of Disease/Phenotypes
Causal Variant Must be Covered in Replication Population
Replication is not possible if SNP is not covered in replication sample

- Expected Linkage
  Disequilibrium Differences
  Across Populations
- Coverage of variants across populations must be present in order to replicate a signal

causal Variant Discovered in original GWAS (pink)

Multi-Ethnic Array

Custom Array

Successfully replicate causal variant

Fail to replicate causal variant

Improved Understanding of Disease/Phenotypes

Variants with large functional effects can be largely population specific*

High prevalence in East Asian populations of CYP2C19 genotype leads to lower Plavix efficacy

Blue = good metabolizers
Orange: poor metabolizers

Large percentage of East Asian population with variant for poor metabolism of Plavix

Opportunity for fine-mapping
Leveraging differential LD patterns enables fine-mapping of the causal SNP variants

Illumina Global Screening Array
*A high powered, economical tool for population scale genomics*

**Universal Genome Wide Association Study (GWAS) Array delivering power across multiple populations**

- Global content enriched for population specific and cross population variation
- High imputation accuracy across the entire allelic frequency spectrum

**Comprehensive selection of clinical research variants**

- >50K clinical research variants spanning breadth of functional variation
- Up to date content reviewed and validated by experts in medical genomics

**Developed for high throughput population screening**

- QC and high value content for sample identification, tracking, and stratification
- Designed on the Infinium® 24 sample format, add up to 50K in custom content

For Research Use Only. Not for use in diagnostic procedures.
Global Screening Array (GSA) Manifest variant count
Predictive, Clinical Research, and QC

- **46,500**
  - Multi-Disease Drop-In
  - User defined custom content or
  - Pre-designed multi-disease drop in

- **9,761**
  - QC
  - Sample tracking and stratification

- **54,739**
  - Clinical Research
  - Up-to-date known clinical associations
  - Pharmacogenomics
  - Well-curated exome content
  - NHGRI-GWAS and HLA content

- **589,656**
  - GWAS backbone
  - Cross-Population and population-specific
  - Enriched for low-frequency variants (1-5%)
  - High imputation accuracy for ALL populations

FINAL Product: 700,656

For Research Use Only. Not for use in diagnostic procedures.
GSA Manifest variant count

**Predictive, Clinical Research, and QC**

- **46,500**
  - User defined custom content OR
  - Pre-designed multi-disease drop in

- **9,761**
  - Sample tracking and stratification

- **54,739**
  - Up-to-date known clinical associations
  - Pharmacogenomics
  - Well-curated exome content
  - NHGRI-GWAS and HLA content

**589,656**

GWAS backbone

- Cross-Population and population-specific
- Enriched for low-frequency variants (1-5%)
- High imputation accuracy for ALL populations

For Research Use Only. Not for use in diagnostic procedures.
Imputation
A New Paradigm Emerges

- Imputation is the New Gold Standard in Array Analysis
- Imputed SNPs have more power than pairwise LD ($r^2$)
  - Clark et al., NATURE GENETICS, 2007
- Imputation increases resolution of the association peak

Key SNP criteria:
- Imputation accuracy: how accurate is the imputed SNP
- Imputation Efficiency: how many SNPs can you impute
- Minor Allele Frequency*
  - Low MAF 1-5% - Key Focus of GSA
  - Common >5%

Gain more power with Imputed SNPs
New Paradigm in GWAS Analysis

- Traditional r² pairwise LD
  - 5.4%
- Direct Risk SNP Typing
  - 10.4%
- Imputed SNPs
  - 8.9%

From: Conjuring SNPs to detect associations. Clark and J. Li NATURE GENETICS | VOLUME 39 | NUMBER 7 | JULY 2007
High Imputation Accuracy Across All Populations

Accurate, economical coverage of the genome

Mean Imputation Accuracy

>1 % Minor allele frequency

<table>
<thead>
<tr>
<th>Population</th>
<th>GSA</th>
</tr>
</thead>
<tbody>
<tr>
<td>AFR</td>
<td>0.95</td>
</tr>
<tr>
<td>AMR</td>
<td>0.97</td>
</tr>
<tr>
<td>EUR</td>
<td>0.97</td>
</tr>
<tr>
<td>EAS</td>
<td>0.96</td>
</tr>
<tr>
<td>SAS</td>
<td>0.96</td>
</tr>
</tbody>
</table>


For Research Use Only. Not for use in diagnostic procedures.
Imputation Increases Genomic Coverage

GSA has greater than 90% coverage in all populations after imputation
Universal Imputation Power Across Populations

all SNPs in 1kGP >1% MAF and >80% imputation accuracy

Thank you and Questions
GSA Manifest variant count
*Predictive, Clinical Research, and QC*

- **46,500**
- **9,761**
- **54,739** Clinical Research
- **589,656**

**Long-term Risk Profiling Variants**
- **Early Detection Markers**
- **Early Treatment Markers**
- **Marker Discovery**
- **Participant Selection for Clinical Trials**

**Sample Classification and QC**

**Pathogenic Variant Validation**


For Research Use Only. Not for use in diagnostic procedures.
Databases Overview

- **ClinVar (135K):**
  - Reports of the relationships among human variations and phenotypes, with supporting evidence.

- **ExAC:**
  - The Exome Aggregation Consortium: exome sequencing data from a variety of large-scale sequencing projects. The data set spans 60706 unrelated individuals sequenced as part of various disease-specific and population genetic studies.

- **NHGRI:**
  - A Catalog of Published Genome-Wide Association Studies

- **eMERGE:**
  - Electronic Medical Records and Genomics (eMERGE) Network
    Combines biorepositories with electronic medical record (EMR) systems for genomic discovery and genomic medicine

- **PharmGKB:**
  - Pharmacogenomics knowledge resource with clinical information including dosing guidelines and drug labels, potentially clinically actionable gene-drug associations and genotype-phenotype relationships
GSA Clinical Research Applications

Variant Validation

GSA Genotyped Sample

Variant Interpretation

Disease and Therapy Marker Research
- Variant Validation
- Family-based Analysis
- Confirm/validate results

Risk Profiling
- Early Detection Markers
- Early Intervention Markers
- Alternative Treatment Markers

For Research Use Only. Not for use in diagnostic procedures.
GSA Pathogenic Variant: APOL1
May Predict Kidney Failure:

About
1 in 5
people
with two copies of
APOL1 renal
risk variants
will develop
kidney disease\(^2\)

These APOL1 variants
account for 70% of non-diabetic
kidney failure in African-Americans\(^2\)

About
4 in 10
African-Americans on dialysis have
kidney failure caused by APOL1\(^2\)

**Sources**


For Research Use Only. Not for use in diagnostic procedures.
Genetic Ancestry Testing
Identification of Y, Mitochondrial and Autosomal variants in a single assay

- **Y chromosome testing:**
  - passed exclusively from father to son
  - can be used to explore ancestry in the direct male line (males only)

- **Mitochondrial DNA testing:**
  - mitochondrial DNA is passed on from mothers,
  - can be used by either sex
  - provides information about the direct female ancestral line

- **Autosomal Single nucleotide polymorphism (SNP) testing:**
  - capture the overall ethnic background of an individual
  - provide an estimate of a person's ethnic background.
  - SNPs can indicate that a person's mixed ancestry
  - 50% African, 25% European, 20% Asian, and 5% unknown

Thank you and Questions
GSA Manifest variant count

Predictive, Clinical Research, and QC

- 46,500
- 9,761
- 54,739 Clinical Research
- 589,656

- Up-to-date known clinical associations
- Pharmacogenomics
- Well-curated exome content
- NHGRI-GWAS and HLA content
Expert Selected Clinical Research Content

Comprehensive set of known and putative clinical variants

**Tier 1**

- Known clinical association
  - Validate Disease Associations
  - Risk profiling
  - Pre-emptive screening
  - PGx Studies (CPIC +)

**Tier 2**

- Strong evidence for clinical association
  - Establish functionality
  - Support putative clinical associations
  - Discover novel associations
Tier 1: Clinical Research Content Selection

*Comprehensive collection of high value variants*

<table>
<thead>
<tr>
<th>Total Tier 1</th>
</tr>
</thead>
</table>

### Variants from GSA consortia
- eMERGE
- PharmGKB*
- Consortia provided

### Variants from ClinVar**
- Include Pathogenic and likely pathogenic
- Exclude all somatic mutations
- Exclude variants missing clinical annotations
- Exclude variants missing review status
- Exclude benign and likely benign

*https://www.pharmgkb.org
GSA: More Than Just a SNP Array

*Ability to design breadth of functional variation*

Illumina technology enables:

- Ability to design based on clinically-relevant CNVs
- >95% of selected variants designed

<table>
<thead>
<tr>
<th>Total</th>
<th>Genic</th>
<th>Intergenic</th>
<th>SNPs</th>
<th>CNVs</th>
<th>Indels</th>
<th>Structural</th>
<th>Unknown</th>
</tr>
</thead>
<tbody>
<tr>
<td>22,760</td>
<td>21,094</td>
<td>1,666</td>
<td>16,545</td>
<td>729</td>
<td>323</td>
<td>4,836</td>
<td>11</td>
</tr>
</tbody>
</table>

For Research Use Only. Not for use in diagnostic procedures.
GSA: Wide Breadth of Highly Penetrant Variants

*Potential to detect >10,000 markers across 20 disease categories*

![Pie chart showing disease categories and marker counts]

- Congenital malformations, deformation and abnormalities: 2113
- Cancer: 2033
- Circulatory system: 1700
- Endocrine metabolic diseases: 1826
- Nervous system: 1460
- Additional 15 disease categories: 1113


For Research Use Only. Not for use in diagnostic procedures.
**GSA: Discover Associations and Establish Functionality**

>27K markers with strong evidence of association

<table>
<thead>
<tr>
<th>Total Tier 2: ~39 K Variants</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Putative disease &amp; phenotype associations</strong></td>
<td><strong>Functionally interesting genes</strong></td>
</tr>
<tr>
<td>~7,000</td>
<td>~32K</td>
</tr>
</tbody>
</table>

**NHGRI-GWAS**

- Genome-wide associations with \( p \leq 5.0 \times 10^{-8} \)
- >700 diseases and traits

**ExAC**

- All ClinVar variants in conserved loci
- All non-tumor samples
- MAF >1%
- Conserved loci with loss of function in functionally interesting genes

*Most novel content not captured in ClinVar will come from ExAC*
Selection of Highest Value Exome Variants

Potential for clinical associations

- **ExAC**: 9,362,318 Variants
- **ClinVar variants in conserved loci**: 6,249
- **ExAC Variants in Clinically Relevant Genes**: 13,446 variants

**Criteria for Selection**

1. Eliminate variants not frequently seen in population
2. Select variants in evolutionary conserved loci
3. Select missense variants below LOF score
4. Select variants that impact ORF integrity
5. Limit variants to gene list (7502 Genes, ~32,329 Variants)
Filtering the Non-ClinVar ExAC Variants

Remove rare alleles & select conserved loci with probable LOF

- Reduce 70,706 Filtered Variants to 32,329 via overlap with 7502 genes
- Adjust final variant number by variant quality cutoff to 13,446
Global Representation of ExAC Variants in GSA Clinical

Novel Exome variants for clinical and predictive research

- A total > 30,000 ExAC variants were targeted on GSA
- Equal distribution of exclusive clinical variants across all populations
- Asian and African populations contribute highest number of exclusive variants

<table>
<thead>
<tr>
<th>1000 Genome Population</th>
<th>ExAC variants in GSA Clinical Content</th>
</tr>
</thead>
<tbody>
<tr>
<td>None</td>
<td>AMR</td>
</tr>
<tr>
<td>Exclusive variants detected only in</td>
<td>4,855</td>
</tr>
<tr>
<td>Total Variants present in selected Population</td>
<td>4,855</td>
</tr>
</tbody>
</table>


For Research Use Only. Not for use in diagnostic procedures.
HLA Content Optimized for Imputation

- 4,000 variants
- Best tag SNP’s based on HLA imputation in available reference populations (European and Japanese)
- HLA collaborators report > 0.80 imputation accuracy for > 96% variants in Europeans

**Adapted from Jia et. al., Plos One 2013**

<table>
<thead>
<tr>
<th>HLA Imputation Accuracy</th>
<th>Europeans</th>
<th>Asians</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>HLA-A</strong></td>
<td>99.1%</td>
<td>98.1%</td>
</tr>
<tr>
<td><strong>HLA-B</strong></td>
<td>96.8%</td>
<td>65.6%</td>
</tr>
<tr>
<td><strong>HLA-C</strong></td>
<td>99.1%</td>
<td>68.8%</td>
</tr>
<tr>
<td><strong>HLA-DQA1</strong></td>
<td>98.5%</td>
<td>96.3%</td>
</tr>
<tr>
<td><strong>HLA-DQB1</strong></td>
<td>99.1%</td>
<td>96.5%</td>
</tr>
<tr>
<td><strong>HLA-DRB1</strong></td>
<td>96.9%</td>
<td>92.3%</td>
</tr>
<tr>
<td>All loci</td>
<td>98.3%</td>
<td>86.4%</td>
</tr>
</tbody>
</table>
Infinium® GSA: Seamless Sample Tracking and Stratification

>9500 variants for sample tracking, QC, and stratification

- Fingerprinting (480)
- Sex determination (3101)
- Ancestry informative (3212)
- Mitochondrial (155)
- Pseudo Autosomal Regions 1 & 2 (535)
- Forensics (173)

GSA: Predefined Multi-Disease Content
Fine mapping of genome-wide significant loci

 Derived from large-scale meta-analysis and exome content

 Predefined 50K bead types* in lieu of custom content

 Leverage established phenotype-specific consortia

- Cardiometabolic
- Autoimmune
- Psychiatric
- Neurological
- Cancer
- Anthropometric

*Available at additional cost

For Research Use Only. Not for use in diagnostic procedures.
Acknowledgements

- Ben Neale Ph.D.  
  Broad Institute
- Eimear Kenny Ph.D.  
  Mount Sinai
- Stephen Chanock Ph.D.  
  NCI
- Mitch Machiela Ph.D.  
  NCI
- Heidi Rehm Ph.D.  
  Harvard Partners
- Jouke Hottenga Ph.D.  
  University of Amsterdam
- Gail Jarvik M.D., Ph.D.  
  University of Washington
- All Consortia Members
Thank you and Questions
Preliminary Content List Disclaimer

- The data and content in these slides is considered preliminary. We want to ensure that everyone is fully aware that this will not be representative of the ‘final’ marker list that will be in the released product but rather contains all genomic locations which have been selected for inclusion on the BeadChip. However, due to additional testing and validation, some of these locations might not be available for typing on the final product.

- The information being shared is considered confidential. The marker list will be available for review under CDA.