Data-agnostic genetic admixture estimation for bioCADDIE indexed cohorts

O. Harismendy, PhD
Overview

• Rationale
  – Human subject cohorts lack racial/ethnical diversity
  – Meta-Data relies on self-reported race and ethnicity
  – Molecular data (SNP, Genome, Exome or RNA sequence) can be used to determine ancestry and quantify admixture

• Goal:
  – Include a diversity score for multiple cohorts indexed by BioCADDIE-DataMed search index

• Potential Impact/utility
  – facilitate sample and cohort selection for secondary research
  – monitor the diversity of publically available datasets
  – identify underserved study areas and populations
Lack of diversity

75% of GWAS (<2010) are performed in European

Rosenberg et al (2010)
HCM misdiagnosis

Penetrance in black would be <1%

Manrai et al (2016)
Simulations showed that the inclusion of even small numbers of black Americans in control cohorts probably would have prevented these misclassifications.

*Manrai et al (2016)*
# Chemotherapy toxicity

<table>
<thead>
<tr>
<th>Drug (Class)</th>
<th>SNP</th>
<th>Toxicity</th>
<th>Study Population</th>
<th>Gene</th>
<th>Location (GRCh38.p7)</th>
<th>MAF by Population (HapMap)</th>
<th>Refeence</th>
</tr>
</thead>
<tbody>
<tr>
<td>thiopurine (antimetabolite)</td>
<td>rs79206939</td>
<td>myelosuppression</td>
<td>East Asian (Korea)</td>
<td>FTO</td>
<td>16:53826140</td>
<td>2.8* 0 0</td>
<td>(52)</td>
</tr>
<tr>
<td>5-FU/FOLFOX</td>
<td>rs16857540</td>
<td>myelosuppression</td>
<td>European (Spain)</td>
<td>NLGN1</td>
<td>3:174182785</td>
<td>0 37.2 15.3</td>
<td>(55)</td>
</tr>
<tr>
<td>cis/carboplatin</td>
<td>rs2838566</td>
<td>hepatotoxicity</td>
<td>East Asian (China)</td>
<td>intergenic</td>
<td>21:44468699</td>
<td>7.1 33 0</td>
<td>(61)</td>
</tr>
<tr>
<td>paclitaxel + epi. (taxane)</td>
<td>rs9501929</td>
<td>neuropathy</td>
<td>European (USA)</td>
<td>TUBB2A</td>
<td>6:3157620</td>
<td>0 28.6 4.7</td>
<td>(59)</td>
</tr>
<tr>
<td>epi/doxo (anthracyclin)</td>
<td>rs229774</td>
<td>cardiotoxicity</td>
<td>diverse (Canada)</td>
<td>RARG</td>
<td>14:83435125</td>
<td>0 11.1 6.6</td>
<td>(64)</td>
</tr>
<tr>
<td>melphalan (alkylating)</td>
<td>rs1469167</td>
<td>oral mucositis</td>
<td>European (USA)</td>
<td>ALDH1A1</td>
<td>9:72942091</td>
<td>0 31.4 2.7</td>
<td>(72)</td>
</tr>
<tr>
<td>epirubicin (anthracyclin)</td>
<td>rs4149639</td>
<td>myelosuppression</td>
<td>East Asian (Japan)</td>
<td>TNFRSF1A</td>
<td>12:6332835</td>
<td>11.4 22 0</td>
<td>(63)</td>
</tr>
<tr>
<td>docetaxel (taxane)</td>
<td>rs3747851</td>
<td>myelosuppression</td>
<td>East Asian (Japan)</td>
<td>DAB2IP</td>
<td>9:121758981</td>
<td>0 15.1 0.4</td>
<td>(63)</td>
</tr>
<tr>
<td>docetaxel (taxane)</td>
<td>rs875858</td>
<td>neuropathy</td>
<td>European (USA)</td>
<td>VAC14</td>
<td>16:70741552</td>
<td>0 0 7.5</td>
<td>(58)</td>
</tr>
<tr>
<td>paclitaxel (taxane)</td>
<td>rs17348202</td>
<td>neuropathy</td>
<td>European (USA)</td>
<td>EPHA4</td>
<td>2:221207458</td>
<td>0 17.7 5.8</td>
<td>(60)</td>
</tr>
<tr>
<td>glucocorticoid</td>
<td>rs2229288</td>
<td>osteonecrosis</td>
<td>diverse (USA)</td>
<td>ZFHX3</td>
<td>16:72794405</td>
<td>0 0 0.5</td>
<td>(65)</td>
</tr>
<tr>
<td>mercaptopurine</td>
<td>rs116855232</td>
<td>dose tolerance</td>
<td>diverse (USA)</td>
<td>NUDT15</td>
<td>13:48045719</td>
<td>9.5 0 0.2</td>
<td>(67)</td>
</tr>
<tr>
<td>cisplatin</td>
<td>rs62280356</td>
<td>ototoxicity</td>
<td>European (USA)</td>
<td>WFS1</td>
<td>4:6274903</td>
<td>0.3 17.6 22.9</td>
<td>(75)</td>
</tr>
<tr>
<td>melphalan (alkylating)</td>
<td>rs1426765</td>
<td>oral mucositis</td>
<td>European (USA)</td>
<td>Intergenic</td>
<td>3:25976116</td>
<td>0 0.8 14.4</td>
<td>(72)</td>
</tr>
<tr>
<td>melphalan (alkylating)</td>
<td>rs6804277</td>
<td>oral mucositis</td>
<td>European (USA)</td>
<td>Intergenic</td>
<td>3:25977271</td>
<td>0 10 12</td>
<td>(72)</td>
</tr>
<tr>
<td>melphalan (alkylating)</td>
<td>rs1940228</td>
<td>oral mucositis</td>
<td>European (USA)</td>
<td>Intergenic</td>
<td>11:103004647</td>
<td>0 6.1 1.7</td>
<td>(72)</td>
</tr>
<tr>
<td>melphalan (alkylating)</td>
<td>rs948695</td>
<td>oral mucositis</td>
<td>European (USA)</td>
<td>Intergenic</td>
<td>11:102990584</td>
<td>0 32.2 1.7</td>
<td>(72)</td>
</tr>
<tr>
<td>oxaliplatin (platinum)</td>
<td>rs10486003</td>
<td>neuropathy</td>
<td>East Asian (Korea)</td>
<td>Intergenic</td>
<td>7:97600466</td>
<td>23.3 0 10.2</td>
<td>(73)</td>
</tr>
</tbody>
</table>

*Mapes et al (2017)*
1000 Genomes Project
### 1000 Genomes Project

#### Super Populations
- East Asian (EAS)
- European (EUR)
- African (AFR)
- Native American (AMR)
- South Asian (SAS)

#### Number of Variants
- 81,443,074 variants
Admixture

Two Colombian genomes

Admixture mapping reveals chromosome 6 association between local AM ancestry and ER+ breast cancer

Fejerman et al (2012)
Breast Cancer Risk

Global and Local ancestry explain the SNP association
The Native American ancestry is protective

<table>
<thead>
<tr>
<th></th>
<th>OR</th>
<th>95% CI</th>
<th>P value</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Model 1</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Global IA ancestry</td>
<td>0.31</td>
<td>0.18–0.54</td>
<td>3.47 × 10^{-5}</td>
</tr>
<tr>
<td><strong>Model 2</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Global IA ancestry</td>
<td>0.55</td>
<td>0.29–1.01</td>
<td>0.056</td>
</tr>
<tr>
<td>Local IA at 6q25</td>
<td>0.55</td>
<td>0.42–0.72</td>
<td>1.87 × 10^{-5}</td>
</tr>
<tr>
<td><strong>Model 3</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Global IA ancestry</td>
<td>0.55</td>
<td>0.30–1.02</td>
<td>0.06</td>
</tr>
<tr>
<td>Local IA at 6q25</td>
<td>0.68</td>
<td>0.50–0.91</td>
<td>0.009</td>
</tr>
<tr>
<td>rs140068132</td>
<td>0.63</td>
<td>0.50–0.80</td>
<td>1.6 × 10^{-4}</td>
</tr>
</tbody>
</table>

*Fejerman et al (2014)*
Outline

• Methods

• Self-Reported vs Molecular Ancestry

• Accuracy of Exome Data

• Summarizing cohort diversity
Study/Cohort selection

• Technical eligibility
  – WGS, WES, RNA-Seq, Genotyping

• Regulatory eligibility
  – General Research use (the best) or
  – No explicit barriers
    • Authorizes ancestry analysis
    • Authorizes analysis outside disease areas
    • OK with IRB exemption

• Not too fragmented
  – Some cohort are meta-cohort with different rules
    • TOPMED = 16 different cohorts and thus authorizations
<table>
<thead>
<tr>
<th>PHS</th>
<th>Name</th>
<th>Category</th>
<th>N</th>
<th>Access Granted</th>
</tr>
</thead>
<tbody>
<tr>
<td>phs000178.v9.p8</td>
<td>TCGA - The Cancer Genome Atlas</td>
<td>GRU</td>
<td>357</td>
<td>Yes*</td>
</tr>
<tr>
<td>phs000677.v1.p1</td>
<td>APOBEC Cytidine Deaminase Mutagenesis Pattern is Widespread in Human</td>
<td>GRU</td>
<td>5427</td>
<td>Yes*</td>
</tr>
<tr>
<td></td>
<td>Cancer</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>phs000147.v3.p1</td>
<td>CGEMS Breast Cancer</td>
<td>GRU</td>
<td>2576</td>
<td>Yes</td>
</tr>
<tr>
<td>phs000935.v1.p1</td>
<td>CSER: The CHOP/UPenn Pediatric Genetic Sequencing Project</td>
<td>GRU</td>
<td>159</td>
<td>Yes</td>
</tr>
<tr>
<td>phs000958.v1.p1</td>
<td>CSER: The MedSeq Project</td>
<td>GRU</td>
<td>100</td>
<td>Yes</td>
</tr>
<tr>
<td>phs000999.v1.p1</td>
<td>CSER: Clinical Sequencing in Cancer: Clinical, Ethical, and Technological Studies</td>
<td>GRU</td>
<td>131</td>
<td>Yes</td>
</tr>
<tr>
<td>phs000673.v1.p1</td>
<td>CSER: Exploring Precision Cancer Medicine for Sarcoma and Rare Cancers</td>
<td>GRU</td>
<td>11</td>
<td>Yes</td>
</tr>
<tr>
<td>(phs001089.v2.p1</td>
<td>CSER: Genomic Diagnosis in Children with Developmental Delay</td>
<td>GRU</td>
<td>835</td>
<td>Yes</td>
</tr>
<tr>
<td>phs000827.v3.p1</td>
<td>CSER: NCGENES (North Carolina Clinical Genomics Evaluation by NextGen Exome Sequencing)</td>
<td>GRU</td>
<td>592</td>
<td>Yes</td>
</tr>
<tr>
<td>phs000750.v1.p1</td>
<td>Genomic Origins and Admixture in Latinos (GOAL)</td>
<td>GRU</td>
<td>271</td>
<td>Yes</td>
</tr>
<tr>
<td>(phs000424.v6.p1</td>
<td>Genotype-Tissue Expression (GTEx) Common Fund Project</td>
<td>GRU</td>
<td>572</td>
<td>Yes</td>
</tr>
<tr>
<td>phs000925.v1.p1</td>
<td>PAGE: The Charles Bronfman Institute for Personalized Medicine (IPM) BioMe Biobank</td>
<td>GRU</td>
<td>13067</td>
<td>Yes</td>
</tr>
<tr>
<td>(phs000442.v1.p1</td>
<td>Drug Resistant Hypertension in African Americans' Exome</td>
<td>GRU</td>
<td>91</td>
<td>NO</td>
</tr>
<tr>
<td>phs000518.v1.p1</td>
<td>NHLBI GO-ESP: Family Studies (Idiopathic Bronchiectasis)</td>
<td>GRU</td>
<td>24</td>
<td>NO</td>
</tr>
<tr>
<td>phs000354.v1.p1</td>
<td>NHLBI GO-ESP: Family Studies (Pulmonary Arterial Hypertension)</td>
<td>GRU</td>
<td>12</td>
<td>NO</td>
</tr>
<tr>
<td>(phs000422.v1.p1</td>
<td>NHLBI GO-ESP: Lung Cohorts Exome Sequencing Project (Asthma)</td>
<td>GRU</td>
<td>191</td>
<td>NO</td>
</tr>
<tr>
<td>(phs000254.v2.p1</td>
<td>NHLBI GO-ESP: Lung Cohorts Exome Sequencing Project (Cystic Fibrosis)</td>
<td>GRU</td>
<td>242</td>
<td>NO</td>
</tr>
<tr>
<td>phs000291.v2.p1</td>
<td>NHLBI GO-ESP: Lung Cohorts Exome Sequencing Project (Lung Health Study of Chronic Obstructive Pulmonary)</td>
<td>GRU</td>
<td>337</td>
<td>NO</td>
</tr>
<tr>
<td>phs000290.v1.p1</td>
<td>NHLBI GO-ESP: Lung Cohorts Exome Sequencing Project (Pulmonary Arterial Hypertension)</td>
<td>GRU</td>
<td>96</td>
<td>NO</td>
</tr>
<tr>
<td>phs001026.v1.p1</td>
<td>CSER: Incorporation of Genomic Sequencing into Pediatric Cancer Care</td>
<td>GRU</td>
<td>125</td>
<td>NO</td>
</tr>
<tr>
<td>phs000812.v1.p1</td>
<td>Characterizing Genetic Susceptibility to Breast and Prostate Cancer - BPC3</td>
<td>GRU</td>
<td>1054</td>
<td>Yes</td>
</tr>
<tr>
<td>phs000306.v4.p1</td>
<td>GENEVA Prostate Cancer</td>
<td>GRU</td>
<td>4642</td>
<td>Yes</td>
</tr>
<tr>
<td>phs000413.v2.p1</td>
<td>St. Jude Children's Research Hospital - Washington University Pediatric Cancer Genome Project (PCGP): Somatic Mutations in Pediatric AML FAB-M7 Subtype by Whole Transcriptome Sequencing</td>
<td>GRU</td>
<td>15</td>
<td>Yes</td>
</tr>
<tr>
<td>phs000414.v1.p1</td>
<td>St. Jude Children's Research Hospital - Washington University Pediatric Cancer Genome Project (PCGP): Whole Genome Sequencing of Core Binding Factor Acute Myeloid Leukemia</td>
<td>GRU</td>
<td>17</td>
<td>Yes</td>
</tr>
<tr>
<td>phs000409.v1.p1</td>
<td>St. Jude Children's Research Hospital - Washington University Pediatric Cancer Genome Project (PCGP): Whole Genome Sequencing of Childhood Medulloblastoma</td>
<td>GRU</td>
<td>93</td>
<td>Yes</td>
</tr>
<tr>
<td>phs000341.v2.p1</td>
<td>St. Jude Children's Research Hospital - Washington University Pediatric Cancer Genome Project: Whole Genome Sequencing of Childhood Hypodiploid Acute Lymphoblastic Leukemia</td>
<td>GRU</td>
<td>241</td>
<td>Yes</td>
</tr>
</tbody>
</table>

26 requested, 12 granted, 7 rejected, 7 pending
Calculate maximum likelihood estimates (MLE) of each reference population using Broyden-Fletcher-Goldfarb-Shannon (BFGS) optimization algorithm
Log-Likelihood model - iAdmix

$q_{ij}$ denote the allele frequency of the $a$ allele at the $i$-th SNP in population $j$

$a_j$ represent the admixture proportion for the $j$-th population and $A=[a_1,a_2,...,a_k]$ be the vector of admixture coefficients

\[
 f_i = \sum_{j=1}^{k} q_{ij} a_j 
\]

weighted allele frequency at SNP $i$ given the allele frequencies and admixture proportions

\[
p(G_i|f_i) = \begin{cases} 
(1 - f_i)^2 & \text{if } G_i = 0 \\
2f_i(1 - f_i) & \text{if } G_i = 1 \\
f_i^2 & \text{if } G_i = 2
\end{cases}
\]

probability if observing the genotype $G_i$ at site $i$

\[
 L(A) = \sum_{i=1}^{n} \ln(Pr(G_i = g_i|f_i))
\]
given vector of admixture proportions, the log-likelihood of the observed genotypes $g$ for an individual

GOAL: Determine the vector $A=[a_1,a_2,...,a_k]$ of admixture proportions that maximizes $L(A)$
• Methods

• Self-Reported vs Molecular Ancestry

• Accuracy of Exome Data

• Summarizing cohort diversity
TCGA self-reported race

Using SNP6 affymetrix array to call molecular ancestry and admixture
Methodology

- Using SNP6 affymetrix array to call molecular ancestry and admixture (iadmix)
- Defining dominant ancestry (fraction > 0.8)
- Defining admixed (max < 0.8)
- 6 possible dominant
  - EUR
  - AFR
  - EAS
  - SAS
  - AMR
  - admixed
Matching

- White: >N EUR
- Black: >N AFR
- Asian: >N EAS+SAS
Dominant (0.8) Ancestries

- **Asian**
  - Admixed
  - EUR
  - AFR
  - EAS
  - SAS
  - AMR

- **Black**
  - Admixed
  - EUR
  - AFR
  - EAS
  - SAS
  - AMR

- **White**
  - Admixed
  - EUR
  - AFR
  - EAS
  - SAS
  - AMR

- **Hispanic**
  - Admixed
  - EUR
  - AFR
  - EAS
  - SAS
  - AMR
Admixed people tend to self report as white, not Hispanic.
Black

black
White
Hispanic

race
- american indian or alaska native
- asian
- black or african american
- native hawaiian or other pacific islander
- not reported
- white

ethnicity
- hispanic or latino
Outline

• Methods

• Self-Reported vs Molecular Ancestry

• Accuracy of Exome Data

• Summarizing cohort diversity
Diverse Exomes

- Selected 100 cases from diverse Race/Ethnicities
- Called variants using FreeBayes pipeline (Marth lab)
- Filtered for dbSNP and high quality/coverage
WXS variants

~20000 dbSNP variants per exome
WXS and SNP array lead to identical diversity estimates
Comparing 1KG Ancestry Estimates

- EUR
- EAS
- AMR
- AFR
- SAS

Graphs showing the comparison of ancestry estimates for different populations.
Comparing Dominant Ancestry

At 80% dominance threshold, ~30% of the subjects are misclassified
Outline

• Methods

• Self-Reported vs Molecular Ancestry

• Accuracy of Exome Data

• Summarizing cohort diversity
Cumulative Fraction by Disease
Dominant (0.8) Ancestries

TCGA

COAD

BRCA

LUAD

- admixed
- EUR
- AFR
- EAS
- SAS
- AMR
Entropy

- Diversity:
  - Normalized Entropy over N primary (>80%) groups

\[
E = 0.92
\]

\[
E = 0.18
\]
Dominant Ancestries & Entropy

TCGA  E=0.75

COAD  E=0.69

BRCA  E=0.81

LUAD  E=0.63

Legend:
- admixed
- EUR
- AFR
- EAS
- SAS
- AMR
Dominant Ancestries & Entropy

- Asian: dominant ancestry, Entropy = 0.50
- Black: dominant ancestry, Entropy = 0.42
- White: dominant ancestry, Entropy = 0.46
- Hispanic: dominant ancestry, Entropy = 0.89

Legend:
- admixed
- EUR
- AFR
- EAS
- SAS
- AMR
## GWAS cancer

### Repositories

- SRA (16,001)
- OmicsDI (9,952)
- BioProject (9,170)
- ArrayExpress (8,979)
- PDB (1,046)
- GEO (443)
- dbGaP (364)
- CIL (340)
- Dryad (309)
- GEMMA (285)

### Disease

- Breast Neoplasms (49)
- Neoplasms (47)
- Lung Neoplasms (34)
- Prostatic Neoplasms (31)
- Lymphoma, Non-Hodgkin (28)
- Melanoma (23)
- Uterine Cervical Neoplasms (20)
- Leukemia, Myeloid, Acute (18)
- Colonic Neoplasms (17)
- Lymphoma, Large B-Cell, Diffuse (17)

### Study Types

- Case Set (75)
- Case-Control (69)
- Cohort (44)
- Tumor vs. Matched-Normal (20)
- Longitudinal (16)
- Family (6)
- Nested Case-Control (4)
- Whole Genome Sequencing (4)
- Control Set (3)
- Longitudinal Cohort (3)

### Study Group

- General Research Use
- Research related to adult diseases and methods (CADM)
- Cancer in all age groups, other diseases in adults only, and methods
- Pancreatic cancer only
- Pancreatic cancer only and GHP institutional certification
- Pancreatic cancer only and JHU institutional certification
- Disease-Specific (Breast, Ovarian, or Endometrial Disease, MDS)
- Cancer in all age groups, other disease in adults only and methods
### GWAS cancer search results

#### Repositories
- SRA (16,001)
- OmicsDI (9,952)
- BioProject (9,170)
- ArrayExpress (8,979)
- PDB (1,046)
- GEO (443)
- dbGaP (364)
- CIL (340)
- Dryad (309)
- GEMMA (285)
- More...

#### Disease
- Breast Neoplasms (49)
- Neoplasms (47)
- Lung Neoplasms (34)
- Prostatic Neoplasms (31)
- Lymphoma, Non-Hodgkin (25)
- Melanoma (23)
- Uterine Cervical Neoplasms (20)
- Leukemia, Myeloid, Acute (18)
- Colonic Neoplasms (17)
- Lymphoma, Large B-Cell, Diffuse (17)

#### Study Types
- Case-Control
- Nested Case-Control

#### Study Group
- General Research Use
- Research related to adult diseases and methods (CADM)
- Cancer in all age groups, other diseases in adults only, and methods
- Pancreatic cancer only
- Pancreatic cancer only and GHP institutional certification
- Pancreatic cancer only and JHU institutional certification
- Disease-Specific (Breast, Ovarian, or Endometrial Disease, MDS)
- Cancer in all age groups, other disease in adults only and methods

#### Ancestry
- Main Ancestry Fraction
- EUR: 0.50
- AFR
- EAS
Conclusions

• Global and Local Ancestry are critical covariates in genetic studies
• Self reported ancestry is not a reliable meta-data
• Exome and SNP arrays provide consistent estimates of ancestry
• Entropy and major/dominant ancestry are concise ways to summarize cohort diversity
Acknowledgments

Jihoon Kim
Max Xu
Lucila Ohno Machado

BioCADDIE team
BioCADDIE funding
Clinical and Translational Research Institute
iDASH cloud