Square wheels: electronic medical records for discovery research in rheumatoid arthritis

Robert M. Plenge, M.D., Ph.D.
October 15, 2009
Clinical Genomics Workshop
October 2009: >30 RA risk loci

Together explain ~35% of the genetic burden of disease

<table>
<thead>
<tr>
<th>Year</th>
<th>Genes</th>
</tr>
</thead>
<tbody>
<tr>
<td>1978</td>
<td>HLA DR4 “shared epitope” hypothesis</td>
</tr>
<tr>
<td>1987</td>
<td>PAD1, PTPN22, CTLA4</td>
</tr>
<tr>
<td>2003</td>
<td>PADI4</td>
</tr>
<tr>
<td>2004</td>
<td>CTLA4</td>
</tr>
<tr>
<td>2005</td>
<td>IL2-IL21</td>
</tr>
<tr>
<td>2007</td>
<td>TNFAIP3, STAT4, TRAF1-C5, IL2-IL21</td>
</tr>
<tr>
<td>2008</td>
<td>CD40, CCL21, CD244, IL2RB, PTPRC, FCGR2A, PRDM1, CD2-CD58</td>
</tr>
<tr>
<td>2009</td>
<td>REL, BLK, TAGAP, TRAF6, PTPRC, FCGR2A, PRDM1, CD2-CD58</td>
</tr>
</tbody>
</table>

Latest GWAS in 25,000 case-control samples with replication in 20,000 additional samples: >10 new loci
How can we collect DNA and detailed clinical data on >20,000 RA patients?
What are the options for collecting clinical data and DNA for genetic studies?
## Options for clinical + DNA

<table>
<thead>
<tr>
<th>design</th>
<th>Clinical data</th>
<th>DNA</th>
<th>Sample size</th>
<th>cost</th>
</tr>
</thead>
<tbody>
<tr>
<td>clinical trial</td>
<td>+++</td>
<td>+++</td>
<td>+</td>
<td>$$$</td>
</tr>
<tr>
<td>registry</td>
<td>++</td>
<td>+++</td>
<td>++</td>
<td>$$</td>
</tr>
<tr>
<td>claims data</td>
<td>+</td>
<td>n/a</td>
<td>+++</td>
<td>$</td>
</tr>
<tr>
<td>EMR</td>
<td>++</td>
<td>+++</td>
<td>+++</td>
<td>$</td>
</tr>
</tbody>
</table>
Partners HealthCare: 4 million patients
Partners HealthCare: linked by EMR
Partners HealthCare: organized by i2b2
• NIH-funded National Center for Biomedical Computing

• Core groups develop and distribute open source software for EMR data

• RA is a ‘Driving Biology Project’ to beta-test the system

www.i2b2.org/disease/arthritis.html
4 million patients

**ICD9 RA and/or CCP checked**
(goal = high sensitivity)

31,171 patients

**Classification algorithm**
(goal = high PPV)

3,585
RA patients

Discarded blood for DNA

Clinical subsets
Linking the Datamart-Crimson

NLP data

Narrative electronic medical record
Natural Language Processing (NLP)

Codified data

codified data (e.g., billing codes)

i2b2 informatics infrastructure

Algorithm to define patients with RA

i2b2 RA-DataMart
30,655 patients
NLP queries
autoantibody status
medication history
codified data
billing codes
laboratory values

RA patients
IDs
13100
65773
23001
12543

Clinical laboratory

sample processing

Match!

FIREWALL

anonymous clinical data
discarded blood sample for DNA

FINAL ANONYMOUS PATIENT CLINICAL DATA
WITH DNA FOR GENETIC STUDIES
Status of *i2b2* Crimson collection

- Over 3,000 samples collected to date
  - *cost = $10 per sample*

- DNA extracted on >2,400 Buffy coats
  - *cost = $20 per sample*
  - >90% had ≥1 ug of DNA
  - >99% had ≥5 ug of DNA after WGA

**genotyping of 384 SNPs (RA risk alleles, AIMs, other) is ongoing...**data soon!
Acknowledgments

Zak Kohane
Susanne Churchill
Vivian Gainer
Kat Liao
Tianxi Cai
Shawn Murphy
Qing Zing
Soumya Raychaudhuri
Beth Karlson
Pete Szolovits
Lee-Jen Wei
Lynn Bry (Crimson)
Sergey Goryachev
Barbara Mawn
& many others!

Namaste!