Toward Precision Medicine: Building a patient centric information commons on common and rare diseases with I2b2/tranSMART
Application to Autism and Phelan McDermid Syndrome

Paul Avillach, MD, PhD
Assistant Professor - Harvard Medical School
Center of Biomedical Informatics
Research Connection – CHIP- Boston Children’s Hospital
Toward Precision Medicine: Building a Knowledge Network for Biomedical Research and a New Taxonomy of Disease
Report from National academy of science, USA, 2011
i2b2

Health care
Health Information System

Browser tools available for Researchers

Clinical Research

1) Analysis tools

2) Structured data from research studies

3) ‘omics data

i2b2

ETL once a week

- DRG
- EHR forms
- EHR reports
- Biology
- Imaging
- Pathology
- Rx

‘omics

Phenotypic

‘omics

ETL
• Integrated platform to support translational research
• Initiated by Johnson & Johnson et Recombinant 7 years ago

• Open-source since January 24th, 2012
• Installed at HEGP Hospital, Paris since May, 2012
• Today, driven and maintained by the tranSMART

http://transmartfoundation.org
Objectives:

1. **Integration** of clinical, biological and ‘omics data in one place – hypothesis free –

2. Generation of **hypothesis** by Clinicians / Researchers
Autism cohorts
Phenotype data

- Simons Simplex Collection (SSC) 2,760
- AGRE 3,300
- Autism Consortium (AC) 525
- BCH i2b2 HER data 16,587
Known relationships between genes

Sequence variation

Copy number variation

Gene expression

+ many more
<table>
<thead>
<tr>
<th>Sample Code</th>
<th>Library Code</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>SSC (928)</td>
<td></td>
<td>Blood (Kunkel-Kohane)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>- Affymetrix Gene ST 1.0 316</td>
</tr>
<tr>
<td></td>
<td></td>
<td>- Affymetrix U133+2 19</td>
</tr>
<tr>
<td></td>
<td></td>
<td>- Illumina HiSeq 154</td>
</tr>
<tr>
<td>LCL (Geschwind)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>- Illumina REF-8 3.0 439</td>
</tr>
<tr>
<td>AC (166)</td>
<td></td>
<td>Blood (Kunkel-Kohane)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>- Affymetrix Gene ST 1.0 117</td>
</tr>
<tr>
<td></td>
<td></td>
<td>- Affymetrix U133+2 21</td>
</tr>
<tr>
<td></td>
<td></td>
<td>- Illumina HiSeq 28</td>
</tr>
<tr>
<td>BCH (386)</td>
<td></td>
<td>Blood (Kunkel-Kohane)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>- Affymetrix Gene ST 1.0 186</td>
</tr>
<tr>
<td></td>
<td></td>
<td>- Affymetrix U133+2 168</td>
</tr>
<tr>
<td></td>
<td></td>
<td>- Illumina HiSeq 32</td>
</tr>
<tr>
<td>AGRE (1,048)</td>
<td></td>
<td>LCL (Geschwind)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>- Illumina REF-8 3.0 1,048</td>
</tr>
</tbody>
</table>
• Static genomic predisposition
  – Goal: enable streamlined analysis of genomic variation at any functional unit resolution
  • Single variant / variant type
  • Single gene
  • Gene set / pathway
  • Regulatory module
  • Cellular system
  • Genomic location / context
  ...

• Measurement types
Raw genotyping array data by cohort

- **SSC (3,184)**
  - Illumina Infinium1M
  - Illumina 1M Duo

- **AC (60)**
  - Affymetrix SNP 6.0

- **AGRE (3,832)**
  - Affymetrix 10K
  - Affymetrix 500K
  - Illumina HumanHap550
  - Illumina Infinium 1M
WHOLE EXOME SEQUENCE data by cohort

- **SSC (2,963)**
  - State 914
    - Illumina GAIIx
    - Illumina HiSeq
  - Eichler 676
    - Illumina GAIIx
    - Illumina HiSeq
  - Wigler 1373
    - Illumina HiSeq
- **AC (381)**
  - Daly 381
    - Illumina HiSeq
- **AGRE (1672)**
  - Walsh 750
    - Illumina HiSeq
  - BI-BCM 922
    - Illumina HiSeq
    - ABI SOLiD
Exome sequence data processing

- Raw Reads
- Map To Reference
- Mark Duplicates
- Indel Realignment
- Base Recalibration
- RR Compression
- Analysis-Ready Reads

Variant calling

- Analysis-Ready Reads
- Joint Variant Calling
- Raw Variants
- Variant Recalibration (separately per variant type)
- Filtered Variants

Variant annotation

- Physical location
  - e.g. Chr:start-end
  - Cytoband
- Gene
  - e.g. Gene name
  - Variant function
- Gene set
  - e.g. Pathway
  - Molecular process
- Predicted variant impact
  - e.g. SIFT
  - PolyPhen
- Conservation
  - e.g. GERP
  - PhyloP
- Population frequency
  - e.g. 1000 Genomes
  - ESP 6500
- Clinical significance
  - e.g. ClinVar
  - OMIM
- Expression patterns
  - e.g. GTEx
  - BrainSpan
- Transcriptional regulation
  - e.g. ENCODE TFBS
  - Histone modifications

Comprehensively annotated variants

Individual genotypes

ANNOVAR

Boston Children's Hospital

HARVARD MEDICAL SCHOOL
TEACHING HOSPITAL
Live DEMO

https://www.youtube.com/watch?v=rUFH697a2n4
**Genome Wide Association Study**
(1 Phenotype compared to ALL SNPs)

- cases (ex: systemic sclerosis)
  - cases DNA
- controls
  - controls DNA

compare ALL SNPs to find differences between cases and controls

**Phenome Wide Association Study**
(1 SNP compared to ALL Phenotypes)

- allele G patients group
- allele A patients group

- allele G patients phenotype
- allele A patients phenotype

compare ALL DIAGNOSIS to find differences between cases and controls
Phenome-Wide Association Studies on a Quantitative Trait: Application to TPMT Enzyme Activity and Thiopurine Therapy in Pharmacogenomics

Antoine Neuraz\textsuperscript{1,2}, Laurent Chouchana\textsuperscript{3}, Georgia Malamut\textsuperscript{4}, Christine Le Beller\textsuperscript{5}, Denis Roche\textsuperscript{6}, Philippe Beaune\textsuperscript{3,6}, Patrice Degoulet\textsuperscript{1,2}, Anita Burgun\textsuperscript{1,2}, Marie-Anne Loriot\textsuperscript{3,6}, Paul Avillach\textsuperscript{1,2*}

1 Biomedical Informatics and Public Health Department, University Hospital HEGP, AP-HP, Paris, France, 2 INSERM UMR S 872 Team 22: Information Sciences to support Personalized Medicine, Université Paris Descartes, Sorbonne Paris Cité, Faculté de Médecine, Paris, France, 3 INSERM UMR S 775, Université Paris Descartes, Sorbonne Paris Cité, Paris, France, 4 Gastroenterology Department, University Hospital HEGP, AP-HP, Paris, France, 5 Pharmacovigilance Center, University Hospital HEGP, AP-HP, Paris, France, 6 Biochemistry, Pharmacogenetics and Molecular Oncology Unit, University Hospital HEGP, AP-HP, Paris, France
PCORnet  $105 million  Start date Jan 2014  18 months

Scientific Advisory Board  
Special Expert Group  

PCORI  
Steering Committee  

Coordinating Center  

CDRN  $7M  
11 CDRN $7M  

PPRN  $1M  
18 PPRN $1M
PCORI PPRN Grant

• Phelan-McDermid Syndrome Data Network
  – PI: Megan O’Boyle, Mother of PMS patient
  – Co-PI: Paul Avillach, MD, PhD
• Total: $1M
• 18 months
• To collect all available patient data from Phelan-McDermid Syndrome (PMS) patients to make meaningful, well-annotated clinical data available to researchers and to share insights with members of the PCORI network
Health care providers

Contact and retrieve EMR clinical notes from their Health care providers

Patients / Parents

Individual patient data entry including clinical notes

Patient ownership and governance of data

Aggregated or individual patient data consultation

Researchers

Omics' data

Secure web Registry

PPRN: Phelan-McDermid Syndrome Data Network (PMS_DN)

Already in place

PCORI - PPRN project

Shrine

Collaboration with Clinical Data Research Networks (CDRN) - PCORI

For example: Scalable Collaborative Infrastructure for a Learning Health System (SCILHS) to find new patients with Phelan-McDermid Syndrome across all their network of 9 Hospitals

Images of Clinical Notes

Optical character recognition

Harvard Medical School Research Computing Private Orchestra: Phelan-McDermid syndrome research environment

Textual Clinical Notes

Natural Language processing

Anonymized curated Clinical data from Clinical notes

Integration of multiple heterogeneous sources of clinical and omics' data using international standards

Exchange interface, with other projects, based on international standards

Clinical data from Registry

Secure web Research platform Clinical and omics' data

Firewall
### Box 1 | Natural language processing

<table>
<thead>
<tr>
<th>Boundary detection</th>
<th>… … ]</th>
<th>[ Fx of obesity but no fx of coronary artery diseases. ]</th>
<th>[ … …</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tokenization</td>
<td></td>
<td></td>
<td>Fx of obesity but no fx of coronary artery diseases.</td>
</tr>
<tr>
<td>Normalization</td>
<td>– – – – – – – – – – – disease_</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Part-of-speech tagging</td>
<td>NN IN NN CC DT NN IN JJ NN NNS</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Shallow parsing</td>
<td>NP PP NP NN NP</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Entity recognition</td>
<td>Obesity Disease or disorder UMLS ID: C0028754 Status: family history Negated: no</td>
<td>Coronary artery disease Disease or disorder UMLS ID: C0010054 Status: family history Negated: yes</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Coronary artery Anatomy UMLS ID: C0205042</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Mrs. Jones is an 84-year-old African-American woman admitted from the emergency department with a complaint of crushing substernal pain... past medical history is significant for a 20-year history of type 2 diabetes mellitus controlled with oral hypoglycaemics. 2 ppd history smoking...

Family history: Sister died from myocardial infarction at 74 years...

Mrs. Jones was discharged on a 1,500 ml fluid restriction, nitroglycerin 0.4 mg/spray 1–2 spray po. Aciphex 20 mg (20 mg tablet DR take 1) PO

Discharge diagnosis: acute MI, diabetes mellitus...

Patient Centric Information Commons (PIC)
PI: Isaac Kohane

RESTful API
Table 4.2: Summary of the different data types available for the NDD PICI and their provenance

<table>
<thead>
<tr>
<th>Patient level databases</th>
<th>Aggregated level databases</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>EHR derived databases</strong></td>
<td><strong>Research cohorts</strong></td>
</tr>
<tr>
<td>Partner HealthCare</td>
<td>Boston Children’s</td>
</tr>
<tr>
<td><strong>Total number of individuals (including family if available)</strong></td>
<td>3,538,474</td>
</tr>
<tr>
<td><strong>Number of patients with a NDD</strong></td>
<td>83,137</td>
</tr>
<tr>
<td><strong>CLINICAL</strong></td>
<td></td>
</tr>
<tr>
<td>EHR Clinical Notes</td>
<td></td>
</tr>
<tr>
<td>EHR ICDs / Procedures</td>
<td></td>
</tr>
<tr>
<td>EHR lab values</td>
<td></td>
</tr>
<tr>
<td>EHR Vital signs</td>
<td></td>
</tr>
<tr>
<td>EHR Drug prescriptions</td>
<td></td>
</tr>
<tr>
<td>Structured clinical data (Cohorts)</td>
<td></td>
</tr>
<tr>
<td>Patient reported clinical data</td>
<td></td>
</tr>
<tr>
<td><strong>BIO</strong></td>
<td></td>
</tr>
<tr>
<td>Biobank data (is there a sample and where?)</td>
<td></td>
</tr>
<tr>
<td><strong>OMICS</strong></td>
<td></td>
</tr>
<tr>
<td>Microarray SNPs</td>
<td></td>
</tr>
<tr>
<td>Gene expression</td>
<td></td>
</tr>
<tr>
<td>Whole Exomes Sequencing</td>
<td></td>
</tr>
<tr>
<td>RNA sequencing</td>
<td></td>
</tr>
<tr>
<td>Curated genetic reports</td>
<td></td>
</tr>
<tr>
<td><strong>NEURO IMAGING</strong></td>
<td></td>
</tr>
<tr>
<td>DICOM Radiology images</td>
<td></td>
</tr>
<tr>
<td>Electroencephalography</td>
<td></td>
</tr>
<tr>
<td>functional Near Infrared Spectroscopy (fNIRS)</td>
<td></td>
</tr>
<tr>
<td>Event Related Potentials (ERP)</td>
<td></td>
</tr>
<tr>
<td>Eye tracking</td>
<td></td>
</tr>
<tr>
<td><strong>FAMILY HISTORY</strong></td>
<td></td>
</tr>
<tr>
<td>Family linkage</td>
<td></td>
</tr>
<tr>
<td><strong>CONSENT</strong></td>
<td></td>
</tr>
<tr>
<td>Details of Patient consent (or no consent) available</td>
<td></td>
</tr>
<tr>
<td><strong>ENVIRONMENT</strong></td>
<td></td>
</tr>
<tr>
<td>Exposure biomarkers</td>
<td></td>
</tr>
<tr>
<td>Self-reported diet</td>
<td></td>
</tr>
<tr>
<td>Self-reported behavior (e.g., smoking, physical)</td>
<td></td>
</tr>
<tr>
<td>Unstructured clinical staff notes</td>
<td></td>
</tr>
<tr>
<td>Socioeconomic status (e.g., education, income)</td>
<td></td>
</tr>
<tr>
<td>Environmental monitoring (ecological/sensor info)</td>
<td></td>
</tr>
</tbody>
</table>
P50: Centers of Excellence in Genomic Science
Neuropsychiatric Genome-Scale and RDOC Individualized Domains (N-GRID)
PI: Isaac Kohane

500 subjects will use the same i2b2/tranSMART infrastructure to store and access all the following datatypes:

- Results of 4h of self-report measures, clinician interview, and neuropsych testing.
- EMR notes and codes
- RDoC-parsing (NLP) of EMR notes
- **Whole-genome** (not just exome) sequencing
- GWAS (illumina psych chip - ~250k SNPs + CNV's) for at least a subset
- RNA-seq (baseline and possibly additional experimental conditions)
- L1000 (expression) data for up to 400+ conditions x multiple cell lines
- illumina expression data x multiple lines (fibroblast, NPC, differentiated neurons)
- nanostring (expression) data x multiple lines (iPS, NPC, differentiated neurons)
- ATAC-seq data x multiple lines
- Additional data from Greenberg lab (electrophys etc)
Autism Cohort

Division of Developmental Medicine
Leonard Rappaport, MD, MS
Ellen Hanson, PhD

BCH Division of Genetics & Genomics
Timothy Yu, MD, PhD
Ingrid Holm, MD, MPH
Stephanie Brewster, MS, CGC
Joanna Reinwald, MS, GC
Frank Jackson

Laboratory of cognitive neuroscience
Charles Nelson, PhD
Vanessa Vogel-Farley
Nicole Coman

Principal Investigators
Isaac Kohane, MD, PhD
Louis Kunkel, PhD
David Margulies, MD
Jonathan Bickel, MD, MS
Paul Avillach, MD, PhD

CBMI / ResCon tranSMART team
Paul Avillach, MD, PhD
Michael McDuffie, MS
Ally Eran, PhD

The Research Connection
Wendy Wolf, PhD
Sarah Savage, MS, CGC
Catherine Clinton, MS, CGC
Tram Tran

Business Intelligence and Clinical Research Informatics
Mohamad Daniar
Nandan Patibandla
Rick Agrella
Paul OByrne
Lynne N. Alley
Gina Bianco

CBMI
Eric D Perakslis, PhD
Alexa T. McCray, PhD
Dennis Wall, PhD
Nathan Palmer, PhD
Sek Won Kong, MD
Finale Doshi-Velez, PhD

Clinical NLP
Guergana Savova, PhD - PI
Chen Lin
Dmitriy Dligach, PhD
Pei Chen
Sameer Pradhan, PhD
Sean Finan
Timothy Miller, PhD

i2b2 / Partners
Shawn Murphy, MD, PhD
Lori Phillips, Ms
Michael Mendis
PCORI PMS_DN team

Megan O’Boyle, PI & Mom of Shannon
• Paul Avillach, MD, PhD, Co-PI, Harvard Medical School
• Liz Horn, PhD, Co-PI, Network Director

PMSF Research director
• Geraldine Bliss, MSc & Mom of

Project Manager
• Andria Cornell Mann

LGC Data Network Specialist
• Rebecca Davis

Family Engagement Specialist
• Jackie Malasky

Harvard Medical School: CBMI
• Sushma Hanawal
• Michael McDuffie, MSc
• Isaac Kohane, MD, PhD
• Eric Perakslis, PhD

Boston Children’s Hospital: cTAKES NLP
• Guergana Savova, PhD
• Pei Chen

Harvard Medical School: IT Infrastructure Support
• Christopher Botka
• David Hummel
• Daniel Lewis
Avillach Lab
http://avillach-lab.hms.harvard.edu

Developers
• Michael McDuffie, MSc
• Sushma Hanawal
• Pei Chen

Research Associates
• Antoine Neuraz, MD, MSc
• Cartik Saravana, PhD

Graduate students
• Maxime Wack, MD
• Claire Hassen-Kodja, MD, MSc
• Samuel Finlayson

Previous members
• Ephi Sachs, MD
• Heike Lehmann

We are hiring now
• Senior Software Developer *2
• Senior Database Developer *2
• Postdocs *2
P50: Centers of Excellence in Genomic Science
Neuropsychiatric Genome-Scale and RDOC Individualized Domains (N-GRID)
PI: Isaac Kohane

Research Domain Criteria Matrix

<table>
<thead>
<tr>
<th>Domain</th>
<th>Units of Analysis</th>
<th>Genes</th>
<th>Molecules</th>
<th>Cells</th>
<th>Circuits</th>
<th>Physiology</th>
<th>Behavior</th>
<th>Self-Reports</th>
<th>Paradigms</th>
</tr>
</thead>
<tbody>
<tr>
<td>Negative Valence Systems</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Acute threat (“fear”)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Potential threat (“anxiety”)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sustained threat</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Loss</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Frustrative nonreward</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Positive Valence Systems</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>