Translating a Trillion Points of Open Data into Diagnostics, Therapies and New Insights in Health and Disease

Atul Butte, MD, PhD
Chief Data Scientist, University of California Health (UC Health)
Director, Bakar Computational Health Sciences Institute, UCSF
Priscilla Chan and Mark Zuckerberg Distinguished Professor
Conflicts of Interest

• Scientific founder and advisory board membership
  – Genstruct
  – NuMedii
  – Personalis
  – Carmenta

• Honoraria for talks
  – Lilly
  – Pfizer
  – Siemens
  – Bristol Myers Squibb
  – AstraZeneca
  – Roche
  – Genentech
  – Warburg Pincus
  – CRG
  – AbbVie
  – Westat

• Past or present consultancy
  – Lilly
  – Johnson and Johnson
  – Roche
  – NuMedii
  – Genstruct
  – Tercica

• Corporate Relationships
  – Ecoeos
  – Helix
  – Ansh Labs
  – uBiome
  – Prevendia
  – Samsung
  – Assay Depot
  – Regeneron
  – Verinata
  – Pathway Diagnostics
  – Geisinger Health
  – Covance
  – Wilson Sonsini Goodrich & Rosati
  – Orrick
  – 10X Genomics
  – GNS Healthcare
  – Gerson Lehman Group
  – Coature Management

• Speakers’ bureau
  – None

• Companies started by students
  – Carmenta
  – Serendipity
  – Stimulomics
  – NunaHealth
  – Praedicat
  – MyTime
  – Flipora
  – Tumbl.in
  – Polyglot
  – Iota Health
  – Ongevity Health
DNA microarrays allow researchers to analyse the expression of a huge number of genes simultaneously.

**GENOMICS**

**Gene data to hit milestone**

With close to one million gene researchers can identify disease

**BY MONYA BAKER**

Praveen Khatri sits in front of an oversize computer screen, trawling for treasure in a sea of genetic data. Entering the search term ‘breast cancer’ into a public repository called the Gene Expression Omnibus (GEO), the postdoctoral researcher retrieves a list of 1,170 experiments, representing nearly 33,000 samples and a hoard of gene-expression data that could reveal previously unseen patterns.

That is exactly the kind of search that led Khatri’s boss, Atul Butte, a bioinformatician at the Stanford School of Medicine in California, to identify a new drug target for diabetes. After downloading data from 130 gene-expression studies in mice, rats and humans, Butte looked for genes that were expressed at higher levels in the diseased animals.

The result is an unprecedented resource that promises to drive down costs and speed up progress in understanding disease. Gene-sequence data are already shared extensively, but expression data are more complex and can reveal which genes are the most active in, say, liver cells, brain cells, or in diseased versus healthy tissue. And because studies often look at many

As of 13 July, the GEO repository at the European Bioinformatics Institute (EBI) in Hinxton, UK. Some time in the next few weeks, the number of deposited sets will top one million (see ‘Data dump’). The result is an unprecedented resource that promises to drive down costs and speed up progress in understanding disease. Gene-sequence data are already shared extensively, but expression data are more complex and can reveal which genes are the most active in, say, liver cells, brain cells, or in diseased versus healthy tissue. And because studies often look at many...
Cancer researchers share data

The Cancer Genome Atlas
- 14 thousand cases
- 39 types of cancers
- 13 types of data: molecular, clinical, sequencing
Genetics researchers share data
227 million substances \times 1.3 million assays

More than a billion measurements within a grid of 300 trillion cells

71 million meet Lipinski 5

1.2 million active substances

Chemical biologists share data
Molecular biologists share data
Even immunologists and trialists can share data!

Download 380+ studies today
Drug repositioning, new patient subsets, digital comparative effectiveness, more!

immport.org

Sanchita Bhattacharya
Zicheng Hu
Elizabeth Thomson
and many more
ImmPort redistributes data from major NIAID-funded programs and more

Data from 380+ trials and studies already released, involving:

- Immune Tolerance Network (ITN)
- Accelerating Medicines Partnership (AMP) in Rheumatoid Arthritis and Lupus (AMP)
- Human Immunology Project Consortium (HIPC)
- Atopic Dermatitis Research Network (ADRN)
- Clinical Trials in Organ Transplantation (CTOT) and in Children (CTOT-C)
- Population Genetics Analysis Program
- Protective Immunity for Special Populations
- HLA Region Genomics in Immune-mediated Diseases
- Modeling Immunity for Biodefense
- Reagent Development for Innate Immune Receptors
- Adjuvant Development Program
- Innate Immune Receptors and Adjuvant Discovery Program
- Maintenance of Macaque Specific Pathogen-Free Breeding Colonies
- Non-human Primate Transplantation Tolerance Cooperative Study Group

- Immunity in Neonates and Infants
- Consortium for Food Allergy Research
- Development of Sample Sparing Assays for Monitoring Immune Responses (U24)
- Asthma and Allergic Diseases Cooperative Research Centers
- HLA and KIR Region Genomics in Immune-Mediated Diseases
- Systems Approach to Immunity and Inflammation
- Immunobiology of Xenotransplantation
- Cooperative Study Group for Autoimmune Disease Prevention
- Informatics Methodology and Secondary Analyses for Immunology Data in ImmPort
- Centers for Medical Countermeasures against Radiation Consortium
- Inner City Asthma Consortium

Collaborations with NIAMS, NCI, Bill and Melinda Gates Foundation, and March of Dimes
<table>
<thead>
<tr>
<th>Announcement Number</th>
<th>Related Announc.</th>
<th>Issuing Organization</th>
<th>Release Date</th>
<th>Opening Date (SF424 Only)</th>
<th>Expiration Date</th>
<th>Activity Code(s)</th>
<th>Title</th>
</tr>
</thead>
<tbody>
<tr>
<td>RFA-AI-18-042</td>
<td>See Related</td>
<td>NIAID</td>
<td>10/05/2018</td>
<td>01/02/2019</td>
<td>02/02/2019</td>
<td>U01</td>
<td>Fc-Dependent Mechanisms of Antibody-Mediated Killing (U01 Clinical Trial Not Allowed)</td>
</tr>
<tr>
<td>RFA-AI-18-026</td>
<td>See Related</td>
<td>NIAID</td>
<td>06/19/2018</td>
<td>11/13/2018</td>
<td>12/14/2018</td>
<td>R01</td>
<td>Modeling and Simulation to Optimize HIV Prevention Research (MS OPR) (R01 Clinical Trial not allowed)</td>
</tr>
<tr>
<td>RFA-AI-18-023</td>
<td>See Related</td>
<td>NIAID</td>
<td>06/14/2018</td>
<td>09/04/2018</td>
<td>10/05/2018</td>
<td>R01</td>
<td>Immune Mechanisms at the Maternal-Fetal Interface (R01 Clinical Trial Optional)</td>
</tr>
<tr>
<td>RFA-AI-18-010</td>
<td>See Related</td>
<td>NIAID</td>
<td>04/05/2018</td>
<td>06/02/2018</td>
<td>07/03/2018</td>
<td>U01</td>
<td>Impact of Initial Influenza Exposure on Immunity in Infants (U01 Clinical Trial Not Allowed)</td>
</tr>
<tr>
<td>PAR-18-712</td>
<td>See Related</td>
<td>NIAID</td>
<td>03/15/2018</td>
<td>05/05/2018</td>
<td>05/08/2021</td>
<td>R01</td>
<td>Investigations on Primary Immunodeficiency Diseases/Inborn Errors of Immunity (R01 Clinical Trial Not Allowed)</td>
</tr>
<tr>
<td>RFA-AI-18-002</td>
<td>See Related</td>
<td>NIAID</td>
<td>03/08/2018</td>
<td>05/22/2018</td>
<td>06/23/2018</td>
<td>U19</td>
<td>Autoimmunity Centers of Excellence, Basic Research Program (U19 Clinical Trial Not Allowed)</td>
</tr>
<tr>
<td>RFA-AI-18-003</td>
<td>See Related</td>
<td>NIAID</td>
<td>03/08/2018</td>
<td>05/22/2018</td>
<td>06/23/2018</td>
<td>UM1</td>
<td>Autoimmunity Centers of Excellence, Clinical Research Program (UM1 Clinical Trial Required)</td>
</tr>
<tr>
<td>RFA-AI-17-040</td>
<td>See Related</td>
<td>NIAID</td>
<td>12/04/2017</td>
<td>02/28/2018</td>
<td>03/29/2018</td>
<td>U19</td>
<td>Cooperative Centers on Human Immunology (U19 Clinical Trial Optional)</td>
</tr>
<tr>
<td>RFA-AI-17-037</td>
<td>See Related</td>
<td>NIAID</td>
<td>10/25/2017</td>
<td>01/22/2018</td>
<td>02/23/2018</td>
<td>R01</td>
<td>Immunity in the Elderly (R01 Clinical Trial Optional)</td>
</tr>
<tr>
<td>RFA-AI-17-034</td>
<td>See Related</td>
<td>NIAID</td>
<td>10/20/2017</td>
<td>01/21/2018</td>
<td>02/22/2018</td>
<td>U01</td>
<td>Maintaining Immunity After Immunization (U01 - Clinical Trial Optional)</td>
</tr>
<tr>
<td>RFA-AI-16-078</td>
<td>See Related</td>
<td>NIAID</td>
<td>11/28/2016</td>
<td>02/15/2017</td>
<td>03/16/2017</td>
<td>U01</td>
<td>Limited Competition: Clinical Trials in Organ Transplantation in Children (CTOT-C): Mechanistic Ancillary Studies (U01)</td>
</tr>
<tr>
<td>RFA-AI-16-027</td>
<td>See Related</td>
<td>NIAID</td>
<td>09/07/2016</td>
<td>02/03/2017</td>
<td>03/04/2017</td>
<td>U19</td>
<td>T Cell Reagent Resource for the Study of Allergic Diseases (U19)</td>
</tr>
</tbody>
</table>
**Part 1. Overview Information**

<table>
<thead>
<tr>
<th>Participating Organization(s)</th>
<th>National Institutes of Health (NIH)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Components of Participating Organizations</td>
<td>National Institute of Allergy and Infectious Diseases (NIAID)</td>
</tr>
<tr>
<td>Funding Opportunity Title</td>
<td>Informatics Methodology and Secondary Analyses for Immunology Data in ImmPort (UH2 Clinical Trial Not Allowed)</td>
</tr>
<tr>
<td>Activity Code</td>
<td>UH2, Exploratory/Developmental Cooperative Agreement Phase I</td>
</tr>
<tr>
<td>Announcement Type</td>
<td>Reissue of PAR-16-253</td>
</tr>
<tr>
<td>Related Notices</td>
<td>• April 08, 2019 - Notice of Change in Expiration Date of PAR-19-229. See Notice NOT-AI-19-055.</td>
</tr>
<tr>
<td>Funding Opportunity Announcement (FOA) Number</td>
<td>PAR-19-229</td>
</tr>
</tbody>
</table>
An easy-to-use data publication service

“hypotheses come and go, but data remain”

Santiago Ramón y Cajal, 1897

Why share your data?

https://datashare.ucsf.edu/stash
Why use our service?

- **Straightforward compliance.** Submit your data to satisfy publisher and funder requirements for preservation and availability with a minimum of effort. Optionally link data to the article and funding organization.

- **Any field. Any format.** Submit data in any file format from any field of research. Share all of the data from a project in one place.

- **Get discovered.** Each landing page and dataset is optimized for search engines. Include any relevant geo-location information to take advantage of our built-in geospatial search.

- **Credit for your work.** Get an informative landing page to facilitate re-use of your data and a citable DOI so you can get credit. Add your ORCID identifier and we will automatically send the information to your ORCID profile.
Google Cloud Public Datasets

Uncover new insights with high-demand public datasets.

View one-pager  Explore public datasets

Explore public datasets in Google Cloud

Google Cloud Public Datasets facilitate access to high-demand public datasets, making it easy for you to access and uncover new insights in the cloud. By analyzing these datasets hosted in BigQuery and Cloud Storage, you can seamlessly experience the full value of Google Cloud with ease.

Google Dataset Search Beta

Search for Datasets

Try boston education data or weather site:noaa.gov

Learn more about including your datasets in Dataset Search.
University of California Announces Collaboration with Janssen to Expand Data Science Research in Healthcare

New Fellowship Program Facilitated by Johnson & Johnson Innovation to Recruit Data Scientists for High-Impact, Data-Driven Healthcare Research

By Laura Kurtzman

The University of California, Berkeley (UC Berkeley), and the University of California, San Francisco (UCSF), today announced a collaboration with Janssen Research & Development, LLC, part of the Janssen Pharmaceutical Companies of Johnson & Johnson (Janssen), to launch a new data science fellowship program that will explore innovative data-driven approaches to improve human health and train the next generation of leaders in the healthcare data sciences. This program will be the first of its kind in the San Francisco Bay Area, which already serves as a worldwide hub for the tech and
10 reasons why to archive and share study data openly

- Reproducibility
- Transparency
- Support public policy
- Return data to the community
- Visibility into failed trials
- Speed results reporting
- Enable learning
- Enable new ventures
- **New science**
- Trust and Believability
Reproducibility

Transparency

Support public policy

Return data to the community
Visibility into failed trials

Non-publication of large randomized clinical trials: cross sectional analysis

Christopher W Jones attending physician1, Lara Handler school of medicine liaison librarian2, Karen E Crowell clinical information specialist2, Lukas G Keil research assistant3, Mark A Weaver assistant professor4, Timothy F Platts-Mills assistant professor5

Enable learning

Speed results reporting

Scientists voice fears over ethics of drug trials remaining unpublished

Almost a third of large clinical trials in the US still not published five years after being finished, scientists write in BMJ

Sarah Boseley
The Guardian, Tuesday 29 October 2013 19:30 EDT

Enable new ventures

Pathwork Diagnostics
An APOBEC cytidine deaminase mutagenesis pattern widespread in human cancers

Steven A. Roberts1, Michael S. Lawrence2, Leszek Klimczak3, Sara A. Grinnan, David Fargion, Petar Stojevska, Adam Klineman, Gregory V. Kryukov4, Scott L. Carter, Gordon Sanders, Shawn Harris1, Rachit B. Shah5, Michael A. Rinnell3, Gad Getz6,7,8 and Dmitry A. Grabov7

Recent studies indicate that APOBEC cytidine deaminases, which convert cytosine to uracil during DNA editing and retrotransposition or retrotransposon mobilization, may induce mutation clusters in humans. Here, we show that throughout cancer genomes, APOBEC-mediated mutagenesis is prevalent and correlates with APOBEC mRNA levels. Mutation clusters in whole-genome and exome data are indicative of a pattern. Applying those criteria to 914 exome and 14 cancer types, mostly from TCGA, showed a significant overlap in mutagenesis patterns in bladder, cervical, and lung cancers, matching 88% of all samples. Within breast cancer, the WGS closely matched tumors with BRCAness, suggesting that this type of mutation is linked with cancer development. The pattern also correlated to cancer-associated obesity, APOBEC-mediated mutagenesis, and genome instability.

Genome instability triggers the double cancer24. Radiation and chemotherapies as adjuncts in the therapy of cancerous diseases, such as leukemia, lymphomas, and carcinomas, can cause genome instability. In lymphomas, for example, the double cancer syndrome, characterized by secondary tumors, is common. This syndrome is likely to be caused by an increased risk of genotoxic damage, leading to tumor recurrence or metastasis. The risk of secondary tumors is increased in patients with lymphomas and carcinomas, and the risk is further increased in patients with double cancers.

The APOBEC mutation pattern was calculated by one-sided Fisher’s exact test comparing the ratio of the number of C-to-T or C-to-G substitutions and complementary G-to-A or A-to-C substitutions that occur in and out of the APOBEC target motif (TCW or WGA) to an analogous ratio for all cytosines or guanines that reside inside and outside of the TCW or WGA motif within a sample fraction of the genome (Benjamini-Hochberg-corrected q value < 0.05). The number of tumor samples in each category is presented in each pie chart in a. Samples with a q value > 0.05 are represented in black. These samples are excluded from the scatter graphs in a,b. Color scales indicate the magnitude of enrichment in a and the number of APOBEC signature mutations in b for samples with q < 0.05. Dashed lines indicate expected effects with random mutagenesis. Cancer types are abbreviated as in TCGA: cervical squamous cell carcinoma and endodermal adenocarcinoma (CESC), bladder urothelial carcinoma (BLCA), head and neck squamous cell carcinoma (HNSC), breast invasive carcinoma (BRCA), lung adenocarcinoma (LUAD), lung squamous cell carcinoma (LUSC), uterine corpus endometrioid carcinoma (UCEC), ovarian serous cystadenocarcinoma (OV), stomach adenocarcinoma (STAD), rectum adenocarcinoma (READ), colon adenocarcinoma (COAD), prostate adenocarcinoma (PRAD), kidney renal clear-cell carcinoma (KIRC) and acute myeloid leukemia (LAML).
The 10,000 Immunome Project: From the control groups of 242 manually curated experiments

Kelly Zalocusky
Sanchita Bhattacharya
@ImmPortDB
Share successful, failed, and so-so data

- Rituximab in ANCA-Associated Vasculitis (RAVE) trial of new approach to the induction of remission
- But even though rituximab was found to be non-inferior than cyclophosphamide, which drug is the right one to use?
Journal reputation is at a critical moment

- How are journals going to respond to professional skeptics?
- How does a journal respond to another journal’s editor criticizing approval?
- How will journals respond to health systems (like University of California) who will now look at drug efficacy in real world clinical data? Will the data match?
- How will journals respond to payers incentivized to challenge expensive drug approvals, who will now want to see the raw data? $Billions riding on these papers.
- How will journals address the family ready to sell their house to pay for a drug for their family member?
- How will journals counter when government officials label them “fake news”?

“Trust us, it works, we’ve looked at the data”?! Really?
Preeclampsia: large cause of maternal and fetal death

- **Incidence**
  - 5-8% of all pregnancies in the U.S. and worldwide
  - 4.1 million births in the U.S. in 2009
  - Up to 300K cases of preeclampsia annually in the U.S.

- **Mortality**
  - Responsible for 18% of all maternal deaths in the U.S.
  - Maternal death in 56 out of every 100,000 live births in US
  - Neonatal death in 71 out of every 100,000 live births in US

- **Cost**
  - $20 billion in direct costs in the U.S annually
  - Average hospital stay of 3.5 days

Linda Liu
Bruce Ling
Matt Cooper
<table>
<thead>
<tr>
<th>Accession</th>
<th>Title</th>
<th>Type</th>
<th>Organism</th>
<th>Assays</th>
<th>Released</th>
</tr>
</thead>
<tbody>
<tr>
<td>E-GEOD-32472</td>
<td>Oxygen induced complication of prematurity: from experimental data to prevention strategy</td>
<td>transcription profiling by array</td>
<td>Homo sapiens</td>
<td>299</td>
<td>01/11/2011</td>
</tr>
<tr>
<td>E-GEOD-27976</td>
<td>Calvarial osteoblast transcriptome analysis identifies genetic targets and extracellular matrix-mediated focal adhesion as potential biomarkers for single-suture craniosynostosis</td>
<td>transcription profiling by array</td>
<td>Homo sapiens</td>
<td>249</td>
<td>04/03/2012</td>
</tr>
<tr>
<td>E-GEOD-46510</td>
<td>New whole blood gene expression profile predictive of preterm birth</td>
<td>transcription profiling by array</td>
<td>Homo sapiens</td>
<td>154</td>
<td>15/05/2014</td>
</tr>
<tr>
<td>E-GEOD-37210</td>
<td>The application of nonsense-mediated mRNA decay inhibition to the identification of breast cancer susceptibility genes</td>
<td>transcription profiling by array</td>
<td>Homo sapiens</td>
<td>143</td>
<td>11/04/2012</td>
</tr>
<tr>
<td>E-TABM-682</td>
<td>Transcription profiling of human decidua basalis to identify pre-eclampsia susceptibility genes</td>
<td>transcription profiling by array</td>
<td>Homo sapiens</td>
<td>104</td>
<td>07/04/2009</td>
</tr>
<tr>
<td>E-GEOD-35574</td>
<td>Differentially expressed microRNAs revealed by molecular signatures of Preeclampsia and IUGR in human placenta</td>
<td>transcription profiling by array</td>
<td>Homo sapiens</td>
<td>94</td>
<td>07/02/2012</td>
</tr>
<tr>
<td>E-GEOD-41336</td>
<td>Cultured Cyto and Syncytio-trophoblast samples exposed to varying degrees of hypoxia (methylation)</td>
<td>methylation profiling by array</td>
<td>Homo sapiens</td>
<td>90</td>
<td>18/01/2013</td>
</tr>
<tr>
<td>E-GEOD-5999</td>
<td>Transcription profiling of human 27 non-methylation single nucleotide polymorphism</td>
<td>transcription profiling by array</td>
<td>Homo sapiens</td>
<td>72</td>
<td>07/11/2008</td>
</tr>
</tbody>
</table>
Need a diagnostic for preeclampsia

Public big data available

March of Dimes Center for Prematurity Research

Data analyzed, diagnostic designed

SPARK grant ($50k)

Life Science Angels, other seed investors ($2 million)

Acquired by Progenity (La Jolla)

@CarmentaBio

progenity.com

bit.ly/carm_prog

Carmenta Bioscience Secures Over $2 Million in Oversubscribed Seed Financing

Camille Samuels Accepts Seat on Carmenta Board of Directors

Press Release: Carmenta Bioscience, Inc. – Wed, April 11, 2018

Business Wire

Progenity Acquires Carmenta Bioscience for Proprietary Preeclampsia Technology; Appoints Matthew Cooper Chief Scientific Officer

PALO ALTO, Calif.--(BUSINESS WIRE)--April 11, 2018--
How Much Does Pharmaceutical Innovation Cost? A Look At 100 Companies

<table>
<thead>
<tr>
<th>Company</th>
<th>Ticker</th>
<th>Number of drugs approved</th>
<th>R&amp;D Spending Per Drug ($Mil)</th>
<th>Total R&amp;D Spending 1997-2011 ($Mil)</th>
</tr>
</thead>
<tbody>
<tr>
<td>AstraZeneca</td>
<td>AZN</td>
<td>5</td>
<td>11,790.93</td>
<td>58,955</td>
</tr>
<tr>
<td>GlaxoSmithKline</td>
<td>GSK</td>
<td>10</td>
<td>8,170.81</td>
<td>81,708</td>
</tr>
<tr>
<td>Sanofi</td>
<td>SNY</td>
<td>8</td>
<td>7,909.26</td>
<td>63,274</td>
</tr>
<tr>
<td>Roche Holding AG</td>
<td>RHHBY</td>
<td>11</td>
<td>7,803.77</td>
<td>85,841</td>
</tr>
<tr>
<td>Pfizer Inc.</td>
<td>PFE</td>
<td>14</td>
<td>7,727.03</td>
<td>108,178</td>
</tr>
<tr>
<td>Johnson &amp; Johnson</td>
<td>JNJ</td>
<td>15</td>
<td>5,885.65</td>
<td>88,285</td>
</tr>
<tr>
<td>Eli Lilly &amp; Co.</td>
<td>LLY</td>
<td>11</td>
<td>4,577.04</td>
<td>50,347</td>
</tr>
<tr>
<td>Abbott Laboratories</td>
<td>ABT</td>
<td>8</td>
<td>4,496.21</td>
<td>35,970</td>
</tr>
<tr>
<td>Merck &amp; Co Inc</td>
<td>MRK</td>
<td>16</td>
<td>4,209.99</td>
<td>67,360</td>
</tr>
<tr>
<td>Bristol-Myers Squibb Co.</td>
<td>BMY</td>
<td>11</td>
<td>4,152.26</td>
<td>45,584</td>
</tr>
<tr>
<td>Novartis AG</td>
<td>NVS</td>
<td>21</td>
<td>3,983.13</td>
<td>83,663</td>
</tr>
<tr>
<td>Amgen Inc.</td>
<td>AMGN</td>
<td>9</td>
<td>3,692.14</td>
<td>33,229</td>
</tr>
</tbody>
</table>

Sources: InnoThink Center For Research In Biomedical Innovation; The Fundamentals via FactSet Research Systems
LINCS aims to create a network-based understanding of biology by cataloging changes in gene expression and other cellular processes that occur when cells are exposed to a variety of perturbing agents, and by using computational tools to integrate this diverse information into a comprehensive view of normal and disease states that can be applied for the development of new biomarkers and therapeutics. By generating and making public data that indicates how cells respond to various genetic and environmental stressors, the LINCS project will help us gain a more detailed understanding of cell pathways and aid efforts to develop therapies that might restore perturbed pathways and networks to their normal states.

- **5,178 compounds**
  - 1,300 off-patent FDA-approved drugs
  - 700 bioactive tool compounds
  - 2,000+ screening hits (MLPCN and others)

- **3,712 genes** (shRNA + cDNA)
  - targets/pathways of FDA-approved drugs (n=900)
  - candidate disease genes (n=600)
  - community nominations (n=500+)

- **15 cell types**
  - Banked primary cell types
  - Cancer cell lines
  - Primary hTERT immortalized
  - Patient derived iPS cells
  - 5 community nominated
Bin Chen
Wei Wei
Li Ma
Bin Yang
Mei-Sze Chua
Samuel So

Gastroenterology, 2017
Need more drugs for more diseases  
Public big data available  
NIH funding  
Data analyzed, method designed  
Company launched, ARRA, StartX, Stanford license, first deal  
Claremont Creek, Lightspeed ($3.5 million)

Venture capital
'digital drug development' company NuMedii snags $3.5 million

NuMedii, Inc. has landed a deal that could validate its digital discovery platform for psoriasis R&D.

Astellas hooks up with NuMedii to continue drug repurposing deal drive

NuMedii, Inc. Announces New Partnership To Discover And Advance New Treatments For Idiopathic Pulmonary Fibrosis
Figure 2: Patient pedigree
The arrow shows the patient. Diagonal lines show relatives who are deceased. Years are age at death or diagnosis. AAA = abdominal aortic aneurysm. ARMD = age-related macular degeneration. ARVD/C = arrhythmogenic right-ventricular dysplasia or cardiomyopathy. CAD = coronary artery disease. CHF = congestive heart failure. HC = hypercholesterolaemia. HTN = hypertension. OA = osteoarthritis. SCD = sudden cardiac death (presumed). VT = paroxysmal ventricular tachycardia.

Credit: Euan Ashley, Russ Altman, Steve Quake, Lancet
Important genome differences “locked up” in publications

Credit: Rong Chen, Optra Systems, and Personalis, Inc.
Collect the “big data” of findings across publications to analyze the “big data” of the genome.

Credit: Rong Chen, Optra Systems, and Personalis, Inc.
Credit: Rong Chen, Optra Systems, and Personalis, Inc.
Maybe the genome can be used to suggest (promote?) preventative health strategies?

Credit: Rong Chen, Alex Morgan, Joel Dudley, Lancet
Need to use genomes to predict disease (2008)

Publications available for curation

Stanford donor funding

Company launched, Stanford license

MDV, Lightspeed, Abingworth ($20 million)

Same 3 plus Wellington Shields ($22 million)

Series C ($33 million)

IPO (2019, $141 million)

---

**Personalis Awarded Contract From VA Million Veteran Program – Whole Genome Sequencing and Data Analysis for Over 1,000 Individuals**

Press Release: Personalis, Inc. – Tue, Mar 12, 2013 12:02 AM EDT

MENLO PARK, Calif.--(BUSINESS WIRE)--

The US Department of Veterans Affairs (VA) has awarded its first contract for whole genome sequencing and data analysis to Personalis, Inc., of Menlo Park, CA. The contract is to secure computing facility and proprietary algorithms, Personalis will analyse samples from several VA sources, including from the Million Veteran Program, against an advanced human reference sequence, annotate both SNV/indel genetic analyses to help confirm sample/data chain of custody. Personalis will use laboratory genetic analysis, including both DNA sequencing and genotyping, Illumina, Inc., of San Diego, CA.

---

**Personalis Announces Closing of Initial Public Offering and Exercise in Full of Over Allotment Option**

June 24, 2019

MENLO PARK, Calif.--(BUSINESS WIRE)--Jun. 24, 2019-- Personalis, Inc. (Nasdaq: PSNL), a leader in advanced genomics for cancer, today announced the closing of its initial public offering of 9,109,725 shares of common stock at a public offering price of $17.00 per share, which includes the exercise in full by the underwriters of their option to purchase up to 1,188,725 additional shares of common stock. All of the shares were sold by Personalis.
Four Big Lessons Learned in Building Big Data Ecosystems

• Sufficient data already exists to impact medicine
  – Diagnostics and drugs from public big data
  – More data is better, but never a reason to wait for more
  – “Retroactive crowdsourcing”: robust findings come from integrated data sets
  – Real success use-cases get data storers, data users, and data contributors excited

• Public and open data is already extremely high quality
  – Should never wait for perfect data, experiment, conditions

• Sticks seem to work better than carrots for sharing
  – Continue exponential growth, ask grantees to share more

• Need more question askers, train students to initiate science with data
  – High school → higher education → career changers