Dr. Atul Butte is the Director of the new Institute for Computational Health Sciences (ICHS) at the University of California, San Francisco, and a Professor of Pediatrics. Dr. Butte trained in Computer Science at Brown University, worked as a software engineer at Apple and Microsoft, received his MD at Brown University, trained in Pediatrics and Pediatric Endocrinology at Children's Hospital Boston, then received his PhD from Harvard Medical School and MIT. Dr. Butte has authored nearly 200 publications, with research repeatedly featured in Wired Magazine, the New York Times and the Wall Street Journal. Dr. Butte is also the principal investigator of ImmPort, the archival and dissemination repository for clinical and molecular datasets funded by the National Institute of Allergy and Infectious Diseases. In 2013, Dr. Butte was recognized by the White House as an Open Science Champion of Change for promoting science through publicly available data. Other recent awards include the 2014 E. Mead Johnson Award for Research in Pediatrics, 2013 induction into the American Society for Clinical Investigation, the 2012 FierceBiotech IT “Top 10 Biotech Techies”, and the 2011 National Human Genome Research Institute Genomic Advance of the Month. Dr. Butte is also a founder of three investor-backed data-driven companies: Personalis, providing clinical interpretation of whole genome sequences; Carmenta, discovering diagnostics for pregnancy complications; and, NuMedii, finding new uses for drugs through open molecular data.
Getting Started With Open Big Data in Biomedical Research
Atul Butte MD, PhD

Conflicts of Interest

- Scientific founder and advisory board membership
  - Genenorth
  - Nucleotide
  - Personalis
  - Camentis
- Honoraria for talks
  - Lilly
  - Pfizer
  - Siemens
  - Bristol Myers Squibb
  - AstraZeneca
  - Roche
  - Genentech
  - Working Fusion
- Past or present consultancy
  - Lilly
  - Johnson and Johnson
  - Roche
  - Nubli
  - Genenorth
  - Terumo
  - Arno Labs
  - Prexidenta
  - Samsung
  - AstraZeneca
  - Regeneron
  - Veridana
  - Pathway Diagnostics
  - Geisinger Health
  - Coventor
  - Wilson Sanders
  - Gable & Askett
  - 20X Genomics
  - Medgenics
  - MD Healthcare
  - Genentech Group
  - Carmentis Management
  - Corporate Relationships
  - Apetana
  - Thomson Reuters
  - Ibm
  - SAP
  - SV Angel
- Speakers’ bureau
  - None
- Companies started by students
  - Carmentis
  - Serendipity
  - NuMedii
  - Stimulomics
  - Nanowave
  - Freadcat
  - MyTime
  - Flipora

October 4, 2015
Getting Started With Open Big Data in Biomedical Research
Atul Butte MD, PhD

Big Data in Biomedicine
Getting Started With Open Big Data in Biomedical Research
Atul Butte MD, PhD

Why data-driven science?

Eight reasons for excitement
1. There is an increasing call for translational medicine: Universities, Congress, NIH, and elsewhere: “What are we getting for our money?”
2. Many tools now exist that enable the large-scale parallel quantitative assessment of molecular state
   - Premier example is RNA expression detection microarray
   - High-bandwidth measurement tools; not just big, but nearly comprehensive
   - Low-cost
     - Quantitating every gene in the genome: ~$200 per sample
     - Measuring 2.5 million human polymorphisms: ~$100 per sample
     - Knocking out every gene in the worm: ~$450 per kit
     - Sequencing entire human genome: $2000

Sequencing Excitement
- Original genome: $3 bl, 13 yrs
- Helicos: $30k genome
- Pacific Biosystems: sequence human genome in 15 minutes
- Run times in minutes at a cost of hundreds of dollars
- 20 TB in 15 minutes
- Complete Genomics: 80 genomes/day
- Ion Torrent and Illumina: ~$1500 per genome

Why data-driven science?

3. Incredible amounts of publicly-available data
   DNA
   - GenBank: Hundreds of organisms have been completely sequenced, including man and mouse; 486,000+ species have had some DNA sequence deposited, 3.3 trillion bases total
   - NCBI dbGAP (genotype and phenotype): 604 genetic studies including more than a million individuals
   RNA
   - GEO has 1,600,000+ samples from 61000+ experiments
   - ArrayExpress has 1,820,000+ samples from 60000+ experiments
   Protein
   - EBI Pride: 54,000+ samples from 3,500 projects
Show me the data!

The potential and power of gene expression analysis using DNA microarrays has led to the widespread use of this technology. These expression or profiling studies (as they are commonly known) are providing a new and unprecedented view of molecular biology of organisms. The real power of medical genomics is in the ability to integrate and analyze data from thousands of studies.


Gene Expression Omnibus

ArrayExpress is a database of functional genomics data. ArrayExpress stores experiment metadata, microarray expression data (gene expression data from microarray experiments) and other functional genomics data. This data is accessible via various user oriented tools and export facilities.


Cancer researchers share data

<table>
<thead>
<tr>
<th>Available Cancer Types</th>
<th># Cases Shipped by BCR</th>
<th># Cases with Data</th>
<th>Date Last Updated (cent/d Penny)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Breast invasive carcinoma (BRCA)</td>
<td>1100</td>
<td>1098</td>
<td>10/24/14</td>
</tr>
<tr>
<td>Ovarian serous cystadenocarcinoma (OVCA)</td>
<td>586</td>
<td>586</td>
<td>10/21/14</td>
</tr>
<tr>
<td>Uterine Corpus Endometrial Carcinoma (UCEC)</td>
<td>540</td>
<td>540</td>
<td>10/27/14</td>
</tr>
<tr>
<td>Kidney renal clear cell carcinoma (KRCC)</td>
<td>536</td>
<td>536</td>
<td>10/27/14</td>
</tr>
<tr>
<td>Gastric adenocarcinoma (GAM)</td>
<td>529</td>
<td>529</td>
<td>10/27/14</td>
</tr>
<tr>
<td>Head and Neck squamous cell carcinoma (HNCC)</td>
<td>528</td>
<td>528</td>
<td>10/27/14</td>
</tr>
<tr>
<td>Lung adenocarcinoma [LUAD]</td>
<td>521</td>
<td>521</td>
<td>10/27/14</td>
</tr>
<tr>
<td>Brain Lower Grade Glioma [LGG]</td>
<td>515</td>
<td>515</td>
<td>10/27/14</td>
</tr>
<tr>
<td>Thyroid carcinoma [THCA]</td>
<td>503</td>
<td>503</td>
<td>10/27/14</td>
</tr>
</tbody>
</table>

Lung squamous cell carcinoma
Prostate adenocarcinoma
Skin Cutaneous Melanoma

170 million substances x 1.1 million assays
122 million meet Lipinski 5
1 million active substances

Chemical biologists share data

Genetics researchers share data

Getting Started With Open Big Data in Biomedical Research
Atul Butte MD, PhD

Molecular biologists share data

**Why data-driven science?**

4. New community of sharing: tools, data, publications, methods
- Journals require this; NIH is starting to require this
- Along with communities comes contention and agreement: “What is the name for this gene?”
- Increasing standardization in names, abbreviations, codes, and file formats
- Where standards have not been reached, there is at least the understanding that it must be reached
- Increasing recognition of reproducibility

**R / Bioconductor**
- Open source version of SPLUS, a well-established statistical system originally from Bell Labs
  - http://www.r-project.com
  - http://www.bioconductor.org

**ScienceInsider**
White House Unveils Long-Awaited Public Access Policy

- Data Underlying Published Research Results Will Be Accessible and Open Immediately. The foundation will require that data underlying the published research results be immediately accessible and open. This was subject to the transition period and a 12-month embargo may be applied.

October 4, 2015
Validation methods are increasingly commoditized.
Getting Started With Open Big Data in Biomedical Research
Atul Butte MD, PhD

**Translational Pipeline**

- Clinical and Molecular Measurements
- Translational Question or Trial
- Statistical/Computational methods
- Validating drug or biomarker

---

**Why data-driven science?**

5. Change in role of the bioinformatician from service provider to question asker
   - Each high-bandwidth measure yields sizeable amount of raw data
   - Distilling raw data and filtering out noise through the proper use of bioinformatics
   - Bioinformatics clearly plays a role in the storage, retrieval, and sharing of measurements, and relating to clinical outcomes
   - However, role for bioinformatics in genomic medicine is now beyond that of providing a service, and in enabling new and interesting questions to be asked in biomedical research
   - A researcher, whether clinical, experimental, or theoretical, can ask questions no one else can ask today, when powered by bioinformatics

---

**Why translational bioinformatics?**

7. It’s incredible how much bioinformatics you need to know just to read the New England Journal of Medicine!
   - "Shrunken centroid", "Unsupervised cluster analyses",
   - "gene-expression signature"
   - "global scaling, or normalization"
   - "q value for each gene represents the probability that it is falsely called significantly deregulated"
   - "class-prediction analyses", "10-fold cross validation"
   - "implemented in the PLINK tool set as a Cochran-Mantel-Haenszel stratified analysis"

8. Paucity of people trained to make use of these resources

---

**Why data-driven science?**

6. Increased funding for this line of work
   - NIH Roadmap started in May 2002; Dr. Zerhouni’s roadmap for medical research in the 21st century
   - "At no other time has the need for a robust, bidirectional information flow between basic and translational scientists been so necessary."
   - Three major themes
     - New Pathways to Discovery: addresses Bioinformatics and Computational Biology
     - Research Teams of the Future: cell biologists and computer programmers working to accelerate movement of scientific discoveries from the bench to the bedside
     - Re-engineering the Clinical Research Enterprise: transforming basic research discoveries into drugs, treatments, methods for prevention.
   - CTSA and Translational Science Award (CTSA): 60 funded at ~$10M
   - ITA mentions “informatics” 38 times
   - Recognition that the problem of translational medicine will not go away without the help of informatics

---

**9 reasons to 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
Getting Started With Open Big Data in Biomedical Research
Atul Butte MD, PhD

Reproducibility
- Repeatability of published microarray gene expression analyses

Return data to the community
- The Scientist
- Do-It-Yourself Medicine

Transparency
- EVOLUTION OF TRANSLATIONAL
OMICS

Support public policy
- BMJ

Visibility into failed trials
- BMJ

Enable learning
- New Science

Enable new ventures
- BMJ

Speed results reporting
- NextBio

Find precision subsets
- ImmPort

Digital comparative effectiveness
- Pathwork Diagnostics

New Science
- An IFN-ε‐citrulline dermatitis model: an adaptive model for trials
- Comparative effectiveness

Digital comparative effectiveness
- Northrop Grumman

Find precision subsets
- IMMPORT

If entry criteria are same, outcome measures are same, and comparable studies, can perform meta-trial

October 4, 2015
Getting Started With Open Big Data in Biomedical Research
Atul Butte MD, PhD

ImmPort redistributes data from major NIAID-funded programs
Data from over 300 trials and studies already released, involving:
- Immune Tolerance Network (ITN)
- 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
- Human Immunology Project Consortium
- 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

New collaboration with The Bill and Melinda Gates Foundation
De-identified raw clinical study data is released to the public along with genetic, gene expression, and flow cytometry measurements, in open formats
Hundreds of user downloads per month

Reanalyzing RAVE
- Rituximab in ANCA-Associated Vasculitis (RAVE) trial of new approach to the induction of remission
  - randomized
  - double-blind
  - double-dummy
  - active-controlled
  - non-inferiority

Study Design
- Cyclophosphamide
- Rituximab
- Placebo
- Steroids
- Steroid taper
- Maximum 3
- Up to 3

Original Article
Rituximab versus Cyclophosphamide for ANCA-Associated Vasculitis

Conclusion:
A standard change was not inferior to daily cyclophosphamide in terms of induction of remission in severe ANCA-associated vasculitis and may be superior to cyclophosphamide.
Reproducible science: Reproduce CD19+ B-cell depletion

- 63 of the 99 patients in the rituximab group (64%) reached the primary end point, as compared with 52 of 98 in the control group (53%).
- The treatment difference of 11% points between the groups met the criterion for noninferiority (P<0.001).

In retrospect, do any measured factors predict response?
Getting Started With Open Big Data in Biomedical Research
Atul Butte MD, PhD

Mazen Nasrallah

Any one experiment does not yield clear disease-causal factors

- One example of a microarray experiment with diabetes and control samples
- 187 genes differentially expressed

Keiichi Kodama

Questioning standardization in science

Some scientists suggest that environmental standardization may lead to spurious findings. The implication from this viewpoint will have to be controversial.

Intersect 130 T2D microarray experiments

Most of the 25000 genes in the genome are positive in few T2D microarray experiments

Keiichi Kodama
Relative frequency of positive RNA microarray experiments (out of 130)

Intersect 130 T2D microarray experiments

The 186 best known drug targets or genes with DNA variants (from GWAS) are positive in more experiments

Close collaboration with Dr. Takashi Kadowaki, Momoko Horikoshi, Kazuo Hara, University of Tokyo

Relative frequency of positive RNA microarray experiments (out of 130)

Gene A changes the most in adipose tissue and islet cell experiments

Gene A is higher in high fat diet
Gene A is expressed in mouse fat infiltrate

Gene A knockout has reduced infiltrate in fat

- Mac-2 stain

Gene A knockout has increased insulin sensitivity

- No change in weight gain

Kodama K, Horikoshi M, ... Maeda S, Kadowaki T, Butte AJ. PNAS, 2012.
Getting Started With Open Big Data in Biomedical Research
Atul Butte MD, PhD

Keiichi Kodama

Paraffin-embedded omental adipose tissue from an obese 57-year-old woman, BMI 36.9 kg/m²

Kodama K, Horikoshi M, ... Maeda S, Kadowaki T, Butte AJ. PNAS, 2012.

Keiichi Kodama

Serum soluble Gene A protein correlates with human HbA1c and insulin resistance

n = 55 non-diabetics
60.3 years of age ± 15, 36 males, 19 females
BMI 23.2 ± 4.3 kg/m²

Kodama K, Horikoshi M, ... Maeda S, Kadowaki T, Butte AJ. PNAS, 2012.

Keiichi Kodama

Keiichi Kodama

Therapeutic antibody against Gene A → reduces fat inflammatory infiltrate in mouse

C57BL6/J fed high-fat diet for 18 weeks
Intraperitoneal injection of rat anti-mouse anti-A antibody (n=8) or isotype control (n=8)
100 μg at day 0 and 50 μg at day 1-7

Kodama K, Horikoshi M, ... Maeda S, Kadowaki T, Butte AJ. PNAS, 2012.

Keiichi Kodama

Keiichi Kodama

Therapeutic antibody against Gene A → reduces glucose

Kodama K, Horikoshi M, ... Maeda S, Kadowaki T, Butte AJ. PNAS, 2012.

Keiichi Kodama

Gene A is CD44 (Hyaluronic Acid Receptor)
Anti-CD44 in development for multiple cancers
CD44 is a complicated receptor


Keiichi Kodama

Anti-CD44 for 4 weeks slows weight gain and reduces intake

Kodama K, ... Butte AJ. Diabetes, 2015 Mar;64(3):867-75.
Anti-CD44 for 4 weeks reduces adipose inflammation and hepatic steatosis

Keichi Kodama
Kyoko Toda
Shojiroh Morinaga
Satoru Yamada

Kodama K, ... Butte AJ. Diabetes, 2015 Mar;64(3):867-75.