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
  - 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
  - Coatue Management

- Corporate Relationships
  - Northrop Grumman
  - Genentech
  - Optum
  - Aptalis
  - Allergan
  - Astellas
  - Thomson Reuters
  - Intel
  - SAP
  - SV Angel
  - Progenity
  - Illumina

- 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-expression data sets publicly available, researchers can identify disease mechanisms and develop novel treatments.

BY MONYA BAKER

Purvesh 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 diabetes-prone animals than in control rodents. One, called T1G3, was a promising candidate.

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

bit.ly/genedata
The Cancer Genome Atlas

- 14 thousand cases
- 39 types of cancers
- 13 types of data: molecular, clinical, sequencing
Genetics researchers share data

<table>
<thead>
<tr>
<th>Study</th>
<th>Embargo Release</th>
<th>Details</th>
<th>Participants</th>
<th>Type of Study</th>
<th>Project</th>
</tr>
</thead>
<tbody>
<tr>
<td>GDFR: Genome Wide Association Study in Familial Parkinson Disease (PD)</td>
<td>Feb 13, 2009</td>
<td>VDA</td>
<td>1991</td>
<td>Case-control</td>
<td>GDFR</td>
</tr>
<tr>
<td>Framingham SHAREs</td>
<td>Version 1: Oct 19, 2008</td>
<td>VDA</td>
<td>14277</td>
<td>Longitudinal</td>
<td>SHAREs</td>
</tr>
<tr>
<td>GAIN: Collaborative Association Study of Psoriasis</td>
<td>Aug 13, 2008</td>
<td>VDA</td>
<td>2875</td>
<td>Case-control</td>
<td>GAIN</td>
</tr>
<tr>
<td>GAIN: Genotyping the 270 HapMap samples for GAIN by Broad</td>
<td></td>
<td>VDA</td>
<td>-</td>
<td>Parent-offspring trios</td>
<td>GAIN</td>
</tr>
<tr>
<td>GAIN: Genotyping the 270 HapMap samples for GAIN by Perlegen</td>
<td></td>
<td>VDA</td>
<td>-</td>
<td>Parent-offspring trios</td>
<td>GAIN</td>
</tr>
<tr>
<td>GAIN: International Multi-Center ADHD Genetics Project</td>
<td>Mar 26, 2008</td>
<td>VDA</td>
<td>2835</td>
<td>Parent-offspring trios</td>
<td>GAIN</td>
</tr>
<tr>
<td>GAIN: Linking Genome-Wide Association Study of Schizophrenia</td>
<td>Version 1: Nov 07, 2008</td>
<td>VDA</td>
<td>5066</td>
<td>Case-control</td>
<td>GAIN</td>
</tr>
<tr>
<td>GAIN: Major Depression: Stage 1 Genomewide Association in Population-Based Samples</td>
<td>Version 2: Dec 03, 2008</td>
<td>VDA</td>
<td>3741</td>
<td>Case-control</td>
<td>GAIN</td>
</tr>
<tr>
<td>GAIN: Search for Susceptibility Genes for Diabetic Nephropathy in Type 1 Diabetes</td>
<td>Version 1: Jun 11, 2007</td>
<td>VDA</td>
<td>1500</td>
<td>Case-control</td>
<td>GAIN</td>
</tr>
<tr>
<td>GAIN: Whole Genome Association Study of Bipolar Disorder</td>
<td>Version 1: Nov 25, 2008</td>
<td>VDA</td>
<td>3261</td>
<td>Case-control</td>
<td>GAIN</td>
</tr>
<tr>
<td>GAW16 Framingham and Simulated Data</td>
<td>Oct 19, 2008</td>
<td>VDA</td>
<td>7130</td>
<td>Longitudinal, population-based</td>
<td>SHAREs</td>
</tr>
<tr>
<td>Genome-wide Association Studies in the Hutterites</td>
<td></td>
<td>VDA</td>
<td>632</td>
<td>Population-based</td>
<td>University of Chicago</td>
</tr>
<tr>
<td>Genome-wide Association Study of Neuroblastoma</td>
<td></td>
<td>VDA</td>
<td>1002</td>
<td>Case-control</td>
<td>COG</td>
</tr>
<tr>
<td>Genome-wide Study in Amyotrophic Lateral Sclerosis and Controls: First Stage Analysis</td>
<td>Jun 26, 2008</td>
<td>VDA</td>
<td>544</td>
<td>Case-control</td>
<td>NINDS</td>
</tr>
<tr>
<td>Ischemic Stroke Genetics Study (ISGS)</td>
<td>Jun 26, 2008</td>
<td>VDA</td>
<td>485</td>
<td>Case-control</td>
<td>NINDS</td>
</tr>
<tr>
<td>Mayo-Peruigen LEAPS (Linked Efforts to Accelerate Parkinson’s Solutions) Collaboration</td>
<td>Mar 03, 2008</td>
<td>VDA</td>
<td>1550</td>
<td>Case-control</td>
<td>UJFF</td>
</tr>
<tr>
<td>NEI Age-Related Eye Disease Study (AREDS)</td>
<td>Jun 11, 2007</td>
<td>VDA</td>
<td>600</td>
<td>Case-control</td>
<td>NEI</td>
</tr>
<tr>
<td>NINDS Parkinson’s Disease</td>
<td>Oct 12, 2007</td>
<td>VDA</td>
<td>535</td>
<td>Case-control</td>
<td>NINDS</td>
</tr>
<tr>
<td>NINDS Parkinsonism Study</td>
<td>Oct 12, 2007</td>
<td>VDA</td>
<td>1263</td>
<td>Case-set</td>
<td>NINDS</td>
</tr>
<tr>
<td>NINDS The Amish Parkinson Disease Study</td>
<td>Jun 26, 2008</td>
<td>VDA</td>
<td>87</td>
<td>Case-set</td>
<td>NINDS</td>
</tr>
<tr>
<td>NINNI Inflammatory Bowel Disease Study</td>
<td></td>
<td>VDA</td>
<td>544</td>
<td>Case-control</td>
<td>NINDS</td>
</tr>
<tr>
<td>NINNI Inflammatory Bowel Disease Study</td>
<td></td>
<td>VDA</td>
<td>1032</td>
<td>Case-control</td>
<td>COG</td>
</tr>
<tr>
<td>SEARCH GWAS Study of Statin-Induced Myopathy</td>
<td></td>
<td>VDA</td>
<td>175</td>
<td>Case-control</td>
<td>University of Oxford</td>
</tr>
<tr>
<td>Study of Irish Amyotrophic Lateral Sclerosis (IALS)</td>
<td></td>
<td>VDA</td>
<td>432</td>
<td>Case-control</td>
<td>NINDS</td>
</tr>
<tr>
<td>The Finland United States Investigation of NIDDM Genetics (FUSION) study</td>
<td></td>
<td>VDA</td>
<td>2335</td>
<td>Case-control</td>
<td>University of Michigan</td>
</tr>
<tr>
<td>Australia-United States Association Study of Dietary Lungs, Ear, Earpathines</td>
<td></td>
<td>VDA</td>
<td>4661</td>
<td>Case-control</td>
<td>NINDS</td>
</tr>
<tr>
<td>PDF</td>
<td></td>
<td>VDA</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
170 million substances x 1.1 million assays

More than a billion measurements within a grid of 190 trillion cells

122 million meet Lipinski 5

1 million active substances

Chemical biologists share data
Molecular biologists share data
Yes, clinical trialists can share data!
Download 300+ studies today
Clinical trials, new patient subsets, digital comparative effectiveness, more

ImmPort Shared Data
Your site for searching and downloading shared data

ImmPort Shared Data enables searching and downloading data from studies funded by NIAID DAIR and DMID, other NIH age. Additional resources include step-by-step data reuse analysis code, the Cell Ontology Visualizer, the Cyto interaction literature mining tool.

Filter Option

SDY557 Noninvasive Monitoring in Kidney Transplantation (CTOT-01)
IMPROVING THE EFFICIENCY AND EFFECTIVENESS OF GENOMIC SCIENCE TRANSLATION

WORKSHOP SUMMARY

INTRODUCTION

Clinical trial data submitted to the agency* covers a marketing decision on the study products has been made (FDA, 2014). The ALLTrials campaign was launched, calling for “all past and present clinical trials to be registered and their full methods and summary results reported” (ALLTrials, 2013). As of December 2014, more than 85,000 people had signed the ALLTrials petition, and 592 organizations had joined ALLTrials (ALLTrials, 2014).

Astellas, Bayer, Boehringer Ingelheim, GlaxoSmithKline, Lilly, Novartis, Roche, Sanofi, Takeda, UCB, and V&W. Healthcare committees committed to sharing clinical trial data through ClinicalTrialsOnlineRequest.com and allowing an independent review panel to make decisions on data requests (ClinicalTrialsOnlineRequest.com, 2014).

The British Medical Journal issued a policy requiring data sharing for clinical trials it publishes (BMJ, 2015).

The Pharmaceutical Research and Manufacturers of America (PhRMA), the European Federation of Pharmaceutical Industries and Associations (EFPIA), and the Biotechnology Industry Organization (BIO) released principles documents signaling their support for sharing clinical trial data (BIO, 2014; PhRMA, 2015).

The National Institute of Allergy and Infectious Diseases (NIAID) made de-identified data from 31 data trials available through the Immunology Database and Analysis Portal (ImmPort) (ImmPort, 2013).

NHG issued a new policy on sharing of genomic data. The new policy outlines and emphasizes the expectation that investigators who obtain informed consent from study participants for potential future use of the participants’ de-identified data for both research and broad sharing, and commit to sharing data no later than the earliest publication of the study results (NHG, 2014).

Numerous approaches and models for sharing clinical trial data are being implemented with varying levels of access control. At one end of the spectrum, ImmPort in the United States and the FreeBIRD website (FreeBIRD, 2014) in the United Kingdom make available some data, de-identified data sets from publicly funded clinical trials (NIAID and CRASH trials, respectively) with minimal restrictive terms; the data sets can be downloaded from the Web upon registration and acceptance of their terms.

APPENDIX

Provide the data to an academic institution, whereas that institution becomes the entity that discovers the data.

The term “advisory” is often used in the disclosure control literature to refer to the role of the individual or entity that is trying to re-identify data subjects. Other terms used are “attacker” and “intruder.” Discussions about the OI having a potential adversary are not intended to paint OI as having malicious objectives. Rather, in the context of a risk assessment, one must consider a number of possible data recipients as being potential adversaries and manage the re-identification risk accordingly.

Data Sharing Models

A number of different ways to provide access to IDP have been proposed and used, such as different advantages and risks (Miller et al., 2013). First, there is the traditional public data release where anyone can get access to the data with no registration or conditions. Examples of such releases include the publicly available clinical trial data from the International Stroke Trial (IST) (Sanderson et al., 2011) and data posted to the Dryad online open access data repository (Dryad, unlisted, Haggard, 2013).

A second form of data sharing, which is more restrictive, occurs when there exists a formal request and approval process to obtain access to clinical trial data, such as the GlaxoSmithKline (GSK) trials repository (Harries, 2012). Mason and Rockhold, 2013; Project Data Sphere (whose focus is on oncology trial data) (Bamshad et al., 2012; Hedd, 2012); the Yale University Open Data Access (TODA) Project, which is initially making trial data from MedCrain available (COEN, 2014; Kromnoda and Kiss, 2013); and the Immunology Database and Analysis Portal (ImmPort. n.d.), which is restricted to researchers funded by the Division of Allergy, Immunology, and Transplantation of the National Institute of Allergy and Infectious Diseases (NIAID), other approved life science researchers, National Institutes of Health employees, and other nonresearcher government employees (ImmPort. n.d.). More recently, pharmaceutical companies have created the ClinicalTrialsDataRequest.com website, which facilitates data requests to multiple companies using one portal. Following this restrictive model, a request can be processed by the study sponsor or by a delegate of the sponsor (e.g., an academic institution).
10 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
• Trust and Believability
Reproducibility

Repeatability of published microarray gene expression analyses

John PA Ioannidis1,2, David BA Mario Falcini3, Cesare Furlanello Michael Nitzberg4, Grier P Page5

Given the complexity of microarrays, studies guidelines of data availability, some of which are general in nature, several public databases allow access to raw data. In this report, we discuss the principles and challenges of reproducibility in microarray gene expression analysis.

Transparency

NIH multisrules for validating key results

In translational science, at least one thing is apparent: reproducibility, a term often used in the context of research, is crucial. In a recent survey, researchers estimated that about one-third of research projects are not reproducible.

Support public policy

BMJ

Roche offers researchers access to all Tamiflu trials

Deborah Cohen

Roche has offered to provide researchers with access to all Tamiflu data, after years of controversy about the effectiveness of the drug in treating H1N1 flu.
Visibility into failed trials

**BMJ**

RESEARCH

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

Christopher W Jones attending physician, Lara Handler school of medicine liaison librarian, Karen E Crowell clinical information specialist, Lukas G Keil research assistant, Mark A Weaver assistant professor, Timothy F Platts-Mills assistant professor

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**

Welcome to the UK Clinical Trials Gateway

All the world’s genomic data
An APOBEC cytidine deaminase mutagenesis pattern widespread in human cancers

Steven A. Roberts, Michael S. Lawrence, Leszek J. Klimczak, Sara A. Grimm, David Fargot, Peter Stojan Adam Klimczak, Gregory V. Kryukov, Scott L. Carter, Gordon Saka, Shawn Harris, R. R. Shah, Michael A. Rosenblum, Gad Getz, and Dmitry A. Garantza

Recent studies indicate that a selecton of APOBEC cytidine deaminases, which convert cytosine to uracil during DNA editing and recombination or retrotransposon mobilization, may induce mutation clusters in human tumors. We show here that throughout cancer genomes APOBEC-mediated mutagenesis is pervasive and correlated with APOBEC mRNA expression levels. Mutations clustered in whole-genome and exome cancer data in the cognate context of APOBEC mutation pattern. Applying those criteria to WES exome from 13 cancer types, mostly using TCGA (ATCGA), showed a significant positive association with a large variety of tumor classes, and survival in cancer patients, suggesting that this type of mutation is common in cancer development. The pattern was also extended to cancer- and general mutation pattern.

Genome instability triggers the development of cancer. Recent studies indicate that APOBEC cytidine deaminases are involved in cancer development. Normal somatic mutations can also be induced in normal somatic cells, and these mutations can be detected in cancer cells. Cytosine deamination is a key step in this process, leading to the formation of uracil and thymine, which can then be recognized and repaired by the cell's DNA repair mechanisms. This process can lead to the formation of clusters of mutations, which can be associated with specific cancer types.

Figure 2: Presence of an APOBEC mutation pattern in exome data sets from different cancer types. (a) Fold enrichment and (b) mutation load (b) of the APOBEC mutation pattern were determined in each of 2,680 whole exome-sequenced tumors representing 14 cancer types. Samples were categorized by the statistical significance of the APOBEC mutation pattern and the frequency of occurrence. The significance of 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 C-to-G 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 q > 0.05 are represented in black. These samples are excluded from the scatter graph 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 mutation generation. Cancer types are abbreviated as in TCGA: cervical squamous cell carcinoma and endocervical adenocarcinoma (CESC), bladder urothelial carcinoma (BLCA), 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).

Hyperlinks to TCGA data sets and references published in this paper are provided in Supplementary Table S3.
The 10,000 Immunome Project: From the control groups of 242 manually curated experiments

Data available in the 10,000 Immunomes Project

Total Samples 42117
Total Distinct Subjects 10344

MEASUREMENT

SUBJECTS

Secreted Proteins 4835
ELISA 4035
Multiplex ELISA 1286

Virus Titer

Virus Neutralization Titer 3609
HAI Titer 2368

Clinical Lab Tests

Complete Blood Count 2639
Comprehensive Metabolic Panel 1684
Fasting Lipid Profile 664

Questionnaire 1422

Cytofluorometry

Flow Cytometry (PBMC) 1415
CyTOF (PBMC) 907
Flow Cytometry (Whole Blood) 583

HLA Type 1093

Gene Expression Array 476
Whole Blood 311
PBMC 165

Kelly Zalocusky
Sanchita Bhattacharya
@ImmPortDB

bit.ly/10kimpdf
http://10kimmunomes.org/
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
<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 Syncyto-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-annotated genes to test their expression</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)

Stock Watch

Carmenta Bioscience Secures Over $2 Million in Oversubscribed Seed Financing

Carmenta Bioscience, Inc. – Wed, April 11, 2018

Camille Samuels Accepts Seat on Carmenta Board of Directors

Business Wire

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

@CarmentaBio

progenity.com

bit.ly/carm_prog
### How Much Does Pharmaceutical Innovation Cost? A Look At 100 Companies

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

*Sources: InnoThink Center For Research In Biomedical Innovation; Tho Fundamentals via FactSet Research Systems*
control food  niclosamide  NEN

Before treatment

After treatment

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)

@NuMedii

NuMedii
Translating Big Data into new medicines

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

The oversubscribed round was led by Claremont Creek Ventures and Lightspeed Venture Partners and included Life Science Angels and others.

NuMedii's data-into-gold approach rolls a wide range of data — from public scientific data bases and other sources — into an algorithm to predict if a compound will translate into medicines that tackle the most deadly diseases.

@NuMedii

FierceBiotechIT

Allergan taps NuMedii's digital platform for psoriasis R&D
October 5, 2015 | By Nick Paul Taylor

Astellas hooks up with NuMedii to continue drug repurposing deal drive
January 15, 2016 | By Nick Paul Taylor

NuMedii, Inc. Announces New Partnership To Discover And Advance New Treatments For Idiopathic Pulmonary Fibrosis

Ron Leuty
Reporter-
San Francisco Business Times
Future speculation

• Open data will democratize biomedical innovation, and that’s a good thing
• Open data is going to lead to more believability
• Digitalization of biomedicine will bring new players into the field
• DNA and other molecular measurements will be routine
• Less privacy, and less concern for privacy
• The entire world will be in continual study → more importance on continually acquired accurate data