Bioinformatics for Discovery and Global Collaborations: Approaches and Lessons Learned

Yuri Quintana, Ph.D.
Director, Global Health Informatics, Beth Israel Deaconess Medical Center
Assistant Professor, Harvard Medical School
Why do we need Bioinformatics?
Rise of Chronic Diseases

- NCDs are the leading causes of death (63%) in all regions except Africa.
- NCDs projected to cost more than US$30 trillion (48% of global GDP in 2010)
- Respiratory diseases and allergies are a major global problem

Percentage of deaths from non-communicable diseases by country, 2013  (DOI: 10.4103/1658-600X.179820 Institute for Health Metrics and Evaluation)
Diversity of Patients

• People are moving from countries and regions (lost health records)
• Most clinical research has been done in USA and Europe (biased sample)
• Personalized treatment optimized to genetics requires global approach

Source: The global flow of people - Migration flows between and within regions for five-year periods [http://www.global-migration.info](http://www.global-migration.info)
Evolution of Putting Research into Practice

We need not only better data and evidence-based science but better learning networks and research communities.

Optimizing Healthcare

- Most clinical systems have incomplete data due to problems with data collection, interoperability, and data sharing
- Many centers have challenges integrating clinical and genomic data due to complexity and cost
- Many centers do not integrate family history or environmental history
- Research has largely been focused on populations in high income countries due to funding sources, biased genetic pool

- Future personalized treatments will need to be tailored to genetics, clinical, family history and environmental factors
What is Bioinformatics?
Biomedical Informatics

Biomedical informatics is an interdisciplinary field that develops analytical methods and software tools for understanding clinical and biological data. It combines elements from the fields of basic sciences, biology, computer science, mathematics, and engineering.

References: Biomedical Informatics: Computer Applications in Health Care and Biomedicine, Shortliffe and Cimino (editors), Springer, 2013
AMIA Board white paper: definition of biomedical informatics and specification of core competencies for graduate education in the discipline. Kulikowski 2012 DOI: 10.1136/amiajnl-2012-001053
Evolution of Field

Population Health
- Public health informatics

Consumer Health
- Consumer health informatics
- E-Health
- Telehealth

Clinical Sciences
- Clinical Informatics
- Health information management
- Imaging informatics
- Bioinformatics

Applied Sciences
- Research informatics

Basic Sciences

Applications

Population Health

- Population Health Databases and Analytics

Consumer Health

- Online Patient Education, Online Communities, Patient Generated Data
- Patient Education, Patient Portals, Mobile Apps, Wearable, VR, SGH
- Physician-Physician Consults, Patient-Physician, Patient-Patient

Clinical Sciences

- Medical Records, Decision Support, Alerts, Mobile Apps, Sensors
- Paper Chase
- Clinical Warehouse, Health Exchanges

Applied Sciences

- Advanced Image Analytics
- PubMed
- Precision Medicine

Basic Sciences

- Genomic Informatics

Timeline:

- 1970
- 1980
- 1990
- 2000
- 2010
- 2020
Contributions to the field by DCI

http://www.hmfpinformatics.org

Connecting healthcare providers to clinical data and literature

“LINC with Tomorrow” Dr. Warner Slack on national TV, 1967

PaperChase, the first search engine for biomedical literature, Dr. Howard Bleich, 1977, precursor to PubMed

The Online Medical Record (OMR), Dr. Charles Safran, 1989

Connecting patients and families to their data and care network

Baby CareLink telemicine application, Dr. Charles Safran, Dr. Jim Gray, et. al. 1996

PatientSite (2000), OpenNotes (2010), patient access to medical records and physician notes, Dr. Tom Delbanco, Jan Walker, Dr. Henry Feldman, et. al.

InfoSAGE Mobile care coordination and communication for elders and families, C Safran, Y, Quintana B Crotty, et. al.
Complexity of Data

Yuri Quintana, Ph.D.
Overview of Bioinformatics
De-Identification Levels

Human Genome Project Timeline

Image Source: NIH https://www.genome.gov/imagegallery/viewimage/?imageID=26964377742
Complexity of Data

- Genome
  - Genomics
- Transcriptome
  - Transcriptomics
- Proteome
  - Proteomics
- Metabolome
  - Metabolomics

DNA Sequencer
- DNA sequencers Manufacturers
  - Roche, Illumina, Life Technologies,
  - Beckman Coulter, Pacific Biosciences

DNA microarray
- Gene expression heat maps
- Affymetrix chips

Mass spectrometer
- Mass spectrum of a peptide
- Orbitrap mass spectrometer

Image Sources: Nature [https://www.nature.com/scitable/topicpage/translation-dna-to-mrna-to-protein-393](https://www.nature.com/scitable/topicpage/translation-dna-to-mrna-to-protein-393) and Wikimedia
The Cost of Sequencing a Human Genome

Networks for Collaboration and Learning with Bioinformatics Data
Global Learning Health System

1. **Global Data Commons**
   - Contribute outcomes and clinical datasets to global community
   - Analyze & develop new precision treatments
   - Discuss with peers and refine treatments
   - Apply treatments in local centers but with common data formats

Overview of Bioinformatics

Yuri Quintana, Ph.D.
Global Pediatric Cancer Network

Cure4Kids.org – Global Education and Collaboration on Pediatric Oncology

POND4Kids.org – Global Shared Clinical Protocols and Outcomes Database


Yuri Quintana, Ph.D.  
Overview of Bioinformatics
CaBIG

Cancer Biomedical Informatics Grid (CaBIG) was a USA government program from 2004-2011 to develop an open source, open access information network called caGrid for secure data exchange on cancer research.

NCI published review of CaBIG that included a long list of problems with the program, and recommended that most projects should be discontinued. Kenneth Buetow (April 1, 2008). "Heading for the BIG Time" (PDF). The Scientist. 22 (4), and

CancerLinQ™ by ASCO

CancerLinQ™ (2015-) aggregates Clinical Records from oncology practices, gives insights from de-identified data on hundreds of thousands of patients, potentially identifying important trends and increasing the confidence of care decisions, visualizes patients’ medical histories in powerful new ways.

As of Dec 2017, 100 oncology practices in 40 states, 12 source systems, 600K cancer Records, 2,500 oncologists

Image Source: [https://cancerlinq.org/](https://cancerlinq.org/)
Alicanto™ – Global Communities of Practice

An Integrated Platform for Education, Communication and Data Sharing

http://www.alicantocloud.com

© Yuri Quintana 2016
Alicanto™ Networks

OPENPediatics™ from Boston Children’s Hospital
https://www.openpediatrics.org

MADCAP Network from Dana-Farber Cancer Center
https://www.madcapnetwork.org
i2b2 and Transmart
eTRIKS is the result of a collaboration between 17 different partners. Each combining their strengths in the development of a platform and services for data staging, exploration and use in translational research. [https://www.etriks.org/](https://www.etriks.org/)
NIH Genomic Data Commons (USA)

Genomic and Clinical Data Sources

- The Cancer Genome Atlas (TCGA)
- TARGET (Therapeutically Applicable Research to Generate Effective Treatments)
- International Cancer Genome Consortium
- NCI clinical trials
- User-submitted studies

Genomic Data Commons (GDC)

1. Import and standardize genomic and clinical data from legacy and current NCI programs
2. Harmonize mapping of sequence data to the most current genome and transcriptome build
3. Implement state-of-the-art methods for derived data:
   - Mutation calls
   - Copy number
   - Structural variants
   - Digital gene expression
4. Maintain data security and manage authorized access
5. Provide data for browsing, download, or analysis on a colocalized computer cluster
6. Open GDC for upload of new cancer genomic data from researchers worldwide for comparison with existing data and sharing

GDC Use Cases

- Identify low-frequency cancer drivers
- Define genomic determinants of treatment response
- Compose clinical trial cohorts sharing targetable genetic lesions

Biobanking and BioMolecular Resource Research Infrastructure (Europe)

Informatics Tools

**CDSKB** | Catalogs and shares clinical decision support implementation artifacts and designs consideration for genomic medicine programs

**eleMAP** | Harmonizes local phenotype data dictionaries to existing metadata and terminology standards

**eRecordCounter** | Provides exploratory data figures for research planning purposes and feasibility assessment

**PheKB** | Offers a collaborative environment to build and validate electronic algorithms to identify characteristics of patients within health data

**PheWAS Catalog** | Functions as a platform for analysis of phenotypes against a single gene variant

**SPHINX** | Operates as a tool for exploring data for hypothesis generation, especially around drug response implications of genetic variation across the eMERGE PGx cohort

**Clinical Report File Formats** | Lists clinical reports and research materials (e.g. consent forms)

**Genotyping Tools** | PennCNV, Biofilter, Biobin, PLATO, ATHENA, Synthesis-View, PheWAS-View, and Phenogram

**Phenotyping Tools** | PheKB, eleMAP, and PheWAS
USA NIH Big Data to Knowledge (BD2K)

The Big Data to Knowledge (BD2K) program is a trans-NIH initiative that was launched in 2013 to support the research and development of innovative and transformative approaches and tools to maximize and accelerate the integration of big data and data science into biomedical research. The BD2K Program also supported initial efforts toward making data sets “FAIR” Findable, Accessible, Interoperable, and Reusable. In its first phase (FY2014-FY2017), BD2K invested $200 million in grant awards to address some major data science challenges and to stimulate data-driven discovery.

Genomics England, with the consent of participants and the support of the public, is creating a lasting legacy for patients, the NHS and the UK economy, through the sequencing of 100,000 genomes. Objectives: 1) To bring benefit to patients 2) To create an ethical and transparent programme based on consent 3) To enable new scientific discovery and medical insights 4) To kickstart the development of a UK genomics industry.

Image Source: https://www.genomicsengland.co.uk
eMERGE Network (USA)

eMERGE is a national network organized and funded by the National Human Genome Research Institute (NHGRI) that combines DNA biorepositories with electronic medical record (EMR) systems for large scale, high-throughput genetic research in support of implementing genomic medicine. 7 clinical sites 2 pediatric sites in 2012. https://emerge.mc.vanderbilt.edu/
All of US - Precision Medicine Initiative (USA)

The mission of the All of Us Research Program is to accelerate health research and medical breakthroughs, enabling individualized prevention, treatment, and care for all of us. Building a repository of data for 1,000,000+ people.

Source: https://allofus.nih.gov
USA Veteran Affairs Million Veteran Program

MVP is a national, voluntary research program funded entirely by the Department of Veterans Affairs Office of Research & Development. The goal of MVP is to partner with Veterans receiving their care in the VA Healthcare System to study how genes affect health. To do this, MVP will build one of the world’s largest medical databases by safely collecting blood samples and health information from one million Veteran volunteers. Data collected from MVP will be stored in a secure manner and will be coded for researchers so that researchers cannot directly identify participants. Researchers will study diseases like diabetes and cancer, and military-related illnesses, such as post-traumatic stress disorder.

Image Source: https://www.research.va.gov/MVP/
Established by BGI-Shenzhen, the China National GeneBank (CNGB) is the first national genebank integrating a large-scale bio-repository and an omics database. It is approved by Chinese government, with the mission of collecting, preserving and exploiting genomics resources, and to build a network fostering global communication and collaboration on biodiversity conservation and genetic resources utilization. In addition, CNGB is supported by BGI’s high-throughput sequencing and bio-informatics capacity, and it will not only provide a repository system for biological collection, but more importantly develop a novel platform to further understand genomic mechanisms of life.

Image Source: [http://www.genomics.cn](http://www.genomics.cn)
Biomedical Network Key Roles

- Chief Scientist
- Biomedical, Life Sciences, Clinical Scientists
- Research Informatics Director
- Clinical Informatics Director
- Biomedical Platforms Architect
- Biomedical Engineers and Programmers
- Clinical and Genomic Classification Experts
- Quality Assurance Engineers and Analysts
- Data Scientists
- Machine Learning/AI Researchers
- Visualization and User Interface Experts
- Research Ethics and Privacy Directors
- Communication (Internal & External) Directors
- Education and Training Directors
- Cybersecurity Experts
- Cloud Computing Architects and Engineers
- External Partnerships Director
Challenges in Implementation

• Lack of Focus (Feature rich, function poor)
• Lack of Adequate Trained Personnel
• Lack of Multidisciplinary Team
• Poor Governance
• Poor Leadership
• Lack of Adequate Funding (but not the biggest problem)
• Collecting and Managing Data Usage Consents
• Data Use Agreements between institutions
• Lack of diversity in patient data sets
• Agreement on Data formats, taxonomies and indexing needs to support scientific questions
• Usability of tools
• De-identification strategy
• Legal and Compliance Regulations
• Sustainable funding models
Clinical and Bioinformatics Training

- Graduate Programs in Bioinformatics, Clinical Informatics and Bio-Engineering
- American Medical Informatics Association and Journals www.amia.org
- IEEE Life Sciences Societies and Conferences lifesciences.ieee.org
- Clinical Informatics Subspecialty Board Examination - American Board of Preventive Medicine (ABPM)
- American Nurses Credentialing Center (ANCC) offers a board certification in Nursing Informatics
- CIIP (Certified Imaging Informatics Professional) certification by ABII (The American Board of Imaging Informatics)
- American Health Information Management Registered Health Information Administrator and Certified Coding Associate
Roadmap

• Select carefully for key leadership positions
• Assemble a multi-disciplinary team
• Involve patients as true partners
• Start with focused and feasible initial plan, but scalable
• Have clear transparency and ethics
References

Bioinformatics

- Biomedical Informatics: Computer Applications in Health Care and Biomedicine. (Health Informatics) 4th ed. 2014 Edition By Edward H. Shortliffe and James J. Cimino(Editors)
- AMIA Board white paper: definition of biomedical informatics and specification of core competencies for graduate education in the discipline. Kulikowski 2012 DOI: 10.1136/amiajnl-2012-001053
- Methods in Biomedical Informatics A Pragmatic Approach. Edited by Neil Sarkar
- Journal of Biomedical Informatics, Springer

Quintana cited work on Global Health Informatics


Training

- Best practices in bioinformatics training for life scientists
- Rosalind http://rosalind.info/
- Bioinformatics Education, Ontario Institute for Cancer Research https://bioinformatics.ca

Bioinformatics Data Networks

- NIH Genomic Data Commons https://gdc.cancer.gov/
- NIH All of US - Precision Medicine Initiative https://allofus.nih.gov
- US Veteran Affairs Million Veteran Program Image Source: https://www.research.va.gov/MVP/
- Transmart Foundation http://transmartfoundation.org and i2b2: Informatics for Integrating Biology & the Bedside https://www.i2b2.org
- Genomics England - 100,000 Genomes Project https://www.genomicsengland.co.uk
- BGI-Shenzhen, the China National GeneBank (CNGB) BGI-Shenzhen, the China National GeneBank (CNGB) http://www.genomics.cn
Related Medical Informatics Specialties

- **Bioinformatics** - Informatics applied to cellular and molecular biology, and genomics.
- **Clinical informatics** - Informatics applied in healthcare or individual health settings. It is related to, such as nursing informatics, dental informatics, pathology informatics.
- **Imaging informatics** - Informatics applied to imaging, systems in health care settings (storage, retrieval, analysis).
- **Consumer health informatics** - Informatics applies to consumer and patient access.
- **Research informatics** - Informatics to support biomedical and health research, with a focus to support translation of knowledge bench to bedside.
- **Public health informatics** - Informatics applied to public health such as health surveillance, reporting, and health promotion.
- **Health information management** - A related field applied to the management of medical records and access.
- **E-Health** – Terms from the 1990s on the application of electronic methods to health care
- **Telehealth** – Study of the distribution of health-related services and information via electronic
- **Clinical Computing** – Early term used by pioneers such as Dr. Warner Slack and Dr. Howard Bleich in 1970s
Contact

Yuri Quintana, Ph.D.
Director, Global Health Informatics
Beth Israel Deaconess Medical Center
Harvard Medical School
Email: yquintan@bidmc.harvard.edu
Web: www.yuriquintana.com
Twitter: www.twitter.com/yuriquintana