Big Data in Biomedicine

Katerina Kechris, PhD
Professor
Biostatistics and Informatics
Computational Bioscience Program

10/2/2019
BIOS 6606

Outline

- Big Data Types
  - –Molecular, Imaging, Clinical, Digital
- Common Themes

Big Data in Biomedicine

1. Molecular
   – sequencing, mass spec
2. Imaging
   – MRI, CT scans
3. Clinical
   – electronic health records
4. Public health
   – digital, apps, social media




Public Health Informatics

mHealth

- Use of mobile/wireless devices to improve health outcomes, health care services, and health research (NIH)
- vs eHealth/telehealth (general)
  - health services & information delivered/enhanced through the Internet/modern technologies

Global & Rural Health

- World Health Organization (WHO)
- mHealth important for underserved areas with widespread mobile phone usage
- Programs to address diseases linked with poverty (e.g., HIV/AIDS, malaria, TB)
- Disseminating information, remote diagnosis, improving patient compliance

Examples of mHealth

WHO Survey

- health call centers
- emergency toll-free telephone services
- managing emergencies and disasters
- mobile telemedicine
- appointment reminders
- community mobilization and health promotion
- treatment compliance
- mobile patient records
- information access
- patient monitoring
- health surveys and data collection
- surveillance
- health awareness raising
- decision support systems.

http://mhealthimpact.ucdenver.edu

mHealthImpact
CREATING AND CURATING DIGITAL HEALTH SOLUTIONS

ABOUT US | OUR WORK | COLLABORATIONS | NEWS | CONTACT | GALLERY

Welcome to the mHealth Impact Lab

mHealth Impact Lab Project Examples

The mHealth Impact Lab has worked with various researchers, clinicians, and enterprises to validate and test digital health solutions within communities – both locally and globally. We assist in various stages of a project's life cycle, including community engagement, design, and evaluation."

We R Native
Evaluating We R Native, a technology-based resource to support mental health among American Indian and Alaska Native Youth

Nudge
Taking Text Messaging to scale to support cardiovascular medication adherence

CR Nudge
Optimizing a mobile application to support cardiovascular rehabilitation

Native WYSE
Women-Young, Strong, and Empowered making CHOICES

Wearable Devices

Chronic Obstructive Pulmonary Disease (COPD)

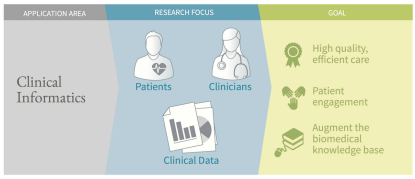
Sensor data
- Activity
- Inhaler Use

ActiGraph

Clinical Informatics

Clinical Informatics

- "the interdisciplinary study of the design, development, adoption and application of IT-based innovations in healthcare services delivery, management and planning." (R. Proctor)



https://www.dbmi.columbia.edu/research/research-areas/clinical-informatics/

Clinical Informatics

- Use of data and information technology
  - deliver health care services
  - improve patients' ability to monitor/maintain health
- Health professionals (medicine, nursing, pharmacy, etc)
- Collect, store, analyze health care data
- Clinical decision support
- Methods and policies for privacy and security

https://www.dbmi.columbia.edu/research/research-areas/clinical-informatics/

Electronic Health Records

J Oncol Pract. 2009 Sep; 5(5): 262–263. PMCID: PMC2790662
doi: 10.1200/JOP.091034

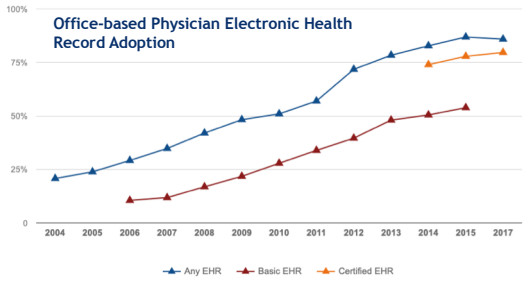
Policy Update: Federal Incentives for the Adoption of Electronic Health Records

Chantal Worzala, PhD

Author information ▶ Article notes ▶ Copyright and License information ▶

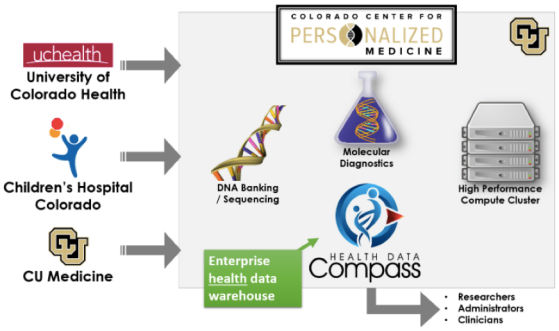
stimulus bill passed by Congress in 2009 to promote conversion of paper records to electronic data

The Office of the National Coordinator for Health Information Technology | Health IT Dashboard

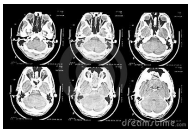


https://dashboard.healthit.gov/

www.ucdenver.edu/about/departments/healthdatacompass



Imaging Informatics



Medical Imaging

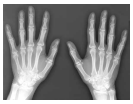
- Different modalities of digital technology to capture medical images
- Used for
  - Diagnosis
  - Monitoring
  - Treatment
  - Evaluating physiology, metabolism and molecular function

Imaging Informatics

- How medical images are used and exchanged throughout complex healthcare systems.
  - Picture archiving and communication system (PACS)
- Image processing & compression
- Data analysis & mining
  - "A survey on deep learning in medical image analysis" (Litjens et al., 2017, *Medical Image Analysis*)
    https://www.sciencedirect.com/science/article/pii/S1361841517301135

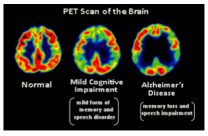
Modalities

- X-ray – send beams through the body, absorbed in different amounts depending on the density



broken bones, lungs, blood vessels, breast (mammography

- Ultrasound - high frequency sound waves; pregnancy, abnormalities heart & blood vessels, pelvis/abdomen organs
- Positron Emission Tomography (PET) - injected radiopharmaceuticals



neurological diseases, cancer, heart conditions & functions (O2 use, glucose metabolism, blood flow)

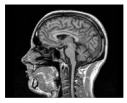
http://www.medicalimaging.org/about-mita/medical-imaging-primer/

Modalities

- Computed Tomography (CT/CAT scan) - 3-d views, multiple X-rays

tumors, injuries, embolism, aneurysms soft tissues, pelvis, blood vessels, lung, brain, heart, abdomen and bones

- Magnetic Resonance Imaging (MRI) - radio waves & magnetic field



blood vessels, breasts, bones/joints, organs in the pelvis, chest and abdomen, brain, spinal injuries, tendon/ligament tears

http://www.medicalimaging.org/about-mita/medical-imaging-primer/

Applications

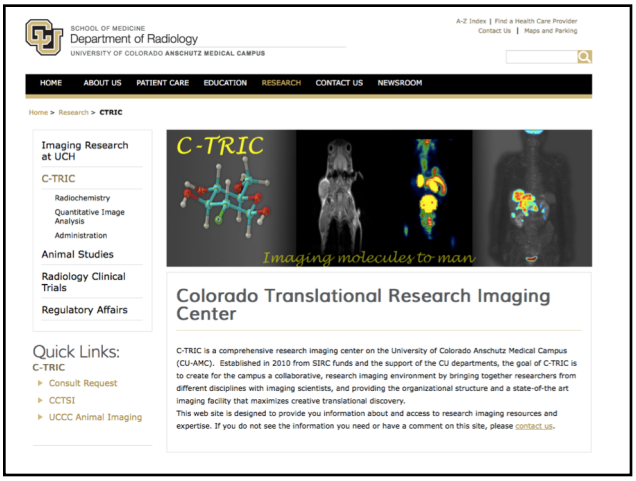
- Biologists study cells and generate 3D confocal microscopy data sets
- Virologists generate 3D reconstructions of viruses from micrographs
- Radiologists identify and quantify tumors from MRI and CT scans
- Neuroscientists detect regional metabolic brain activity from PET and functional MRI scans

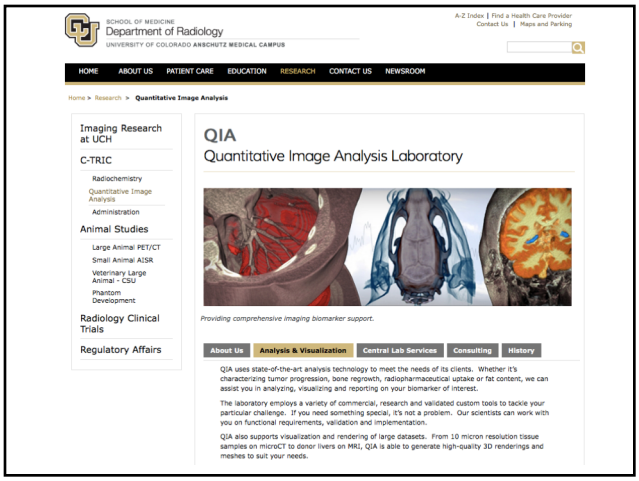
https://mipav.cit.nih.gov

Image Analysis Problems

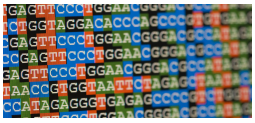
- Segmentation – delineating different organs
- Classification – determining classes - e.g. types of leukocytes or tumors
- Registration – comparing different modalities/patients
- Reconstruction – making 3D-measurments
- Measuring flow – e.g. inside aorta
- Reconstructing flow fields – e.g. inside the heart
- Integrating with other data – e.g. clinical, genetic

(Anders Heyden Lund University)



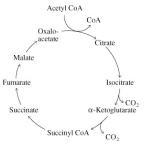


Bioinformatics



Molecular Big Data

- Used to measure & study
  - DNA, RNA, proteins
  - Modifications (e.g., methylation)
  - Structure (e.g., protein, RNA)
  - Molecular interactions, biological networks
  - Metabolic/enzymatic processes
  - Regulation
  - Cellular organization
  - Evolution
  - Biodiversity



High-throughput Technologies

- Arrays, sequencing, mass spectrometry
- Comprehensive profiling
  - Genome, transcriptome, proteome, metabolome, epigenome, etc.
- Integration – multiple profiles

- Study processes/mechanisms (e.g., DNA repair, transcription, pathways)
- Find associations with phenotype, behavior or disease

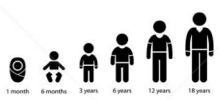
Example Projects

Animal model:

Alcohol Abuse & Dependence

Observational study:

Chronic Obstructive Pulmonary Disease (COPD)



Prospective Cohort:

Obesity & Diabetes

Alcohol Abuse & Dependence:

Animal Model

Integration of miRNA, mRNA, genetics & behavioral phenotypes

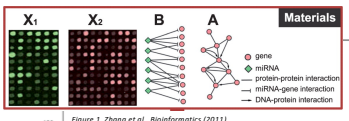


Figure 1, Zhang et al., Bioinformatics (2011)

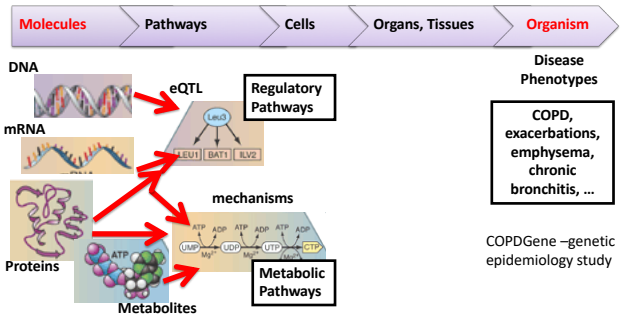
Ethanol response & intake

Study pathways in the brain associated with predisposition to ethanol response & intake

Recombinant Inbred Strains

R01 AA021131 (K. Kechris)

COPD and Pathways

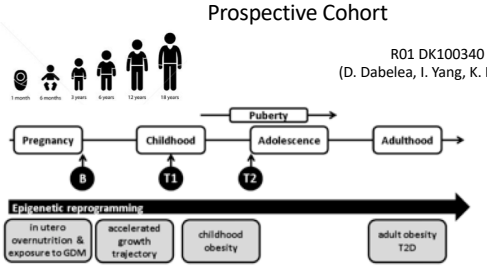


COPDGene –genetic epidemiology study

Obesity & Diabetes:

Prospective Cohort

R01 DK100340
(D. Dabelea, I. Yang, K. Kechris)

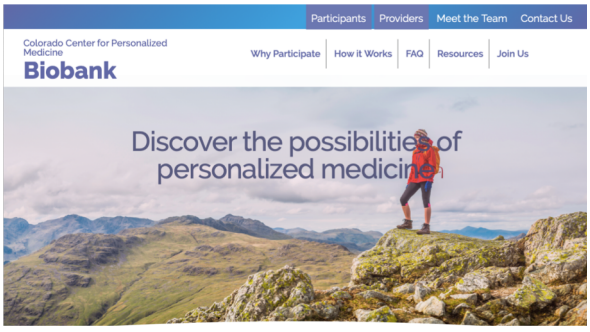


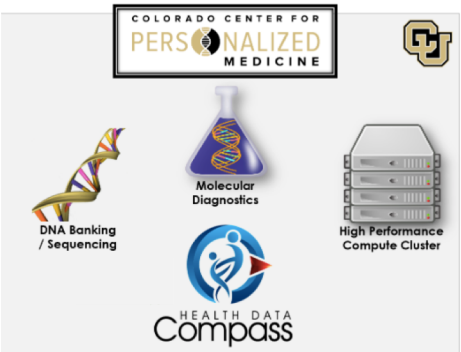
Gestational diabetes mellitus (GDM)

Integrate genetic, epigenetic, exposure and outcome data @ T1 & T2, to identify methylation changes associated with exposure to GDM & how exposure ⇔ methylation ⇔ fat mass.

Cores @ AMC

University of Colorado School of Medicine
Biological Mass Spectrometry Facility





Common Themes with Big Data

Big Data

- 3 V's: Volume, variety, and velocity of data
- Computational Infrastructure
  - Storage
  - Parallel computing
  - High performance computing
  - Cloud computing
  - Access & security

Common Big Data Tasks

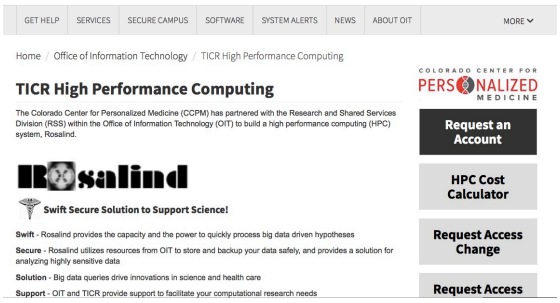
- storage
- retrieval
- sharing
- analysis
- security
- cleaning

Themes

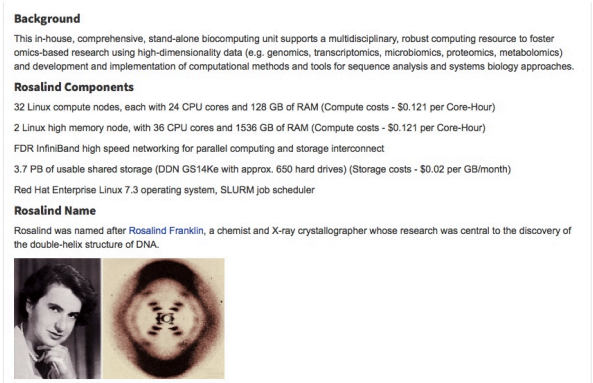
1. Computing Resources
2. Privacy/security
3. Data Processing
4. Multiple Testing
5. IID not always true
6. Data Mining

1. Computing Resources - Campus

Translational Informatics and Computational Research (TICR)



https://www1.ucdenver.edu/offices/office-of-information-technology/ticr-high-performance-computing



Background

This in-house, comprehensive, stand-alone biocomputing unit supports a multidisciplinary, robust computing resource to foster omics-based research using high-dimensionality data (e.g. genomics, transcriptomics, microbiomics, proteomics, metabolomics) and development and implementation of computational methods and tools for sequence analysis and systems biology approaches.

Rosalind Components

32 Linux compute nodes, each with 24 CPU cores and 128 GB of RAM (Compute costs - $0.121 per Core-Hour)

2 Linux high memory node, with 36 CPU cores and 1536 GB of RAM (Compute costs - $0.121 per Core-Hour)

FDR InfiniBand high speed networking for parallel computing and storage interconnect

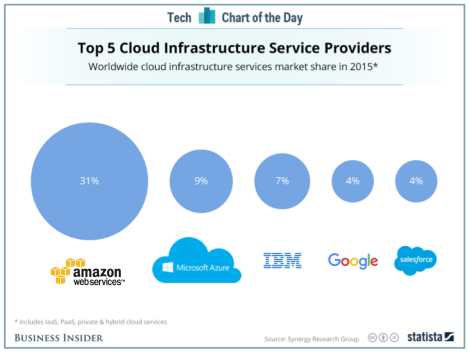
3.7 PB of usable shared storage (DDN GS14Ke with approx. 650 hard drives) (Storage costs - $0.02 per GB/month)

Red Hat Enterprise Linux 7.3 operating system, SLURM job scheduler

Rosalind Name

Rosalind was named after Rosalind Franklin, a chemist and X-ray crystallographer whose research was central to the discovery of the double-helix structure of DNA.

Computing Resources - Cloud



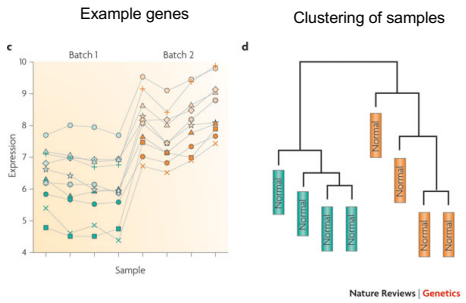
2. Privacy & Security

- Preserving patient privacy
- Authentication - securing access, protecting identities users, ensuring users claims
- Access control
- De-identification/masking
- Data encryption – preventing unauthorized access of sensitive data by translating data

3. Data Processing & QC

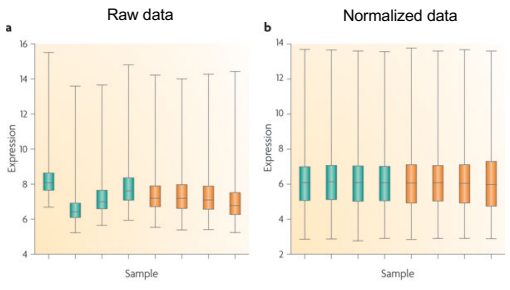
- Cleaning of data (e.g., entries in EHR)
- Evaluating QC metrics (e.g., signal to noise ratio in imaging)
- Assessing technical variability (e.g., high-throughput technologies)
  - Batch adjustment (technical effects)
  - Normalization (to compare different samples)
  - Filtering
    - features (genes, metabolites, etc) with low signal, outliers
    - bad samples (technology failed)

Adjusting for batches



Irizarry et al., (2007) Nature Methods 2:345-239

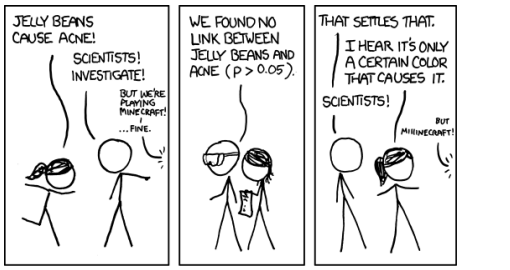
Normalizing the data



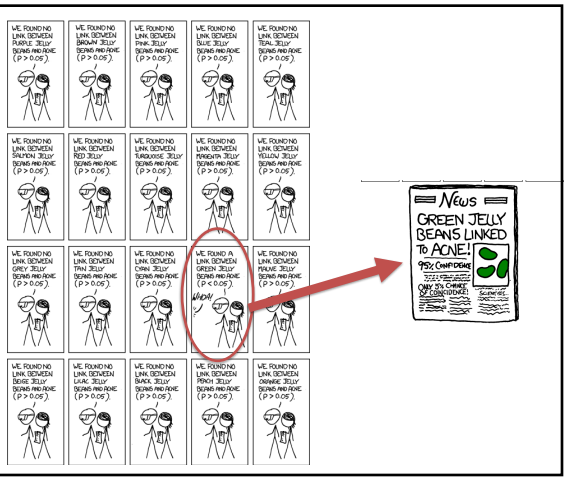
Each boxplot shows expression of all genes profiled

Irizarry et al., (2007) Nature Methods 2:345-239

4. Multiple Testing



https://xkcd.com/882/



Multiple Testing

- Testing 20,000 genes for differential expression.
- Suppose that the null hypothesis is true for each gene.
  - What is the null hypothesis for each gene?
- Apply p-value < 0.01
  - How many times do we expect to incorrectly reject the null hypothesis (i.e., observed p-value ≤ .01)?

FDR – Alternative to p-value

- False discovery rate (FDR) accounts for multiple comparisons
  - proportion of "significant" tests that are false positives
- Calculated for each test (e.g., gene)
  - Is dependent on the distribution of the other test results (e.g., other genes)
- For 5% FDR threshold – it is expected that 5% of significant genes are false positives.

5. Independent and Identically Distributed (IID) not always true

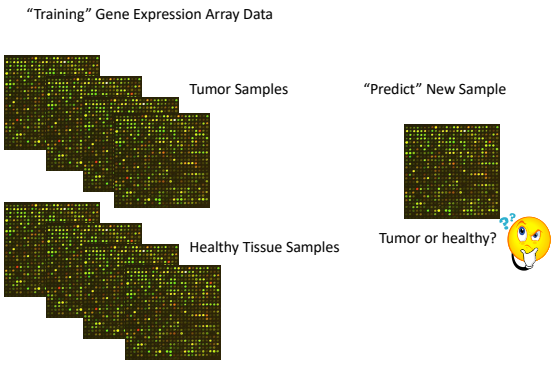
- Interactions
  - e.g., signaling pathway, protein-protein interactions
- Spatial dependencies
  - e.g., voxels in image, local genome structure
- Multiple records for same subject
  - e.g., multiple visits in EHR, time course

6. Data Mining

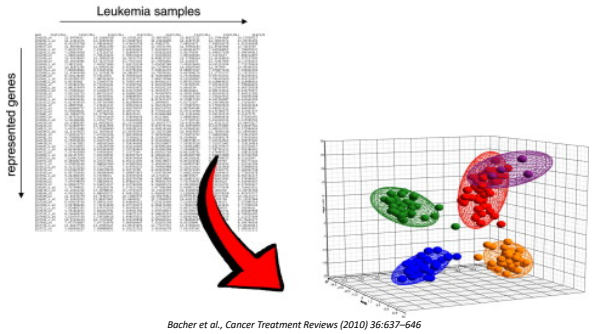
- Finding patterns in large high-dimensional data

1. Classification/Prediction – use data to discriminate between groups
2. Clustering/Subtyping – use data to find new groups

Classification



Clustering



Bacher et al., Cancer Treatment Reviews (2010) 36:637–646

Classes @ UCD – General

- More statistics!
  - BIOS 6611/12 Biostatistical Methods
  - BIOS 6660 Analysis of Genomic Data using R and Bioconductor
- Computing
  - BIOS 6640 Python and R in Data Science
- Machine Learning/Data Mining
  - Downtown CS department (Farnoush Banaei-Kashani)
  - Downtown Math department (e.g., "Reading Course: Deep Learning: A new application to genetics", Audrey Hendricks)

Classes @ UCD – Data Specific

Genomics/Bioinformatics

- MOLB 7900 - Practical Computational Biology for Biologists: Python
- MOLB 7910 - Practical Computational Biology for Biologist: R

Imaging

- ANAT 6205 Imaging and Modeling

Health Informatics

- CLSC6800 Intro to Health Information Technology
- CLSC6820 Management of Healthcare Information Technology
- CLSC6080 Database Management Systems
- NURS6603 Health Systems Management
- NURS6284 Digital Health Tools
- NURS6286 Foundations Informatics

mHealth

- CBHS 6628 Technology-Based Health Promotion
- CBHS 6670: Methods for Development and Evaluation of Technology Based Health Promotion Programs