Bioinformatics Workshop

Session #1

Course Introduction and Prerequisites

Register here

https://redcap.link/bfx

Chris Miller and Jason Walker

ICTS – Precision Health Function

Support transdisciplinary research
- Match clinicians and investigators
- Review grants
- Fund Precision Health Innovation awards

Develop common workflows
- Genomic consent and return of results
- Enable access to large genomic data sets
- Develop infrastructure to speed discovery translation

Educate the community
- Expand access to educational research programs
- Educate scientific and clinical community in precision health
- Engage broader STL community in genomic research

Leadership Team

Megan Cooper, MD, PhD

Chris Gurnett, MD, PhD

Contact Precision Health Program Scientist

Jenny McKenzie, PhD
j.mckenzie@wustl.edu

Precision Health Led Projects

- Pilot funding
  - Precision Health Innovation Awards

- Return of Results (ROR) for research participant
  - Genetic counseling, processes for returning results

- Genomic database access and submission
  - UK Biobank, All of Us Research Program, dbGaP, AnVIL, SRA

- Institutional Genomic Consent
  - One Protocol One Consent, Electronic Biobank

Providing Support For

- Core Services
  - WU Biological Therapy Core Facility (BTCF), McDonnell Genome Institute (MGI)

- Informatics Tools for Precision Health
  - Bioinformatics Workshop (BFX), pVAC, CIViC,

- Communications and Outreach
  - Women in Innovation and Technologies (WIT) program, EQUALIZE program through OTM

- Educational Opportunities
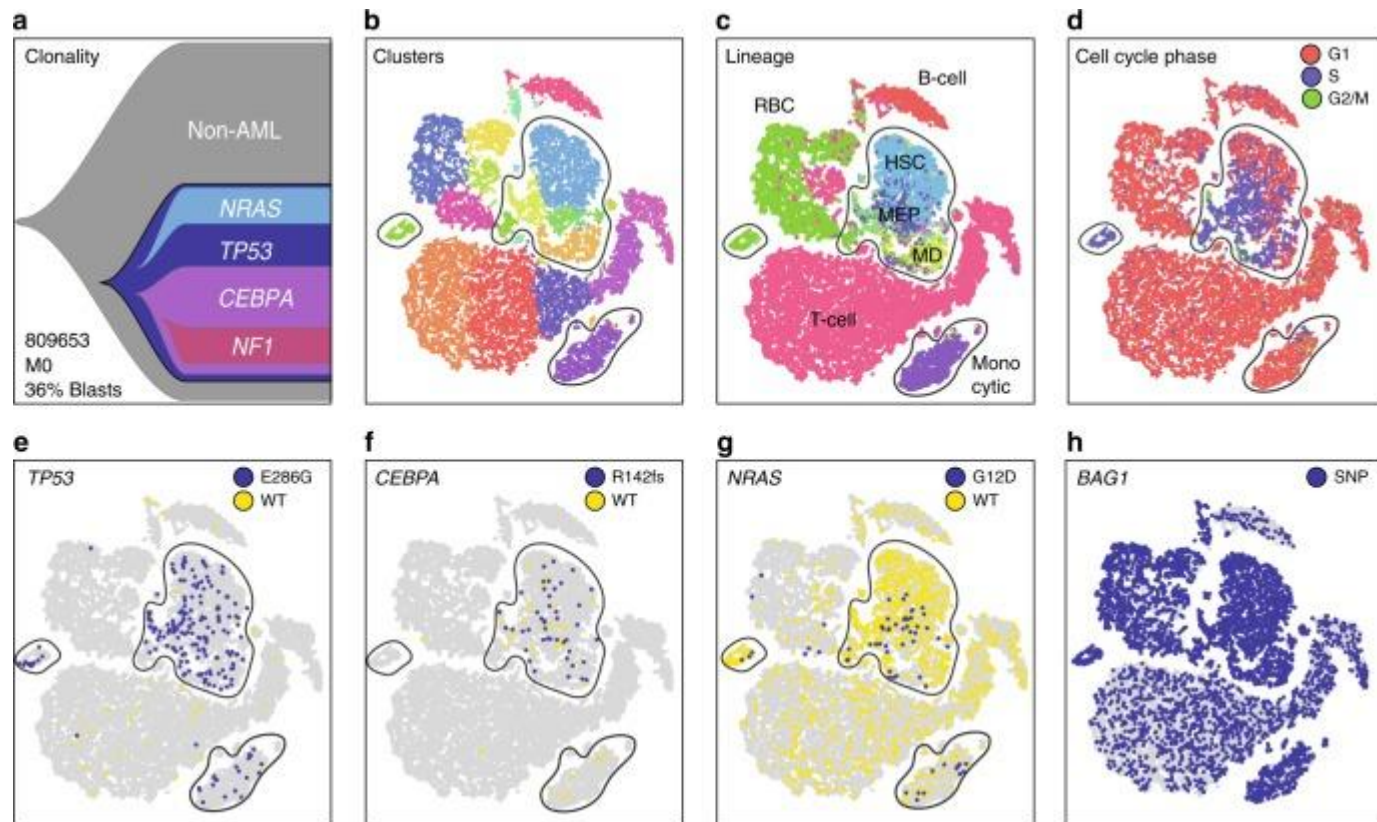 - Precision medicine pathway, Bioinformatics Workshop (BFX), Genomics in Medicine

Institute of Clinical and Translational Sciences

Providing cloud credits and expertise

Why learn bioinformatics?

- Biology is now a quantitative discipline - especially genomics

Cost per Genome

a Clonality

Non-AML

NRAS
TP53
CEBPA
NF1

809653
M0
36% Blasts

b Clusters

c Lineage

RBC

B-cell

HSC

MEP

MD

T-cell

Mono
cytic

d Cell cycle phase

G1
S
G2/M

e *TP53*

E286G
WT

f *CEBPA*

R142fs
WT

g *NRAS*

G12D
WT

h *BAG1*

SNP

Why learn bioinformatics?

- Biology is now a quantitative discipline - especially genomics

- Skills in programming, statistics, and visualization help you get the most out of your data

People who need complex data analysis

People who know how to do complex data analysis

Why learn bioinformatics?

- Biology is now a quantitative discipline - especially genomics

- Skills in programming, statistics, and visualization help you get the most out of your data

- This course aims to teach you the theory and practice of computational biology, with a focus on genomics but lessons that apply broadly

Goals:

- To empower you to improve and expedite your research

- To expose you to new ideas and techniques that may advance your research program

Who we are, and why you should trust us

Jason Walker, M.S.

Informatics group leader,
McDonnell Genome Institute

Chris Miller, Ph.D.

Assistant Professor
Division of Oncology

Over 30 years of combined experience in Bioinformatics and Computational Biology

Other Lecturers/Organizers include:

Malachi Griffith   Obi Griffith   Jennifer Foltz-Stringfellow
Susanna Kiwala   Todd Wylie   Many others!

Don't trust your data

Trusting your data

Property	Value	Accuracy
Mean of x	9	exact
Sample variance of $x : \sigma^2$	11	exact
Mean of y	7.50	to 2 decimal places
Sample variance of $y : \sigma^2$	4.125	±0.003
Correlation between x and y	0.816	to 3 decimal places
Linear regression line	$y = 3.00 + 0.500x$	to 2 and 3 decimal places, respectively
Coefficient of determination of the linear regression : R^2	0.67	to 2 decimal places
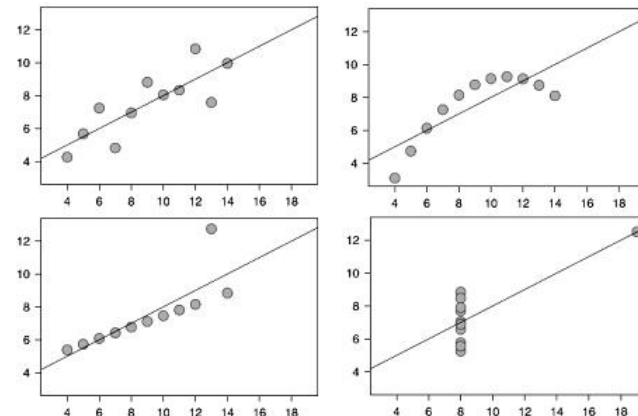
Datasaurus Dozen
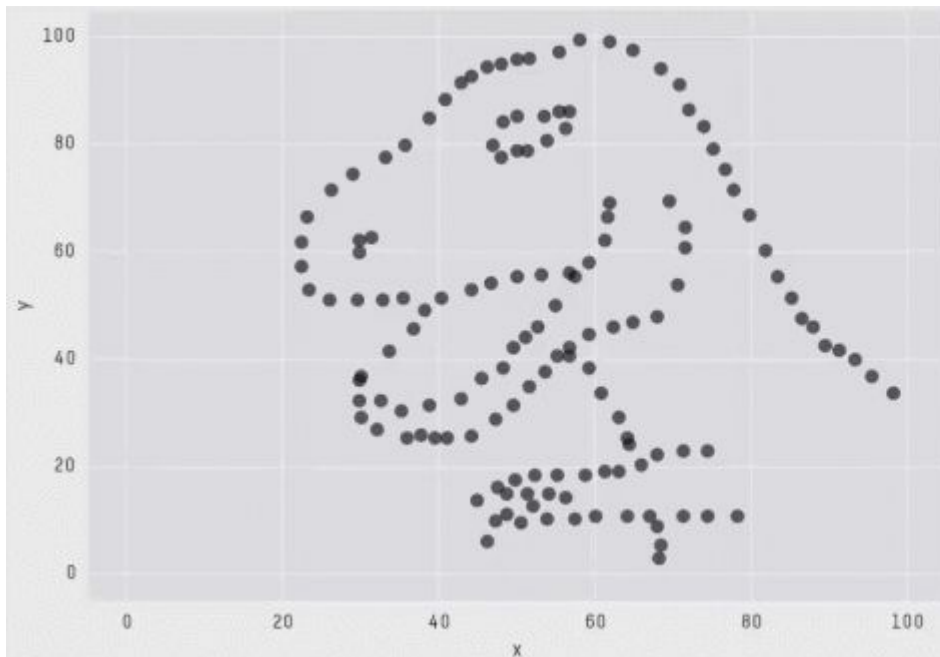
X Mean: 54.2659224
Y Mean: 47.8313999
X SD   : 16.7649829
Y SD   : 26.9342120
Corr. : -0.0642526

"Analyzing your data means inherently distrusting your data until you have exhausted yourself into giving up and trusting it."

-Aaron Quinlan

Course structure

- Pair an introduction to a biological or technical concept with some of the tools needed to analyze it

- Next week:
    - Sequence data generation
    - How to read, manipulate, and run quality control on sequence data

Prerequisites

- You do not need to know all of these things on Day 1
    - Tutorials and documentation in today's notebook

- But you do need to get moving!
    - Hands-on learning is *essential*. If you can't follow along and do the assignments, you will not get the most out of this course

Access to compute

- Laptop
  - limited resources
  - one time cost

- Compute cluster
  - Thousands of cores, Terabytes of RAM
  - Yearly fee ($100), storage costs
  - Shared usage (queues)

- Cloud
  - Virtually unlimited resources
  - Pay for usage

Cluster computing

- Large compute cluster
  - Dozens of servers
  - Thousands of cores
  - Terabytes of RAM

- How do we give everyone access efficiently and fairly?

LSF - Load Sharing Facility

- Binds different machines together into one common pool

- Jobs can be run on any machine

- Need to specify job requirements
    - How many CPUs
    - How much RAM
    - How much tmp (sometimes)
    - What kind of environment to run in

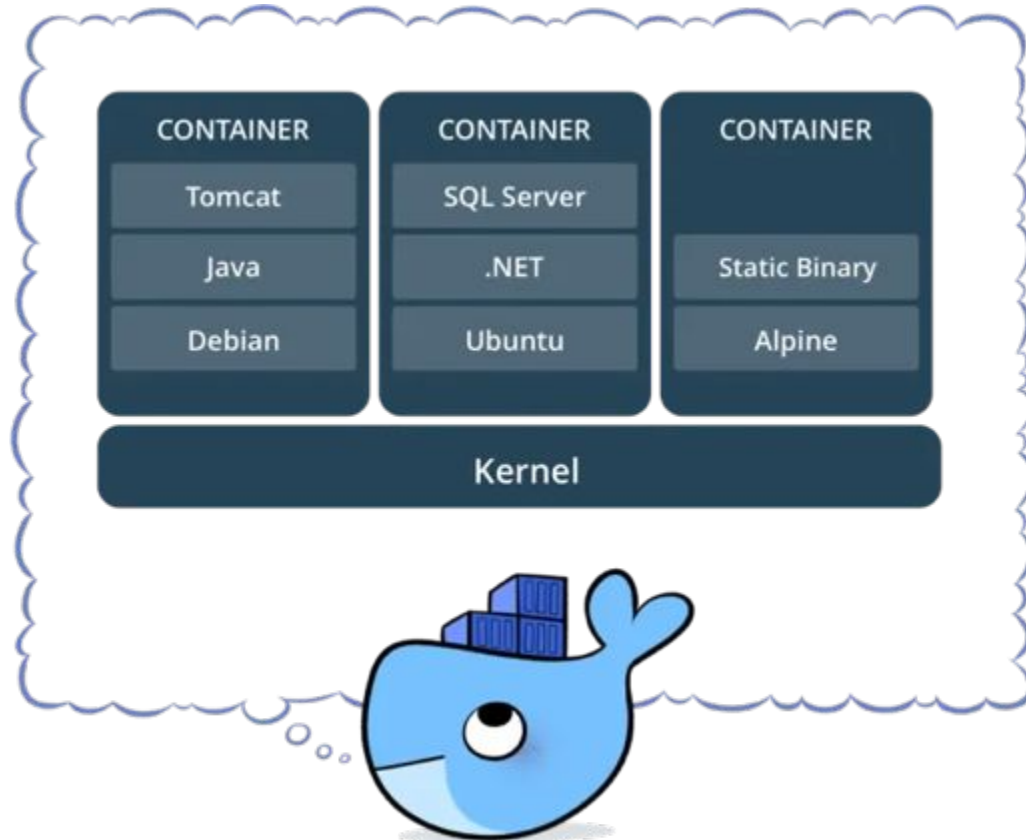- Cloud environments have similar dispatch systems

Computing Environments

- Laptop
    - You administer
    - You control completely

Computing Environments

- Laptop
    - You administer
    - You control completely

- Shared compute cluster
    - A sysadmin or group administers it
    - You control very little

Docker containers

Computing Environments

- Laptop
    - You administer
    - You control completely

- Shared compute cluster
    - A sysadmin or group administers it
    - You control very little

- Docker (containers)
    - Sysadmins handle the hardware
    - You control the software almost completely

Finding docker images

- Search engines
  - "docker bedtools"

- Repositories - Bioconda/Quay.io/Dockerhub
  - "docker mosdepth quay"

- Slack - ask in #docker

- Building your own

Homework

- Will not be turned in or graded

- Will be useful for understanding subsequent lectures
-
- Posted on the course page

  https://github.com/genome/bfx-workshop