Building predictive models of disease

Icahn School of Medicine at Mount Sinai

August 11, 2013
Technologies Emerging to Observe the Functioning of Single Molecules in Real Time

- Replicating DNA
- Transcribing Genes (RNA)
- Translating RNA into Proteins
These technologies are enabling scoring of very large-scale, high-dimensional data on individuals for low cost.

- Modified and unmodified DNA
- Modified and unmodified coding and non-coding RNA
- Phosphorylated and unphosphorylated proteins
- Metabolites
That promise to enable the construction of molecular networks that define the biological processes that comprise living systems.
Single molecule real time observation system
Building a new, super high resolution microscope

Confinement to $20 \times 10^{-21}$ liters

Science, Vol 299, Jan 31 2003, pp682-686
Multiplexing
In addition, many significant variants cannot be detected with short read technologies: Long reads needed to complete the picture.
Assaying internal molecular states and microenvironments is critically important, but what about the external macroenvironment

Your own DNA

Your microbiome DNA

DNA from bugs in the environment
Listeria Outbreak in the U.S. (September 2011)

Deaths from tainted cantaloupes rise to 15

By the CNN Wire Staff
updated 5:31 PM EST, Fri September 30, 2011
Cholera Outbreak in Haiti (October 2010)
Rapid identification of the Haiti cholera outbreak strain
**Timeline**

**Initiating the project (6-7 Nov)**
- 6 Nov: Matt calls
- 7 Nov: PacBio decides to go

**Sample prep and sequencing (8-12 Nov)**
- 8 Nov: Matt’s group cultures samples
- 10 Nov: Matt’s group sends DNA
- 11 Nov: Sequencing begins
- 12 Nov: Sequencing of 5 genomes complete

**Analysis (13-15 Nov)**
- Data QC
- Assembly and variation detection
- Building phylogenetic trees
- Annotating structure variation regions

**Writing the paper (15-19 Nov)**
- PacBio first draft (16 Nov)
- Refined with editor input (18 Nov)
- Submitted to NEJM (19 Nov)

**Provisional acceptance (20-24 Nov)**
- Reviews received 22 Nov
- Decision indicating intent to publish 24 Nov

**Formal acceptance (25 Nov – 1 Dec)**
- Multiple revisions with editor (25-30 Nov)
- Official acceptance (1 Dec)

**Paper Published!! (9 Dec)**
Single nucleotide variations also unambiguously positioned the Haitian cholera strain next to South Asia strains.
We have previously demonstrated the ability of PacBio SMRT sequencing (long reads one of the hallmarks) to complete genomes.
...and demonstrated utility of long reads to organize mutation information from tumors to validate targets and identify resistance haplotypes

**LETTER**

doi:10.1038/nature11016

**Validation of ITD mutations in FLT3 as a therapeutic target in human acute myeloid leukaemia**

Circular Consensus Sequence

1.4 kbps, ~99% accuracy CCS reads

CCS reads with single ITD sequences

CCS reads with multiple ITD sequences

ITD sequences

Count alternative codons at positions for residue E608, F691, D835 and Y842.
PacBio RS Quick Turnaround Time Provides a Path for Real Time Pathogen Surveillance
Diverse Range of Specific and Non-Specific Viruses Targeted

- Specific targets included (but not limited to):
  - Human Rhinovirus
  - Human Respiratory Syncytial Virus
  - Human Metapneumovirus
  - Human Coronavirus
  - Human Parainfluenza Virus
  - Influenza A viruses
  - Paramyxovirinae
  - Respirovirus
  - Rubulavirus
  - Pneumovirinae
  - Parechovirinae
  - Enterovirus
  - Rotavirus
  - Caliciviridae
  - Astroviridae
  - Adenovirus
  - Human Herpesvirus 5
  - Human Herpesvirus 3

- Examples of specific viruses:
  - Rhinovirus: Common Cold
  - Herpes viruses: HHV5 Mononucleosis, HHV3 Chickenpox/Shingles
  - Astrovirus: Gastroenteritis (in children)
  - Adenovirus: Upper respiratory tract infection
  - Influenza A: Flu
  - Parainfluenza virus: Lower respiratory tract infection
  - Rotavirus: Severe diarrhoea (“stomach flu”)
Enabling a More Comprehensive Understanding of Personal Genomes Via the Construction of a Disease Weather Map

- Prepare Samples
  - Convert RNA to cDNA
  - Purify DNA
  - Fragment DNA
  - Add 3’ A Tail
  - Ligate Adapters

Environmental Swabs
- Train Stations (BART, Cal Train)
- Airports (SFO, SJC, OAK)
- Emergency Rooms
- Sewage Treatment Plants
- Universities

Sequence

Project findings onto map (detections and trends)

Assemble, Identify, and Quantify Genomes
Starting with low-resolution maps as a proof of concept.
Detection of Viral Pathogens from Inanimate Surfaces

- High-traffic surfaces at Pacific Biosciences:
  - Front door handle
  - Common laboratory bench top
  - Break room refrigerator door handle
  - Slide projector remote control
  - Lavatory toilet flush handle
  - Lavatory door handle
  - Laboratory telephone handle
  - Cubicle desk surface
  - Money

Sampled every week for a period of one month
Anonymous donors, submitting nasopharyngeal swabs every two weeks over ~2.5 months
Sequencing Read Distribution from Sewage

- Common virus causing common cold; hand foot mouth disease; aseptic meningitis
- Oyster-associated non-bacterial gastroenteritis
- Common virus infecting tobacco plants
- Common in field-grown bell, hot, and ornamental pepper species
- Common virus infecting tomato and pepper plants

- Lactococcus phage Q54 1
- Lactococcus phage BK5-T 1
- Lactococcus phage TP901-1 1
- Lactococcus phage 1706 2
- Pseudomonas phage LIT1 1
- Lactococcus phage KSY1 2
- Pleurotus ostreatus virus 1 1
- Melon necrotic spot virus 9
- Carnation mottle virus 2
- Enterobacteria phage GA 2
- Acinetobacter phage AP205 2
- Enterobacteria phage Qbeta 1
- Human enterovirus B 1
- Porcine kobuvirus swine/K-30-HUN/200
- Aichi virus 4
- Southern bean mosaic virus 1
- Ryegrass mottle virus 1
- Rubus chlorotic mottle virus 1
- Eggplant mosaic virus 1
- Turnip yellow mosaic virus 1
- Nemesia ring necrosis virus 1
- Potato virus S 1
- Clover yellow mosaic virus 7
- Potato virus X 1
- White clover mosaic virus 8
- Pepino mosaic virus 1
- Garlic virus D 1
- Garlic virus A 2
- Garlic virus C 1
- Cucumber green mottle mosaic virus 1
- Tobacco mosaic virus 8
- Pepper mild mottle virus 12
- Tomato mosaic virus 12
- Paprika mild mottle virus 2
- Bell pepper mottle tobamovirus 3
- Cucumber mottle virus 1
- Rehmannia mosaic virus 1
- Brome mosaic virus 1
- Cucumovirus 1
- Astrovirus MLB1 1
- Escherichia phage K1H2

bacteria

eukaryota

ssRNA viruses

dsDNA viruses

archaea
Metagenomic Hit Distribution for Inanimate Surfaces and Nasal Swabs

Surfaces:

- fridge door
- desk surface
- toilet flush handle

Nasopharyngeal swabs:

- human
- poultry (chicken & turkey)
- pig
- bacteria
- virus - sometimes!
- protozoa
- not assigned

Wash Your Hands!!!
Detection of Viral Pathogens – Influenza on Surfaces

- Multiple strains identified
  - H1N1
  - H3N2
  - H2N2, H5N1 (rarely)
- Some samples with multiple strains present
- Some strains only occurred over single sampling period
  - H5N1 – 1
  - H2N2 – 2

<table>
<thead>
<tr>
<th>Sampling period</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>front door 1</td>
<td>H1N1</td>
<td>H1N1/H2N2</td>
<td>H1N1</td>
<td>H1N1</td>
</tr>
<tr>
<td>fridge door 2</td>
<td>H1N1</td>
<td>H1N1/H3N2</td>
<td>H1N1/H3N2</td>
<td>H1N1/H3N2</td>
</tr>
<tr>
<td>projector remote 1</td>
<td>H1N1/H3N2</td>
<td>H1N1</td>
<td>H1N1</td>
<td>H1N1/H3N2</td>
</tr>
<tr>
<td>projector remote 2</td>
<td>H1N1/H2N2</td>
<td>H1N1</td>
<td>H1N1</td>
<td>H1N1</td>
</tr>
<tr>
<td>restroom door 1</td>
<td>H1N1</td>
<td>H1N1/H3N2</td>
<td>H1N1/H3N2</td>
<td>H1N1/H3N2</td>
</tr>
<tr>
<td>restroom door 2</td>
<td>H1N1</td>
<td>H1N1/H3N2</td>
<td>H1N1/H3N2</td>
<td>H1N1/H3N2</td>
</tr>
<tr>
<td>toilet flush handle 1</td>
<td>H1N1/H3N2/H5N1</td>
<td>H1N1</td>
<td>H1N1/H3N2</td>
<td>H1N1/H3N2</td>
</tr>
<tr>
<td>desk 2</td>
<td>H1N1</td>
<td>H1N1</td>
<td>H1N1</td>
<td>H1N1</td>
</tr>
<tr>
<td>desk 3</td>
<td>H1N1/H3N2/H5N1</td>
<td>H1N1</td>
<td>H1N1/H3N2</td>
<td>H1N1/H3N2</td>
</tr>
<tr>
<td>desk 8</td>
<td>H1N1/H3N2</td>
<td>H1N1</td>
<td>H3N2/H1N1</td>
<td>H1N1/H3N2</td>
</tr>
<tr>
<td>$1*</td>
<td>H1N1</td>
<td>H1N1</td>
<td>H3N2/H1N1</td>
<td>H1N1/H3N2</td>
</tr>
</tbody>
</table>
The New Generation of Microbe Hunters

HOT SPOTS Researchers hope to use rapid DNA sequencing of microbes to make “disease weather maps” of large areas, like a hypothetical disease outbreak in the Bay Area, above.

LOCAL TESTING On a smaller scale, colors show increasing amounts of flu virus measured at different locations and surfaces inside the Pacific Biosciences building during flu season.
The ultimate is to get to a high resolution map: What’s going on in your neighborhood?
Coming soon to a neighborhood near you?

Mom, can I go over to Billy’s house after school?
My god, she’s going to check disease weather map

Just a minute, dear…
Give me the street-level view of Billy’s house please
Disease Weather Map: Street Level View

**Long-term Metagenomic Report**
- Viral pathogens (1yr)
  - No diarrheal viruses
  - Cold viruses detected 20 days
  - Influenza detected 3 days
- Disease susceptibility profile
  - Bacterial composition indicates pro-athero, pro-diabetes environment
  - Food viruses detected indicate high-fat Western diet, minimal vegetables

**Daily Metagenomic Report**
- Current pathogen composition within 95% CI of year mean
- No cold, flu, etc. viruses detected
Disease weather map saves the day

Okay, Johnny, you can go over to Billy’s house but come home for dinner.

Uhhhhgggg ....
Multiscale measures of patients now available through efforts like Mount Sinai’s Biobank (>25,000 *identified* patients and growing fast)
Big Data Warehouses at Medical Centers like Mount Sinai Contain Virtually All Facts And Transaction Records For Millions of Patients

Institute for Personalized Medicine at Mount Sinai
Topological network of Mt. Sinai clinical data reveals diabetes populations stratified by phenotypes.

Each node represents one or more similar patients (by clinical phenotype).
Comparison of genetic variation between A and B unveils CACNA1E as a discriminating SNP
Comparison of genetic variation between A and B unveils CACNA1E (calcium channel gene) as a discriminating SNP.
Significance of calcium clinical measure differentiates across the network’s diabetic subregions
Integrating data to build predictive models of complex disease and drug response phenotypes

Integrate to Construct Causal Probabilistic Networks

DNA Variations

Protein-Protein Interactions and Protein Complexes

Gene Expression

GCN4

ILV6

DNA-Protein Binding

LEU2

Nature Reviews Drug Discovery 8:286-295 (2009)
By sampling populations and measuring traits, we can establish associations between traits of interest.

Monitor disease and molecular traits in populations.

Putative causal gene

Disease trait
While correlations between molecular and clinical data have demonstrated great utility for classification,

**Netherlands Cancer Institute / Rosetta study to predict survival in breast cancer using large-scale expression signatures**

*NEJM 2002: Vol 347, No. 25*
Molecular state information on its own only gives you correlation; does not elucidate causal connections.
Establishing causality

Classic Approach: Direct, hard hitting, artificial perturbations of genes

Establishing causality:

Insulin Causal Gene

Insulin Levels
DNA variation as the perturbation source enables for causal inferences to be made

Linking molecular and disease variation with DNA Variation in populations
Leveraging DNA variation as a systematic perturbation source

Given this triangle relationship, we can mathematically infer the most likely relationship.
The networks we need to construct should reflect behavior of the system.

- Human and experimental animal populations that are *coherent*
- Example
  - BTBR/ob (insulin resistance → T2D) vs. B6/ob (insulin resistance → no T2D)
  - 1000 animals: 500 F2s from a 4 week time point (BTBR/ob not sick at this point), 500 from 12 week time point (BTBR/ob sick and dying at this point)
  - Isolate islets, hypothalamus, white adipose tissue, liver, and skeletal muscle from all animals
  - Clinically characterize, genotype, profile, and analyze!
Networks can be linked to human forms of disease to help motivate relevance for human treatments

- 4 different parts of brain
- Liver
- Macrophages*
- Blood monocytes
- Omental and Mediastinal Visceral Fat
- Skeletal Muscle
- Carotid and coronary Lesion
- Atherosclerotic Arterial wall
- Atherosclerosis-Free Artery Wall (IMA)
- Subcutaneous Fat

AND

Thousands of individuals now genotyped and profiled

Isolate DNA and genotype

Genotype

AND

Isolate RNA and sequence or profile
Networks facilitate direct identification of genes that are causal for disease
Prospective validation, again, is the gold standard for these types of predictions.
In fact, 8 of the first 9 genes we attempted to validate from our first set of predictions validated

<table>
<thead>
<tr>
<th>Gene</th>
<th>Model</th>
<th>Fat-related traits</th>
<th>Lipid traits and glucose</th>
</tr>
</thead>
<tbody>
<tr>
<td>Zfp90</td>
<td>tg</td>
<td>Increased fat/muscle growth, body weight, total fat pad mass, total fat/body weight, retroperitoneal fat pad, mesenteric fat pad, subcutaneous fat pad</td>
<td>Decreased total cholesterol, HDL, unesterified cholesterol, triglyceride, and glucose</td>
</tr>
<tr>
<td>Gas7</td>
<td>tg</td>
<td>Decreased fat/muscle growth, body weight, gonadal fat pad, total fat pad mass, decreased mesenteric fat</td>
<td>Decreased total cholesterol, HDL (females)</td>
</tr>
<tr>
<td>Gpx3</td>
<td>tg</td>
<td>Decreased fat/muscle growth (males)</td>
<td>Decreased total cholesterol, HDL (females)</td>
</tr>
<tr>
<td>Lactb</td>
<td>tg</td>
<td>Increased fat/muscle growth</td>
<td></td>
</tr>
<tr>
<td>Me1</td>
<td>ko</td>
<td>Decreased body weight</td>
<td></td>
</tr>
<tr>
<td>Gyk</td>
<td>ko</td>
<td></td>
<td>Increased free fatty acids, decreased glucose</td>
</tr>
<tr>
<td>Lpl</td>
<td>ko (het)</td>
<td>Increased fat/muscle growth, total fat pad mass, total fat/body weight, mesenteric fat pad, subcutaneous fat pad</td>
<td>Increased triglycerides, decreased total cholesterol and unesterified cholesterol</td>
</tr>
<tr>
<td>C3ar1</td>
<td>ko</td>
<td>Increased fat/muscle growth, gonadal fat pad, subcutaneous fat pad</td>
<td></td>
</tr>
<tr>
<td>Tgfbr2</td>
<td>ko (het)</td>
<td>Decreased fat/muscle growth</td>
<td></td>
</tr>
</tbody>
</table>
Building networks over multiple tissues simultaneously
From the networks we identify the modules as functional units of the system and then construct networks based these functional units as nodes.
Using this process we can identify causal regulators of networks as a path to target disease (a way to target networks)

- Angiogenesis-enriched Module
- Vascularization-enriched Module
- Cell Cycle Module

- Test all sub-networks over the entire system to identify those that are causal for disease
- One sub-network in particular stood out:
  - 75% of genes supported as causal for athero lesions in this cross (117 of 157 genes) fall in this sub-network \( (p = 5.6 \times 10^{-3}) \)
  - Near 50% of genes supported as causal for obesity (366 of 735 genes) are in this sub-network \( (p = 3.78 \times 10^{-235}) \)
  - Near 45% of genes supported as causal for diabetes traits (263 of 571 genes) are in this sub-network \( (p = 1.5 \times 10^{-133}) \)
  - Sub-network contains hypertension genes
  - All enrichments seen across sexes and multiple tissues

Nature 452:429-435
Models used to interpret targets under development

Novel phosphatase under development at Merck for T2D

Predictions derived from the predictive models

- Increases fat mass
- Negatively impacts hypertension genes

GOOD
- Lowers glucose
- Raises insulin

BAD
Predictions are great, but only meaningful if they are validated.
But wait, the network also shows PPM1L and PPARG (target of Avandia) in a causal relationship.

Network Predicts:
- Avandia will lower glucose
- Avandia will make you fat
- Avandia will increase cardiovascular risk

Validation 2 years later:

Overall, for congestive heart failure, the estimated risk was higher compared to control and statistically significant for both drugs [rosiglitazone and pioglitazone].

The New York Times
Research Ties Diabetes Drug to Heart Woes

By GARDNER HARRIS

Hundreds of people taking Avandia, a controversial diabetes medicine, needlessly suffer heart attacks and heart failure each month, according to confidential government reports that recommend the drug be removed from the market.
The models can also be mined for new targets without all of the liabilities.

Novel orphan GPCR with no known function

Lowers glucose
Raises insulin
Lowers fat mass
Lowers lipids
Pharmacologic validation of the network predictions around P2ry14

- LDL C
- Gas5
- Mylc2p
- Mtif2
- Psrc1
- Hist3h2
- Cuedc1
- Leptin
- Glucose
- Sned1
- Rasd1
- Gbp1
- Dusp3
- Insulin

Graph showing the network with experimental data:
- % BW
- Placebo
- Low Dose
- Med Dose
- High Dose
- Normal Diet Control
# Acknowledgements

## Pacbio
- Steve Turner
- Andrew Kasarskis
- **NGAT Group**
  - Jonas Korbach
  - Tyson Clark
  - Khai Luong
  - Kristi Spittle
- Robert Sebra
- Seven Lin
- Eric Olivarez
- Kevin Travers

## Informatics
- Jason Chin
- Aaron Klammer
- Michael Brown

## Mount Sinai
- Andrew Kasarskis
- Jun Zhu
- Bin Zhang
- Michael Linderman
- Gaurav Pandey
- Rui Chang
- Shaun Purcell
- Ali Bashir
- Joel Dudley

## Partners
**Sage Bionetworks**
- Stephen Friend
- Et al.

**UCSF**
- Neil Shah
- Cathy Smith
- W. Michael Korn
- Joe De Risi Lab

**UCLA**
- Lusis Lab

**Harvard**
- Matt Waldor lab
- John Mekalanos lab
- [U of Maryland](#)
- Dave Rasko lab
- [U of Wisconsin](#)
- Attie Lab

**NEB**
- Rich Roberts
- Tamas Vincze
- [U of Minnesota](#)
- Gang Fang
- Vipin Kumar

---

*Note: [U of Maryland](#) and [U of Wisconsin](#) links are placeholders for actual links*