Biology: From Wet to Dry

David Heckerman
Microsoft Research

Emergence of a Fourth Research Paradigm

Thousand years ago – **Experimental Science**

- Description of natural phenomena

Last few hundred years – **Theoretical Science**

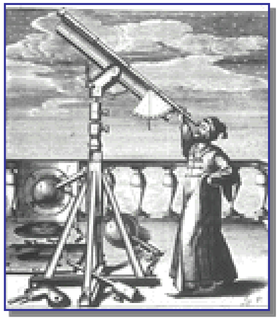
- Newton’s Laws, Maxwell’s Equations…

Last few decades – **Computational Science**

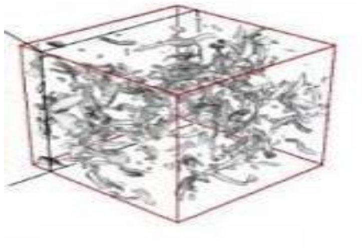
- Simulation of complex phenomena

Today – **Data-Intensive Science**

- Scientists overwhelmed with data sets from many different sources
  - Captured by instruments
  - Generated by simulations
  - Generated by sensor networks



$$\left(\frac{\dot{a}}{a}\right)^2 = \frac{4\pi G\rho}{3} - \mathrm{K}\frac{c^2}{a^2}$$



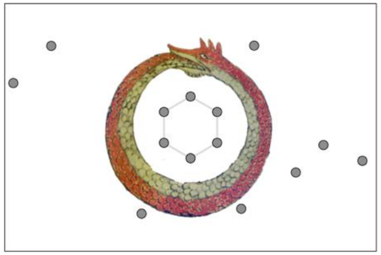
eScience is the set of tools and technologies to support data federation and collaboration

- **For analysis and data mining**
- **For data visualization and exploration**
- **For scholarly communication and dissemination**

(With thanks to Jim Gray)

Biology: From Wet to Dry

- Old days: Creative one-off wet-lab experiments



- Recent days: Assembly line experiments (DNA, RNA, proteins), collaboration possible

- Now: Can do the real science ourselves without the wet lab

eScience Research Group

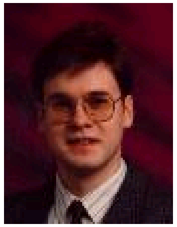
Jonathan Carlson

Bob Davidson

David Heckerman

Carl Kadie

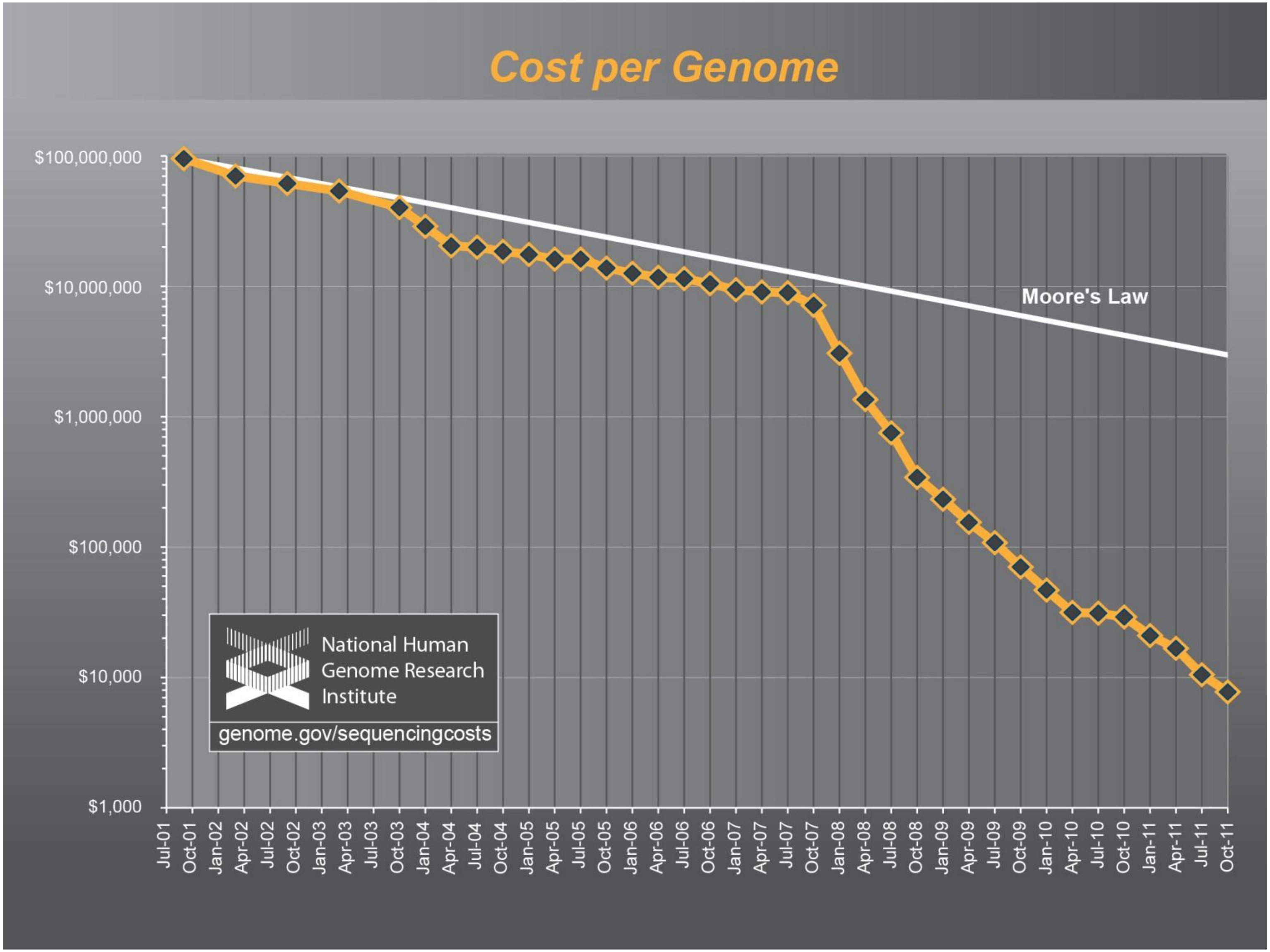
Jennifer Listgarten



Nebojsa Jojic

Christoph Lippert

The Genomics Revolution



Personalized Medicine

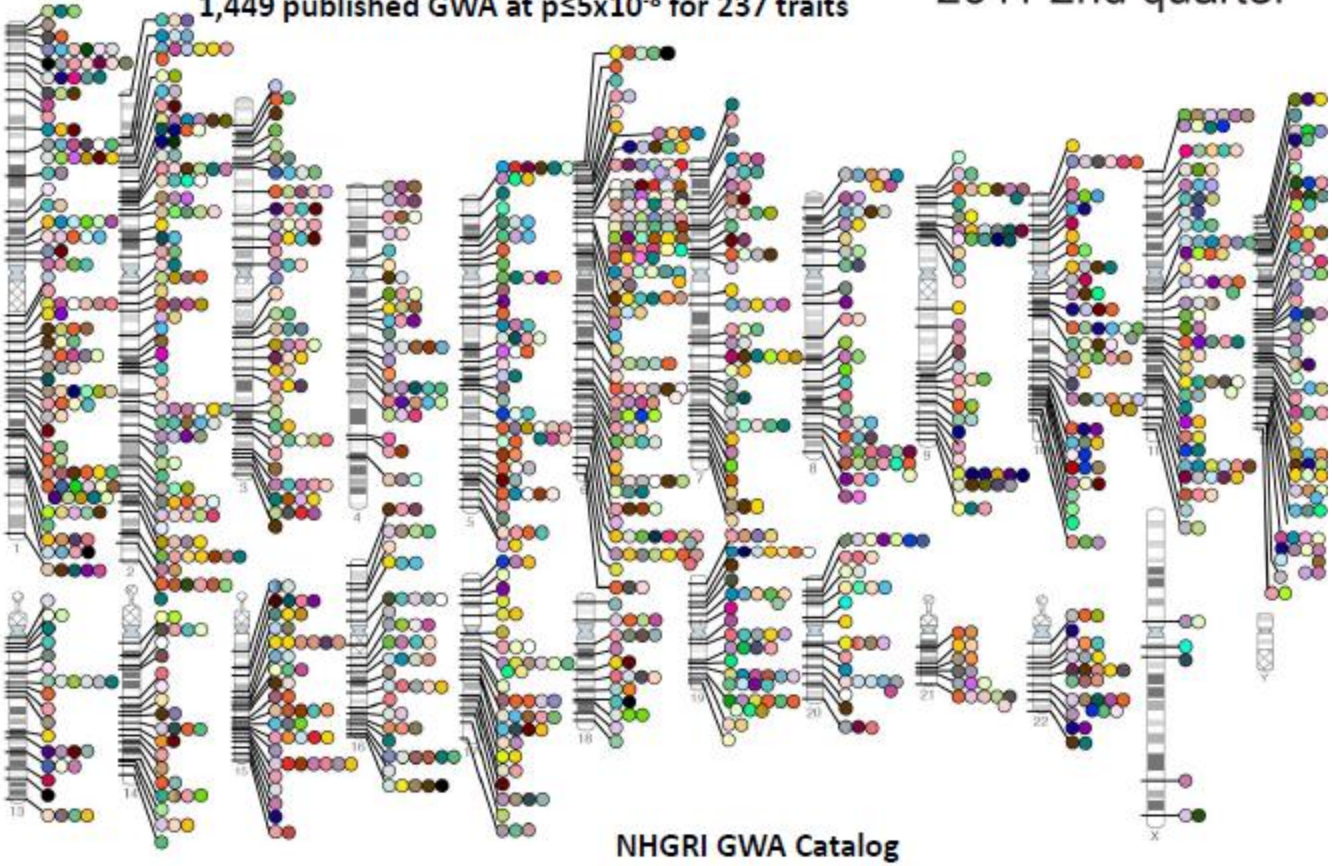
Identify genetic markers (SNPs) associated with

- Getting a disease
- Reacting badly to a drug
- Reacting favorably to a drug

Identifying genetic causes of disease (Genome-Wide Association Studies, GWAS)

Published Genome-Wide Associations through 06/2011, 1,449 published GWA at p≤5×10^{-8} for 237 traits

2011 2nd quarter



NHGRI GWA Catalog
www.genome.gov/GWAStudies

Example

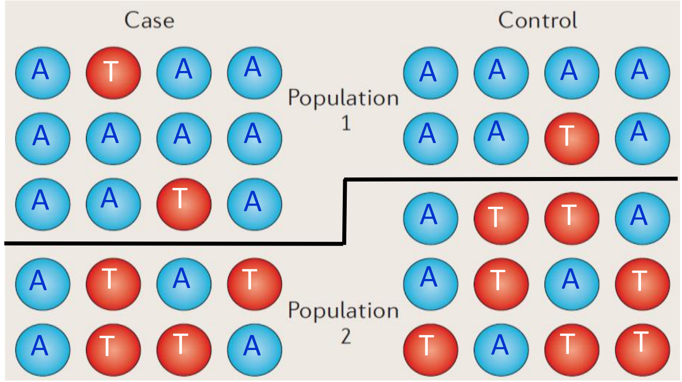
- ALS (Lou Gehrig's disease): Found a single DNA change that accounts for about a third of all familial disease in Europe (Traynor et al.; *Neuron* Sept 2011)

GWAS issues

- Much of the hanging fruit has been picked
- Remaining signals are weak and scattered across the genome
- To pick up these signals, we need lots of data
  - deCode
  - 23andMe
  - Kaiser
- Large data → confounding
  - Multiple ethnicities
  - Closely related individuals

Challenge: Confounding factors
(Advanced machine learning required)

- Suppose the set of cases has a different proportion of ethnicity X from control.
- Suppose we use linear regression to look for SNP-phenotype correlations.
- Then genetic markers that differ between X and other ethnicities in the study, Y, will appear artificially to be associated with disease.
- Problem gets worse with more data.

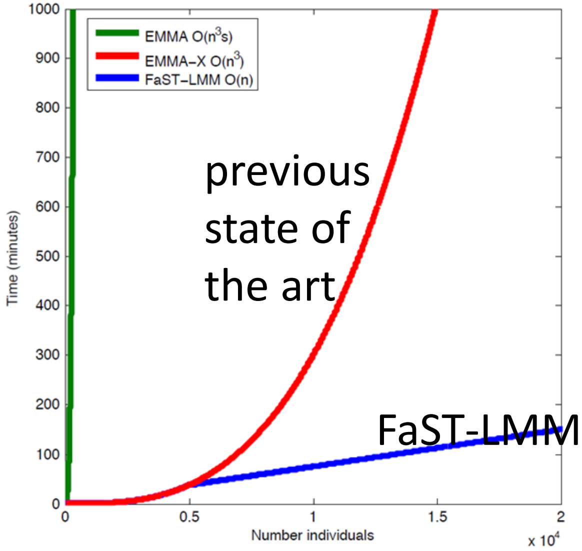


FaST-LMM: Factored Spectrally Transformed Linear Mixed Models

- Best algorithms for GWAS use linear mixed models
- But these have $O(N^3)$ runtime and $O(N^2)$ memory use; N<5,000
- FaST-LMM has $O(N)$ runtime and memory use; N>100,000; much more signal
- Requires number of SNPs used to estimate similarity among individuals to be less than N
- Results are more accurate than standard approach!

nature | **methods** September 2011

nature | **methods** June 2012

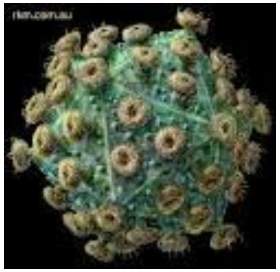


previous state of the art

FaST-LMM

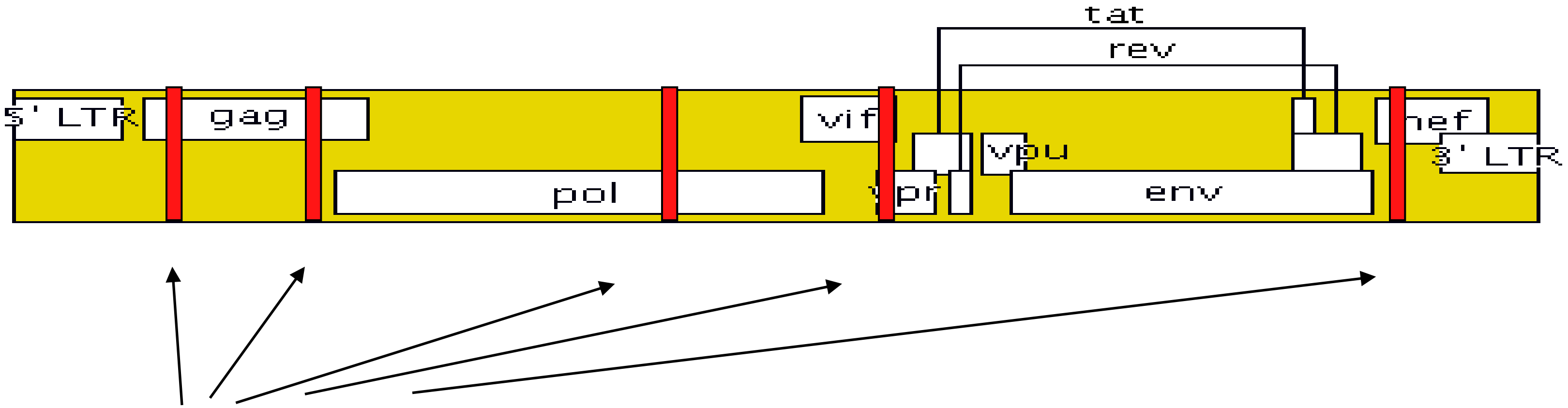
Vaccine design

- Spammers mutate their messages to work around filters
- Solution: Go after the weak link

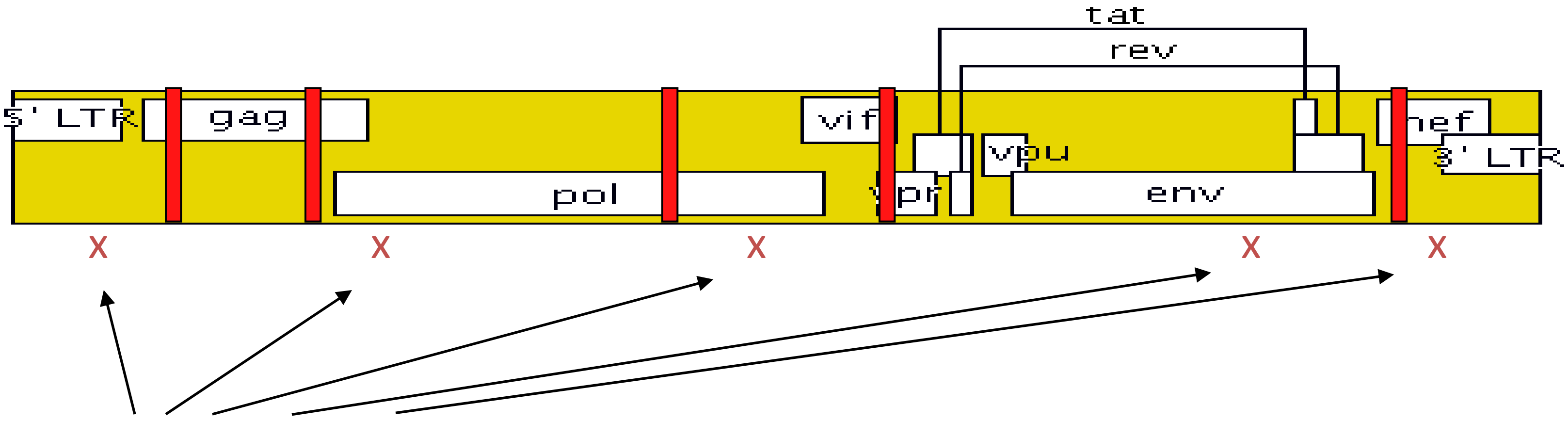


- HIV mutates to avoid attack by immune system
- Solution: Go after the weak link

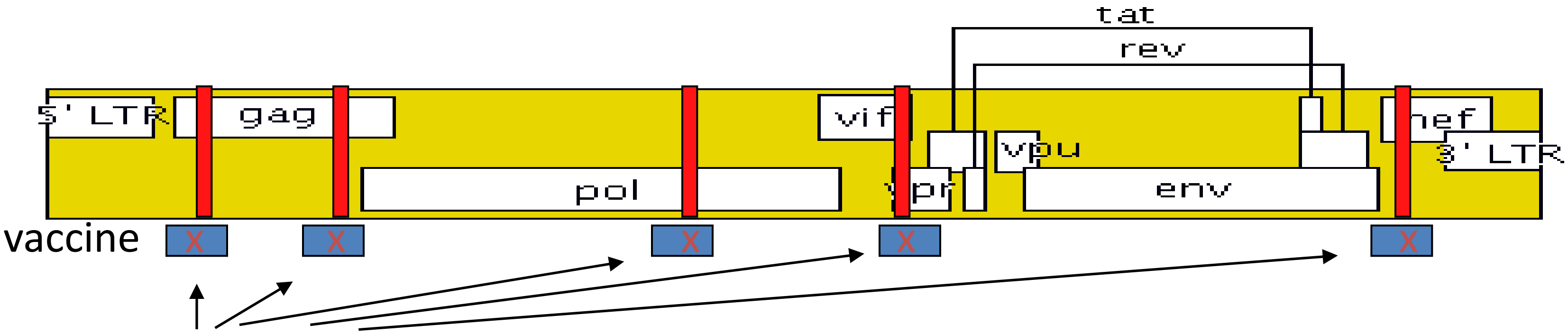
Hypothesis: Certain parts of HIV are critical to its function



If HIV mutates within these epitopes, it becomes less or non-functional



Left to its own devices, our immune system attacks at random spots (“epitopes”)

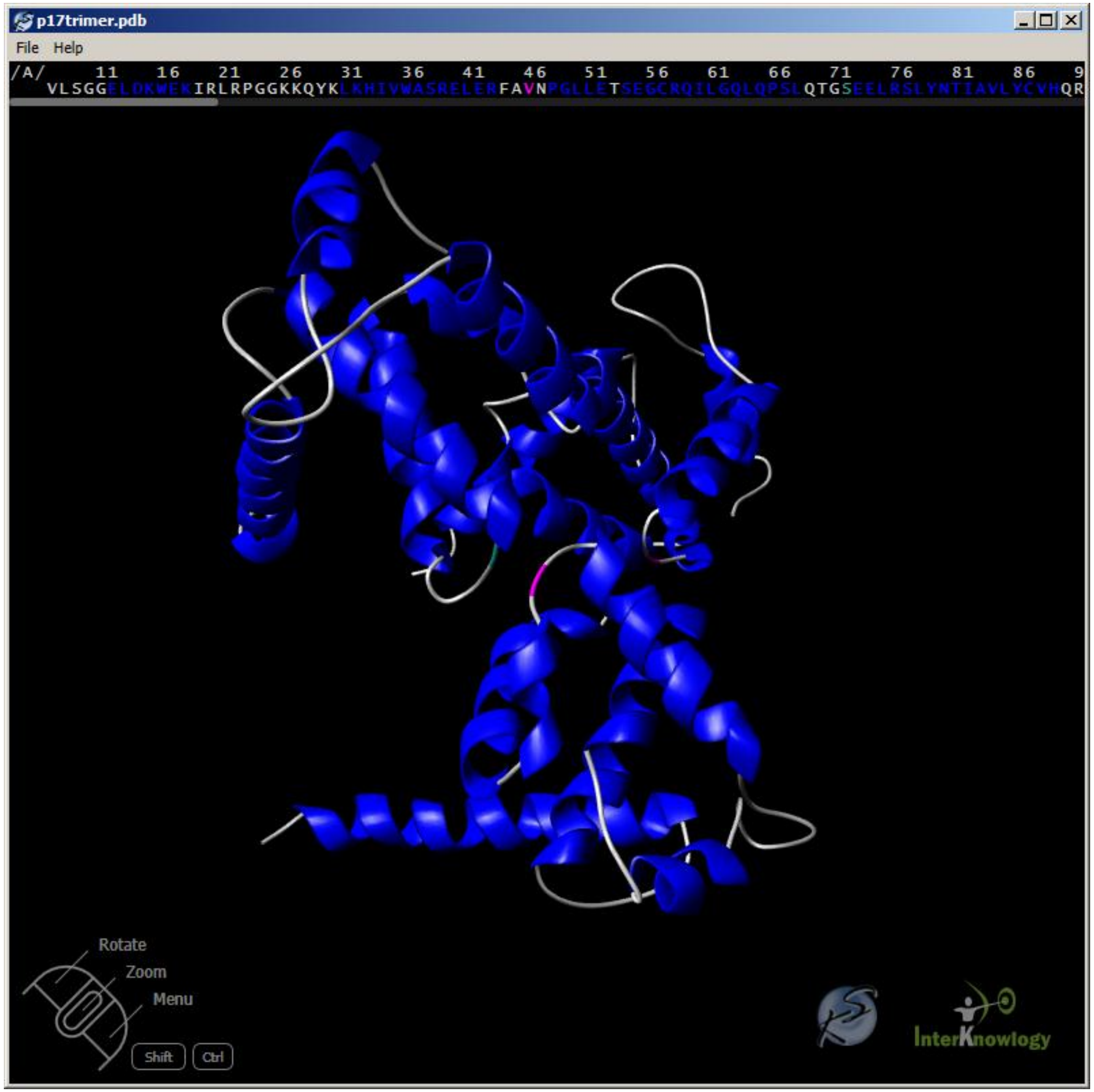


A focused vaccine can show immune system where to attack

Work with Bruce Walker at Harvard, we have identified a half dozen weak points.
Simple machine learning.

Challenge: There are hundreds of different immune system types

Finding vulnerable spots



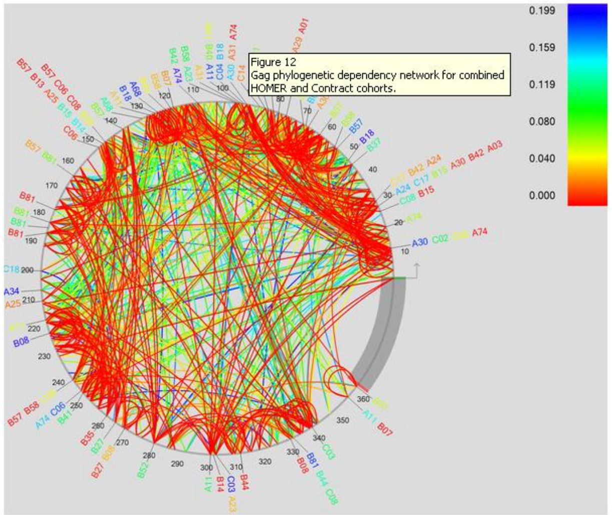
```
/A/     11   16   21   26   31   36   41   46   51   56   61   66   71   76   81   86   9
    VLSGGELDKWEKIRLRPGGKKQYKLKHIVWASRELERFAVNPGLLETSEGCRQILGQLQPSLQTGSEELRSLYNTIAVLYCVHQR
```

Finding vulnerable pairs with machine learning

- Basic idea: Watch how HIV mutates in an individual under natural attack from the immune system
- Challenge: Individuals are not infected with the same sequence; noise
- Solution: PhyloD, a machine learning algorithm that accounts for differences in the sequences
- Demo

- Published in *Science*, March 2007
- Now used by dozens of HIV research groups
- We've published 32 papers; over 1000 citations

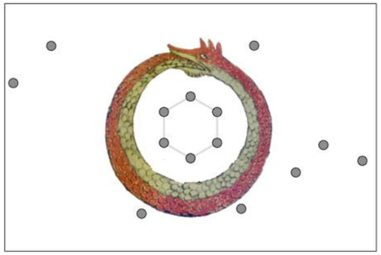
- Another important discovery: Natural killer cells also attack HIV (*Nature* 2011)



PhyloD.Net on cover of *PLoS Comp Bio*, Nov 2008
Carlson, Kadie, & Heckerman et al.

Biology: From Wet to Dry

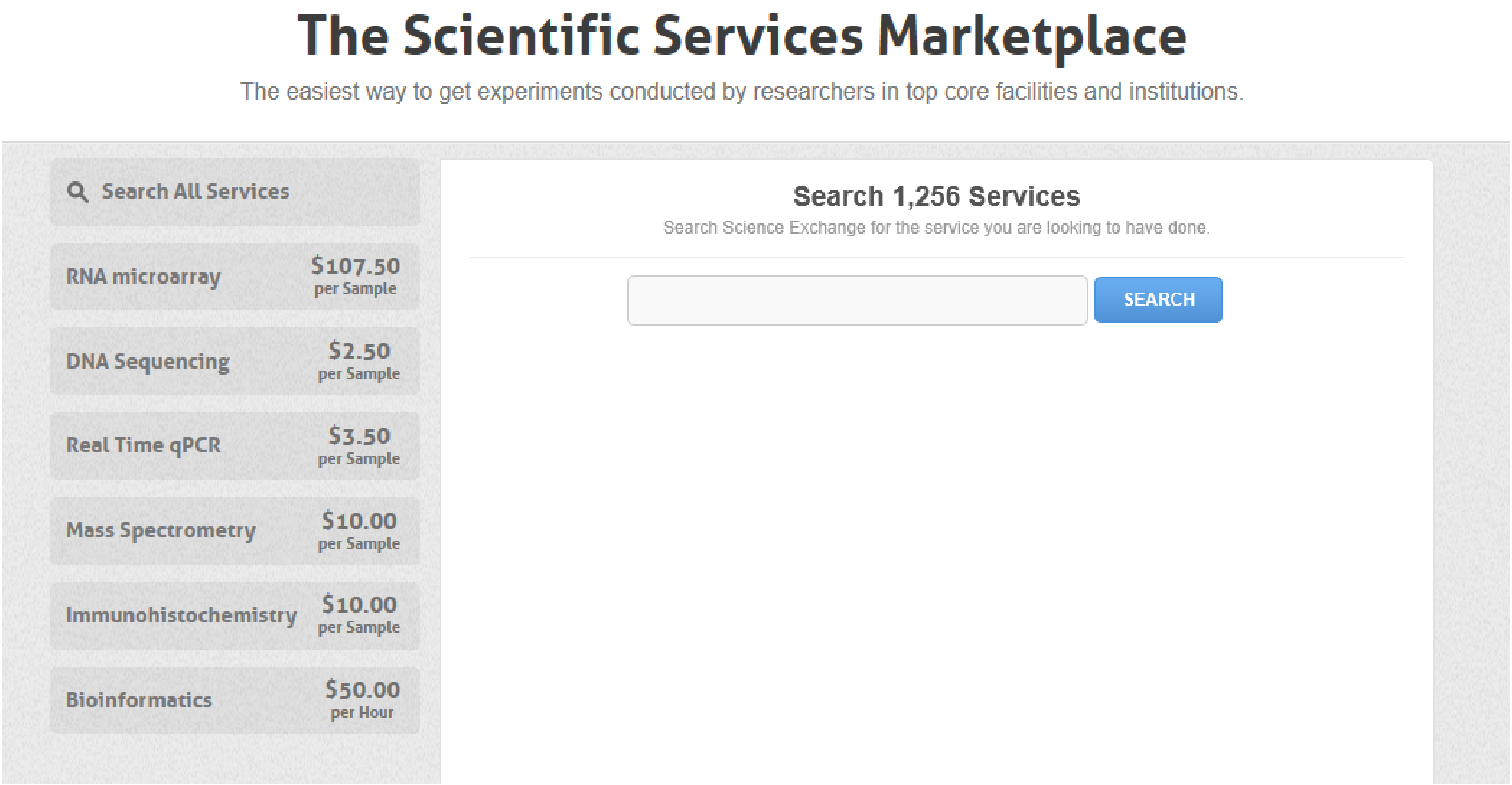
- Old days: Creative one-off wet-lab experiments



- Recent days: Assembly line experiments (DNA, RNA, proteins), collaboration possible

- **Now: Can do the real science ourselves without the wet lab**

Data as a commodity



The Scientific Services Marketplace

The easiest way to get experiments conducted by researchers in top core facilities and institutions.

RNA microarray	$107.50 per Sample
DNA Sequencing	$2.50 per Sample
Real Time qPCR	$3.50 per Sample
Mass Spectrometry	$10.00 per Sample
Immunohistochemistry	$10.00 per Sample
Bioinformatics	$50.00 per Hour

Search 1,256 Services

Search Science Exchange for the service you are looking to have done.

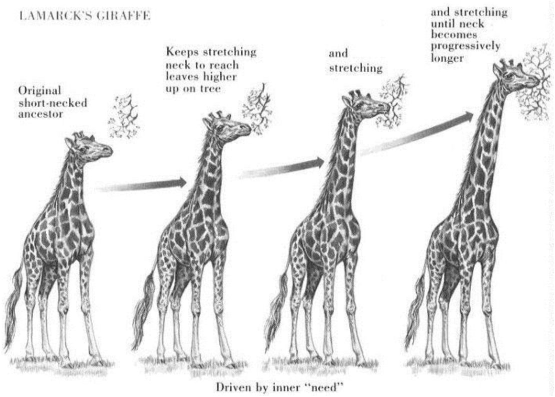
Genomics: Data is already there

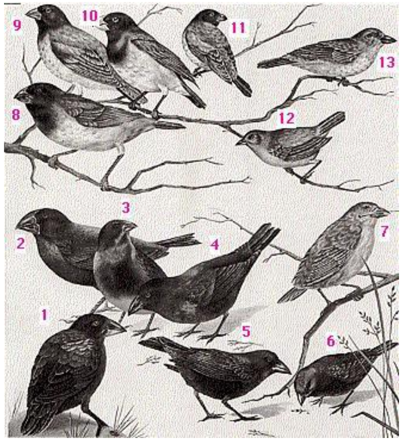
- Genome/epi-genome interactions
- Finding new uses for approved drugs
- Coronary artery disease

Genome/epi-genome interaction

Lamark: Environment → ? → Traits

Darwin: ? → Traits


LAMARCK'S GIRAFFE
Original short-necked ancestor
Keeps stretching neck to reach leaves higher up on tree
and stretching
and stretching until neck becomes progressively longer
Driven by inner "need"

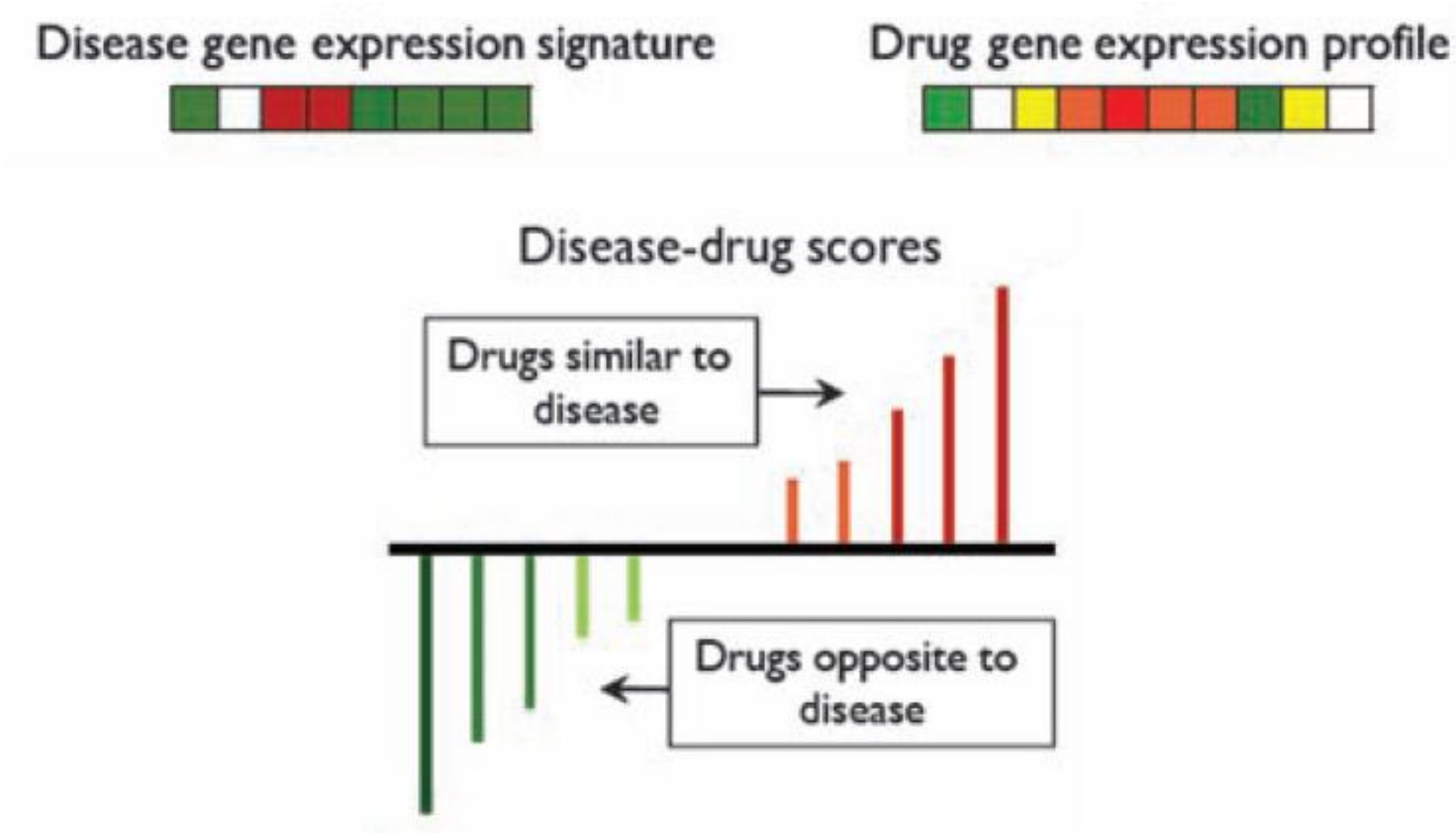


They were both right: Genome and epi-genome

Listgarten et al.: Using public data, showed how genome influences epi-genome

Finding new uses for approved drugs

Butte lab, Science 2011


Disease gene expression signature
Drug gene expression profile
Disease-drug scores
Drugs similar to disease
Drugs opposite to disease

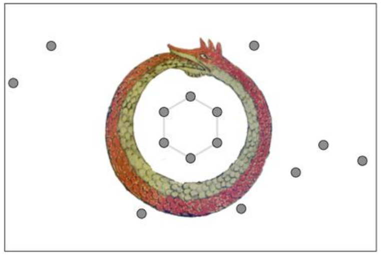
Identified Cimetidine (for ulcers) as useful in treatment of lung adenocarcinoma

Moondog project with Azure

- Wellcome Trust data for seven common diseases
- With FaST-LMM and Azure, can look at all SNP pairs (about 60 billion of them)
- 400 compute years; 20 TB output
- Found new interactions in coronary artery disease

Biology: From Wet to Dry

- Old days: Creative one-off wet-lab experiments



- Recent days: Assembly line experiments (DNA, RNA, proteins), collaboration possible

- Now: Can do the real science ourselves without the wet lab

Questions