Big data and new models needed to study DNA variation in evolution and cancer

David Haussler, UC Santa Cruz
The G10K Community of Scientists
The Genome 10K Community Goal:
To understand how complex animal life evolved through changes in DNA and use this knowledge to become better stewards of the planet.

• Collect samples and sequence at least 10,000 different vertebrate species, bank fibroblast cell lines and make iPS lines for > 1,000 species. Currently ~350 genomes and dozens of iPS lines from various labs.

• Annotate genomes, map and interpret genetic differences between species, and compute the evolutionary record of genetic changes on each lineage

• Correlate with ecologic, biologic and geologic data for deep study of vertebrate diversity, biology, evolution, and for species conservation
Grand scientific challenge of vertebrate molecular evolution

Reconstruct the evolutionary history of each base in the genomes of the living species

- Recognize functional elements from patterns of negative and positive selection
- Find the origins of evolutionary innovations specific to the human and other lineages
Early look at some evolutionary differences in human neurodevelopment
Differentiating stem cells into neurons to discover specific regulatory changes.

**Pax6**: Marker for young neuro-precursors

**Tbr1**: Marker for Cortical layer VI neurons
Differences in gene expression during early neural development between rhesus and human

- Neural genes are defined as genes having 5 fold higher expression after neural differentiation compared to their expression in embryonic stem cells.

- Between 160-300 genes are >2-fold differentially expressed between human and rhesus for each week of development.
All genes with a dynamic expression pattern during human and/or rhesus cortical neuron differentiation (-11,000)

Cluster 1
- Embryonic Stem Cell Genes
- Early Neuron Differentiation Genes
- Late Neuron Differentiation Genes
- Mature Cortical Neuron Genes

Cluster 2
- High during Human ESC neural differentiation
  - Functional annotation (Top 5 GO-term categories)
    - # genes
    - P-value
      - RNA processing: 73, 2.7E-8
      - Translation: 49, 1.8E-6
      - mRNA processing: 45, 3.2E-5
      - mRNA splicing: 41, 4.8E-5
      - mRNA metabolic process: 48, 7.1E-5

Cluster 5
- Prolonged expression in human neurospheres
  - Functional annotation (Top 5 GO-term categories)
    - # genes
    - P-value
      - Regulation of cell development: 11, 1.8E-4
      - Regulation of Nervous System Development: 10, 5.1E-4
      - Regulation of Neurogenesis: 9, 8.7E-4
      - Regulation of Neuron Differentiation: 8, 1.1E-3
      - Negative Regulation of Cell Differentiation: 10, 1.2E-3
Genome-wide gene profiling by RNA-seq, ChIP-seq & DNaseI-seq

ESC Neurosphere

RNA-seq

OCT4

p300 ChIP-seq

DNaseI-seq

OCT4: An Embryonic Stem Cell-specific enhancer

Frank Jacobs
Genome-wide gene profiling by RNA-seq, ChIP-seq & DNaseI-seq

ESC Neurosphere RNA-seq

ESC_fj74 nh1 RNA-seq

TBR1

p300C hiP seq

ESC Neurosphere

ESC_D02_17 lg DNaseI-seq

ESC_D05_21 lg DNaseI-seq

ESC Neurosphere

TBR1: A Long range Cortical Neuron-specific enhancer

Frank Jacobs
HES5: differentially expressed & regulated
General differences observed

- Increased expression of genes involved in cell proliferation during early human neurodevelopment
- Genes associated with neural differentiation are delayed in human relative to rhesus, prolonging process
- Challenging to find specific substitutions and rearrangements that account for the differences
- Once we find them, using new technology we can make selective changes in the genomes of the cells in cell culture and study the effects
Mathematical Foundations for Comparative Genomics
One kind of graph unifies key data structures in comparative genomics

- A graph theoretic model that allows for the trinity of 
duplications, substitutions 
and rearrangements, 
generalising many 
parsimony problems

Sequence Graphs

Genome rearrangement theory

Multiple sequence alignment with insertions and deletions

Sequence graphs are a simple construction kit to describe genome variation.
Segments of DNA are attached in different ways in different genomes.

Variation exists even within a single genome representation, as represented in a De Bruijn graph (a kind of sequence graph).
Sequence graphs include both the breakpoint graph and bi-directed graph formalisms.

DNA-labeled arrows are sequences.

Colored lines are bonds.

A green and blue genome for green and blue genomes.

Break point graph.
History graphs add descent edges to sequence graphs

Colored arrows are DNA sequences

Horizontal black lines are bonds

Lightning bolts are substitutions

Dotted lines are descent edges
Stochastic Models of Genome Evolution: the Jukes-Cantor model of base substitution

The state transition diagram for the Jukes-Cantor model of DNA evolution:

\[ R = \begin{pmatrix} -3r & r & r & r \\ r & -3r & r & r \\ r & r & -3r & r \\ r & r & r & -3r \end{pmatrix} \]

The probabilities of specific substitutions in time \( t \)

\[ P^t = e^{Rt} = I + Rt + \frac{(Rt)^2}{2} + \frac{(Rt)^3}{6} + \ldots \]
The spectral decomposition of the rate matrix is

\[ R = \beta_0 E_0 + \beta_1 E_1 + \cdots + \beta_{N-1} E_{N-1} \]

where the betas are the eigenvalues and \( E_0, \ldots, E_{N-1} \) are mutually orthogonal projection matrices. The probabilities of specific state changes in time \( t \) are given by the matrix

\[ P^t = e^{Rt} = I + Rt + \frac{(Rt)^2}{2} + \frac{(Rt)^3}{6} + \ldots \]

\[ = \sum_{d=0}^{N-1} E_d + \sum_{d=0}^{N-1} t\beta_d E_d + \sum_{d=0}^{N-1} \frac{(t\beta_d)^2}{2} E_d + \sum_{d=0}^{N-1} \frac{(t\beta_d)^3}{6} E_d + \ldots \]

\[ = e^{t\beta_0} E_0 + e^{t\beta_1} E_1 + \cdots + e^{t\beta_{N-1}} E_{N-1} \]
For Jukes-Cantor, the eigenvalues are 0 and -4r, and the (integer-valued !) projection matrices are

\[
E_0 = \frac{1}{4} \begin{pmatrix}
1 & 1 & 1 & 1 \\
1 & 1 & 1 & 1 \\
1 & 1 & 1 & 1 \\
1 & 1 & 1 & 1
\end{pmatrix}, \quad E_1 = \frac{1}{4} \begin{pmatrix}
3 & -1 & -1 & -1 \\
-1 & 3 & -1 & -1 \\
-1 & -1 & 3 & -1 \\
-1 & -1 & -1 & 3
\end{pmatrix}
\]

Plugging these into the general formula we get

\[
P^t = E_0 + e^{-4rt} E_1
\]

\[
= \frac{1}{4} \begin{pmatrix}
1 + 3e^{-4rt} & 1 - e^{-4rt} & 1 - e^{-4rt} & 1 - e^{-4rt} \\
1 - e^{-4rt} & 1 + 3e^{-4rt} & 1 - e^{-4rt} & 1 - e^{-4rt} \\
1 - e^{-4rt} & 1 - e^{-4rt} & 1 + 3e^{-4rt} & 1 - e^{-4rt} \\
1 - e^{-4rt} & 1 - e^{-4rt} & 1 - e^{-4rt} & 1 + 3e^{-4rt}
\end{pmatrix}
\]
Whole genomes change by 2-break rearrangements

State space of all genome configurations for 2 genes

Here we restrict to circular chromosomes
For this case of 2-gene genomes, the rate matrix for 2-break rearrangements is

\[
R = \begin{pmatrix}
-2r & r & r \\
r & -2r & r \\
r & r & -2r
\end{pmatrix}
\]

The spectral decomposition has integer-valued projection matrices like the Jukes-Cantor model, and gives

\[
P_2^t = e^{Rt} = \frac{1}{3} \begin{pmatrix}
1 + 2e^{-3rt} & 1 - e^{-3rt} & 1 - e^{-3rt} \\
1 - e^{-3rt} & 1 + 2e^{-3rt} & 1 - e^{-3rt} \\
1 - e^{-3rt} & 1 - e^{-3rt} & 1 + 2e^{-3rt}
\end{pmatrix}
\]
For 3-gene genomes, there are 15 states

3 types of transitions: 0, 1 and 2 ops
For \( n \)-gene genomes, there are \((2n-1)(2n-3) \ldots (1)\) states. The general model of evolution of \( n \)-gene genomes by 2-break rearrangements is a random processes on matchings, explored in many areas:

1. Diaconis and Holmes (mixing times),
2. Saxl (group representation theory),
3. MacDonald and James (symmetric functions and zonal polynomials),
4. Chillag (generalized circulants),
5. Saw and Takemura (multivariate statistics, Wishart distributions),
6. Godsil (association schemes),
7. Krieg, Bump (Hecke algebras),
8. Thrall (Lie groups).
A **homogeneous space** is a set $X$ (e.g. the state space of a Markov process) and a group $G$ that acts on $X$. When states are matchings on $\{1,2, \ldots, 2n\}$ (i.e. $n$-gene genomes), $G$ is naturally the group $S_{2n}$ of permutations of $\{1,2, \ldots, 2n\}$. For a permutation $\pi$ and state

$$x = \{\{i_1, i_2\}, \ldots, \{i_{2n-1}, i_{2n}\}\}$$

the action of $\pi$ changes $x$ to

$$\pi x = \{\{\pi(i_1), \pi(i_2)\}, \ldots, \{\pi(i_{2n-1}), \pi(i_{2n})\}\}$$
random walk on $X$ by action of group $G$
Let the state $x_0$ be an arbitrary origin. The **stabilizer subgroup** $H = H_n$ is the subgroup of actions in $G$ that leave $x_0$ fixed. For matchings, $H$ is the hyperoctahedral group of symmetries of the $n$-cube. States in $X$ are cosets of $G = S_{2n}$ w.r.t. $H$.

We write $X = G/H$. This is why

$$|X| = \frac{|S_{2n}|}{|H_n|} = \frac{(2n)!}{n!2^n} = (2n - 1)(2n - 3) \cdots (1)$$
In homogeneous space $X = G/H$, the group $G$ acts naturally on pairs of states

$$\pi(x, y) = (\pi x, \pi y)$$

The orbital of $(x, y)$ is \[ \{(\pi x, \pi y) : \pi \in G\} \]

All state pairs in the same orbital are said to have the same difference. Thus, each orbital defines a difference in a difference set $D$. In the case of the discrete Fourier space,

$$D = \{- (n-1), - (n-2), \ldots, -1, 0, 1, \ldots, n-1\}.$$
The difference between two \( n \)-gene genomes is a partition of the integer \( n \). So \( D = \) set of partitions of \( n \).

For example, if \( n = 3 \), then \( D = \{(1,1,1),(2,1),(3)\} \).
In a **symmetric random walk** on $X$ the probability is the same for all transitions with the same difference. The dynamics are defined by a function on the difference set $D$. The theory can be generalized to all complex functions on $D$. We call these **radial functions**. A radial function on $D$ induces a unique function on $X$ and $G$.

For radial functions $f$ and $g$, here viewed as functions on the group $G$, we define their **convolution** as

$$(f \ast g)(\gamma) = \sum_{(\alpha, \beta): \gamma = \alpha \beta} f(\alpha)g(\beta)$$

This becomes the usual notion of convolving the effect of one random action followed by another when $f$ and $g$ are probability distributions.)
A homogeneous space $X = G/H$ is a **Gelfand space** if convolution of radial functions is commutative, i.e.

$$f * g = g * f$$

In this case $(G,H)$ is said to be a **Gelfand pair**. (Same Israel Gelfand that Bernard quoted.)

The Jukes-Cantor space, the discrete Fourier space $\{0, \ldots, n-1\}$, and the space of $n$-gene genomes are all Gelfand spaces.
The **Fourier Transform** is a linear mapping that $\Phi$ converts convolution into multiplication.

Think of a radial function as a $|D|$-dimensional vector. Then the Fourier transform $\Phi$ is defined by a matrix whose rows are a special orthogonal set of radial functions $\{\phi_d : d \in D\}$ called **normalized spherical functions**. The Fourier transform is written

$$\hat{f} = \Phi \overline{f}$$

where $\hat{f}$ is the Fourier transform of $f$ and $\overline{f}$ is the complex conjugate of $f$. For the Fourier state space

$$\phi_d(k) = e^{i2\pi kd/n}$$
We say that the Fourier transform converts convolution into multiplication because for any radial functions $f$ and $g$,

$$f * g = \sum_{d \in D} \hat{f}_d \hat{g}_d \phi_d$$

Gelfand spaces are precisely the homogenous spaces where there is a well-defined Fourier transform of the simple type we have described. There are only a few infinite families of discrete Gelfand pairs on the permutation group, so we are lucky to get one for genome rearrangements.
The spectral decomposition is associated with the inverse Fourier transform

\[ f = \sum_{d \in D} \hat{f}_d \phi_d \]

The radial functions \( f \) and \( \{\phi_d : d \in D\} \) are represented as matrices, and the Fourier coefficients \( \hat{f}_d \) play the role of eigenvalues.
As an example, for the Jukes Cantor case, as $|D|$-dimensional vectors (functions on $D$), the normalized spherical functions are $(1,1)^T$ and $(3, -1)^T$. Equivalently, these can be represented by $|X|$-by-$|X|$ matrices, which turn out to be the projection matrices in the spectral decomposition.

$$E_0 = \frac{1}{4} \begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \end{pmatrix} \quad E_1 = \frac{1}{4} \begin{pmatrix} 3 & -1 & -1 & -1 & -1 \\ -1 & 3 & -1 & -1 & -1 \\ -1 & -1 & 3 & -1 & -1 \\ -1 & -1 & -1 & 3 & -1 \\ -1 & -1 & -1 & -1 & 3 \end{pmatrix}$$
Because of the conversion of convolution to multiplication, if you convolve $f$ with itself $i$ times, you get

$$f \ast f \ast \cdots \ast f = \sum_{d \in D} \hat{f}^i_d \phi_d$$

By Taylor expansion you can get any analytical function of convolution powers, e.g. an exponential.
Thus, if $f$ is taken from a radial rate matrix $R$ (i.e. rate depending only on differences in $D$) and $t$ is any amount of time, the matrix of probabilities of state changes over various differences is

$$P^t = \sum_{d \in D} e^{t \hat{f}_d} \phi_d$$

This generalizes the spectral decomposition method for Jukes-Cantor to a broad set of state spaces.
The Fourier transform for a general Gelfand space can be expressed as a matrix whose columns are the unnormalized spherical functions. For example, for the Jukes Cantor case, the normalized spherical functions are \((1,1)^T\) and \((3, -1)^T\) so the Fourier transform matrix is

$$
\Phi = \begin{pmatrix} 1 & 3 \\ 1 & -1 \end{pmatrix}
$$

Wonderful thing: for a Gelfand space in which the difference is symmetric, all the coefficients of the Fourier transform are integers.
For the case of n-gene genomes (matchings), the Fourier transform has an integer-valued matrix indexed by the partitions of $n$. The first few transform matrices are:

$n = 2$
\[ \Phi = \begin{pmatrix} 1 & 2 \\ 1 & -1 \end{pmatrix} \]

$n = 3$
\[ \Phi = \begin{pmatrix} 1 & 6 & 8 \\ 1 & 1 & -2 \\ 1 & -3 & 2 \end{pmatrix} \]

$n = 4$
\[ \Phi = \begin{pmatrix} 1 & 12 & 12 & 32 & 48 \\ 1 & 5 & -2 & 4 & -8 \\ 1 & 2 & 7 & -8 & -2 \\ 1 & -1 & -2 & -2 & 4 \\ 1 & -6 & 3 & 8 & -6 \end{pmatrix} \]
There is no known computationally tractable closed-form formula for the integers in the Fourier transform matrix for matchings.

Nevertheless, genome evolution by 2-break rearrangements is a special case of an extensive and beautiful theory (symmetric Gelfand spaces)

Including duplications, gains and losses complicates the model considerably
Comparative Genomics in Cancer
In cancers driven by a single mutation, like BRAF V600 in metastatic melanoma, targeted drugs can give spectacular results.

![Before initiation of vemurafenib](image1)

![15 weeks on vemurafenib](image2)

Roche
But combination or immunotherapies will be required to prevent relapse, just as in the treatment of HIV AIDS.
Some motivations for large-scale application of comparative genomics in cancer

- Bring data to research and insights to clinical practice
- Learn to link phenotypes, including clinical outcomes, to underlying molecular aberrations
- Create the infrastructure to select patient populations for targeted clinical trials, and to enable a new kind of global rapid learning cycle that complements targeted trials
- Gain a mechanistic, molecular level understanding of the etiology of disease and mechanisms of resistance to treatment

All these require statistical power
Genomes are the key to the future of cancer treatment
The Cancer Genome Atlas: 10,000 tumors from 20 adult cancers

TCGA Sequencing Centers

The Broad Institute

BC

Wash U

Harvard

USC

UNC

Baylor
CANCER GENOMICS HUB

- Total Cost ~ $100/year/genome at 50K genomes
- Houses genomes from all major NCI projects
- Planned 5 PB, Scalable to 20 PB

- FISMA compliant
- 1st NIH Trusted Partner
- COTS hardware
- High availability
- CentOS, standard Linux tools
- General Parallel Filesystem
- Dual RAID 6
- Co-location opportunities

CGHub at San Diego Supercomputer Center
Current Stats

716,000 total files downloaded

10,462 TB transferred

495 TB data
43,000 files

2-4 Gb/s typical downloads in aggregate outbound from CGHub
Future Requires Global Network of Hubs
Different Requirements for 1M Genomes

- Different types of data interactions:
  - Support both research and clinical practice
  - Compute within a provided cloud
  - Separately URIed, metadata-tagged parts of a single patient file supporting 3rd party mashups and tools

- Harmonized portable consents, sample donor has fine-grained control of who can access their data parts, trusts the security provided

- APIs, not file formats. 3rd parties must be able to build on it: goal to enable research and clinical analysis, not usurp it

- Benchmarking so all can use system to improve methods, e.g. variant calling
Possible Genome Commons Architecture

**Read Layer**
- Genomic sequence data

**Variation Layer**
- Sequence database (BAMs) ~100 petabytes
  - Variant calling
- Variation database (VCF) ~1 petabyte
  - Variation Analysis

**Interpretation Layer**
- Sequence graphs
  - Functional impact analysis
- Interpretation database (relational) ~1 terabyte
  - Queries

**Clinical data**
What would it cost to store and analyze 1M Cancer Genomes in 2014?

- Our estimate is ~ $50/genome/year in 2014 to store and analyze 1M whole genomes (~ 100 petabytes, 2 months of YouTube growth)
  - 25,000 disks and 100,000 processor cores
  - Including operating costs: space, electricity, operators
  - Including 2\textsuperscript{nd} center to protect against disasters
- Note that cancer is the high water mark for global genome commons requirements, requirements for other diseases are smaller, less complex, assuming cancer includes full germline and somatic cell analysis.

Dave Patterson, www.eecs.berkeley.edu/Pubs/TechRpts/2012/EECS-2012-211.html
Extracting molecular state from raw DNA reads

chr2 : 29,064,107

OV-0751 Somatic Reads

chr2 : 28,500,054

Tandem Duplication Size = 564,053 bp

Zack Sanborn, now at Five3 Genomics
Completely solved problem? Not yet. Given the same raw sequence (BAM) files, different mutation calling pipelines do not completely agree.

**Point mutations called in tumor TCGA-13-0725**

<table>
<thead>
<tr>
<th></th>
<th>Called by 2 other centers</th>
<th>Called by at least 1 other</th>
</tr>
</thead>
<tbody>
<tr>
<td>Broad</td>
<td>62%</td>
<td>85%</td>
</tr>
<tr>
<td>UCSC</td>
<td>74%</td>
<td>89%</td>
</tr>
<tr>
<td>WUSTL</td>
<td>63%</td>
<td>82%</td>
</tr>
</tbody>
</table>

Still work to do to harden mutation-calling software, even for point mutations.

UCSC, Broad are leading a series of TCGA/ICGC international benchmark challenges. Visit cghub.ucsc.edu for TCGA Benchmark 4.
Even more differences in calling structural changes

- 2 Glioblastoma samples. Circle plot shows amplifications, deletions, inter/intra chromosomal rearrangement

- These 2 samples have 23/25 top Broad, 21/29 top UCSC events
In 11/16 WGS TCGA glioblastoma cases similar events lead to homozygous loss of CDKN2A/B

<table>
<thead>
<tr>
<th>One Copy Deleted by</th>
<th>Other Copy Deleted by</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>5 GBMs</strong></td>
<td>Arm-Level loss of chr9p (via inter-chrom translocation)</td>
</tr>
<tr>
<td>Focal Loss</td>
<td>Arm-Level loss of chr9p (mechanism unknown)</td>
</tr>
<tr>
<td><strong>3 GBMs</strong></td>
<td>Complete loss of chr9</td>
</tr>
<tr>
<td>Focal Loss</td>
<td>Complex event</td>
</tr>
<tr>
<td><strong>2 GBMs</strong></td>
<td></td>
</tr>
<tr>
<td>Focal Loss</td>
<td></td>
</tr>
<tr>
<td><strong>1 GBM</strong></td>
<td></td>
</tr>
<tr>
<td>Focal Loss</td>
<td></td>
</tr>
<tr>
<td><strong>5 GBMs</strong></td>
<td>No loss detected</td>
</tr>
<tr>
<td>No loss detected</td>
<td>No loss detected</td>
</tr>
</tbody>
</table>
• **Chromothripsis**: DNA replication process get confused for a period or DNA is shattered into pieces by some high energy event when chromosome is in condensed state

• DNA repair mechanisms try to stitch genome back together

• Can generate rearrangements, losses, and circular “double minute” chromosomes
DM from another GBM tumor. We estimate 20% of GBMs have oncogenic DMs.

Validation by FISH

Zack Sanborn, Cameron Brennan
Highlights from analysis of 500 GBMs

FIGURE LEGEND

DNA Copy

Focal del

Broad del

Broad amp

Focal amp

Wildtype

SNV

Mutated

Structural Variant

Variant

TCGA GBM Analysis Working Group
Tumors have metagenomes: mixture of clones resulting from somatic selection of subclones

One can use sequence graphs for analysis of cancer metagenomes
Algebraic/Combinatorial Approach to Comparative Metagenomics

Flows:

Alternating and simple flows:
Duplication – raw data

Detected Breakend

Primary Copy-Number Signal
Duplication – model from data

Single duplication event (Copy number change + Breakend)

Red = creation/duplication
Deletion – raw data

(No breakend detected)
Deletion – model from data

Single deletion event

Suggested novel breakend creation

Blue = removal/deletion
Finally, key is interpretation of genomics data at the pathway level. Curated and/or Collected: Reactome, KEGG, Biocarta, NCI-PID, Pathway Commons. TCGA Glioblastoma Analysis.
The Age of Opportunity for the Study of Genetics and Medicine

- **#1 infrastructure issue** is to achieve statistical power by aggregating information. We must head off the development of genomic information silos

- **#1 interpretive challenge** is to accurately read a genome and model effects of genetic changes on molecular pathways and phenotypes

- **We must accelerate biomedical research and improve clinical practice by building new global platforms for storage, exchange and analysis of molecular and phenotypic information**
Some Current Collaborators

Collaborators

- Dave Patterson group, UC Berkeley
- David Altshuler, Charles Sawyers, Mike Stratton, Betsy Nabel, Brad Margus, Karen Kennedy, Tom Hudson
- Richard Durbin, Sanger Centre
- Broad Institute, Wash U., Baylor
- The Cancer Genome Atlas and its labs, esp. GBM analysis working group
- Stand Up To Cancer and its labs
- Intl. Cancer Genome Consortium and its labs
- Chris Benz, Buck Institute
- Laura Van’t Veer, Laura Esserman, Joe Costello, Eric Collisson, Margaret Tempero, UCSF
- UCSC Storage Systems Group
- Joe Gray, Paul Spellman, OHSU