needles in stacks of needles


martin krzywinski
http://mkweb.bcgsc.ca

canada’s michael smith genome sciences centre
bc cancer research center

canada vancouver
in each of our $\sim 10^{13}$ cells

is a complete genome of $3 \cdot 10^9$ base pairs
changing any of the bases
in any of the cells
can lead to disease
genome alterations are like spelling mistakes
genome alterations are like spelling mistakes

our biology is robust against many changes
genome alterations are like spelling mistakes

our biology is robust against many changes

but if we accumulate too many of them
genome alterations are like spelling mistakes

our biology is robust against many changes

but if we accumulate too many of them

our abilities to adapt and recover will be
genome alterations are like spelling mistakes

our biology is robust against many changes

but if we accumulate too many of them

our abilities to adapt and repair will be
THE CHALLENGE

to understand
the genetic basis of disease

to create
better diagnostics and therapies

to improve
patients’ outcomes and quality of life
GENETIC INSTABILITY IS A DRIVER FOR DIVERSITY IN CANCER

Figure 1 adapted from Yates et al. Nature Reviews Genetics 13:795 (2012).
efficient algorithms

graphs and networks

clustering

text mining

visualization
efficient algorithms
FIND DIFFERENCES IN GENOMES

graphs and networks
ASSEMBLE GENOME SEQUENCE

clustering
FIND PATTERNS IN GENE EXPRESSION

text mining
DISCOVER BIOLOGICAL RELATIONSHIPS

visualization
DNA Blueprint for Fetus Built Using Tests of Parents

By ANDREW POLLACK
Published: June 6, 2012 | 252 Comments

For the first time, researchers have determined virtually the entire genome of a fetus using only a blood sample from the pregnant woman and a saliva specimen from the father.

The accomplishment heralds an era in which parents might find it easier to know the complete DNA blueprint of a child months before it is born.

That would allow thousands of genetic diseases to be detected prenatally. But
DNA is not a blueprint
THIS IS DNA

taaccctaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaaccctaaa}

sequence at start of human chromosome 1
DNA DOES NOT DIRECTLY DESCRIBE THE ORGANISM
life is the emergent property of biochemical reactions

\[
A \rightarrow B
\]
DNA encodes the enzymes that catalyze these reactions

enzyme

A $\rightarrow$ B
there are millions of reactions
IF A HOUSE HAD DNA...
...A LIST OF TOOLS THAT MAKE THE TOOLS TO MAKE THE HOUSE
DNA CHANGES ARE HARD TO DECIPHER FUNCTIONALLY
molecular cellular mechanisms are profoundly interconnected with many multi-function components
ICDM 2012 Brussels

Micro RNA

Expression regulation

Gene ~20,000

Exon ~9 exons/gene

DNA 3x10^9 base pairs

RNA

Increased expression

Alternate transcripts

~6 transcripts/gene

Protein

Reaction

C → C'

4x

A → A'

B → B'
EPIGENOME — “REST OF THE GENOME”

individual genomes
many types of structural variations are possible

calcul
difficult to assess

individual genomes

≈3 million differences
efficient algorithms
FIND DIFFERENCES IN GENOMES

graphs and networks
ASSEMBLE GENOME SEQUENCE

clustering
FIND PATTERNS IN GENE EXPRESSION

text mining
DISCOVER BIOLOGICAL RELATIONSHIPS

visualization
we learn about the genome by sequencing it

and aligning it to the human reference sequence

align 8 billion 35 base reads to 3 billion base reference

~100-fold coverage

SPACED SEEDS (MAQ)  BURROW-WHEELER (BOWTIE)

Reference genome (> 3 gigabases)

Short read

Extract seeds

Position N

Position 2

CTGC CGTA AACT AATG

Position 1

ACTG CCGT AAAC TAAT

ACTG **** AAAC ****

**** CCGT **** TAAT

**** **** AAAC TAAT

ACTG CCGT **** ****

**** CCGT AAAC ****

Index seed pairs

Six seed pairs per read/fragment

Seed index (tens of gigabytes)

ACTG **** AAAC ****

**** CCGT **** TAAT

ACTG **** **** TAAT

**** CCGT AAAC ****

Look up each pair of seeds in index

Hits identify positions in genome where spaced seed pair is found

Confirm hits by checking **** positions

Reference genome (> 3 gigabases)

Short read

Concatenate into single string

Burrows-Wheeler transform and indexing

Bowtie index (~2 gigabytes)

Look up 'suffixes' of read

Hits identify positions in genome where read is found

Convert each hit back to genome location

Figure 1 from Trapnell et al. Nature Biotechnology 27:455 (2009).
LIMITATIONS

for efficiency, number of mismatches is limited
e.g. 2 for BWT aligner Bowtie

BWT PERFORMANCE

50x faster than spaced seeds methods

25 million 35 base reads per hour per CPU (2009)
4 hours to align per 1X coverage
1.3 Gb memory footprint

WE ALWAYS FIND GENOME CHANGES

Overlap of observed changes between AML tumor genome and other reference genomes. Millions of single base changes (SNPs).

false positives · natural variation
passenger mutations · driver mutations

Adapted from Fig 1 in Ley et al. DNA sequencing of a cytogenetically normal acute myeloid leukaemia genome. Nature 456:66 (2008).
DECISION TREES HELP CLASSIFY SNPS

3,813,205 tumour SNVs (Maq15)

- 2,647,695 well supported SNVs (decision tree)

- 2,584,418 present in skin (SNPs)

- 63,277 tumour-specific SNVs

- 31,632 new SNVs

- 20,440 in non-genic regions

- 11,192 SNVs in genic regions

- 10,735 intronic

- 241 SNVs in coding sequence

- 181 SNVs predicted to alter gene function (non-synonymous and splice junctions)

- 14 validated as germline SNVs (SNPs)

- 8 validated as somatic SNVs (acquired mutations)

- 152 validated as wild type (false positives)

- 60 synonymous

- 31,645 in dbSNP/Watson/Venter

- 216 in UTR

- 7 unable to be validated (technical failures)

# You Can’t Publish a Single Genome Anymore

## Table 2. Whole-Genome and Whole-Exome Sequencing Statistics

<table>
<thead>
<tr>
<th>Statistic</th>
<th>Whole-Exome Capture</th>
<th>Whole Genome</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tumor/normal pairs sequenced</td>
<td>159</td>
<td>24</td>
</tr>
<tr>
<td>Total tumor Gb sequenced</td>
<td>1,031.6</td>
<td>4,946.0</td>
</tr>
<tr>
<td>Median fold tumor target coverage (range)</td>
<td>91 (51–201)</td>
<td>69 (25–103)</td>
</tr>
<tr>
<td>Median normal fold target coverage (range)</td>
<td>92 (62–141)</td>
<td>36 (28–55)</td>
</tr>
<tr>
<td>Median somatic mutation rate per Mb in target territory (range)</td>
<td>6.8 (0.3–94.7)</td>
<td>13.3 (4.5–55.3)</td>
</tr>
<tr>
<td>Median number of coding mutations per patient (range)</td>
<td>216 (1–3,512)</td>
<td>323 (63–2,279)</td>
</tr>
<tr>
<td>Median number of nonsynonymous mutations per patient (range)</td>
<td>167 (1–2,721)</td>
<td>248 (53–1,770)</td>
</tr>
<tr>
<td>Median number of transcribed noncoding mutations per patient (range)</td>
<td>187 (13–2,559)</td>
<td>18,314 (4,632–100,707)</td>
</tr>
<tr>
<td>Total number of structural rearrangements</td>
<td>n/a</td>
<td>2,349</td>
</tr>
<tr>
<td>Total number of frame-preserving genic rearrangements</td>
<td>n/a</td>
<td>71</td>
</tr>
<tr>
<td>Total number of frame-abolishing genic rearrangements</td>
<td></td>
<td>235</td>
</tr>
<tr>
<td>Median number of genes powered at 20% exonic territory (range)</td>
<td>15,647 (15,046–16,019)</td>
<td>16,905 (10,136–16,952)</td>
</tr>
<tr>
<td>Median number of genes powered at 50% exonic territory (range)</td>
<td>6,788 (6,078–7,402)</td>
<td>8,771 (2,634–8,863)</td>
</tr>
</tbody>
</table>

Selected sequencing statistics for 183 WES and WGS cases. "Tumor Target Territory" refers to the exonic territory targeted by the exome capture bait set reported by (Fisher et al., 2011) and used in this study. The "Whole-Exome Capture" column does not include data on 23 cases analyzed by both WES and WGS.

---

VARIETY OF GENE MUTATION PROFILES ACROSS SAMPLES

VARIETY OF MUTATIONS ACROSS GENES

**U2AF1**

- S34F
- G217I

**RBM10**

- E4K
- R6H
- E67*
- R157fs
- Y206*
- R230*
- I316F
- Y580F
- Y596_splice
- R885L
- E810*
- V846_splice

**ARID1A**

- S446fs
- Q479*
- Q481*
- Q501*
- P946fs
- W1545*
- E1726*
- P526L
- E1777K
- E1779K
- G1848W
- N956Y
- W1023L
- S1465C
- G1996D
- E2000V
- G2283D

**Gene Domains**

- Zinc fingers
- RRM domain
- RS domain
- G-patch
- LXXLL motif
- ARID domain

oncoprint

http://www.cbioportal.org/public-portal/faq.jsp
YOU CAN'T PUBLISH 100 GENOMES ANYMORE: 1000 GENOMES PROJECT

Summary of 1000 Genomes Project phase I data

<table>
<thead>
<tr>
<th>Feature</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Autosomes Samples</td>
<td>1,092</td>
</tr>
<tr>
<td>Total raw bases (Gb)</td>
<td>19,049</td>
</tr>
<tr>
<td>Mean mapped depth</td>
<td>5.1</td>
</tr>
<tr>
<td>No. sites overall</td>
<td>36.7 M</td>
</tr>
<tr>
<td>Novelty rate *</td>
<td>58%</td>
</tr>
<tr>
<td>No. synonymous/non-synonymous/nonsense</td>
<td>NA</td>
</tr>
<tr>
<td>Average no. SNPs per sample</td>
<td>3.60 M</td>
</tr>
<tr>
<td>Indels</td>
<td>1.38 M</td>
</tr>
<tr>
<td>Novelty rate *</td>
<td>62%</td>
</tr>
<tr>
<td>No. inframe/frameshift</td>
<td>NA</td>
</tr>
<tr>
<td>Average no. indels per sample</td>
<td>344 K</td>
</tr>
<tr>
<td>Genotyped large deletions</td>
<td>13.8 K</td>
</tr>
<tr>
<td>Novelty rate *</td>
<td>54%</td>
</tr>
<tr>
<td>Average no. variants per sample</td>
<td>717</td>
</tr>
</tbody>
</table>

* Compared with dbSNP release 135 (Oct 2011), excluding contribution from phase I 1000 Genomes Project (or equivalent data for large deletions).
efficient algorithms
FIND DIFFERENCES IN GENOMES

graphs and networks
ASSEMBLE GENOME SEQUENCE

clustering
FIND PATTERNS IN GENE EXPRESSION

text mining
DISCOVER BIOLOGICAL RELATIONSHIPS

visualization
DE NOVO ASSEMBLY

the first human genome was assembled in 2001

it is now common to assemble genomes *de novo* (from their reads)

**Figure 1** Bridges of Königsberg problem. (a) A map of old Königsberg, in which each area of the city is labeled with a different color point. (b) The Königsberg Bridge graph, formed by representing each of four land areas as a node and each of the city’s seven bridges as an edge.

Box 1 Origin of de Bruijn graphs

In 1946, the Dutch mathematician Nicolaas de Bruijn became interested in the ‘superstring problem’\textsuperscript{12}: find a shortest circular ‘superstring’ that contains all possible ‘substrings’ of length \( k \) \((k\text{-mers})\) over a given alphabet. There exist \( n^k \) \( k\text{-mers} \) in an alphabet containing \( n \) symbols: for example, given the alphabet comprising \( A, T, G \) and \( C \), there are \( 4^3 = 64 \) trinucleotides. If our alphabet is instead 0 and 1, then all possible 3-mers are simply given by all eight 3-digit binary numbers: 000, 001, 010, 011, 100, 101, 110, 111. The circular superstring 0001110100 not only contains all 3-mers but also is as short as possible, as it contains each 3-mer exactly once. But how can one construct such a superstring for all \( k \)-mers in the case of an arbitrary value of \( k \) and an arbitrary alphabet? De Bruijn answered this question by borrowing Euler’s solution of the Bridges of Königsberg problem. Briefly, construct a graph \( B \) (the original graph called a de Bruijn graph) for which every possible \((k - 1)\text{-mer}\) is assigned to a node; connect one \((k - 1)\text{-mer}\) by a directed edge to a second \((k - 1)\text{-mer}\) if there is some \( k\)-mer whose prefix is the former and whose suffix is the latter (Fig. 2). Edges of the de Bruijn graph represent all possible \( k\)-mers, and thus an Eulerian cycle in \( B \) represents a shortest (cyclic) superstring that contains each \( k\)-mer exactly once. By checking that the indegree and outdegree of every node in \( B \) equals the size of the alphabet, we can verify that \( B \) contains an Eulerian cycle. In turn, we can construct an Eulerian cycle using Euler’s algorithm, therefore solving the superstring problem. It should now be apparent why the ‘de Bruijn graph’ construction described in the main text, which does not use all possible \( k\)-mers as edges but rather only those generated from our reads, is also named in honor of de Bruijn.

Figure 2  De Bruijn graph. The de Bruijn graph \( B \) for \( k = 4 \) and a two-character alphabet composed of the digits 0 and 1. This graph has an Eulerian cycle because each node has indegree and outdegree equal to 2. Following the blue numbered edges in order from 1 to 16 traces an Eulerian cycle 0000, 0001, 0010, 0110, 1100, 1001, 0010, 0101, 1011, 0111, 1111, 1110, 1101, 1010, 0100, 1000. Recording the first character (in boldface) of each edge label spells the cyclic superstring 000110010111101.
DE BRUIJN GRAPHS FOR ASSEMBLY

Vertex: k-mers
Edge: pairwise alignments

Hamiltonian cycle
Visit each vertex once (harder to solve)

Eulerian cycle
Visit each edge once (easier to solve)

Genome: ATGGCGTGCA ATGGCGT

ERROR CORRECTION — MANY ANSWERS ARE POSSIBLE

Courtesy of Shaun Jackman and the ABySS (short read assembler) team.
EXPLORING GENOME ASSEMBLIES — ABYSS EXPLORER

clustering

FIND PATTERNS IN GENE EXPRESSION
HIERARCHICAL CLUSTERING

CLUSTERING METHODS

k-means & SOM better than hierarchical
complete better than single linkage
Euclidian distance for log ratio data
Pearson correlation for absolute data

CORRELATION IS THE NEW CAUSATION

Most Random Gene Expression Signatures Are Significantly Associated with Breast Cancer Outcome

David Venet¹, Jacques E. Dumont², Vincent Detours²,³*

¹ IRIDIA-CoDE, Université Libre de Bruxelles (ULB), Brussels, Belgium, ² IRIBHM, Université Libre de Bruxelles (ULB), Campus Erasme, Brussels, Belgium, ³ WELBIO, Université Libre de Bruxelles (ULB), Campus Erasme, Brussels, Belgium

Abstract
Bridging the gap between animal or in vitro models and human disease is essential in medical research. Researchers often suggest that a biological mechanism is relevant to human cancer from the statistical association of a gene expression marker (a signature) of this mechanism, that was discovered in an experimental system, with disease outcome in humans. We examined this argument for breast cancer. Surprisingly, we found that gene expression signatures—unrelated to cancer—of the effect of postprandial laughter, of mice social defeat and of skin fibroblast localization were all significantly associated with breast cancer outcome. We next compared 47 published breast cancer outcome signatures to signatures made of random genes. Twenty-eight of them (60%) were not significantly better outcome predictors than random signatures of identical size and 11 (23%) were worst predictors than the median random signature. More than 90% of random signatures >100 genes were significant outcome predictors. We next derived a metagene, called meta-PCNA, by selecting the 1% genes most positively correlated with proliferation marker PCNA in a compendium of normal tissues expression. Adjusting breast cancer expression data for meta-PCNA abrogated almost entirely the outcome association of published and random signatures. We also found that, in the absence of adjustment, the hazard ratio of outcome association of a signature strongly correlated with meta-PCNA (R² = 0.9). This relation also applied to single-gene expression markers. Moreover, >50% of the breast cancer transcriptome was correlated with meta-PCNA. A corollary was that purging cell cycle genes out of a signature failed to rule out the confounding effect of proliferation. Hence, it is questionable to suggest that a mechanism is relevant to human breast cancer from the finding that a gene expression marker for this mechanism predicts human breast cancer outcome, because most markers do. The methods we present help to overcome this problem.

Editor: Isidore Rigoutsos, Jefferson Medical College/Thomas Jefferson University, United States of America
Received April 27, 2011; Accepted September 7, 2011; Published October 20, 2011
Copyright: © 2011 Venet et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.
Funding: DV was funded by the IRSIB Brussels Region-Capitale ICT-Impulse 2006 program 'IntSilico wet lab'. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.
Competing Interests: The authors have declared that no competing interests exist.
* E-mail: vdetours@ulb.ac.be
CORRELATION IS THE NEW CAUSATION

Most Random Gene Expression Signatures Are Significantly Associated with Breast Cancer Outcome

David Venet¹, Jacques E. Dumont², Vincent Detours²,³*

¹IRIDIA-CoDE, Université Libre de Bruxelles (ULB), Brussels, Belgium, ²IRIBHM, Université Libre de Bruxelles (ULB), Campus Erasme, Brussels, Belgium, ³WELBIO, Université Libre de Bruxelles (ULB), Campus Erasme, Brussels, Belgium

Abstract

Bridging the gap between animal or in vitro models and human disease is essential in medical research. Researchers often suggest that a biological mechanism is relevant to human cancer from the statistical association of a gene expression marker (a signature) of this mechanism, that was discovered in an experimental system, with disease outcome in humans. We examined this argument for breast cancer. Surprisingly, we found that gene expression signatures—unrelated to cancer—of the effect of postprandial laughter, of mice social defeat and of skin fibroblast localization were all significantly associated with breast cancer outcome. We next compared 47 published breast cancer outcome signatures to signatures made of random genes. Twenty-eight of them (60%) were not significantly better outcome predictors than random signatures of identical size and 11 (23%) were worst predictors than the median random signature. More than 90% of random signatures >100 genes were significant outcome predictors. We next derived a metagene, called meta-PCNA, by selecting the 1% genes most positively correlated with proliferation marker PCNA in a compendium of normal tissue expression. Adjusting breast cancer expression data for meta-PCNA abrogated almost entirely the outcome association of published and random signatures. We also found that, in the absence of adjustment, the hazard ratio of outcome association of a signature strongly correlated with meta-PCNA ($R^2 = 0.9$). This relation also applied to single-gene expression markers. Moreover, >50% of the breast cancer transcriptome was correlated with meta-PCNA. A corollary was that purging cell cycle genes out of a signature failed to rule out the confounding effect of proliferation. Hence, it is questionable to suggest that a mechanism is relevant to human breast cancer from the finding that a gene expression marker for this mechanism predicts human breast cancer outcome, because most markers do. The methods we present help to overcome this problem.

Editor: Isidore Rigoutsos, Jefferson Medical College/Thomas Jefferson University, United States of America
Received April 27, 2011; Accepted September 7, 2011; Published October 20, 2011
Copyright: © 2011 Venet et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.
Funding: DV was funded by the IRSB Brussels Region-Capitale ICT-Impulse 2006 program 'InSilico wet lab'. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.
Competing Interests: The authors have declared that no competing interests exist.
* E-mail: vdetours@ulb.ac.be

CORRELATION IS THE NEW CAUSATION

Most Random Gene Expression Signatures Are Significantly Associated with Breast Cancer Outcome

David Venet¹, Jacques E. Dumont², Vincent Detours²,³*

¹ IRIDIA-CoDE, Université Libre de Bruxelles (ULB), Brussels, Belgium, ² IRIBHM, Université Libre de Bruxelles (ULB), Campus Erasme, Brussels, Belgium, ³ WELBIO, Université Libre de Bruxelles (ULB), Campus Erasme, Brussels, Belgium

Abstract

Bridging the gap between animal or in vitro models and human disease is essential in medical research. Researchers often suggest that a biological mechanism is relevant to human cancer from the statistical association of a gene expression marker (a signature) of this mechanism, that was discovered in an experimental system, with disease outcome in humans. We examined this argument for breast cancer. Surprisingly, we found that gene expression signatures—unrelated to cancer—of the effect of postprandial laughter, of mice social defeat and of skin fibroblast localization were all significantly associated with breast cancer outcome. We next compared 47 published breast cancer outcome signatures to signatures made of random genes. Twenty-eight of them (60%) were not significantly better outcome predictors than random signatures of identical size and 11 (23%) were worst predictors than the median random signature. More than 90% of random signatures >100 genes were significant outcome predictors. We next derived a metagene, called meta-PCNA, by selecting the 1% genes most positively correlated with proliferation marker PCNA in a compendium of normal tissues expression. Adjusting breast cancer expression data for meta-PCNA abrogated almost entirely the outcome association of published and random signatures. We also found that, in the absence of adjustment, the hazard ratio of outcome association of a signature strongly correlated with meta-PCNA ($R^2 = 0.9$). This relation also applied to single-gene expression markers. Moreover, $>50\%$ of the breast cancer transcriptome was correlated with meta-PCNA. A corollary was that purging cell cycle genes out of a signature failed to rule out the confounding effect of proliferation. Hence, it is questionable to suggest that a mechanism is relevant to human breast cancer from the finding that a gene expression marker for this mechanism predicts human breast cancer outcome, because most markers do. The methods we present help to overcome this problem.

Surprisingly, we found that gene expression signatures—unrelated to cancer—of the effect of postprandial laughter, of mice social defeat and of skin fibroblast localization were all significantly associated with breast cancer outcome.
our genome is the result of a single on-going Monte Carlo simulation

EVOLUTION

general rules are elusive
data flood¹  tsunamis²  deluges³
surging oceans⁴
avalanches⁵  icebergs⁶
landslide⁷  earthquakes⁸
explosions⁹
The $1,000 genome, the $100,000 analysis?

Elaine R Mardis

Having recently attended the Personal Genomes meeting at Cold Spring Harbor Laboratories (I was an organizer this year), I was struck by the number of talks that described the use of whole-genome sequencing and analysis to reveal the genetic basis of disease in patients. These patients included a child with irritable bowel disease, a child with severe combined immunodeficiency, two siblings affected with Miller syndrome, and several with cancers of different types. Although each presenter emphasized the rapidity with which these data can now be obtained, it is clear that the acquisition of such information is but a small step toward its proper analysis and application. The next phase, as we all recognized, involves the development of cost-effective means by which these data can be analyzed and integrated into routine patient care. The most pressing need in this regard is for high-quality tools to assemble and interpret whole-genome sequences. The amount of data generated by these procedures is so large that it cannot be handled by human experts, and it is becoming evident that the tools currently available are insufficient for the task.

A second area of difficulty that became apparent is the need for standardization of these approaches. Currently, different laboratories use different strategies to map, assign, and sequence the human genome. It is clear that without standardization, the interpretation of data from different laboratories will be difficult, if not impossible. It is essential that the scientific community work together to develop consensus guidelines for the analysis of whole-genome sequences.

One source of difficulty in using resequencing approaches for diagnosis centers on the need to improve the quality and completeness of the human reference genome. In terms of quality, it is clear that the clone-based methods used to map, assign a minimal tiling path, and sequence the human reference genome did not yield a properly assembled or contiguous sequence equally across all loci. Lack of proper assembly is often due to

It has become extremely hard and costly to pinpoint and understand what we already know.

“Without structure, data are mere babble.”

In 1986, Swanson proposed that Raynaud’s syndrome symptoms can be mitigated by fish oil.

He connected facts by reading disjoint sets of literature.

He again made the connection between magnesium and migraine headaches.

**Argument 1 - migraine literature**

*Calcium channel blockers can prevent migraine attacks.*

**Argument 2 - magnesium literature**

*Magnesium is a natural calcium channel blocker.*
INTEGRATIVE BIOLOGY THROUGH TEXT-MINING

Figure 1 | **Categories of text-mining solutions.** The diagram gives an overview of the different categories of situations in which text mining is applied. Document retrieval is the initial step and leads to the collection of documents for a given query. The other solutions target the identification and evaluation of information that is explicitly stated in the documents.

GENE NAME RECOGNITION AND IDENTIFICATION

Sonic Hedgehog was found to induce HOXA6 expression in NIH-3t3 cells

1. Identify parts of speech for individual words

Sonic/ADJ Hedgehog/N was/V found/V to/PREP induce/V HOXA6/N expression/N in/PREP NIH-3t3/N cells/N

2. Identify semantic classes of phrases using an ontology

GENE[Sonic/ADJ Hedgehog/N] was/V found/V to/PREP induce/V GENE(HOXA6/N) expression/N in/PREP NIH-3t3/N cells/N

3. Apply templates/regular expressions to find relevant patterns, such as

GENE[*] induce [GENE] expression

GENE[Sonic/ADJ Hedgehog/N] was/V found/V to/PREP induce/V GENE(HOXA6/N) expression/N in/PREP NIH-3t3/N cells/N

4. Insert a new fact into the database

[class=induction-of-expression, by=Sonic Hedgehog, of=HOXA6]
ATTACK OF THE SYNONYMS

BRCA1
BRCA-1
BRCA 1
IRIS
PSCP
BRCAI
BRCC1
RNF53
PPP1R53
RING finger protein 53
protein phosphatase 1, regulatory subunit 53
breast cancer 1, early onset

http://www.genecards.org/cgi-bin/carddisp.pl?gene=BRCA1&search=brca1
ATTACK OF THE SYNONYMS

FAT1 FAT tumor suppressor homolog 1
Entrez ID 2195
FAT, ME5, CDHF7, CDHR8, hFat1
tumor suppression, bipolar disorder

CD36 thrombospondin receptor
Entrez ID 948
FAT, GP4, GP3B, GPIP, CHDS7, PASIV, SCARB3, BDPLT10
atherosclerosis, insulin resistance

The science is serious — not the genes

**Stranded At Second:** A fruit fly that dies, usually in the second larval stage of development.

**Agoraphobic:** A fruit fly with larvae that look normal but never crawl out of the egg shell.

**Groucho Marx:** A fruit fly that produces an excess of facial bristles.

**Cheap Date:** A fruit fly that expresses high sensitivity to alcohol.

**Out Cold:** A fruit fly that loses coordination when the temperature drops.

**Kenny:** A fruit fly without this gene dies in two days, named for the South Park character who dies in each episode.

**Ken and Barbie:** Fruit flies that fail to develop external genitalia.

**I'm Not Dead Yet (INDY):** These fruit flies live longer than usual. Reference to Monty Python's The Holy Grail.
BioGraph: unsupervised biomedical knowledge discovery via automated hypothesis generation

Anthony ML Liekens, Jeroen De Knijff, Walter Daelemans, Bart Goethals, Peter De Rijk and Jurgen Del-Favero

Public knowledge bases are integrated to connect genes, diseases and proteins in a weighted heterogeneous network

adapted from Figure 1 in Liekens A. et al. Genome Biology 12:R57 (2011).
TEXT-MINING METHODS
LITERATURE IS STILL LARGELY COMPOSED AND PUBLISHED OPAQUELY

However, despite very significant investment and a massive rise in access to scientific information, our community continues to be beset by propositions and manifestos on the practice of scholarly publishing.

“We are committed to change and innovation that will make science more effective.”

Brussels Declaration on Scientific, Technical and Medical Publishing
"...we plan to automatically publish the logical descriptions of automated experiments."

"What remain to be determined are the limits of automation."

CYTOGENETIC KARYOTYPING

International Standard for Cytogenetic Nomenclature (ISCN)

46,XY

47,XY,+21

47,XY,+3,t(14;18)(q32;q21)

49,XY,+X,der(1)t(1;8)

(p36.21;q24.13),t(2;10)

(p11.2;q10),+der(10)t(2;10)

(p11.2;p10),+7,[dup(7)(q34)],t

(14;18)(q32;q21)[cp7]
efficient algorithms
FIND DIFFERENCES IN GENOMES

graphs and networks
ASSEMBLE GENOME SEQUENCE

clustering
FIND PATTERNS IN GENE EXPRESSION

text mining
DISCOVER BIOLOGICAL RELATIONSHIPS

visualization
GENOME BROWSER MODEL
STRUCTURAL CHANGES ARE HARD TO SHOW FOR ONE GENOME

STRUCTURAL CHANGES ARE HARD TO SHOW FOR MULTIPLE GENOMES

we can no longer afford to show the full data sets

only meaningful differences

...or even only differences of differences
And that’s why we need a computer.
And that’s why we need a human.
HAIRBALLS AND NETWORK HAIRBALLS

both are visualizations of a complex system

The apparent banding pattern of the yellow nodes is an artefact of the graph layout algorithm. Importantly, the layout algorithm was not informed by type of supporting evidence and therefore does not explain the evident separation of blue and red edges.

Figure 2 and caption quote from Rual et al., *Nature* 437(7062):1173-8.
MOST LAYOUTS CANNOT BE COMPARED

HIVE PLOTS — WWW.HIVEPLOT.COM

periodic parallel-coordinate plots of topological properties
FUNCTION IS NOT RELATED TO GENOMIC POSITION
PHYSICAL COORDINATES ARE NATURAL, BUT LIMITING

instead, use functional coordinates clustered by data profile

Motivation

Genome browsers are ideal for viewing local regions of interest. But they do not provide a global overview of these regions.

Purpose of Spark: achieve meaningful overview and detail by focusing on regions of interest

Step 1: Pre-processing

Step 2: Clustering

Step 3: Interactive visualization

SPARK

**SEQUENCE MOTIFS**

HEM13  
CCCATTGTTCTC

HEM13  
TTTCTGGTTCTC

HEM13  
TCAATTTGTAG

ANB1   
CTCATTGTTGTC

ANB1   
TCCATTGTTCTC

ANB1   
CCTATTGTTCTC

ANB1   
TCCATTGTTCGT

ROX1   
CCAAATTGTTTTG

**YCHAATTGTTCTC**

**Table:**

<table>
<thead>
<tr>
<th>Nucleotide</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0.02700000010</td>
</tr>
<tr>
<td>C</td>
<td>0.464100000505</td>
</tr>
<tr>
<td>G</td>
<td>0.00001800112</td>
</tr>
<tr>
<td>T</td>
<td>0.422087088261</td>
</tr>
</tbody>
</table>

**Figure 1** ROX1 binding sites and sequence motif.  
(a) Eight known genomic binding sites in three *S. cerevisiae* genes. (b) Degenerate consensus sequence. (c,d) Frequencies of nucleotides at each position. (e) Sequence logo showing the frequencies scaled relative to the information content (measure of conservation) at each position. (f) Energy normalized logo using relative entropy to adjust for low GC content in *S. cerevisiae*.

SEQUENCE LOGOS — VISUAL JARGON

EXPLORATION / COMMUNICATION

to explore data, use effective visual encodings

to communicate concepts and patterns, use effective design
The EVEREST PowerWall at Oak Ridge National Laboratory, in Tennessee, is a computer visualization facility. EVEREST stands for Exploratory Visualization Environment for Research in Science and Technology. The 9-meter-wide, 2.4-meter-tall screen can display 35 million pixels of information and is now being used as a tool to model climate change.

http://spectrum.ieee.org/energy/nuclear/slideshow-a-nuclear-family-vacation/0
bad encoding doesn’t mean the end of the world ... maybe

Recent observations from satellites and ground stations suggest that atmospheric ozone levels for March in the Arctic were approaching the lowest levels in the modern instrumental era. http://earthobservatory.nasa.gov/IOTD/view.php?id=49874
CONSEQUENCES OF INAPPROPRIATE ENCODING

NYT did not use the figure – because information lost in b/w

Jonathan Corum (personal communication).
CONSEQUENCES OF INAPPROPRIATE ENCODING

use tone instead of hue

Ozone (Dobson Units)

110  220  330  440  550
LUMINANCE EFFECT — THE LIER IN THE HEAT MAP

Same colour looks different

Different colour looks the same

* These rectangles have the same colour but look different

Figure 1 adapted from Yates et al. *Nature Reviews Genetics* 13:795 (2012).