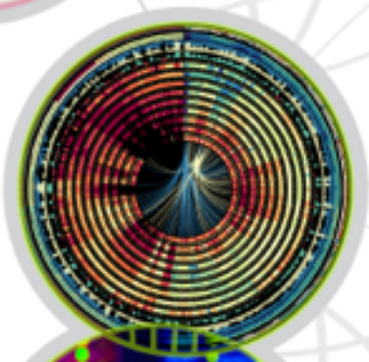
GENOMICS

INNOVATION

CANADA'S MICHAEL SMITH
GENOME
SCIENCES
CENTRE



INFORMATICS

SEQUENCING

COMPUTING

genomics + data mining

needles in stacks of needles*

* title is drawn from Cooper *et al.* Needles in stacks of needles: finding disease-causal variants in a wealth of genomic data. *Nature Reviews Genetics* 12: 629 (2011).

martin krzywinski

http://mkweb.bcgsc.ca

canada's michael smith genome sciences centre

bc cancer research center

vancouver canada

in each of our ~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 aganist mnay chgnaes

genome alterations are like spelling mistakes

our biology is robust aganist mnay chgnaes

but if we aucculamte too mnay of tehm

genome alterations are like spelling mistakes

our biology is robust aganist mnay chgnaes

but if we aucculamte too mnay of tehm

our ailbtiies to apadt and reiapr will be

genome alterations are like spelling mistakes

our biology is robust aganist mnay chgnaes

but if we aucculamte too mnay of tehm

our ailbtiies to apadt and reiapr will be

oeehvmlewrd

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

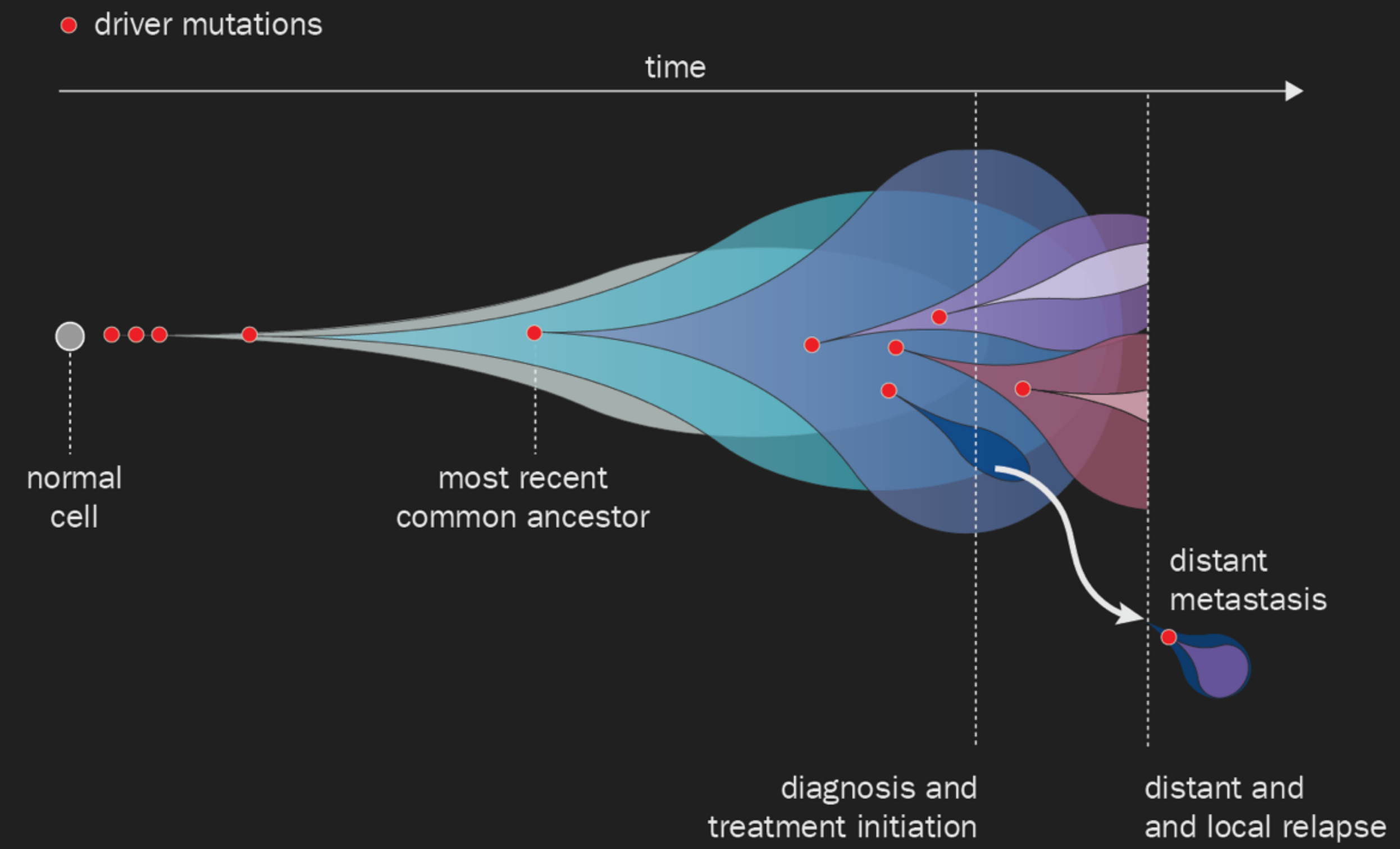


Figure 1 adapted from Yates et al. *Nature Reviews Genetics* 13:795 (2012).

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

The New York Times

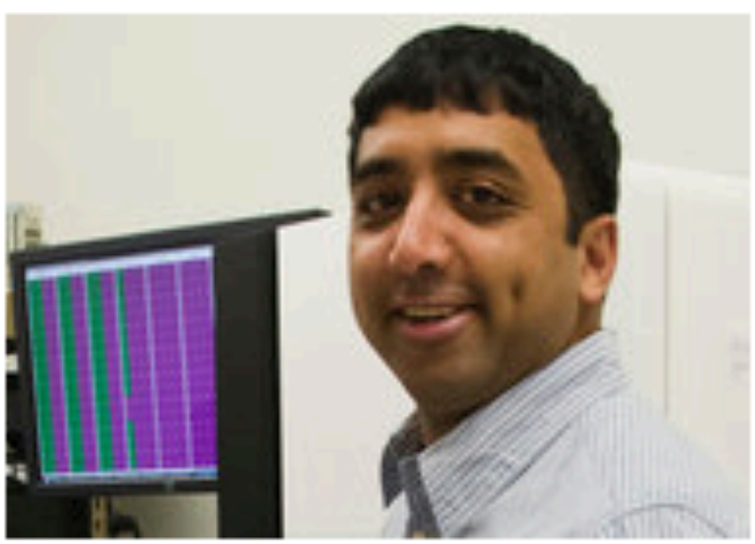
Health

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.



Clare McLean/University of Washington

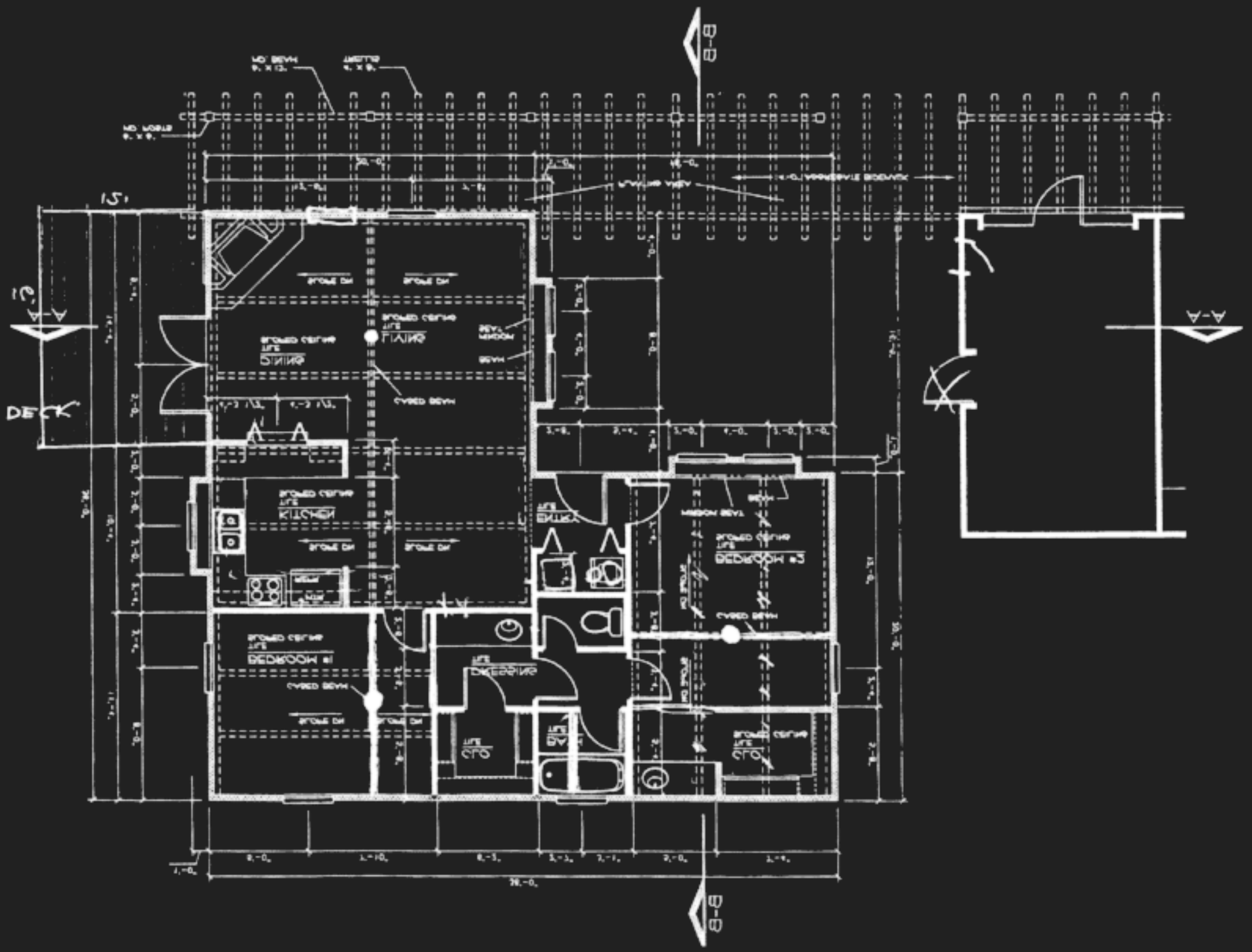
The accomplishment heralds an era in which parents might find it easier to know the complete DNA blueprint of a child months before it is born.

That would allow thousands of genetic diseases to be detected prenatally. But

http://www.nytimes.com/2012/06/07/health/tests-of-parents-are-used-to-map-genes-of-a-fetus.html

DNA is not a blueprint

THIS IS A BLUEPRINT



http://www.hsmagazine.net/2012/08/blueprinthome-first-impressions-reviewew/

THIS IS DNA

taaccctaaccctaaccctaaccctaaccctaaccctaaccctaacccta
accctaaccctaaccctaaccctaaccctaaccctaaccctaaccctaac
cctaacccaaccctaaccctaaccctaaccctaaccctaaccctaacccc
taaccctaaccctaaccctaaccctaacctaaccctaaccctaaccctaa
ccctaaccctaaccctaaccctaaccctaaccccctaaccctaaccctaaa
ccctaaaccctaaccctaaccctaaccctaaccctaaccccaaccccaac
cccaaccccaaccccaaccccaaccctaaccccctaaccctaaccctaacc
ctaccctaaccctaaccctaaccctaaccctaaccctaaccccctaacccc
taaccctaaccctaaccctaaccctaaccctaaccctaaccccctaaccct
aaccctaaccctaaccctcgcggtaccctcagccggcccgcccgcccggg
tctgacctgaggagaactgtgctccgccttcagagtaccaccgaaatctg
tgcagaggacaacgcagctccgccctcgcggtgctctccgggtctgtgct
gaggagaacgcaactccgccgttgcaaaggcgcgccgcgccggcgcaggc
gcagagaggcgcgccgcgccggcgcaggcgcagagaggcgcgccgcgccg
gcgcaggcgcagagaggcgcgccgcgccggcgcaggcgcagagaggcgcg
ccgcgccggcgcaggcgcagagaggcgcgccgcgccggcgcaggcgcaga
cacatgctagcgcgtcggggtggaggcgtggcgcaggcgcagagaggcgc
gccgcgccggcgcaggcgcagagacacatgctaccgcgtccaggggtgga
ggcgtggcgcaggcgcagagaggcgcaccgcgccggcgcaggcgcagaga
cacatgctagcgcgtccaggggtggaggcgtggcgcaggcgcagagacgc
aagcctacgggcgggggttgggggggcgtgtgttgcaggagcaaagtcgc
acggcgccgggctggggcgggggggagggtggcgccgtgcacgcgcagaaa

sequence at start of human chromosome 1

DNA DOES NOT DIRECTLY DESCRIBE THE ORGANISM

Genome Research, October 2004 *http://genome.cshlp.org/content/14/10a.cover-expansion*

life is the emergent property of biochemical reactions

$$A \rightarrow B$$

DNA encodes the *enzymes* that catalyze these reactions

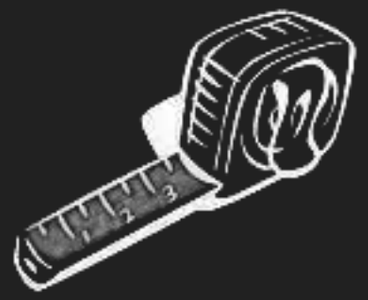
enzyme

$$A \longrightarrow B$$

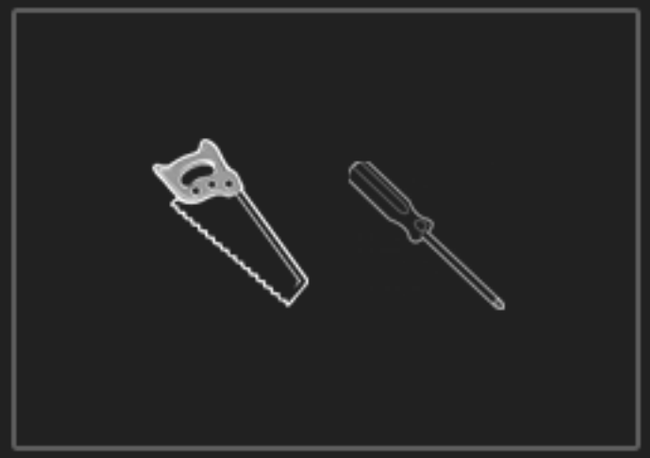
there are millions of reactions

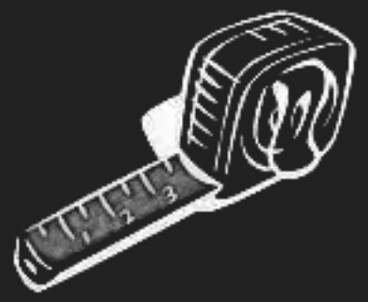
Biochemical Pathways, by Gerhard Michal, 1974, published by Boehringer Mannheim.

IF A HOUSE HAD DNA...

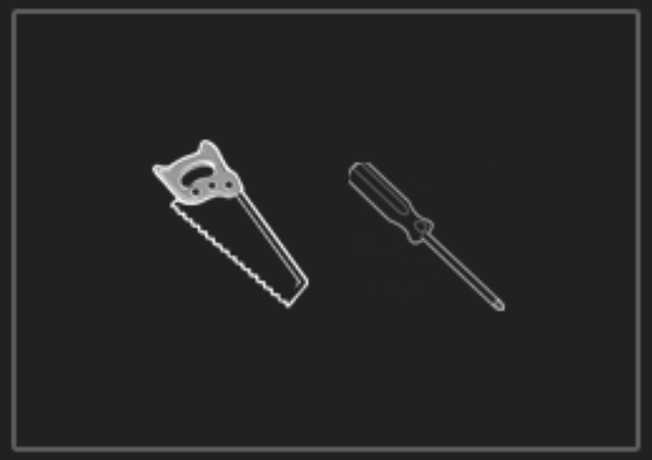


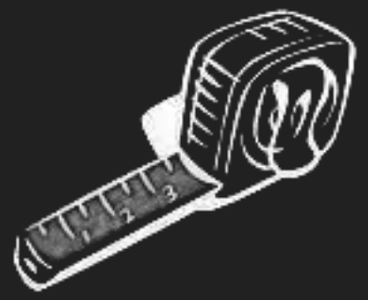
...A LIST OF TOOLS THAT MAKE THE TOOLS TO MAKE THE HOUSE





DNA CHANGES ARE HARD TO DECIPHER FUNCTIONALLY





molecular cellular mechanisms
are profoundly interconnected

with many multi-function components

gene
~20,000

3×10^9
base pairs

DNA

RNA

protein

A → A'

reaction

gene
~20,000

exon
~9 exons/gene

3×10^9
base pairs

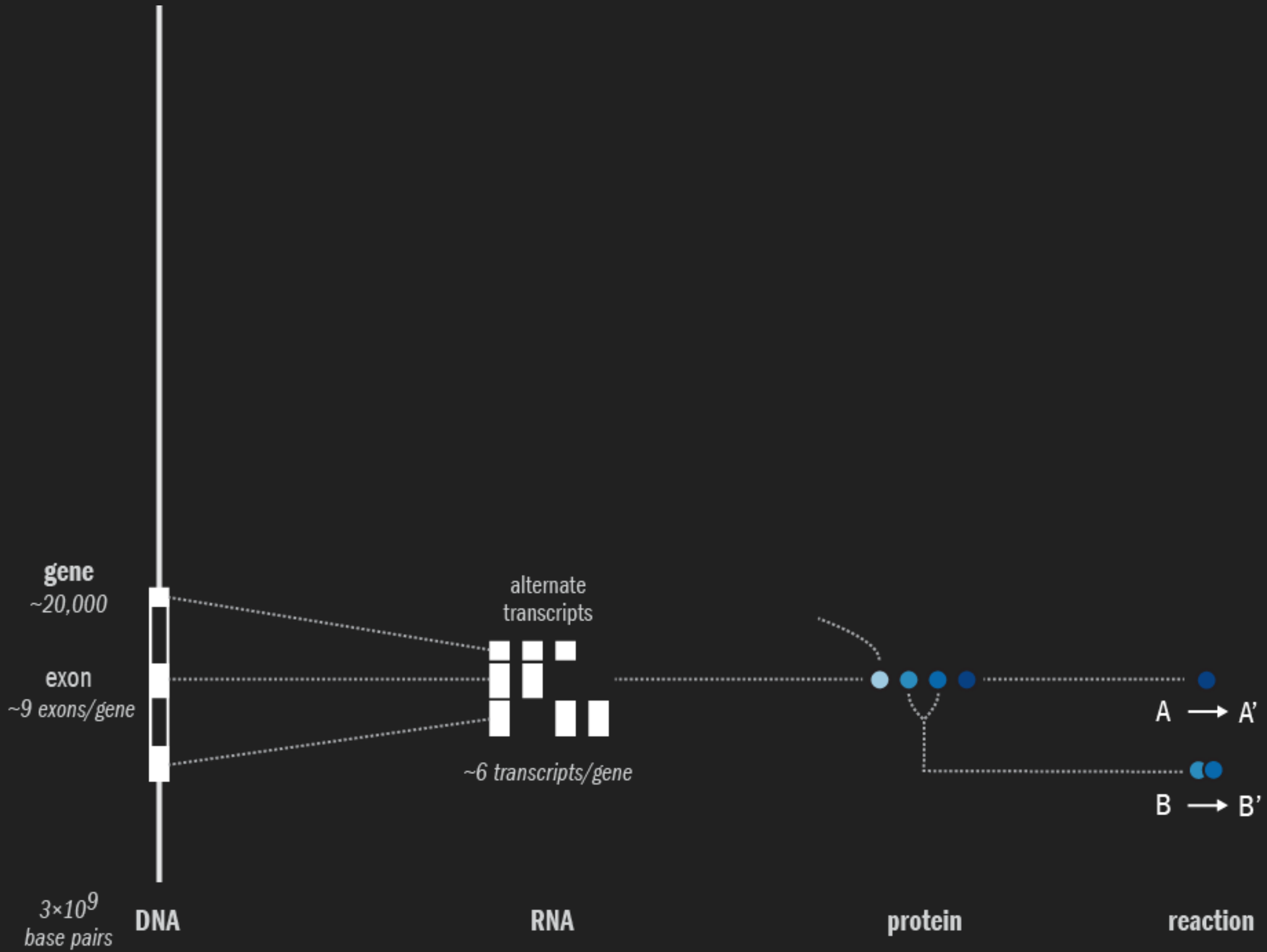
DNA

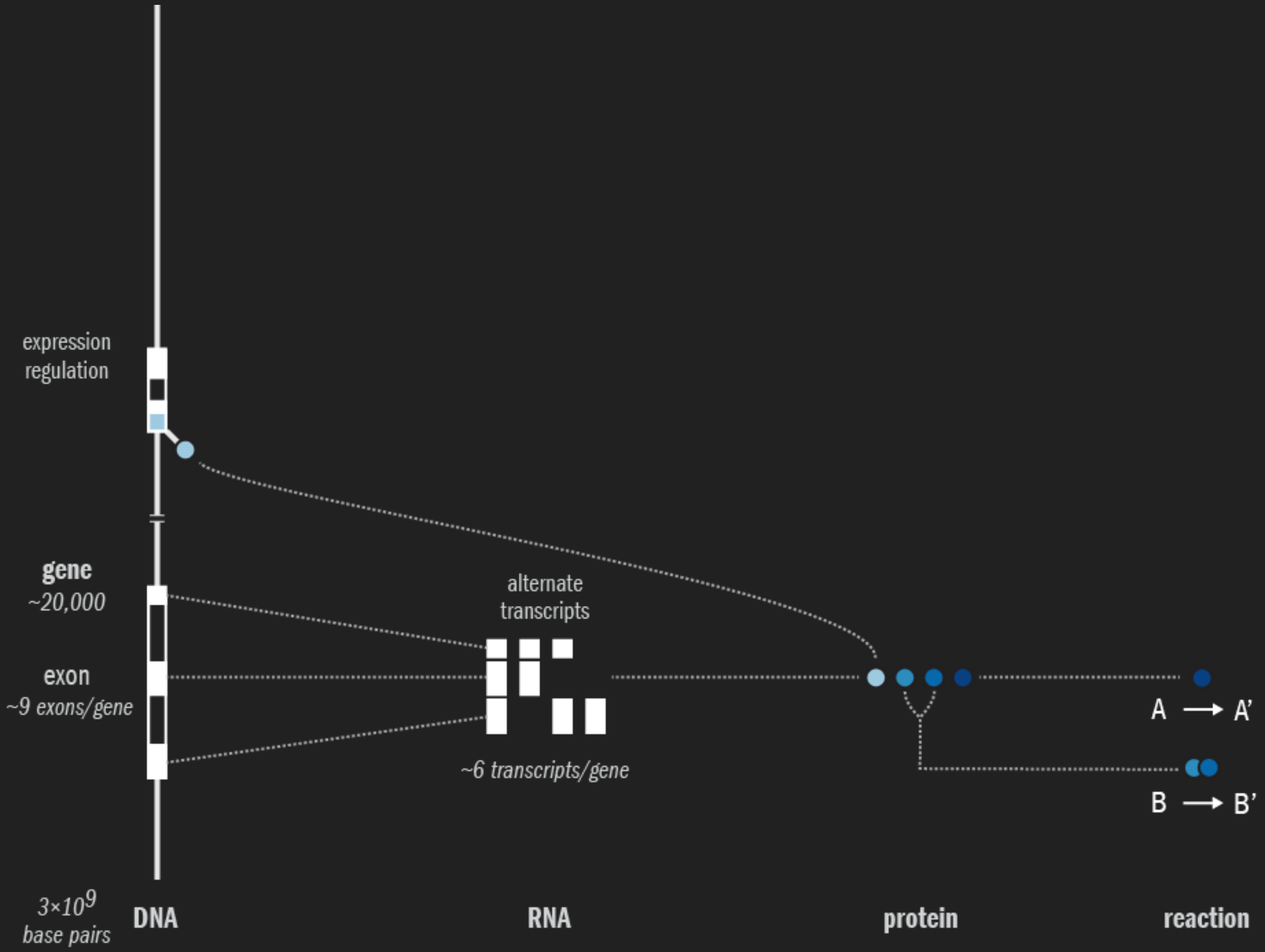
RNA

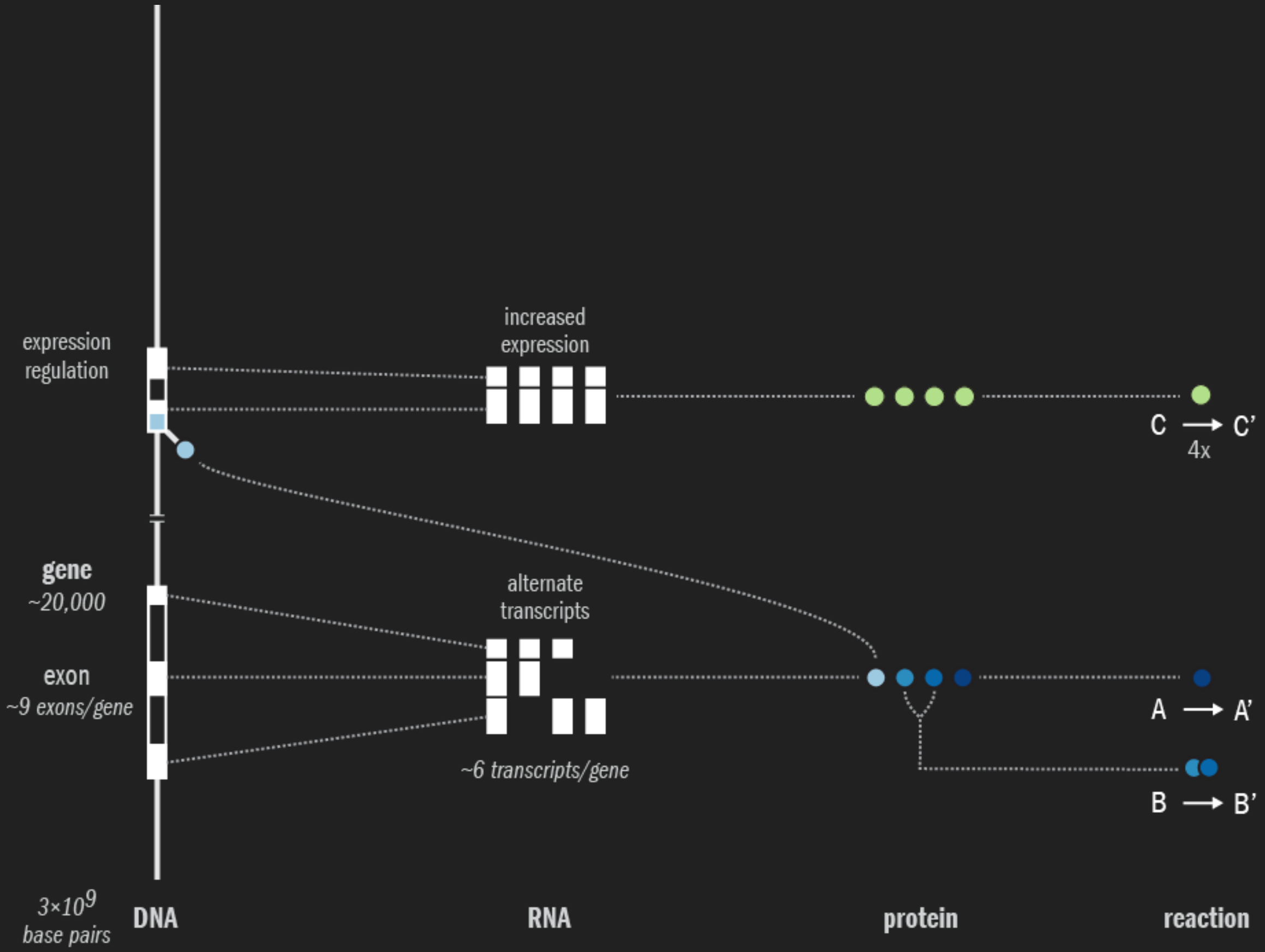
protein

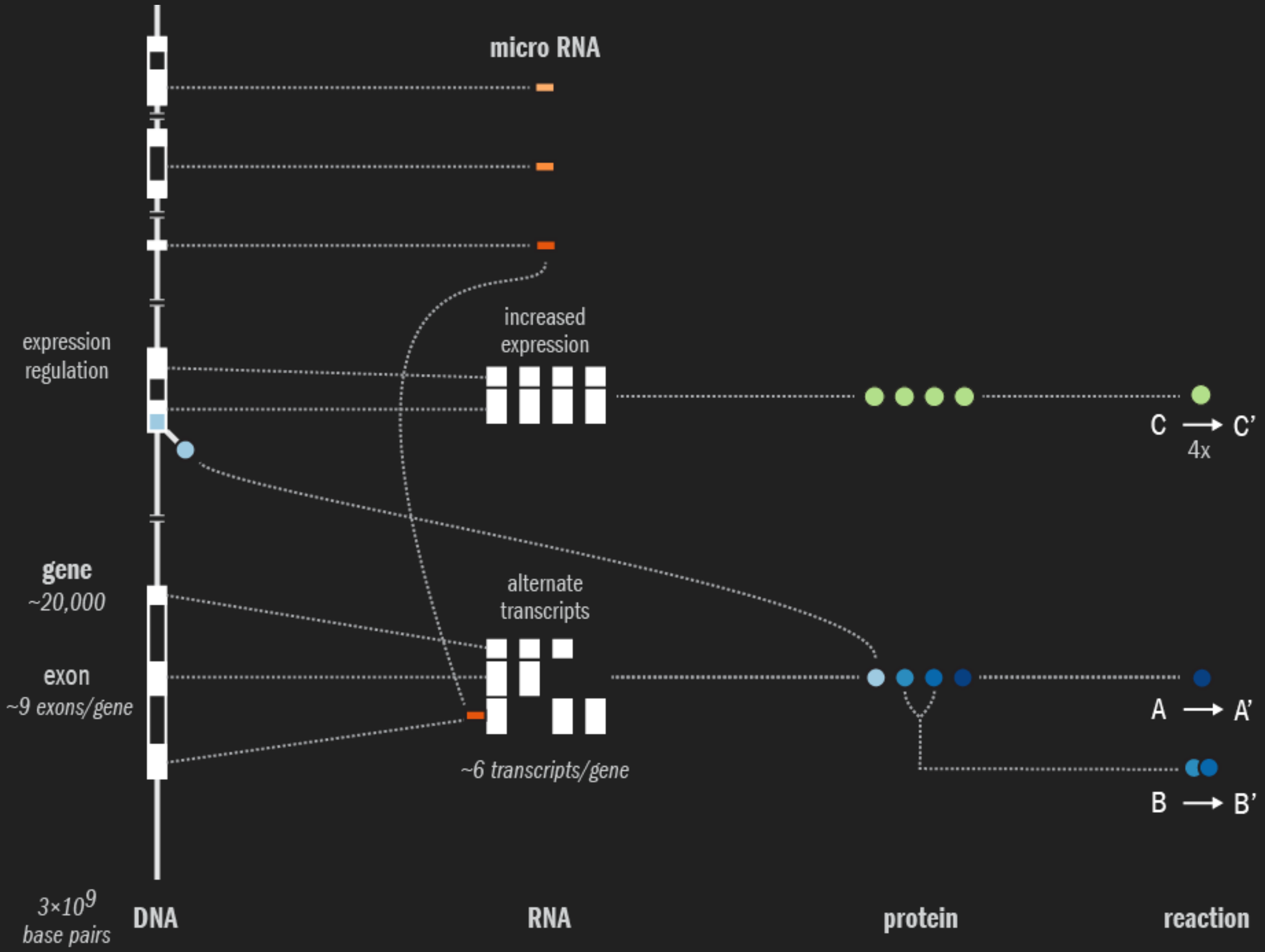
A → A'

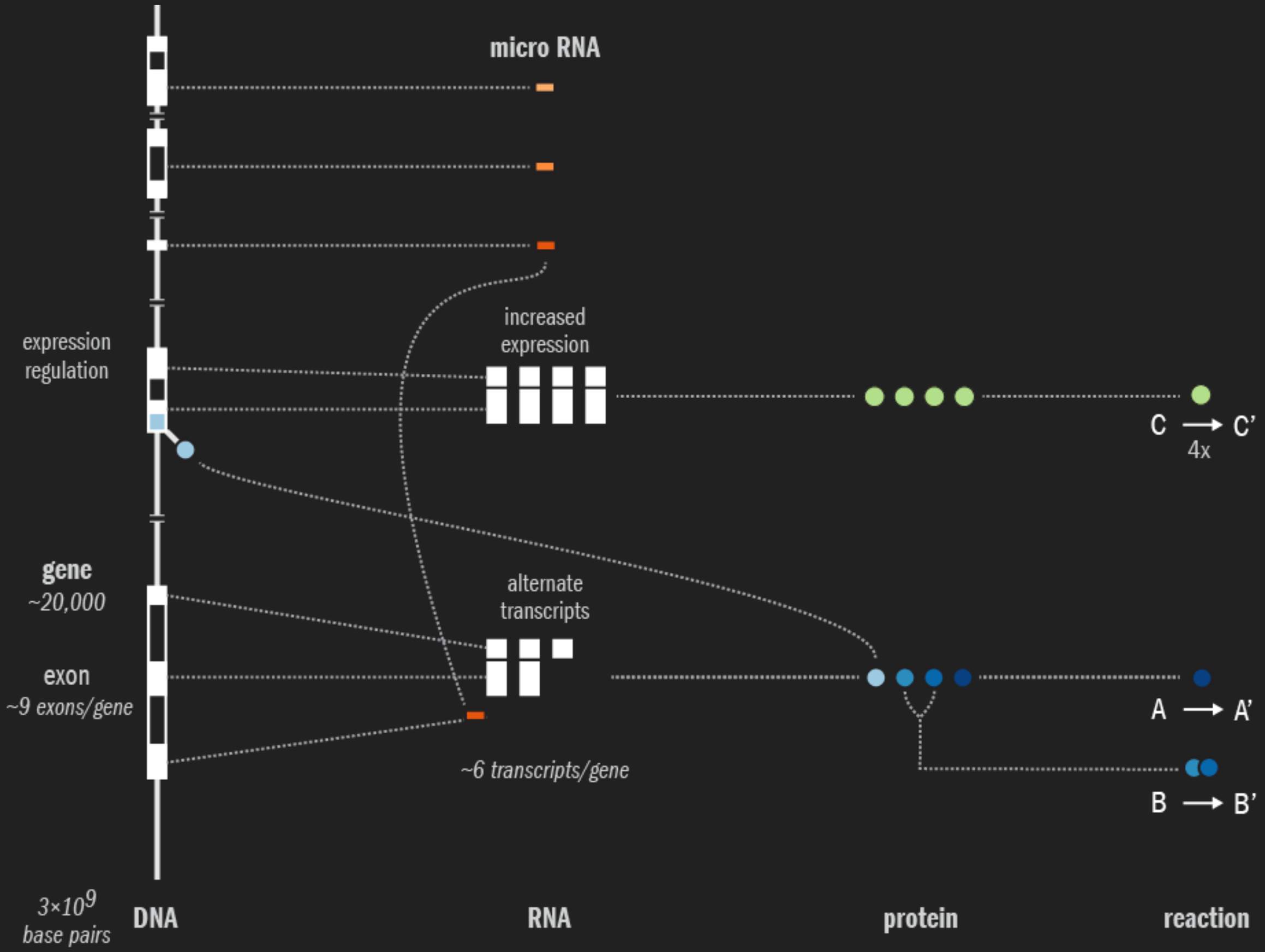
reaction

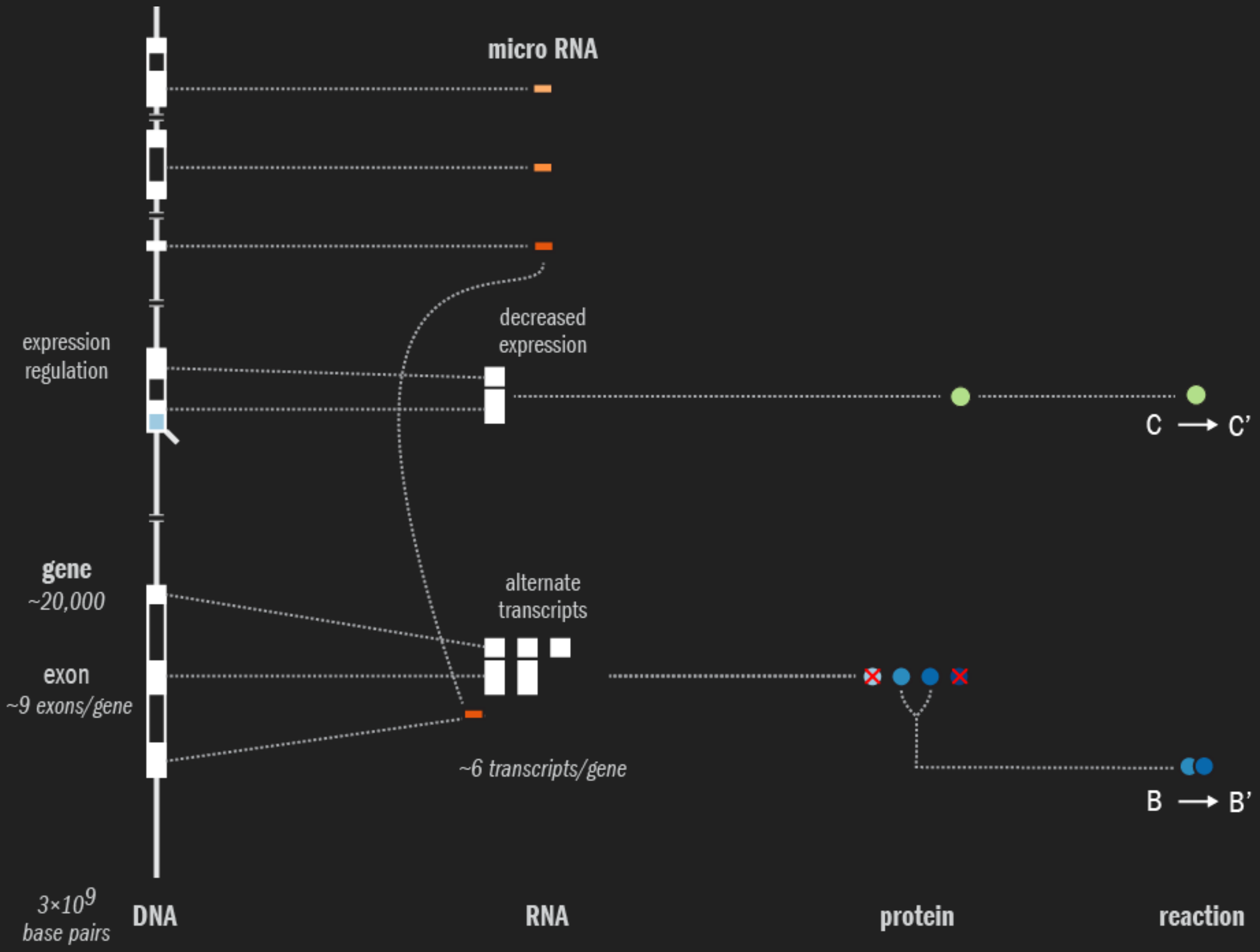

gene
~20,000
exon
~9 exons/gene
3×10^9
base pairs
DNA
alternate
transcripts
~6 transcripts/gene
RNA
protein
A → A'
B → B'
reaction


expression
regulation
gene
~20,000
exon
~9 exons/gene
alternate
transcripts
~6 transcripts/gene
A → A'
B → B'
3×10^9
base pairs
DNA
RNA
protein
reaction

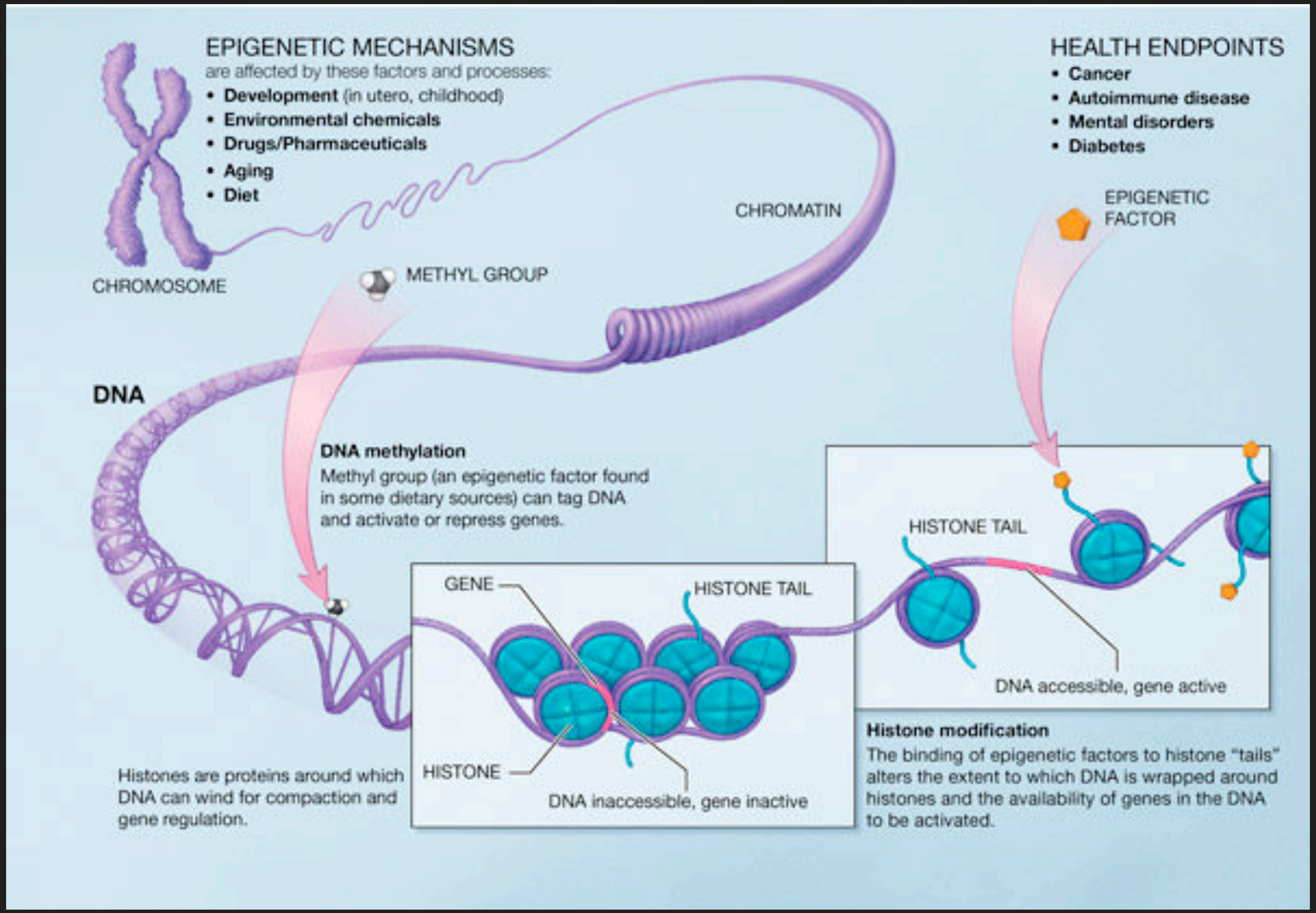

expression
regulation
increased
expression
C → C'
4x
gene
~20,000
exon
~9 exons/gene
alternate
transcripts
~6 transcripts/gene
A → A'
B → B'
3×10^9
base pairs
DNA
RNA
protein
reaction


micro RNA
expression regulation
increased expression
C → C'
4x
gene
~20,000
exon
~9 exons/gene
alternate transcripts
~6 transcripts/gene
A → A'
B → B'
3×10^9 base pairs
DNA
RNA
protein
reaction

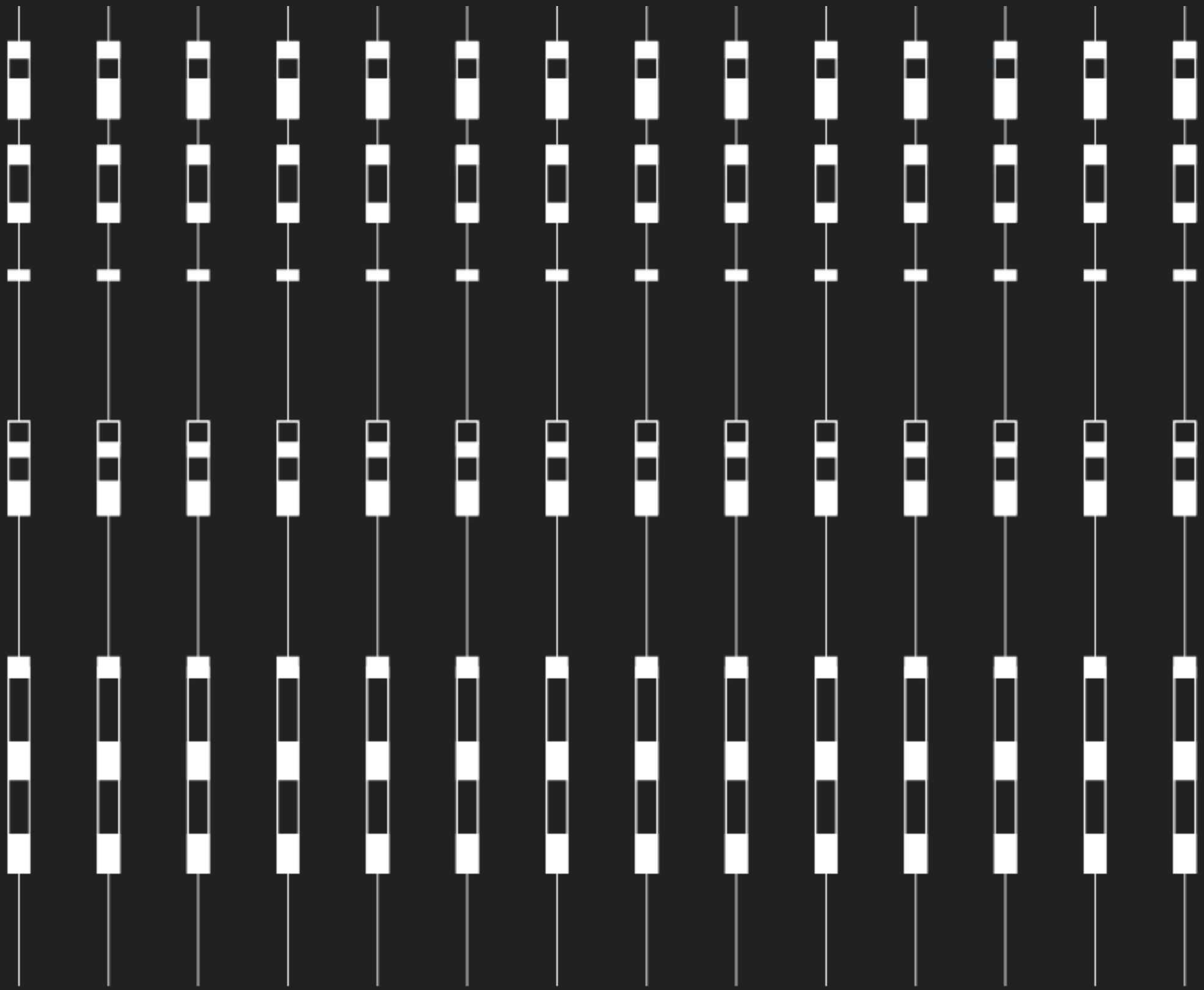

micro RNA
expression regulation
increased expression
C → C'
4x
gene
~20,000
exon
~9 exons/gene
alternate transcripts
~6 transcripts/gene
A → A'
B → B'
3×10^9 base pairs
DNA
RNA
protein
reaction


micro RNA
expression regulation
decreased expression
C → C'
gene
~20,000
exon
~9 exons/gene
alternate transcripts
~6 transcripts/gene
B → B'
3×10^9 base pairs
DNA
RNA
protein
reaction

EPIGENOME — “REST OF THE GENOME”



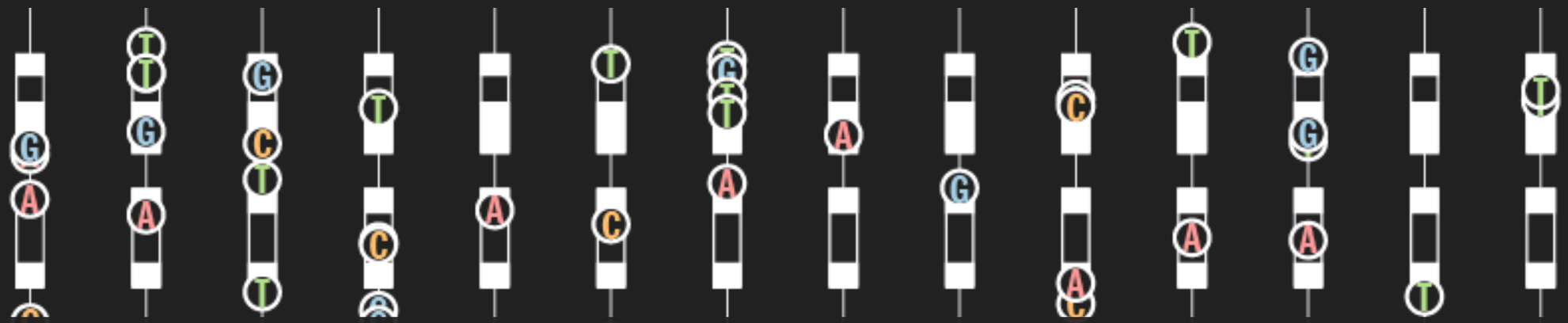
Citation: Laura, B. (2008) Epigenomics: The new tool in studying complex diseases. Nature Education 1(1)



individual genomes

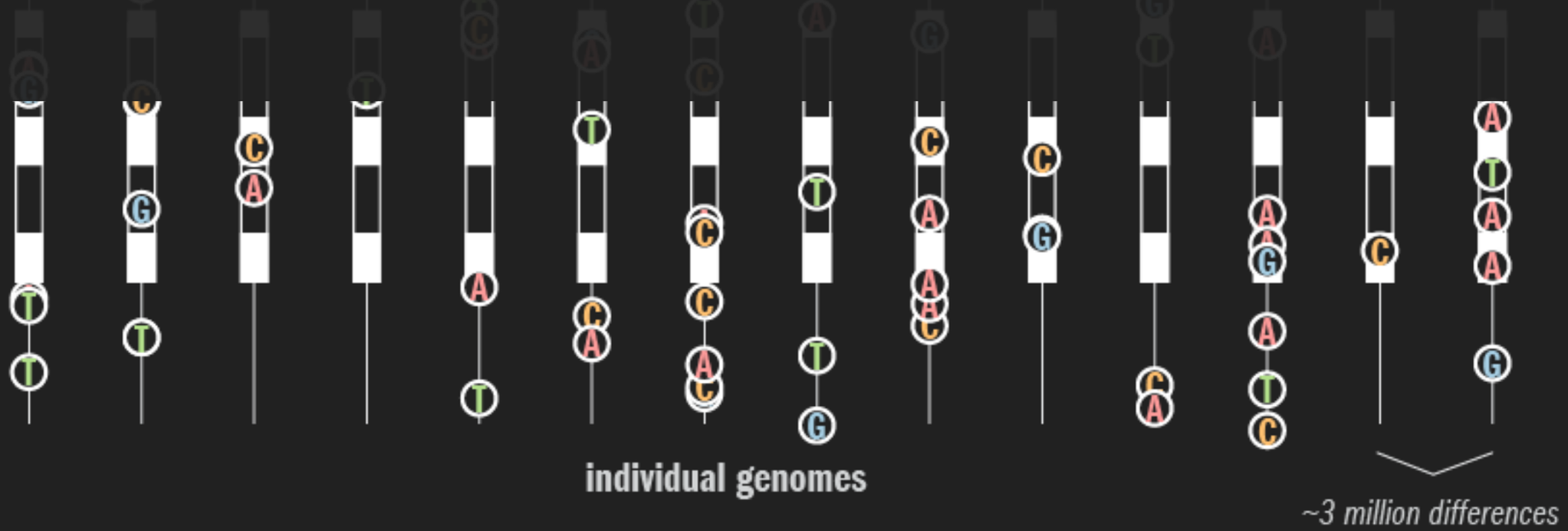
individual genomes

~3 million differences



many types of structural variations are possible

their functional consequences are difficult to assess



efficient algorithms

FIND DIFFERENCES IN GENOMES

graphs and networks

ASSEMBLE GENOME SEQUENCE

clustering

FIND PATTERNS IN GENE EXPRESSION

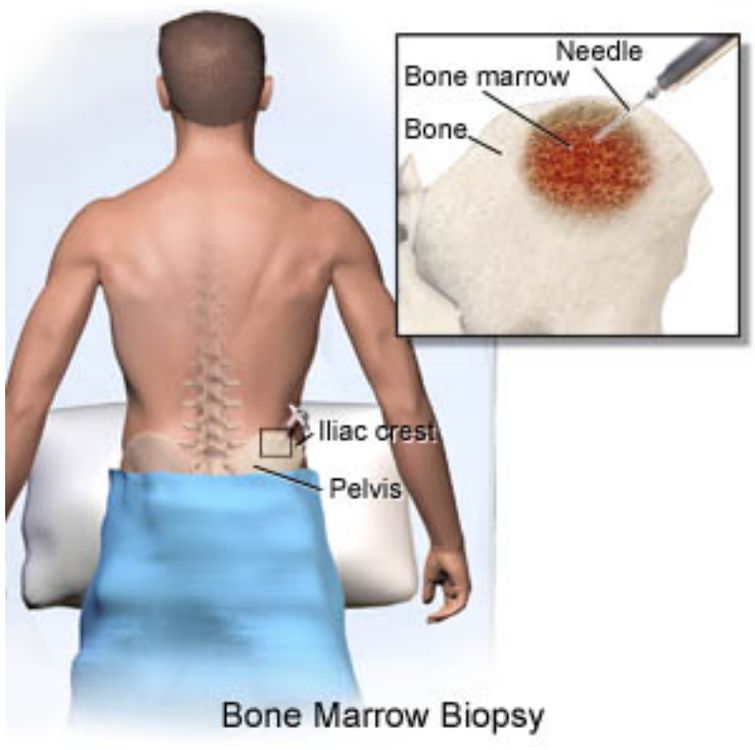
text mining

DISCOVER BIOLOGICAL RELATIONSHIPS

visualization

GENOME SEQUENCING

we learn about the genome by sequencing it



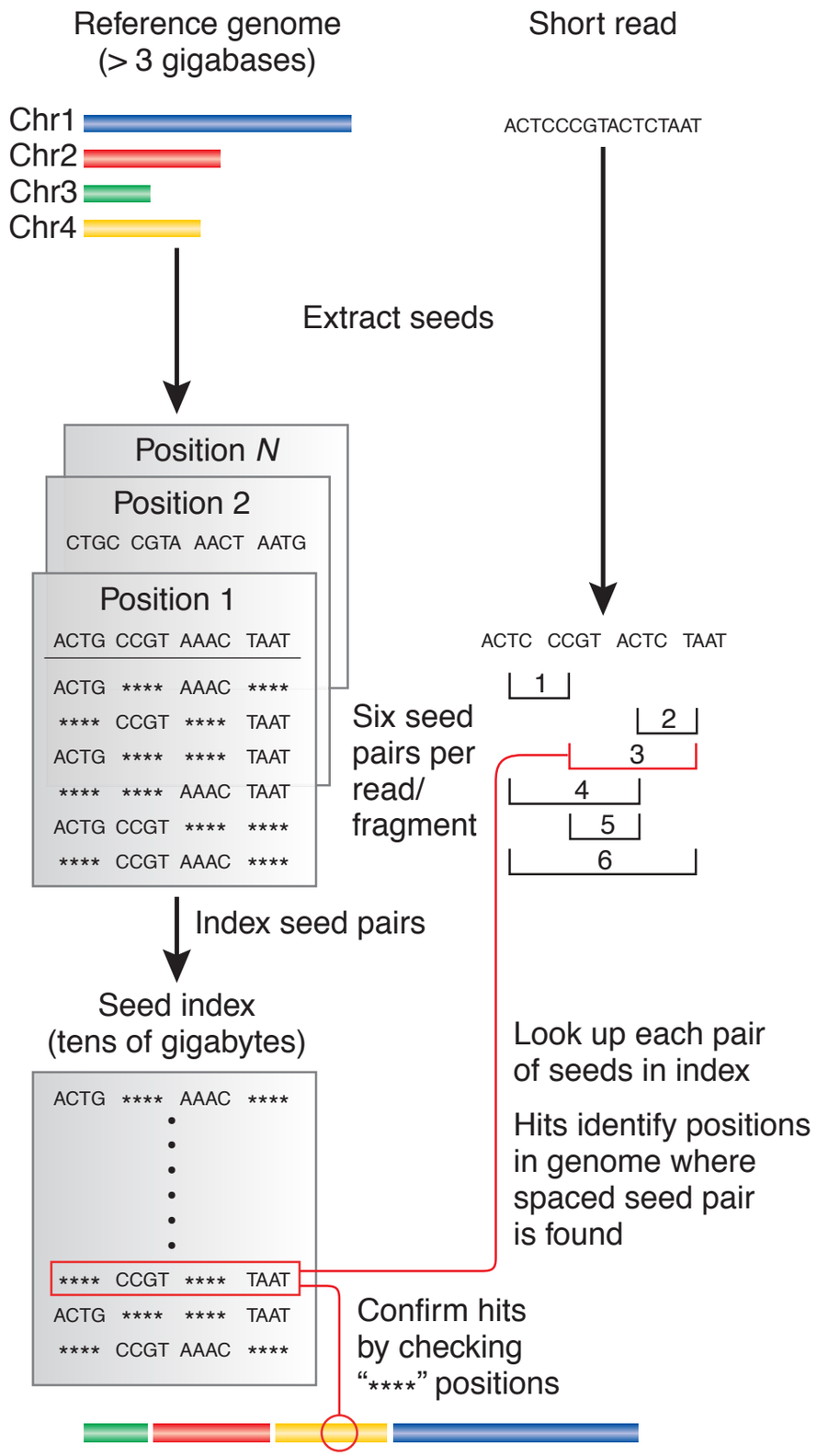
```
GCAGGGTCCGGGGCCCAGTTAAGGGCTCCCCCTCC
AGGTCGGCCTGCACCTCCCCTGCTCTGTGCCAGTG
GCTGAGGCCGAGGCCTGGCCATTGCCCTCCTGGTC
ATTTGAGTCCAGAAGCCAAACGTCTACAGTCGTGT
GATGGACTGCCCAGGCCATGGGGGGCTCTGAATGA
GGCGAGGTGGGCTGCGCGTCTGCAAGAAAGTGCAT
GCAGAGGGAAGCTGGGGTCCACCGCTGGTGAGCGG
CAGAACCGGCCGTGGCCCACCCCGAGACGGAGGCG
TTCTGTGCTCTGGAGGCTTTGGGCAGCTGTCAGGC
AGAGCAGATGGAGGCGTGAGGAGGCGGCTCCGGGG
GCCCTCACAGGGGAGGCCGAGGGGGAGGGCAGGAG
GGTGCAGGAGGCCCGTGGAACTTGGAGGGCTCTGT
CCCAGGGCGGGGGCAGCTCCAAGGCCTCGGGCTTG
GGACTTGGGGGTTGGGGGTGGTCAGAGCATTGTGG
GCTGGGGCTTCCCAGAGGGTGAGGTGTCTGCTGGG
AGGCGTCCCCAGGCCTCCAGCCCAGCCCCGGTTCC
AGGGTATCAATTGCCTGCTGGGAAACCTCAGGGTG
CCTGCCCCTGACACTCCTGGCCCTGGGCTCCCCCC
...
```

and aligning it to the human reference sequence

- align 8 billion 35 base reads to 3 billion base reference
- ~100-fold coverage

Ley et al. DNA sequencing of a cytogenetically normal acute myeloid leukaemia genome. *Nature* 456:66 (2008).

SPACED SEEDS (MAQ)



BURROW-WHEELER (BOWTIE)

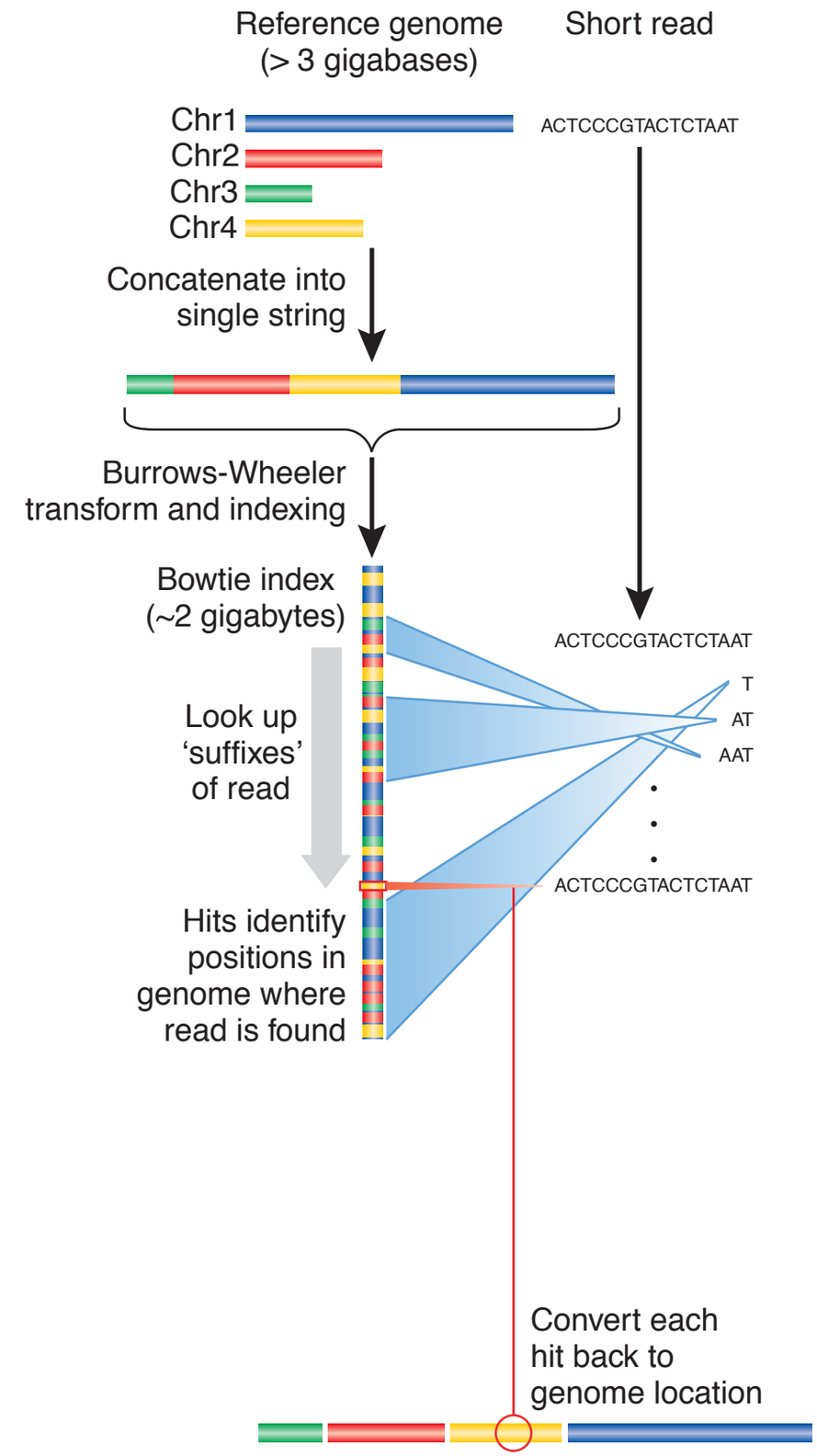


Figure 1 from Trapnell et al. *Nature Biotechnology* 27:455 (2009).

LIMITATIONS

for efficiency, number of mismatches is limited

e.g. 2 for BWT aligner Bowtie

BWT PERFORMANCE

50x faster than spaced seeds methods

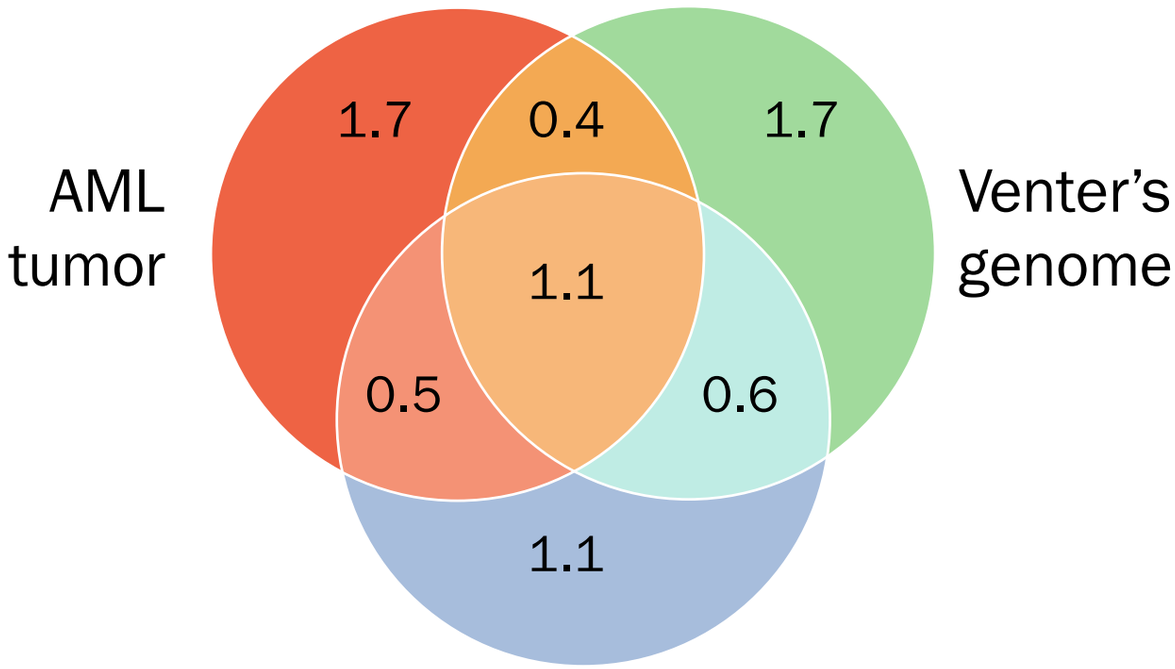
25 million 35 base reads per hour per CPU (2009)

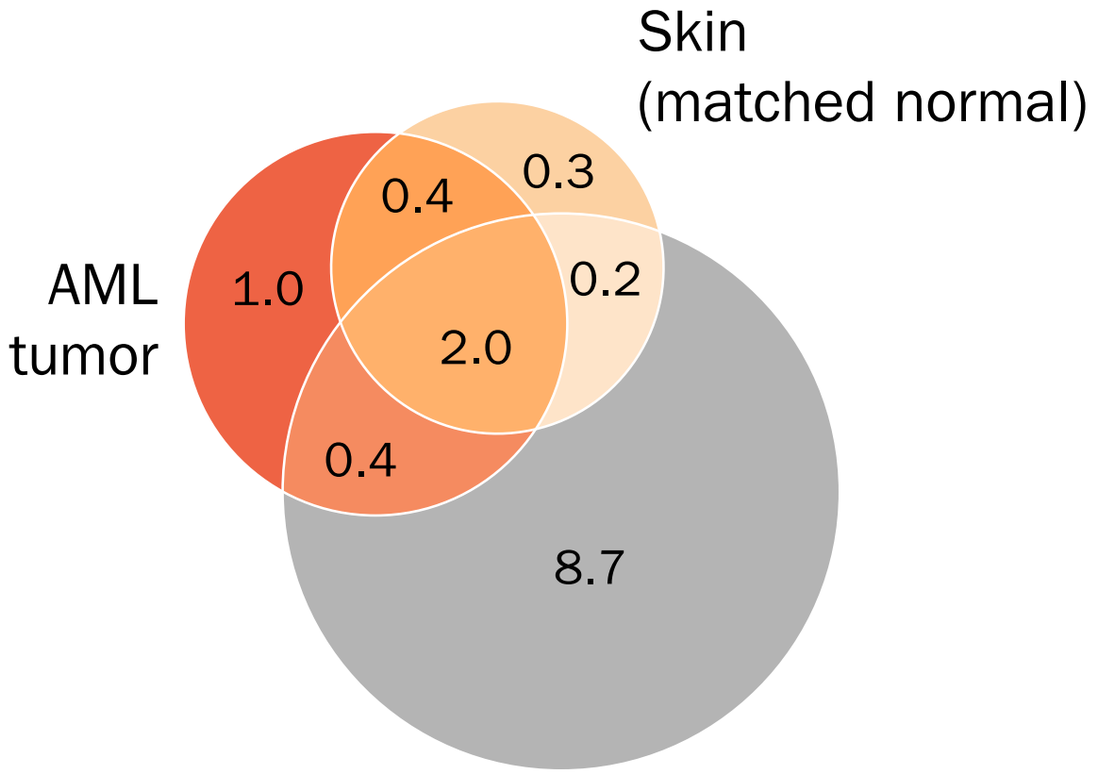
4 hours to align per 1X coverage

1.3 Gb memory footprint

Trapnell et al. *Nature Biotechnology* 27:455 (2009).

WE ALWAYS FIND GENOME CHANGES



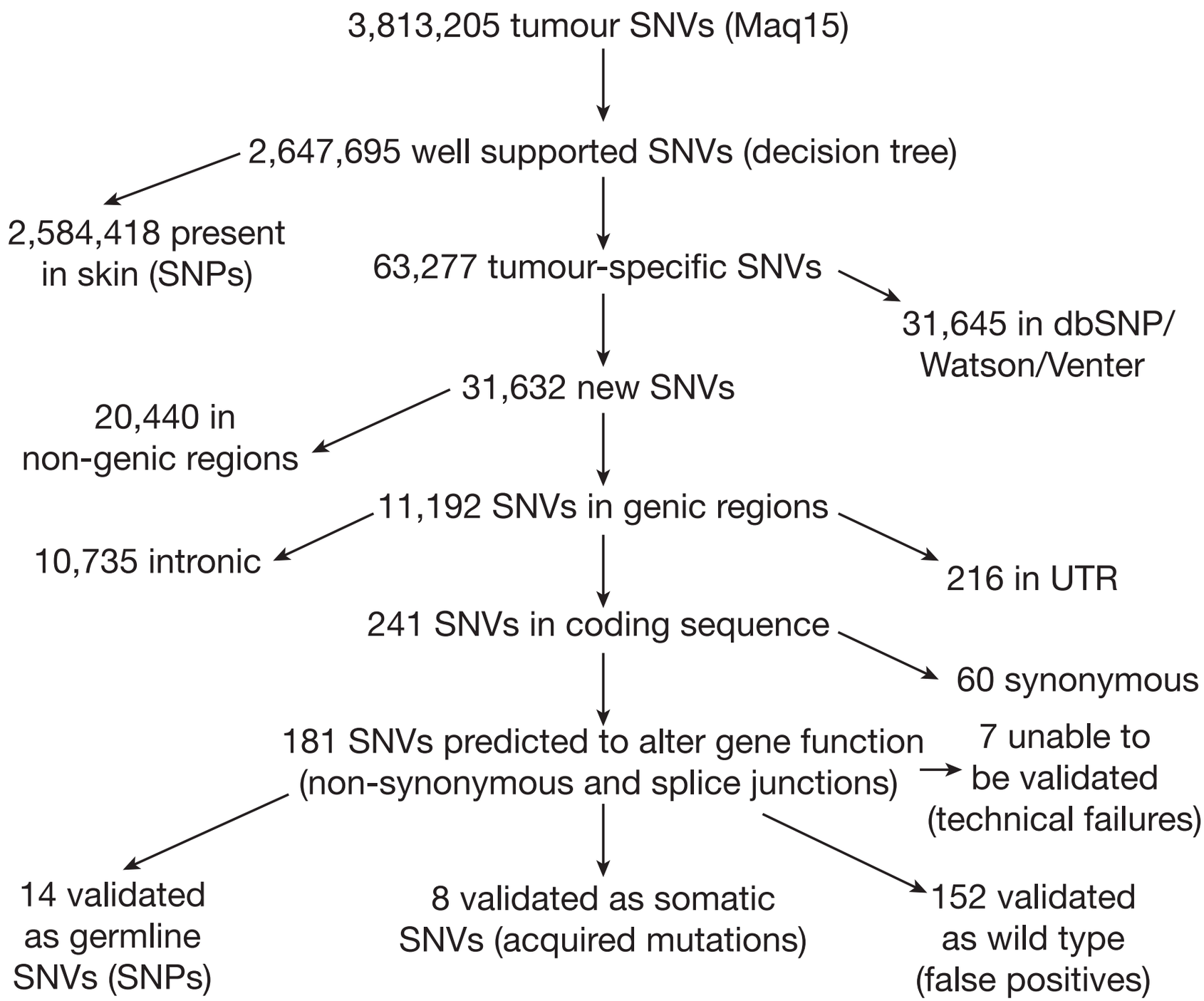


Overlap of observed changes between AML tumor genome and other reference genomes. Millions of single base changes (SNPs).

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



Ley et al. DNA sequencing of a cytogenetically normal acute myeloid leukaemia genome. *Nature* 456:66 (2008).

YOU CAN'T PUBLISH A SINGLE GENOME ANYMORE

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

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

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.

183 tumors

6 Tb of sequence

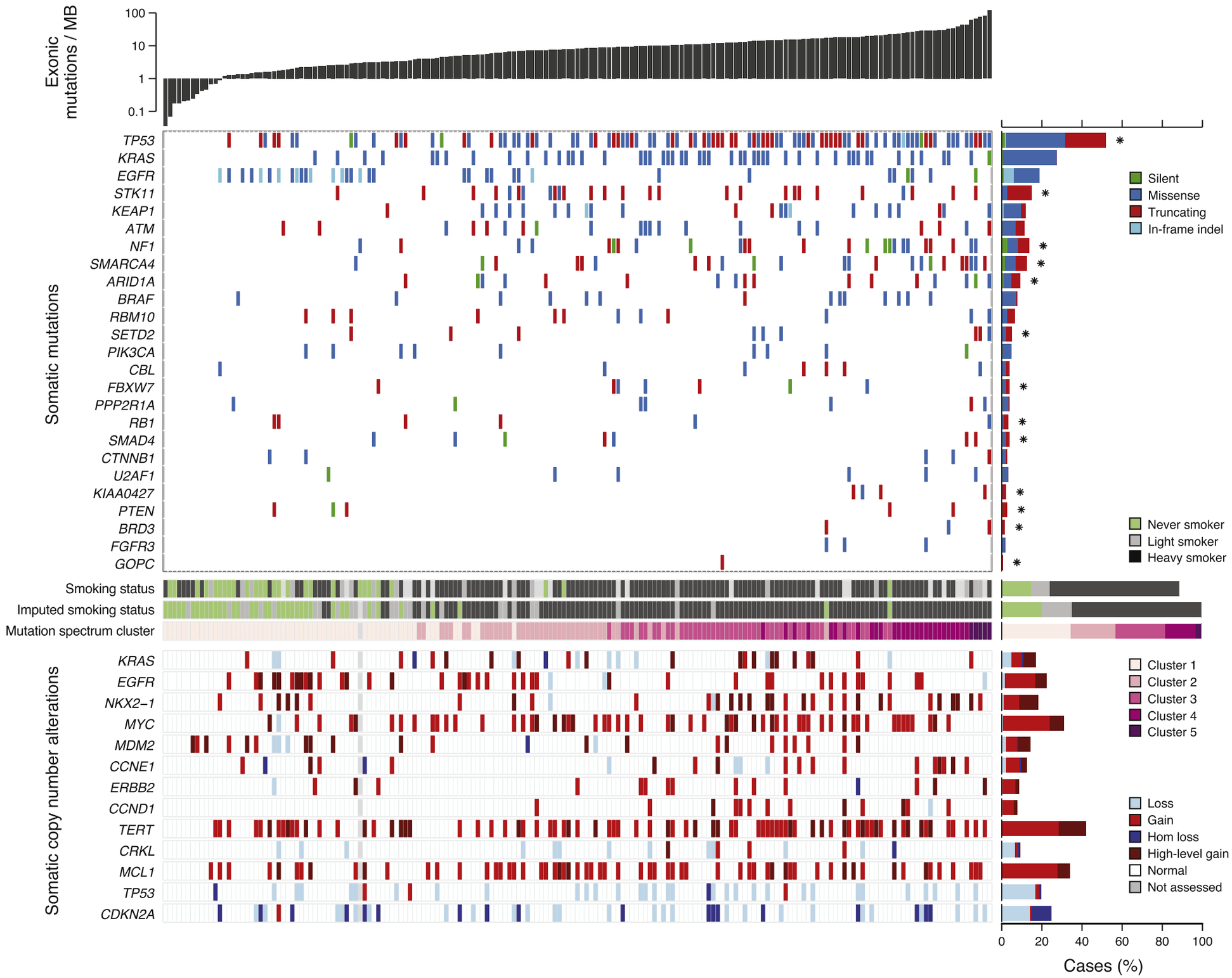
13 mutations/Mb

323 coding mutations per patient

2,350 structural rearrangements

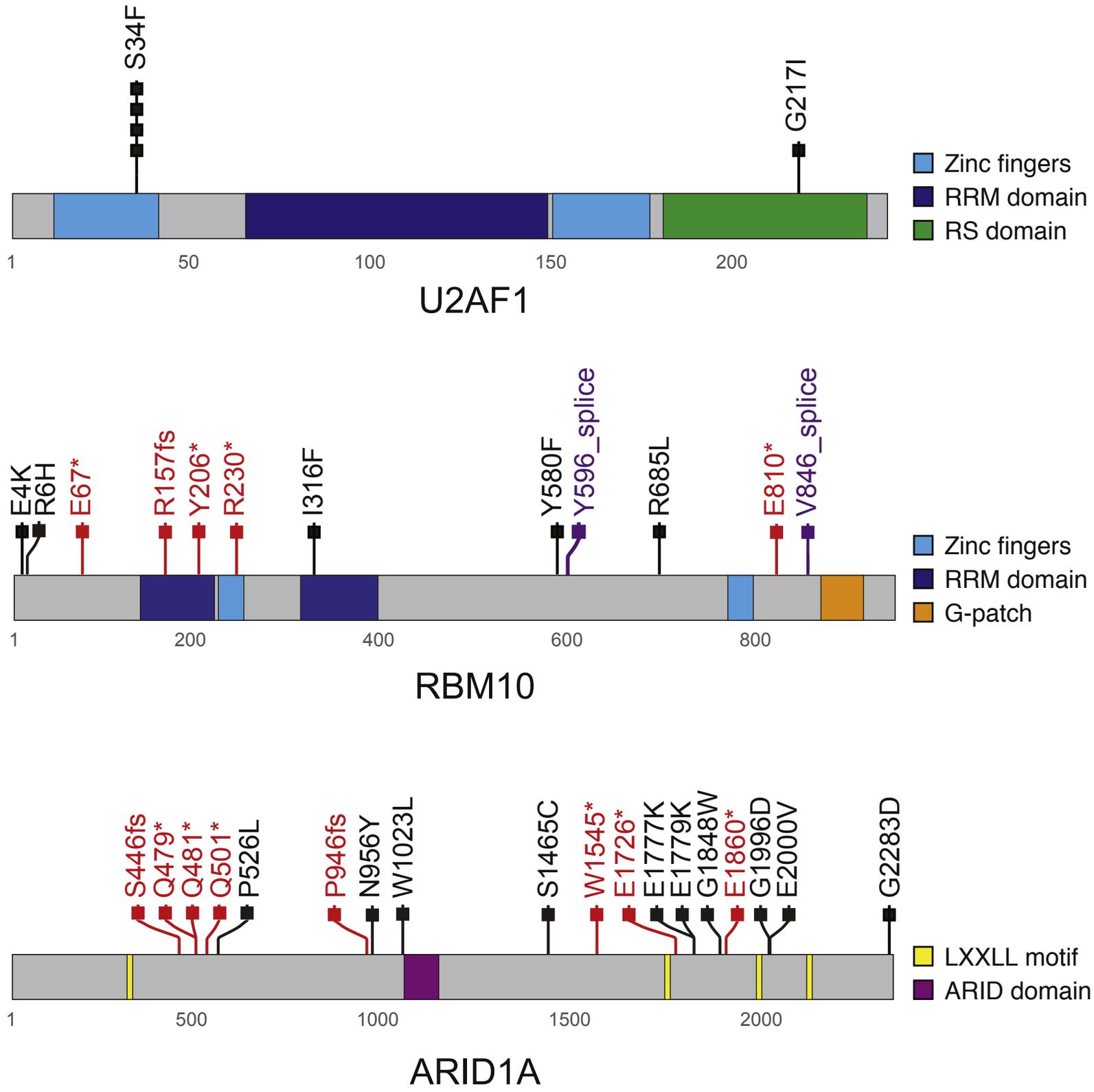
Imielinski M. *et al. Cell* 150:1107 (2012).

VARIETY OF GENE MUTATION PROFILES ACROSS SAMPLES



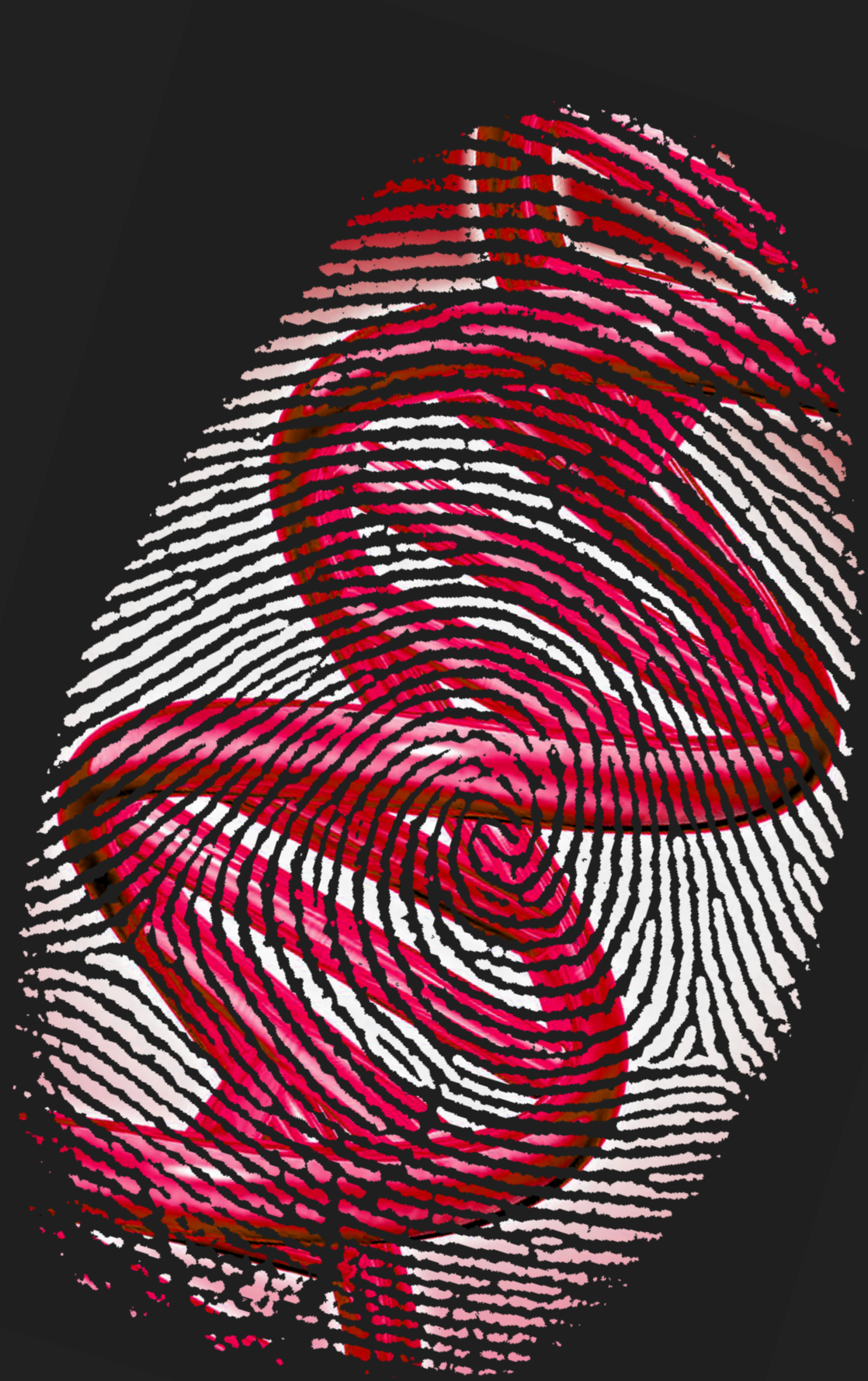
Imielinski M. et al. Cell 150:1107 (2012).

VARIETY OF MUTATIONS ACROSS GENES



Imielinski M. et al. Cell 150:1107 (2012).

oncoprint



http://www.cbioportal.org/public-portal/faq.jsp

Memorial Sloan-Kettering Cancer Center

cBio Cancer Genomics Portal

Visualize, analyze, discover.

HOME | TUTORIALS | NEWS | FAQ | DATA SETS | ABOUT | WEB API | R/MATLAB | NETWORKS

TCGA-BK-A0CC FEMALE, 69 years old
Uterine Corpus Endometrioid Carcinoma (TCGA, Provisional), Serous, Stage III, Grade 3

More about this patient
Living (10 months), DiseaseFree (10 months)

Summary | Mutations | Copy Number Alterations | Pathology Report

Mutations of interest (4 of 35) Search:

Gene	Protein Change	Type	Cohort	COSMIC	FIS	3D
TP53	*Q331**	NS		16		
PPP2R1A	*S256F*	MS		11	H	3D
FAT1	*E314**	NS				
EPHA7	*H408Q*	MS			M	3D

Show all 35 mutations — Show 25 per page

CNA of interest (17 of 385) Search:

Gene	CNA	% in Cohort
MCL1	AMP	
PSMD4	AMP	
ZNF687	AMP	
PI4KB	AMP	
RFX5	AMP	
PIP5K1A	AMP	
CCNE1	AMP	
BCL7C	AMP	
CTF1	AMP	
FBXL19-AS1	AMP	
FBXL19	AMP	
ORAI3	AMP	
SETD1A	AMP	
HSD3B7	AMP	
STX1B	AMP	
CEBPA	AMP	
EGFR	AMP	

Show all 385 CNAs — Show 25 per page

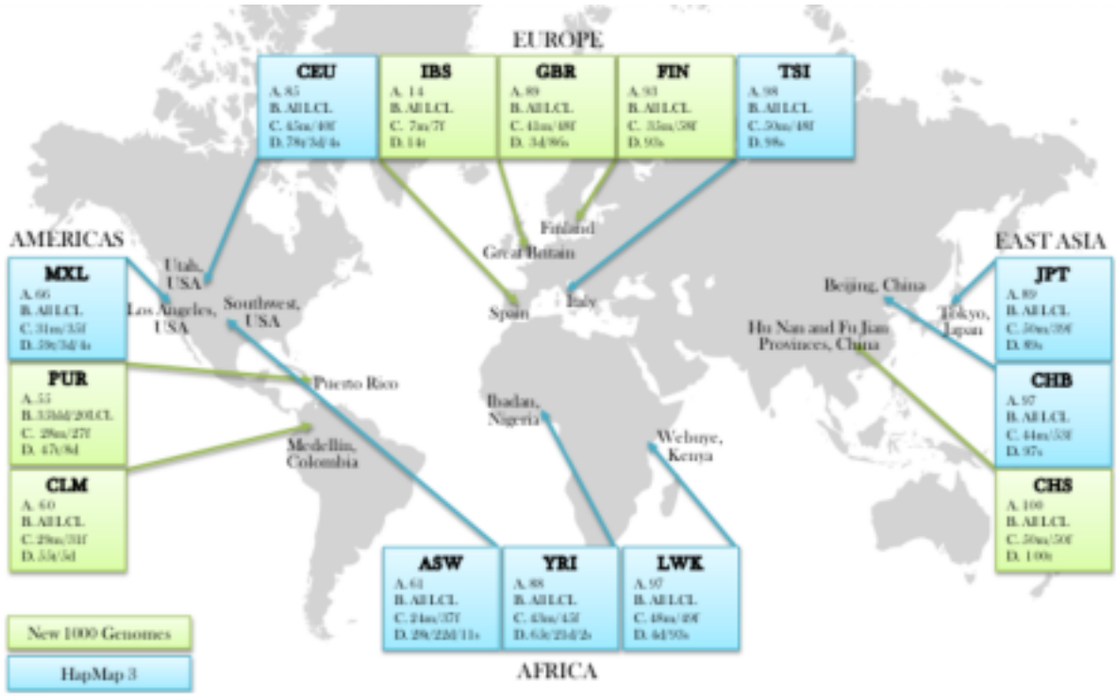
cBio@MSKCC | MSKCC | TCGA
Questions and feedback: cbioportal@googlegroups.com.

http://www.cbioportal.org/public-portal/tumormap.do?case_id=TCGA-BK-A0CC&cancer_study_id=ucec_tcga

YOU CAN'T PUBLISH 100 GENOMES ANYMORE:
1000 GENOMES PROJECT

Summary of 1000 Genomes Project phase I data

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



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

Genomes Project Consortium. *Nature* 467:1061 (2010).

efficient algorithms
FIND DIFFERENCES IN GENOMES

graphs and networks
ASSEMBLE GENOME SEQUENCE

clustering
FIND PATTERNS IN GENE EXPRESSION

text mining
DISCOVER BIOLOGICAL RELATIONSHIPS

visualization

Drew Shenman, New Jersey - The Newark Star Ledger

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.

Compeau P.E. et al. Nature Biotechnology 29:987 (2011).

DE BRUIJN GRAPH

Box 1 Origin of de Bruijn graphs

In 1946, the Dutch mathematician Nicolaas de Bruijn became interested in the 'superstring problem'[12]: find a shortest circular 'superstring' that contains all possible 'substrings' of length k (k-mers) over a given alphabet. There exist n^k k-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 $(k-1)$-mer is assigned to a node; connect one $(k-1)$-mer by a directed edge to a second $(k-1)$-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.

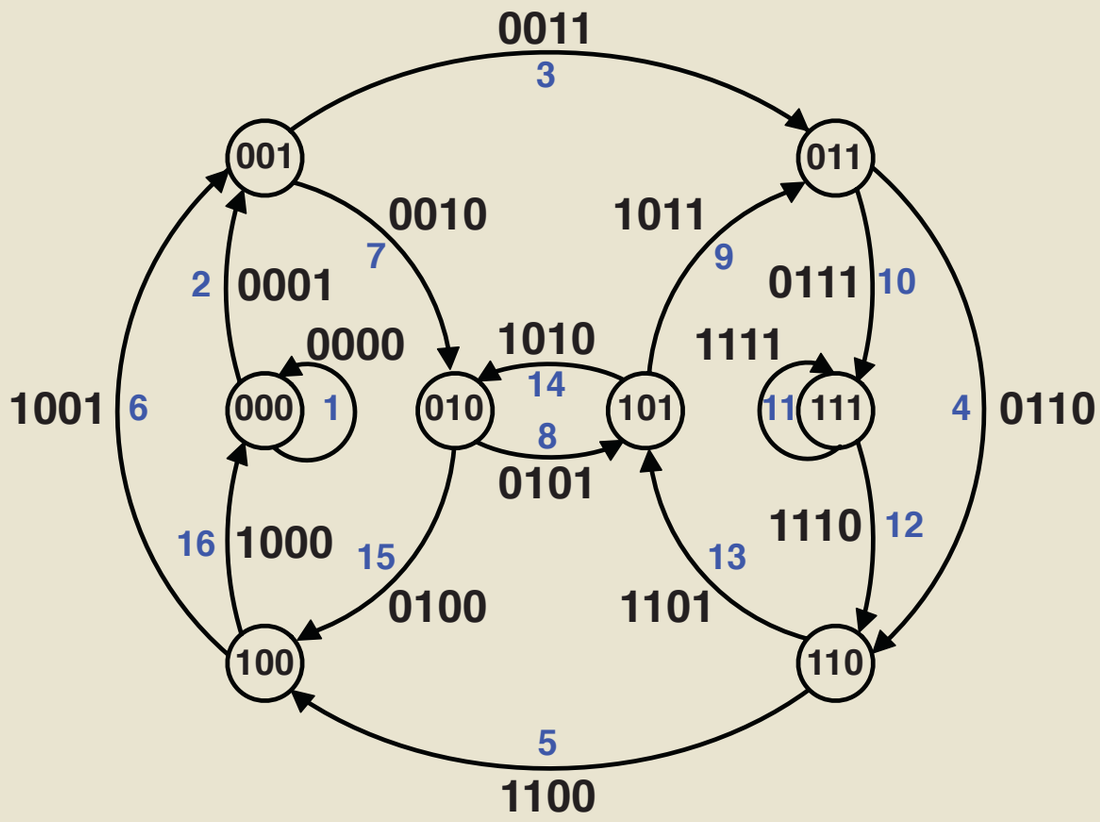
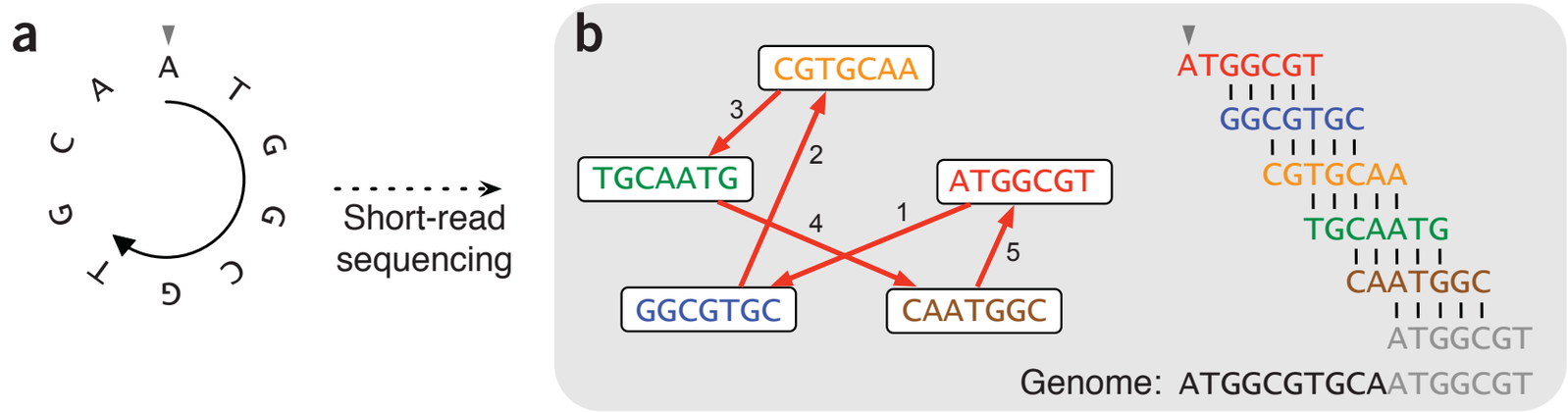


Figure 2 De Bruijn graph. The de Bruijn graph B for $k = 4$ and a two-character alphabet composed of the digits 0 and 1. This graph has an Eulerian cycle because each node has indegree and outdegree equal to 2. Following the blue numbered edges in order from 1 to 16 traces an Eulerian cycle **0**000, **0**001, **0**011, **0**110, **1**100, **1**001, **0**010, **0**101, **1**011, **0**111, **1**111, **1**110, **1**101, **1**010, **0**100, **1**000. Recording the first character (in boldface) of each edge label spells the cyclic superstring **0000110010111101**.

Compeau P.E. et al. *Nature Biotechnology* 29:987 (2011).

DE BRUIJN GRAPHS FOR ASSEMBLY

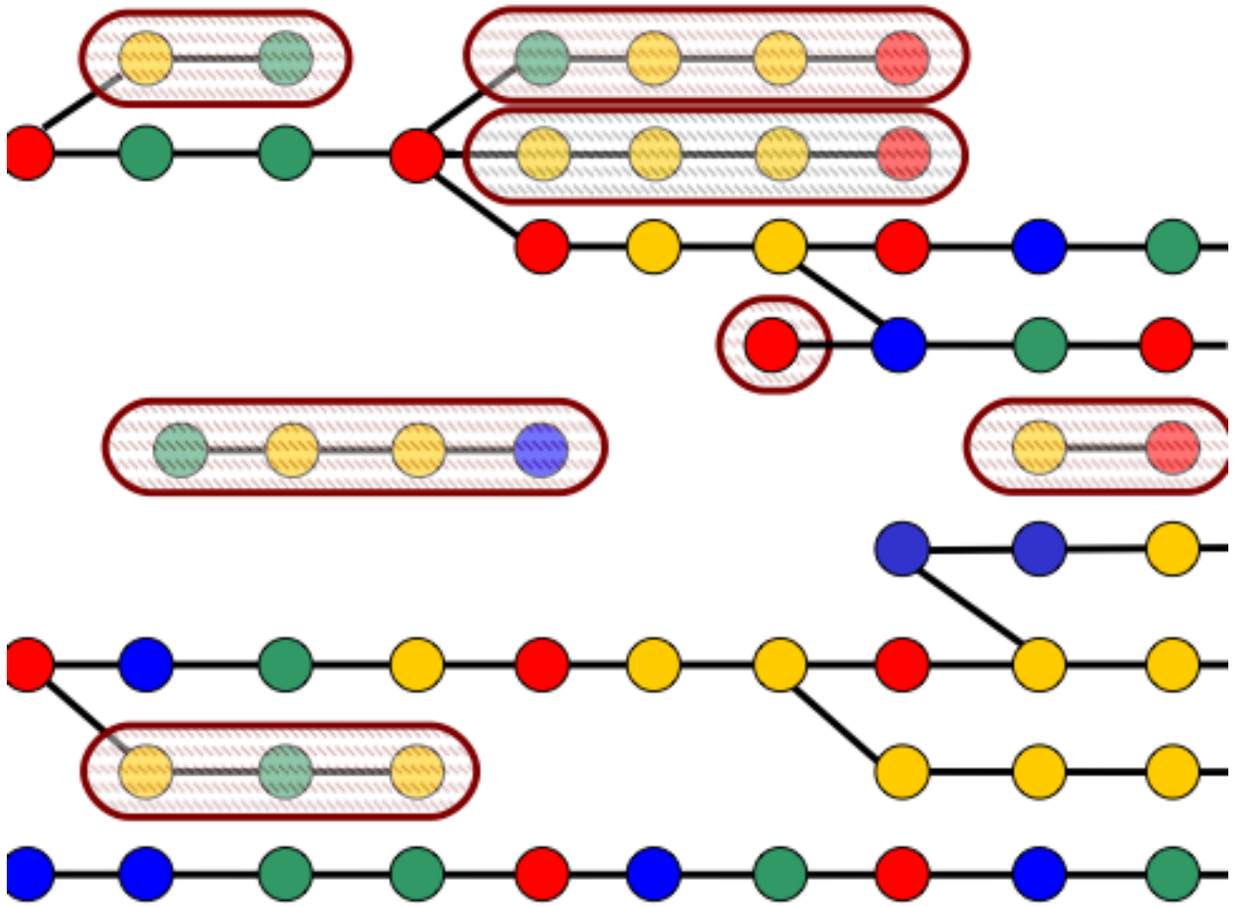


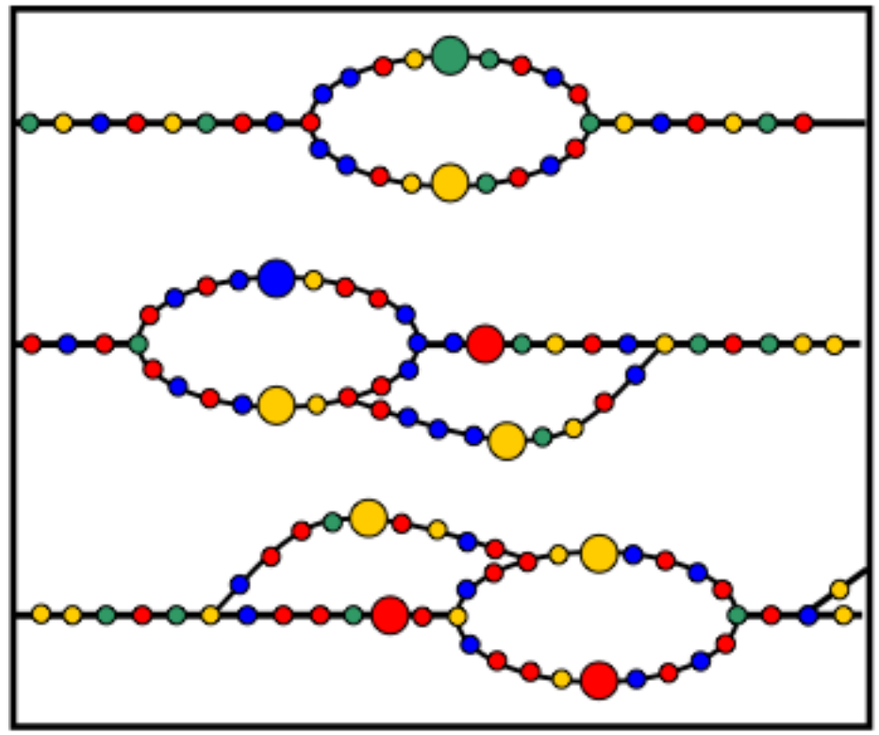
Hamiltonian cycle
Visit each vertex once
(harder to solve)

Eulerian cycle
Visit each edge once
(easier to solve)

Compeau P.E. et al. *Nature Biotechnology* 29:987 (2011).

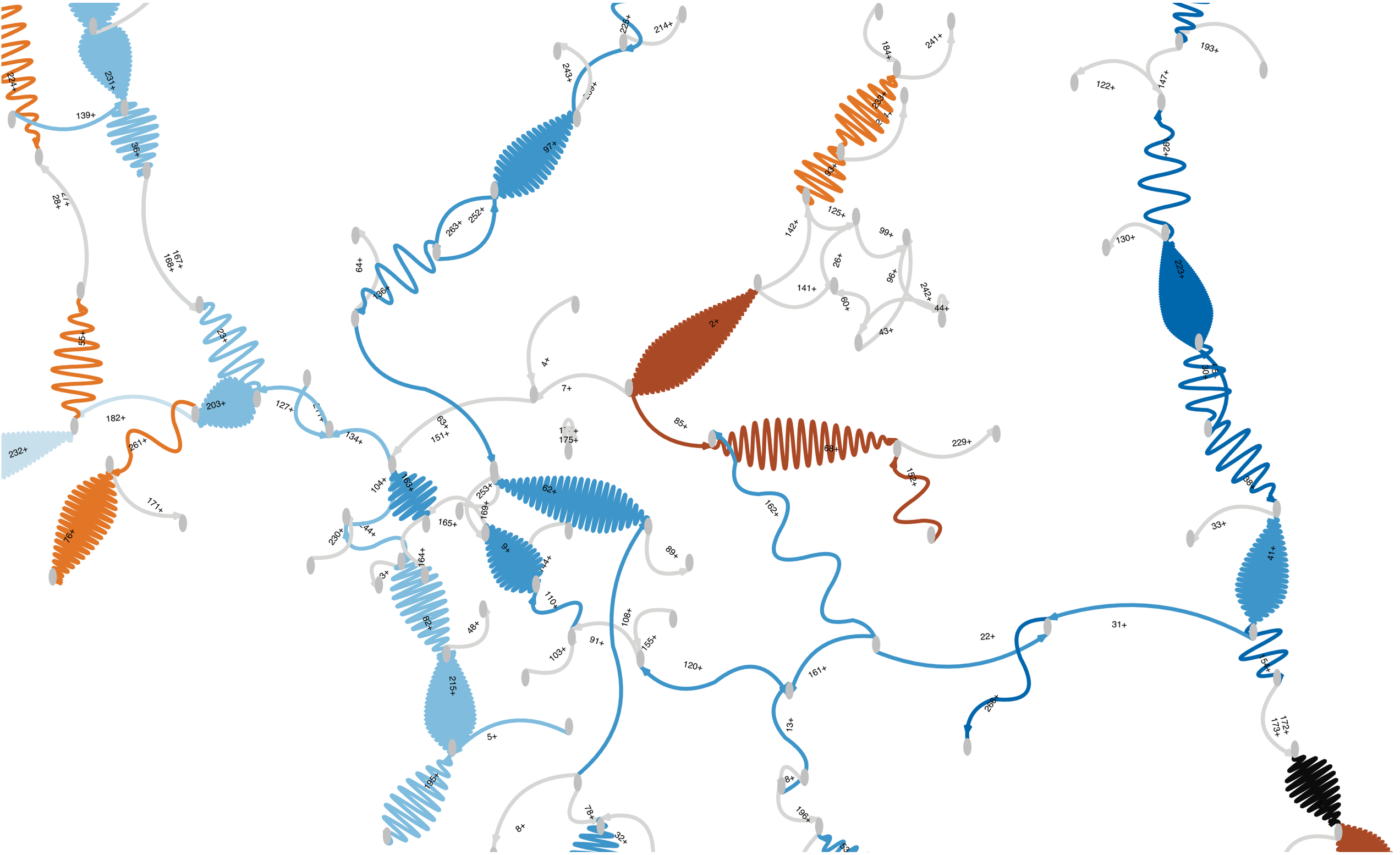
ERROR CORRECTION — MANY ANSWERS ARE POSSIBLE





Courtesy of Shaun Jackman and the ABySS (short read assembler) team.
Simpson J. *et al. Genome Res* 19:1117 (2009).

EXPLORING GENOME ASSEMBLIES — ABYSS EXPLORER



Nielsen C. et al. InfoVis (2009).

efficient algorithms

FIND DIFFERENCES IN GENOMES

graphs and networks

ASSEMBLE GENOME SEQUENCE

clustering

FIND PATTERNS IN GENE EXPRESSION

text mining

DISCOVER BIOLOGICAL RELATIONSHIPS

visualization

micro RNA

expression regulation

increased expression

C → C'
4x

gene
~20,000

exon
~9 exons/gene

alternate transcripts

~6 transcripts/gene

A → A'

B → B'

3×10^9 base pairs

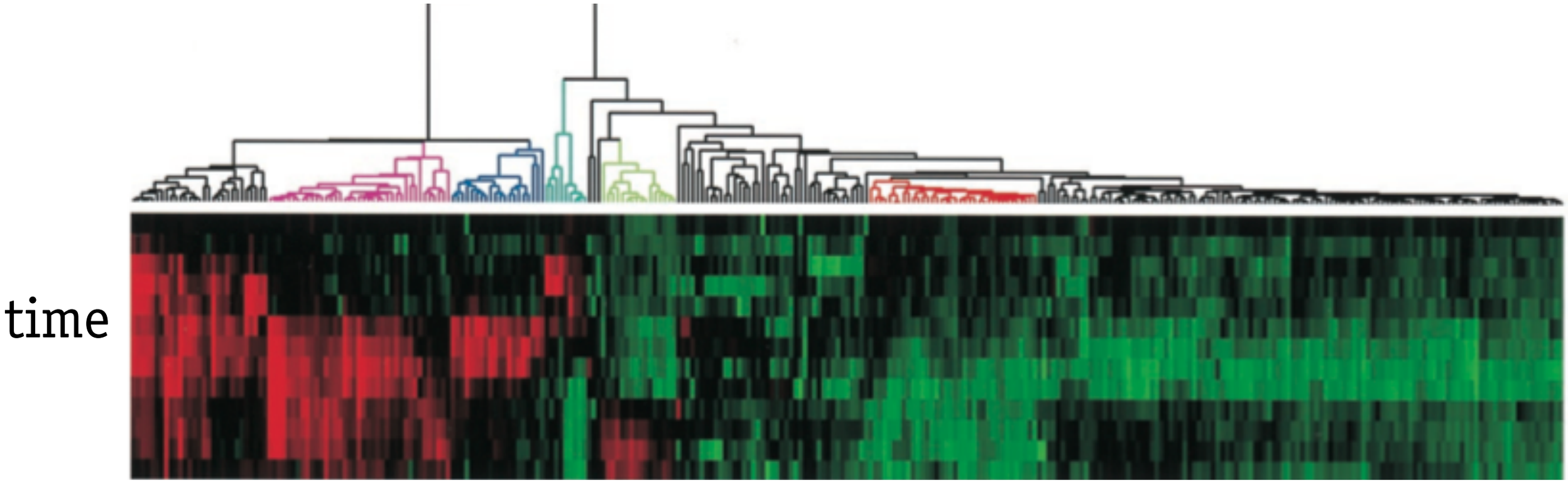
DNA

RNA

protein

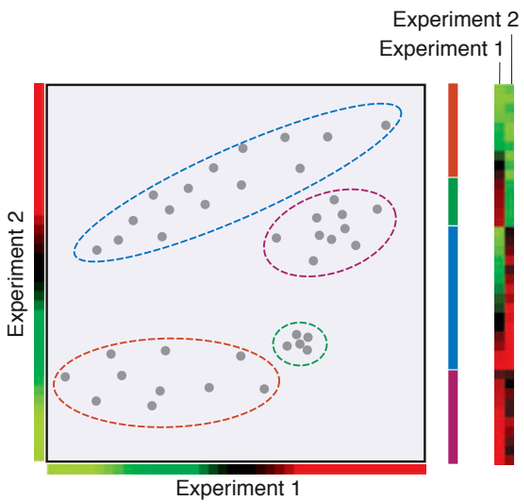
reaction

HIERARCHICAL CLUSTERING

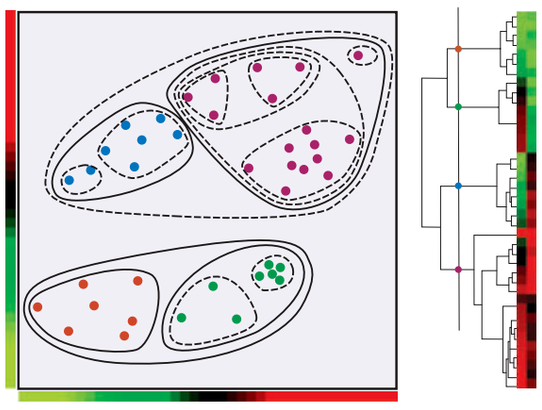


Sharov AA, Dudekula DB, Ko MS (2005) Genome-wide assembly and analysis of alternative transcripts in mouse. *Genome Res* 15: 748-754.

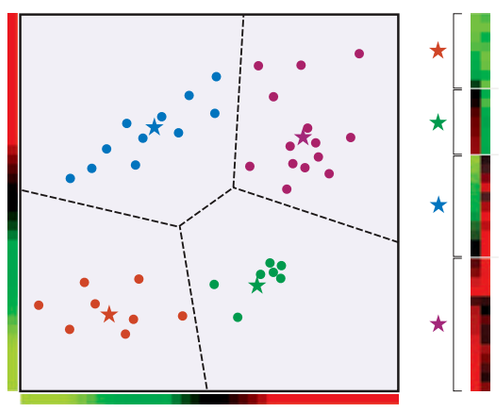
CLUSTERING METHODS



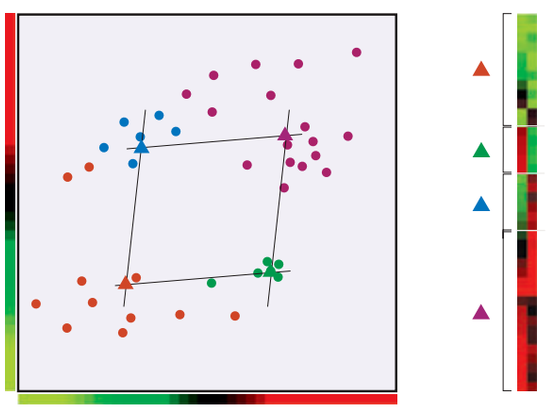
Hierarchical



k-means



SOM



k-means & SOM better than hierarchical

complete better than single linkage

Euclidian distance for log ratio data

Pearson correlation for absolute data

D'haeseleer P. *Nature Biotechnology* 23:1499 (2005).

CORRELATION IS THE NEW CAUSATION

OPEN ACCESS Freely available online

PLoS COMPUTATIONAL BIOLOGY

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

David Venet[1], Jacques E. Dumont[2], Vincent Detours[2,3]*

1 IRIDIA-CoDE, Université Libre de Bruxelles (U.L.B.), Brussels, Belgium, **2** IRIBHM, Université Libre de Bruxelles (U.L.B.), Campus Erasme, Brussels, Belgium, **3** WELBIO, Université Libre de Bruxelles (U.L.B.), 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.

Citation: Venet D, Dumont JE, Detours V (2011) Most Random Gene Expression Signatures Are Significantly Associated with Breast Cancer Outcome. PLoS Comput Biol 7(10): e1002240. doi:10.1371/journal.pcbi.1002240

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

Funding: DV was funded by the IRSIB 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

Venet D. et al. PLoS Computational Biology 7:e1002240 (2011).

CORRELATION IS THE NEW CAUSATION

OPEN ACCESS Freely available online

PLoS COMPUTATIONAL BIOLOGY

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

David Venet[1], Jacques E. Dumont[2], Vincent Detours[2,3]*

1 IRIDIA-CoDE, Université Libre de Bruxelles (U.L.B.), Brussels, Belgium, **2** IRIBHM, Université Libre de Bruxelles (U.L.B.), Campus Erasme, Brussels, Belgium, **3** WELBIO, Université Libre de Bruxelles (U.L.B.), 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.

Citation: Venet D, Dumont JE, Detours V (2011) Most Random Gene Expression Signatures Are Significantly Associated with Breast Cancer Outcome. PLoS Comput Biol 7(10): e1002240. doi:10.1371/journal.pcbi.1002240

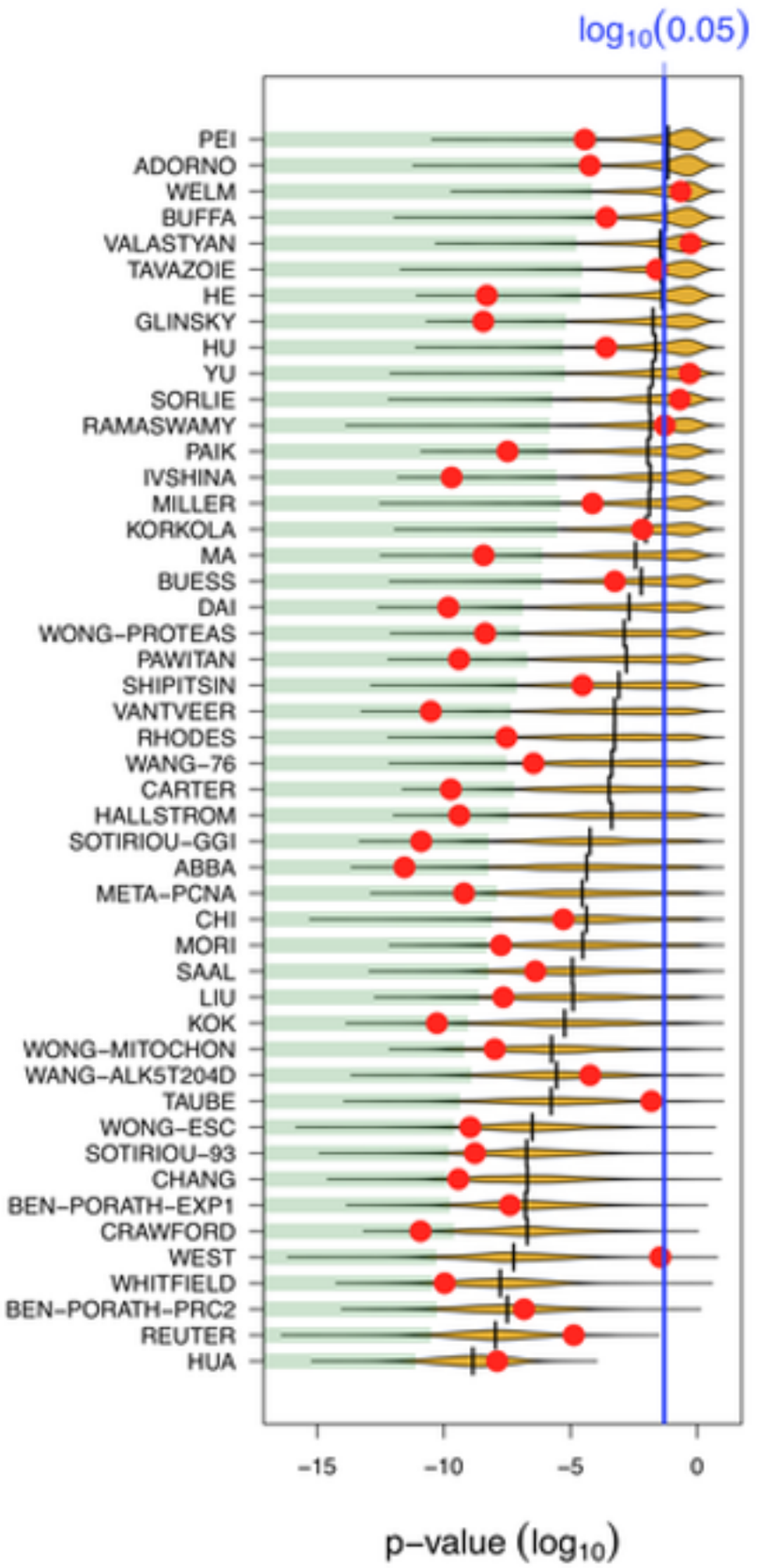
Editor: Isidore Rigoutsos, Jefferson Medical College/Thomas Jefferson University, United States of America

Received April 27, 2011; **Accepted** September 7, 2011; **Published** October 20, 2011

Funding: DV was funded by the IRSIB 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



Venet D. et al. PLoS Computational Biology 7:e1002240 (2011).

CORRELATION IS THE NEW CAUSATION

OPEN ACCESS Freely available online

PLoS COMPUTATIONAL BIOLOGY

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

David Venet[1], Jacques E. Dumont[2], Vincent Detours[2,3]*

1 IRIDIA-CoDE, Université Libre de Bruxelles (U.L.B.), Brussels, Belgium, **2** IRIBHM, Université Libre de Bruxelles (U.L.B.), Campus Erasme, Brussels, Belgium, **3** WELBIO, Université Libre de Bruxelles (U.L.B.), 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.

Citation: Venet D, Dumont JE, Detours V (2011) Most Random Gene Expression Signatures Are Significantly Associated with Breast Cancer Outcome. PLoS Comput Biol 7(10): e1002240. doi:10.1371/journal.pcbi.1002240

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

Funding: DV was funded by the IRSIB 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

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.**

Venet D. et al. PLoS Computational Biology 7:e1002240 (2011).

our genome is the result of a single
on-going Monte Carlo simulation

EVOLUTION

general rules are elusive

efficient algorithms

FIND DIFFERENCES IN GENOMES

graphs and networks

ASSEMBLE GENOME SEQUENCE

clustering

FIND PATTERNS IN GENE EXPRESSION

text mining

DISCOVER BIOLOGICAL RELATIONSHIPS

visualization

data flood[1]

tsunamis[2]

deluges[3]

surging oceans[4]

avalanches[5]

icebergs[6]

landslide[7]

earthquakes[8]

explosions[9]

1. Andrade M et al. Curr Opin Biotechnol 8:675 (1997). 2. Wurman RS. Information Architects (1997). 3. Hess K et al. Trends Biotechnol 19:463 (2001), Editorial Nat Biotechnol 26:1099 (2008). 4. Dubitzky W. Brief Bioinform 10:343 (2009). 5. Antezana E et al. Brief Bioinform 10:392 (2009). 6. Hodgson C. Nat Biotechnol 19:BE44 (2001). Howe D et al. Nature 455:47 (2008). 7. Attwood T et al. Biochem J 424:317 (2009). 8. Whilbanks J. CTWatchQuarterly (2007). 9. Diehn M. et al. Nucleic Acids Res 31:219 (2003).

Mardis *Genome Medicine* 2010, **2**:84
http://genomemedicine.com/content/2/11/84

Genome Medicine

MUSINGS

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 required for it to occur. I therefore offer the following as food for thought.

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

Mardis E.R. Genome Medicine 2:84 (2010).

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

“Without structure,
data are mere babble.”

Attwood et al. Biochem J. 424:317 (2009).

UNDISCOVERED PUBLIC KNOWLEDGE

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.

Swanson D.R. *Perspectives in Biology and Medicine* 30:7 (1986)
Swanson, D.R. *Bull Med Libr Assoc* 78:29 (1990).

INTEGRATIVE BIOLOGY THROUGH TEXT-MINING

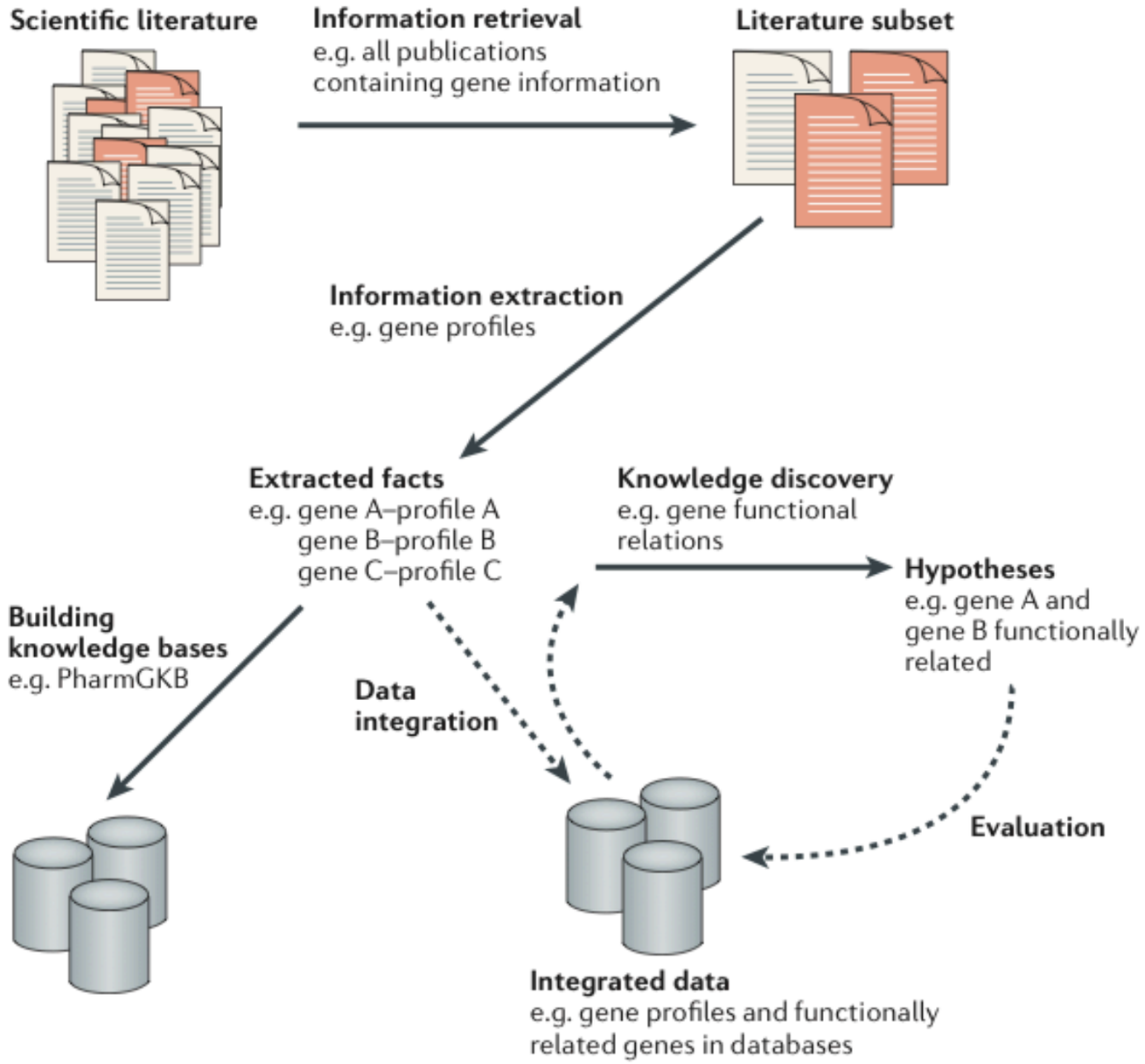
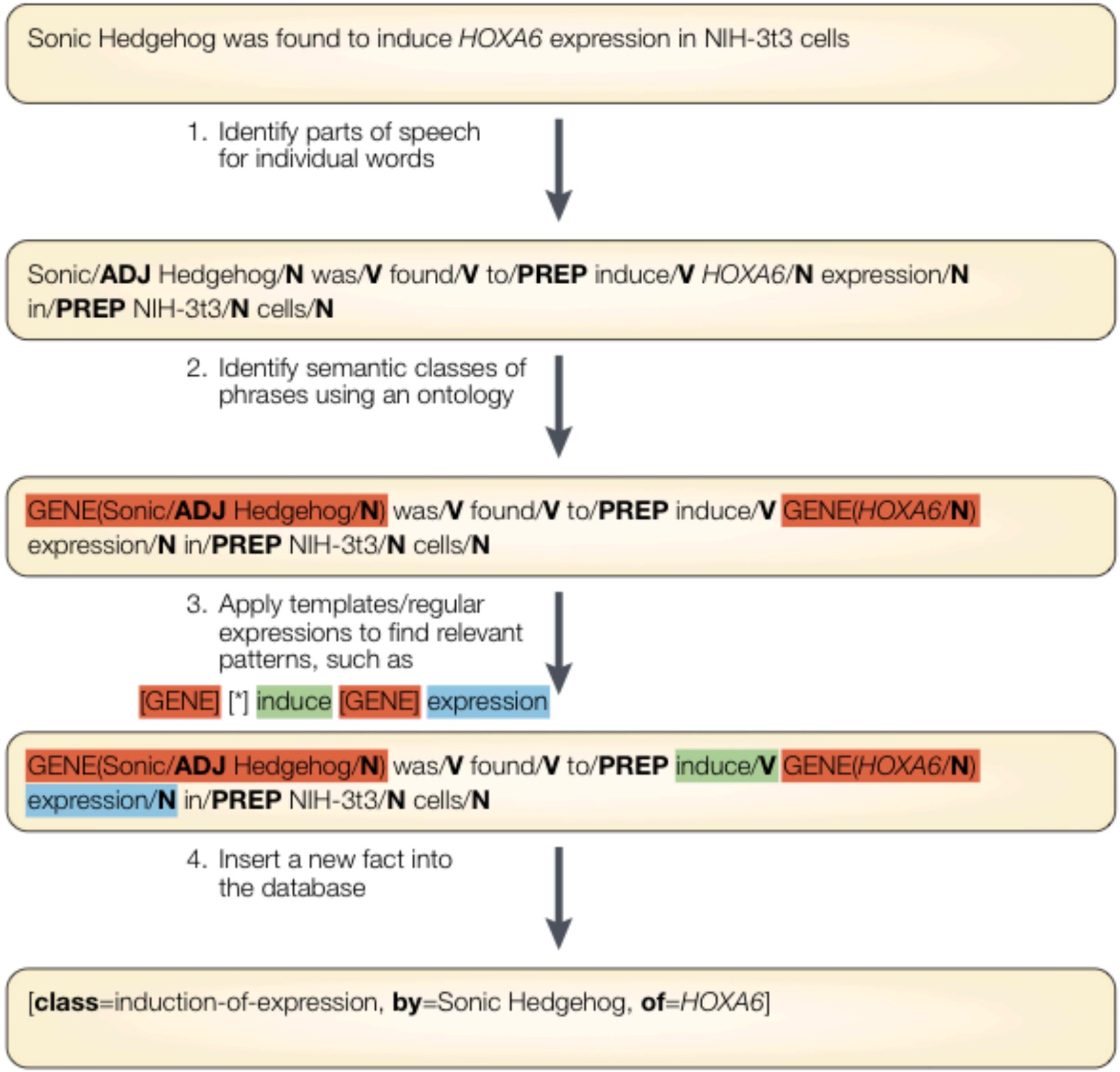


Figure 1 | **Categories of text-mining solutions.** The diagram gives an overview of the different categories of situations in which text mining is applied. Document retrieval is the initial step and leads to the collection of documents for a given query. The other solutions target the identification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

Rebholtz-Schuhmann et al. *Nature Reviews Genetics* 13:829 (2012).

GENE NAME RECOGNITION AND IDENTIFICATION



Yandell et al. *Nature Reviews Genetics* 3:601 (2002).

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

Cohen et al. *PLoS Computational Biology* 4:e20 (2008).
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, GPIV, CHDS7, PASIV, SCARB3, BDPLT10
atherosclerosis, insulin resistance

Cohen et al. PLoS Computational Biology 4:e20 (2008).

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.

http://www.npr.org/templates/story/story.php?storyId=100468532

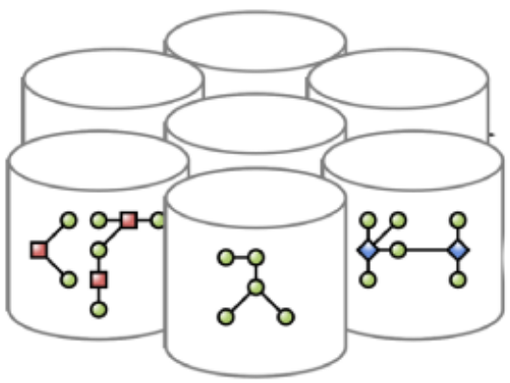
Liekens *et al. Genome Biology* 2011, **12**:R57
http://genomebiology.com/2011/12/6/R57

SOFTWARE Open Access

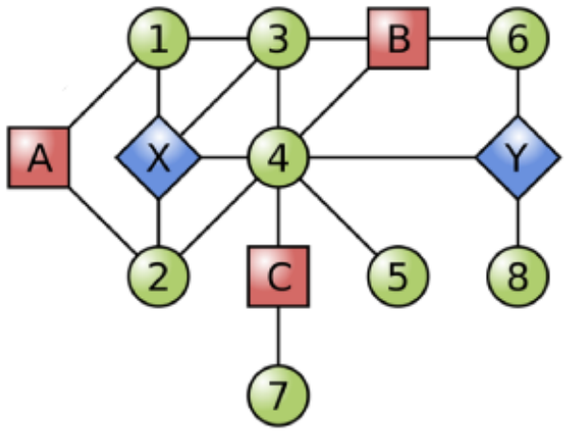
BioGraph: unsupervised biomedical knowledge discovery via automated hypothesis generation

Anthony ML Liekens[1*], Jeroen De Knijf[2], Walter Daelemans[3], Bart Goethals[2], Peter De Rijk[1] and Jurgen Del-Favero[1]

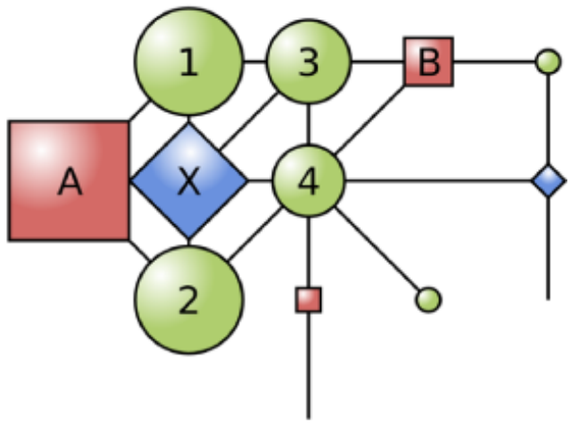
public knowledge bases



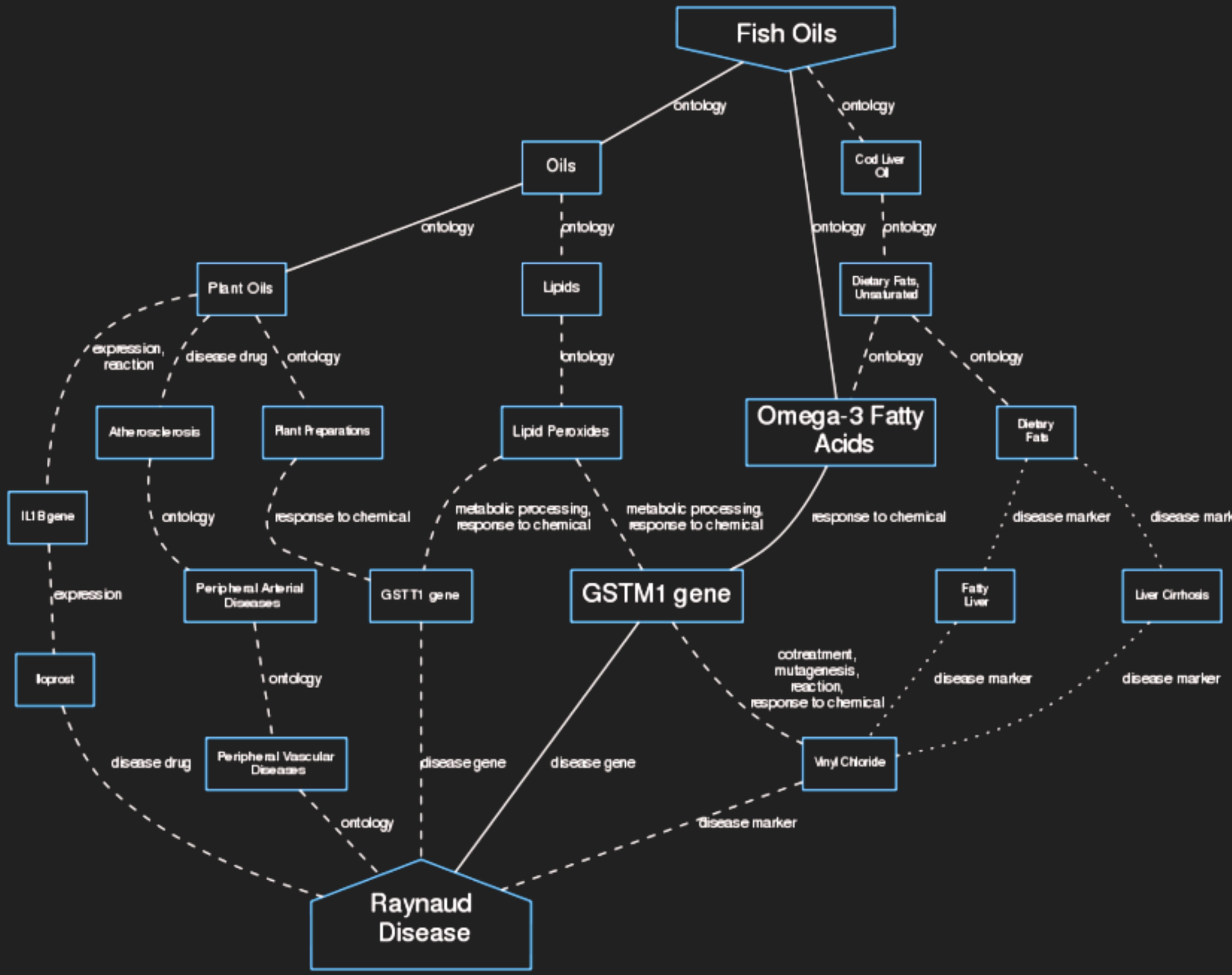
are integrated to connect genes, diseases and proteins



in a weighted heterogeneous network



adapted from Figure 1 in Liekens A. et al. *Genome Biology* 12:R57 (2011).



Liekens A. et al. *Genome Biology* 12:R57 (2011).
http://biograph.be/graphs/3362-7362.dot.png

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.

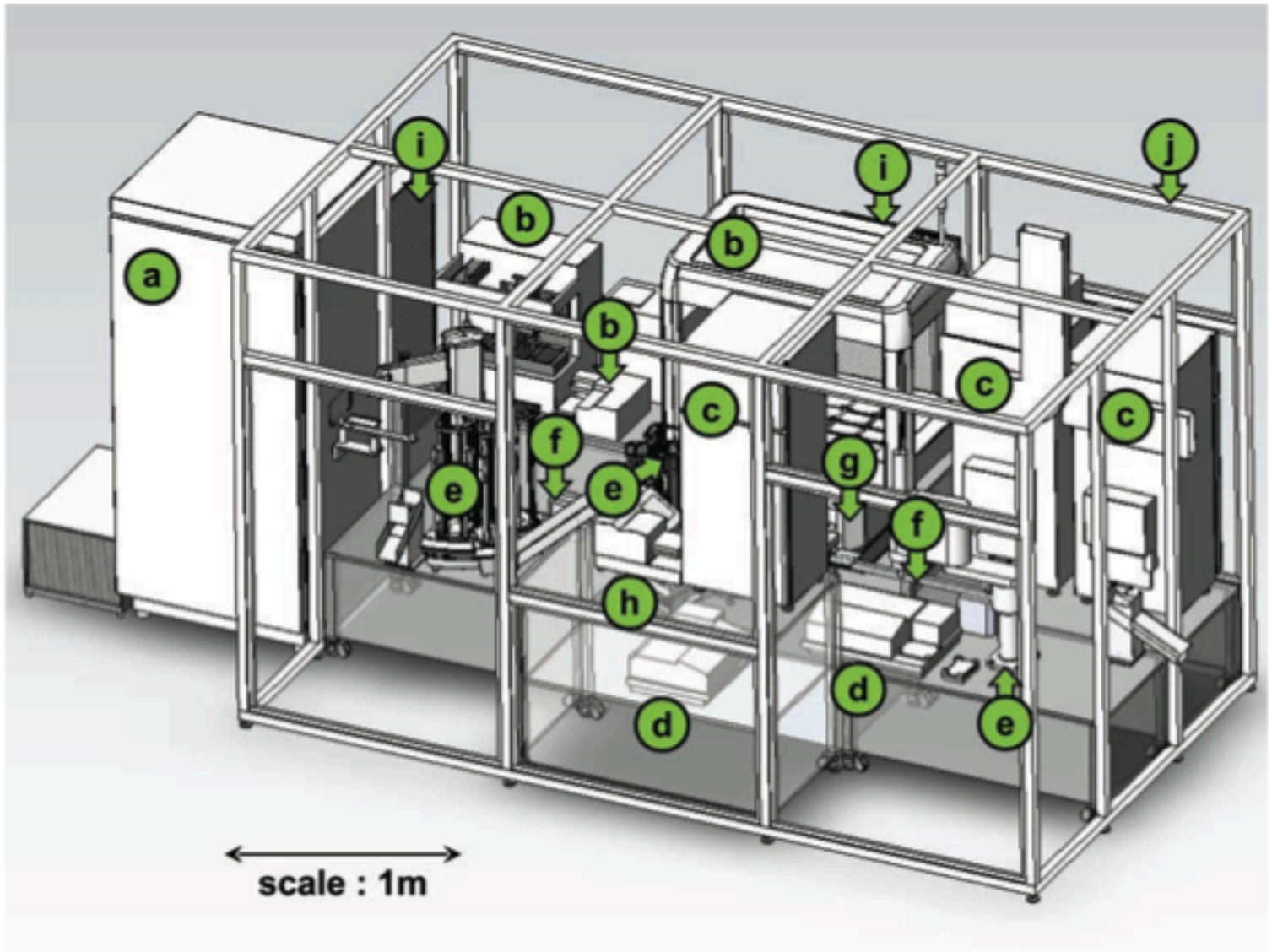
"We are committed to change and innovation that will make science more effective."

Brussels Declaration on Scientific, Technical and Medical Publishing

Akademie Verlag
American Chemical Society
American Institute of Physics
Blackwell Publishing
British Medical Journal Group
Carocci Editore
C. G. Edizioni Medico Scientifiche
Cambridge University Press
Carl Hanser Verlag
Clueb
De Agostini Editore
De Agostini Scuola
Editoriale Folini
Egea
Edinburgh University Press
Elsevier
Elsevier Masson
E. Schweizbart'sche Verlagsbuchhandlung Science Pub
Federico Motta Editore
Institute of Physics Publishing
Gebr. Bomtraeger Science Publishers
Georg Olms Verlag
Georg Thieme Verlag
Groupe de Boeck
Guerini e Associati
John Wiley & Sons
Lippincott Williams & Wilkins
Macmillan Publishers
Multi-Science Publishing Co. Ltd
Nature Publishing Group
Oldenbourg Verlag
Oxford University Press
Portland Press
Provestia Publishing House
Primula Edizioni
Royal Society of Chemistry
S. Hirzel Verlag
Sage Publications
Springer Science+Business Media
Taylor & Francis Group
The McGraw Hill Companies (Milano)
The University of Chicago Press
Utet (Torino)
Weidmannsche Verlagsbuchhandlung
Zanichelli Editore

The Automation of Science

Ross D. King,[1]* Jem Rowland,[1] Stephen G. Oliver,[2] Michael Young,[3] Wayne Aubrey,[1] Emma Byrne,[1] Maria Liakata,[1] Magdalena Markham,[1] Pınar Pir,[2] Larisa N. Soldatova,[1] Andrew Sparkes,[1] Kenneth E. Whelan,[1] Amanda Clare[1]



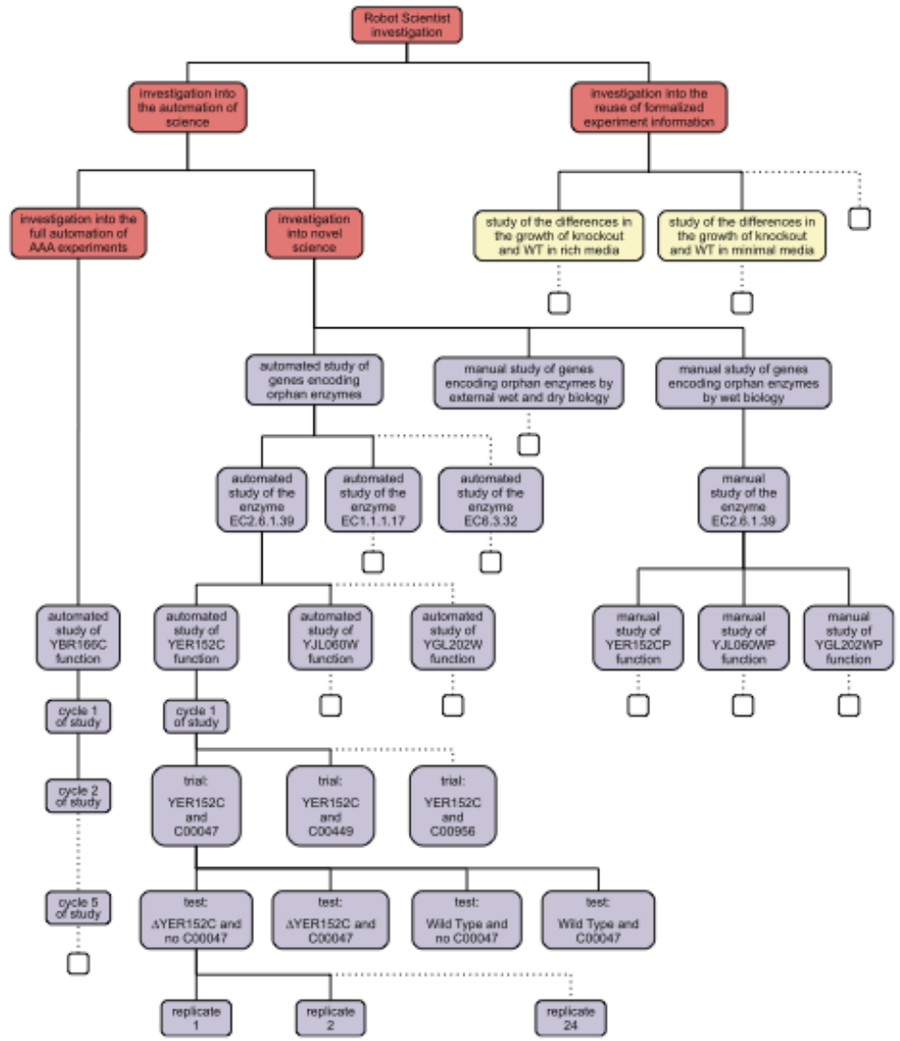
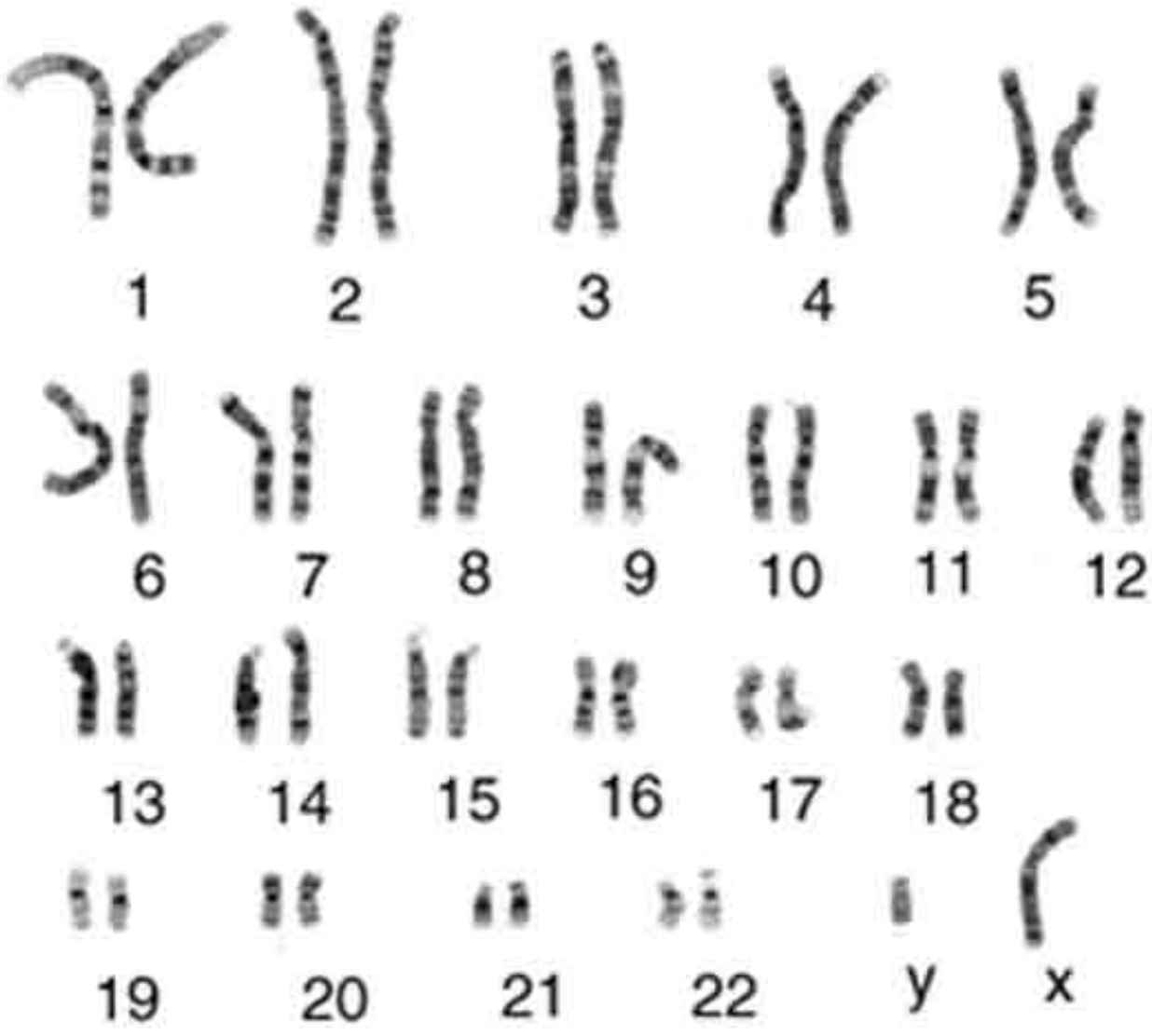


Fig. 3. Structure of the Robot Scientist investigation (a fragment). It consists of two main parts: an investigation into the automation of science and an investigation into the reuse of formalized experiment information. The top levels involve AI research (red), which requires research in functional genomics (blue) and systems biology (yellow). Each level of research unit (studies, cycles, trials, tests, and replicates) is characterized by a specific set of properties (fig. S3) (*16*). Such a nested structure is typical of many scientific experiments, where the testing of a top-level hypothesis requires the planning of many levels of supporting work. What is atypical in Adam's work is the scale and depth of the nesting.

"...we plan to *automatically publish* the logical descriptions of automated experiments."
"What remain to be determined are the limits of automation."

King et al. *Science* 324:85 (2009).

CYTOGENETIC KARYOTYPING



http://www.biotechnologyonline.gov.au/popups/img_karyotype.html

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

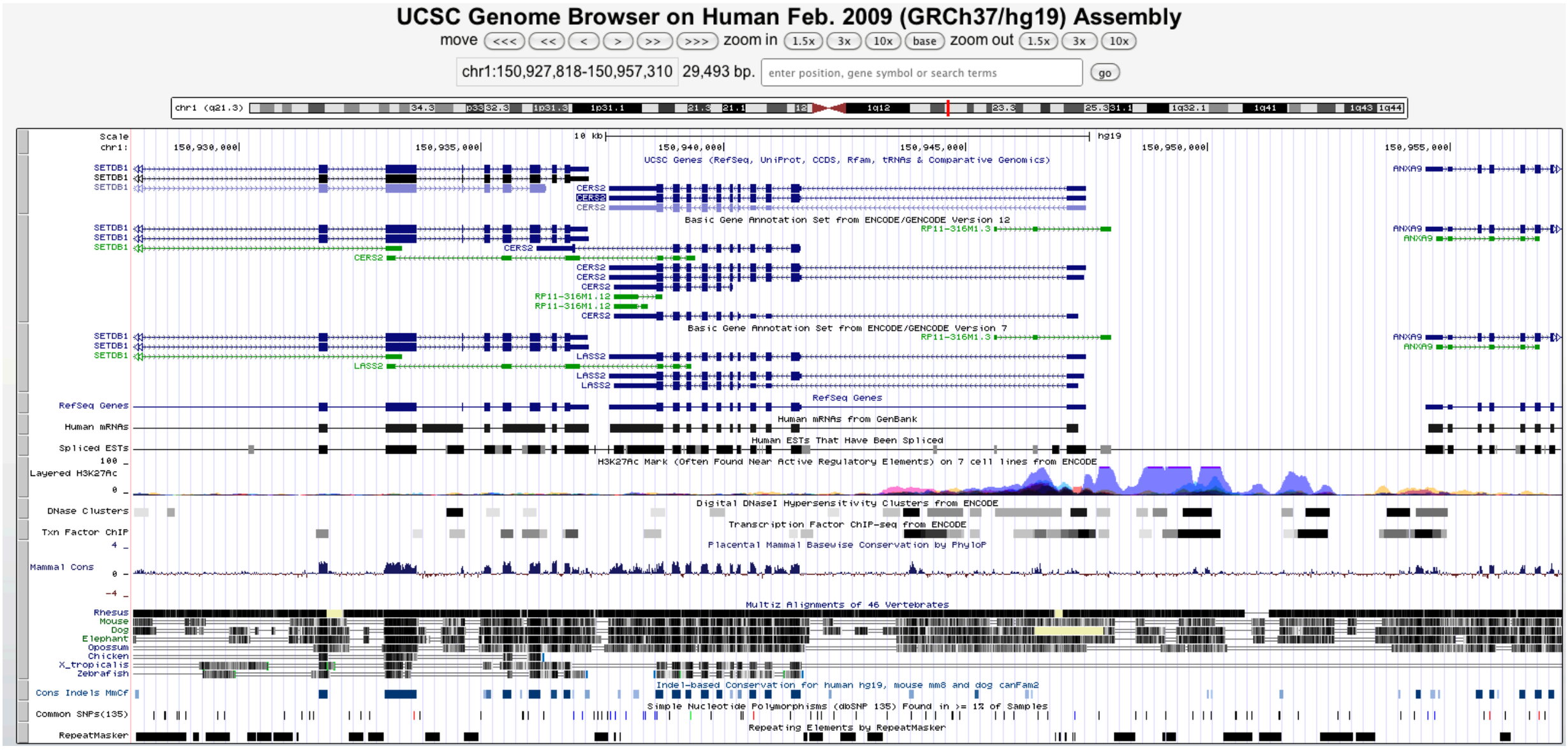
FIND PATTERNS IN GENE EXPRESSION

text mining

DISCOVER BIOLOGICAL RELATIONSHIPS

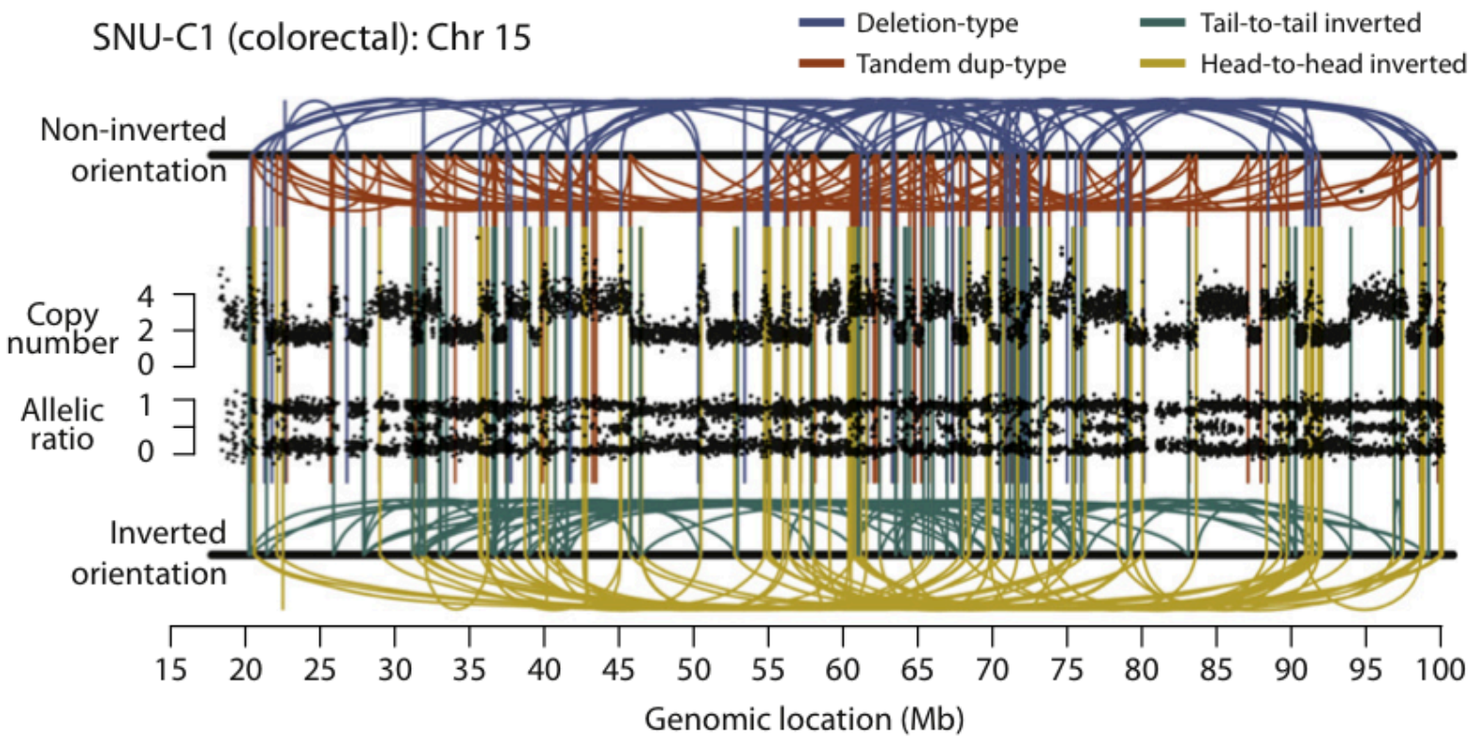
visualization

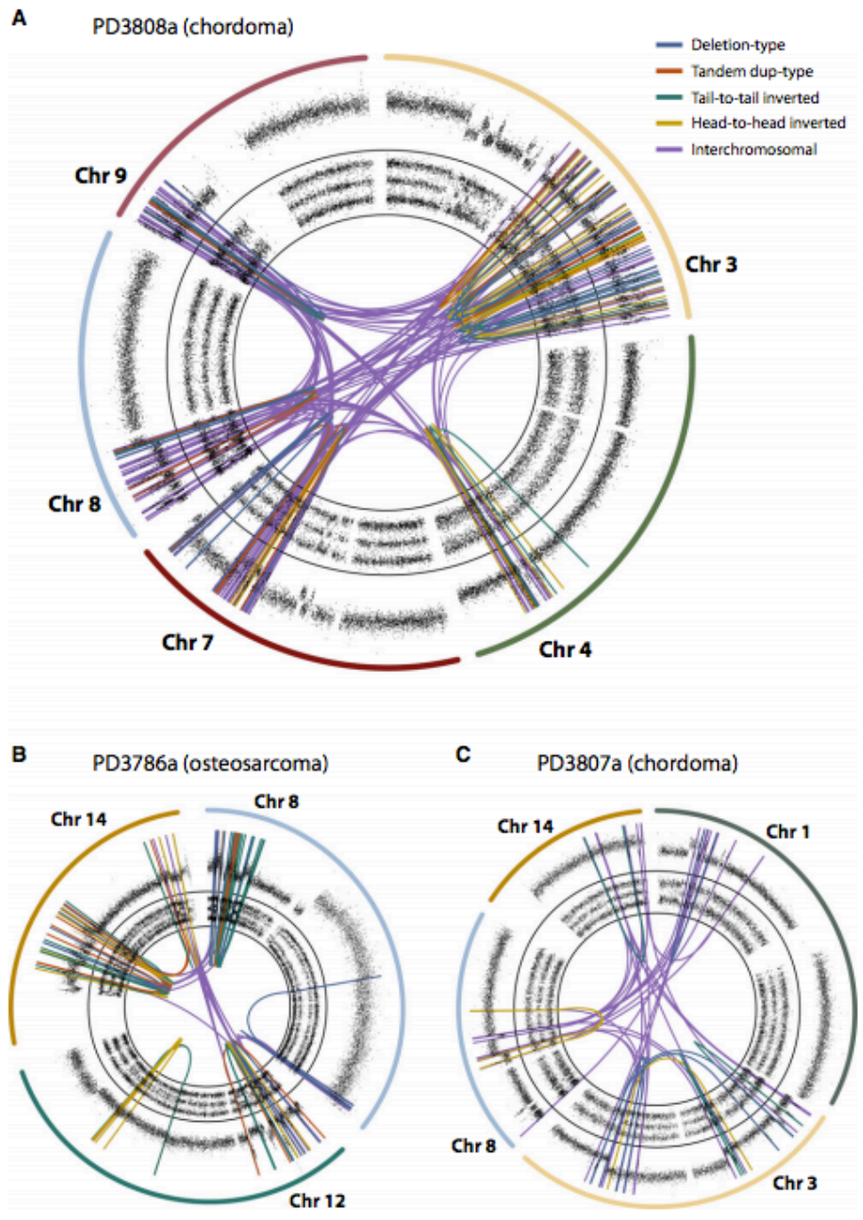
GENOME BROWSER MODEL



UCSC Genome Browser, http://genome.ucsc.edu

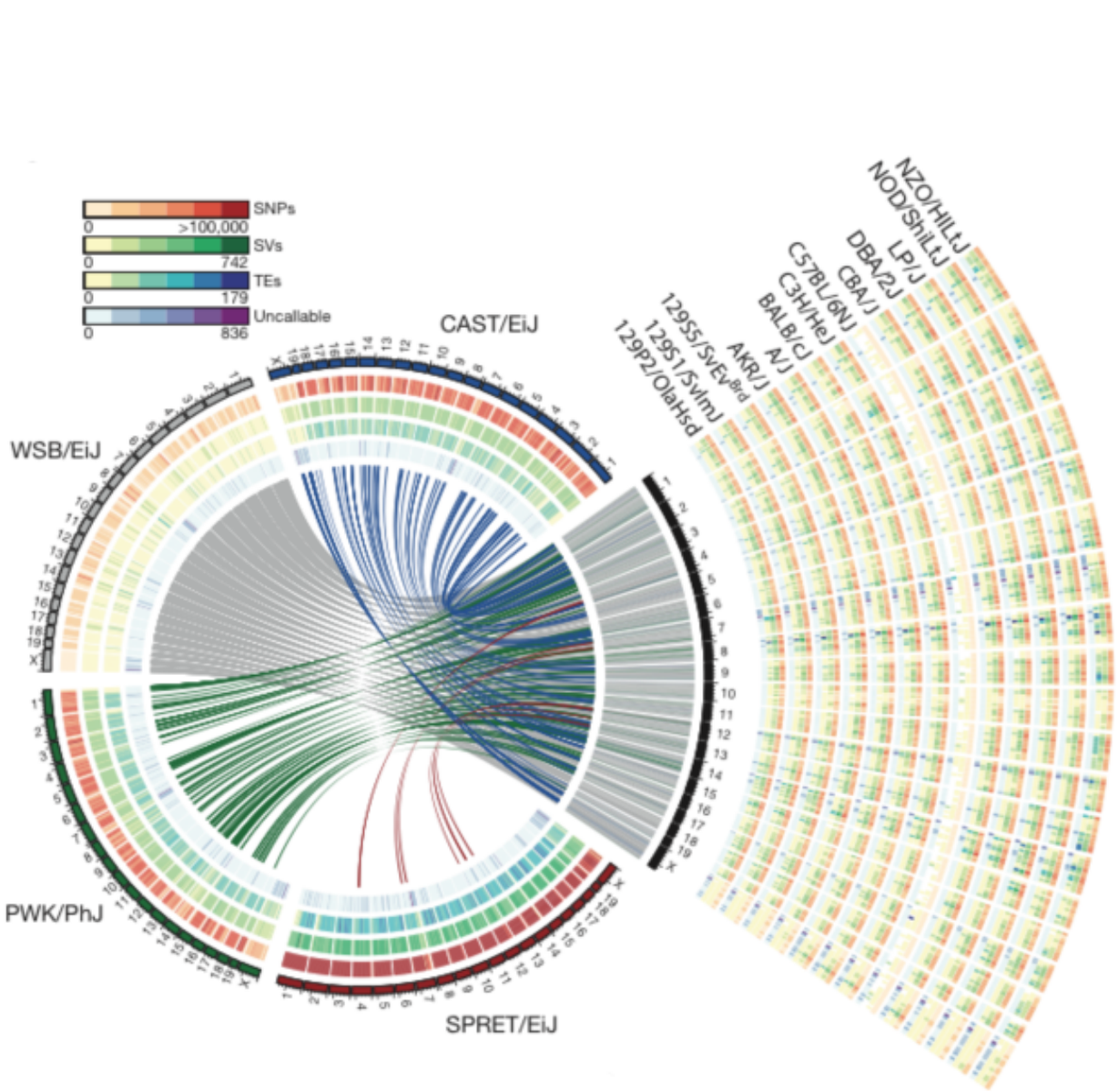
STRUCTURAL CHANGES ARE HARD TO SHOW FOR ONE GENOME

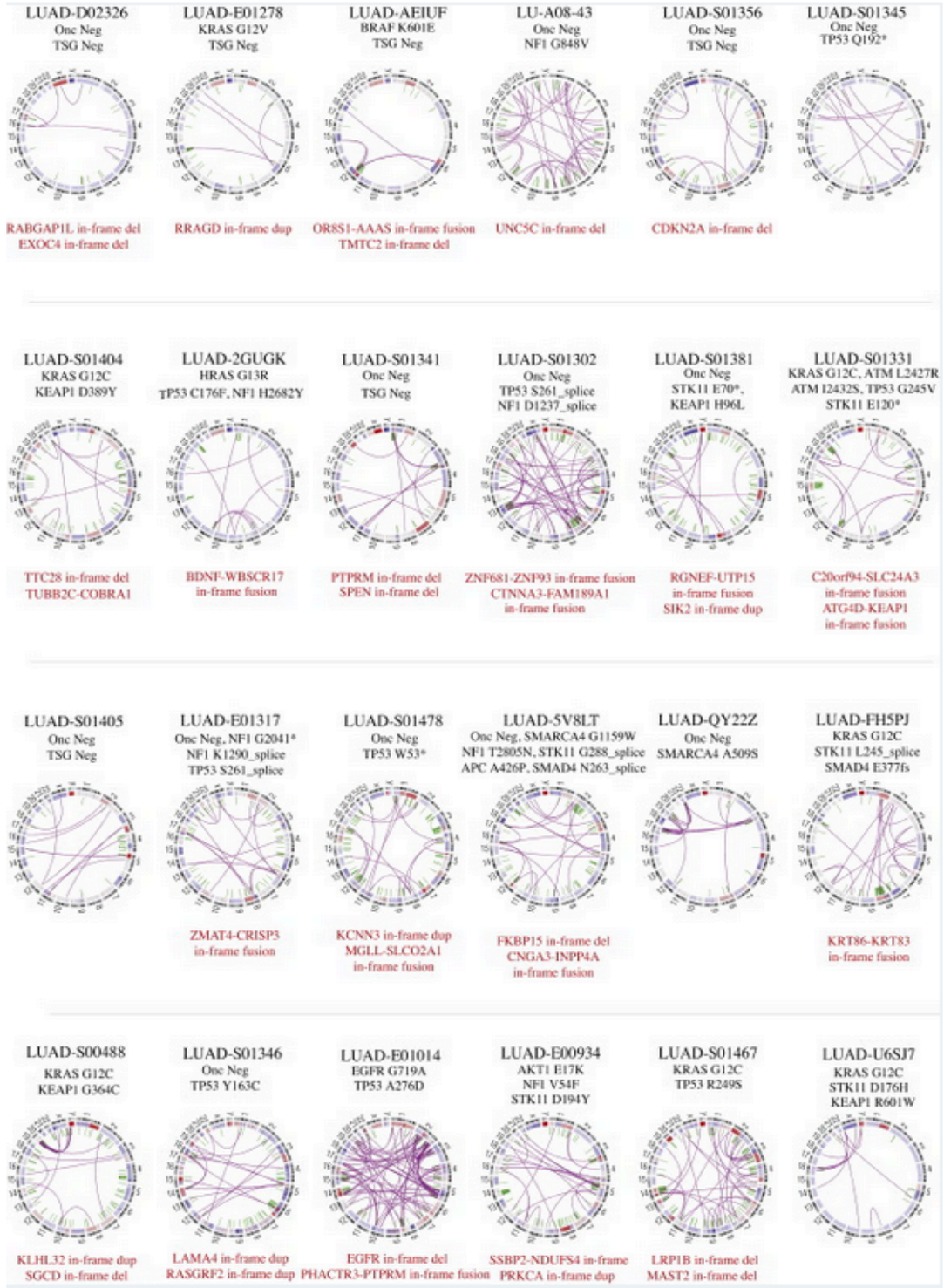




Stephens et al. *Cell* 144:27 (2011).

STRUCTURAL CHANGES ARE HARD TO SHOW FOR MULTIPLE GENOMES



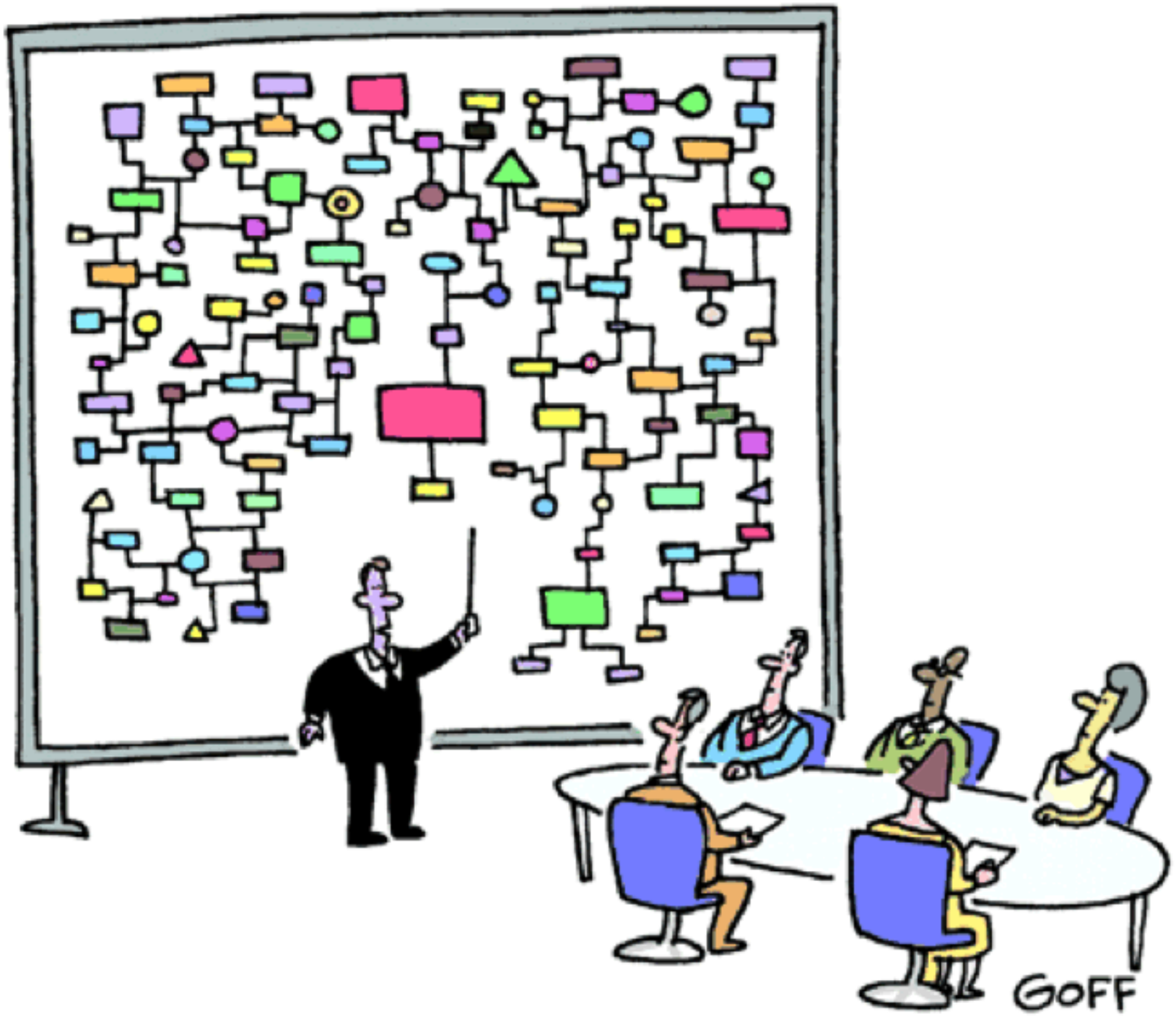


View of 17 mouse genomes. Keane et al. Nature 477:289 (2011).
Rearrangement signatures of adenocarcinomas. Imielinski M et al. Cell 150:1107 (2012).

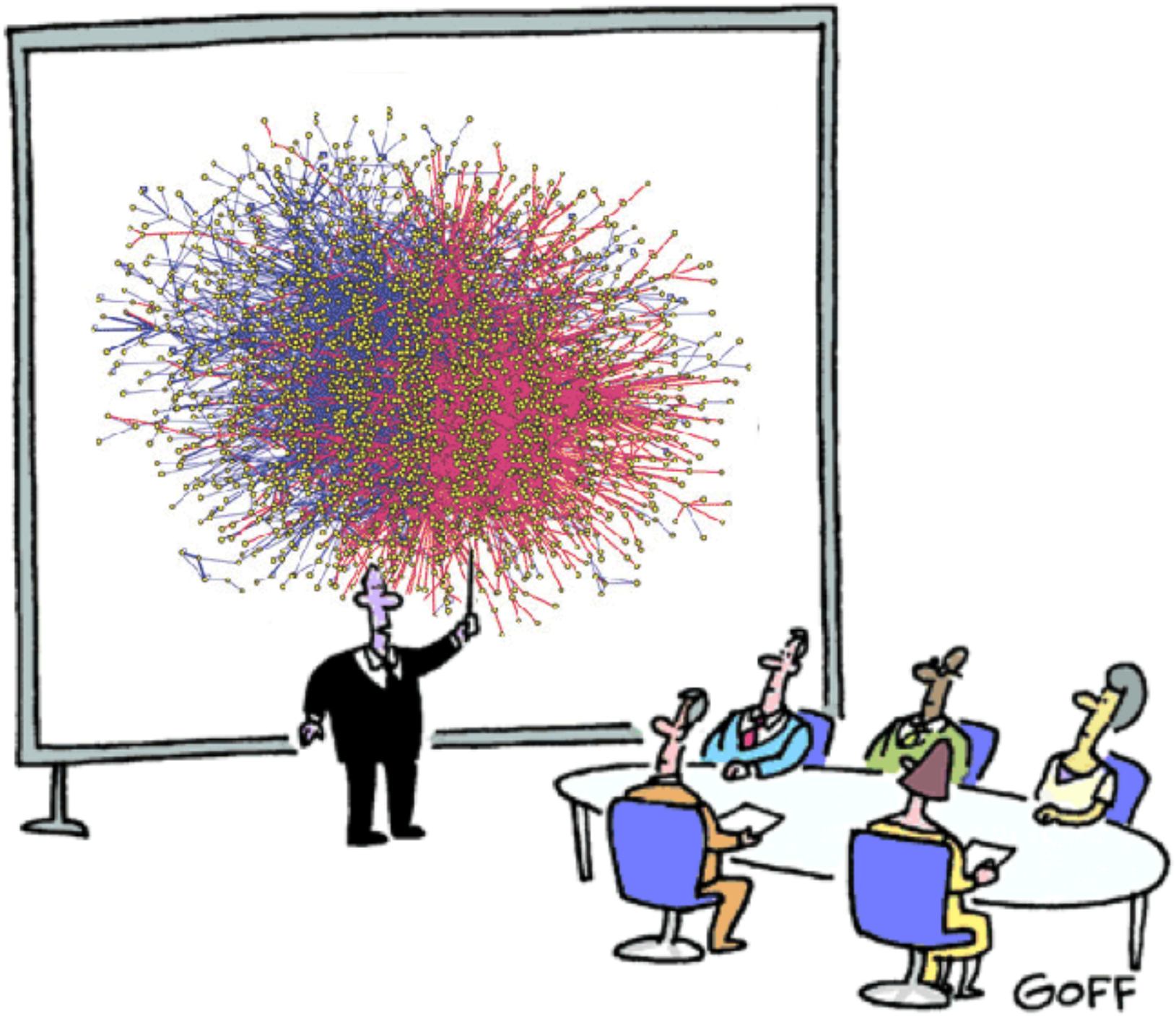
we can no longer afford to show the full data sets

only meaningful differences

...or even only differences of differences



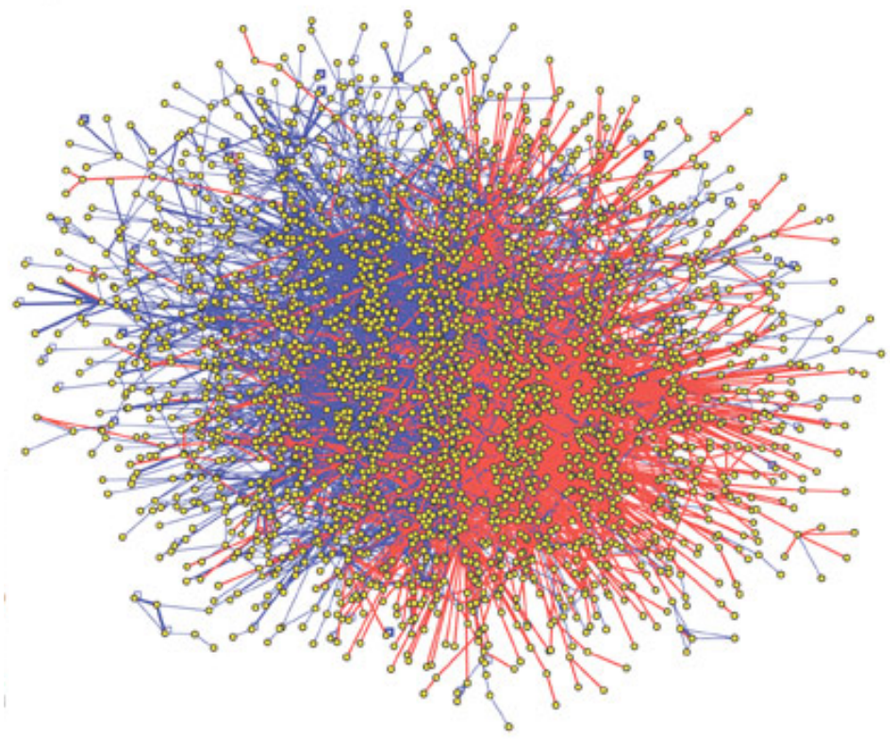
And that’s why we need a computer.



And that’s why we need a human.

HAIRBALLS AND NETWORK HAIRBALLS

both are visualizations of a complex system



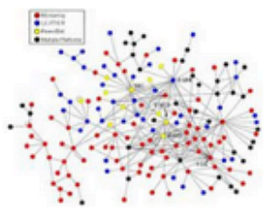
The **apparent banding pattern** of the yellow nodes **is an artefact** of the graph layout algorithm. Importantly, the layout algorithm was **not informed** by type of supporting evidence and therefore **does not explain** the evident **separation of blue and red** edges.

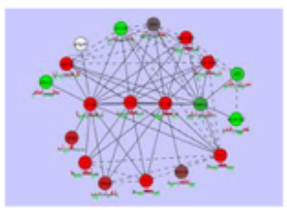
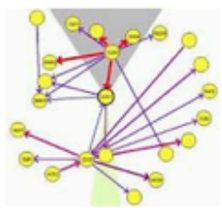
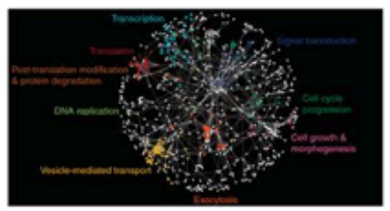
Miller W et al 2008. Sequencing the nuclear genome of the extinct woolly mammoth. *Nature* 456(7220): 387-390.
Figure 2 and caption quote from Rual et al., *Nature* 437(7062):1173-8.

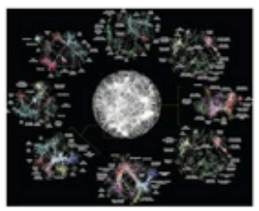
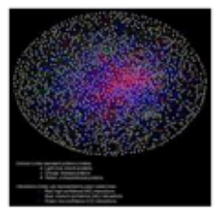
Google gene interaction network

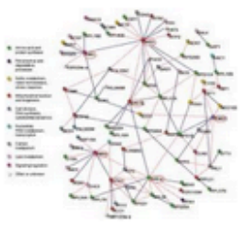
Martin Krzywinski 1 + Share

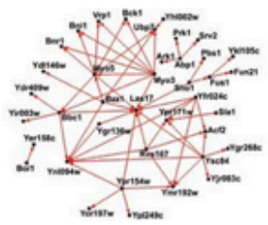
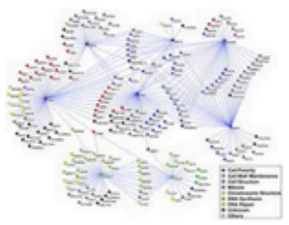
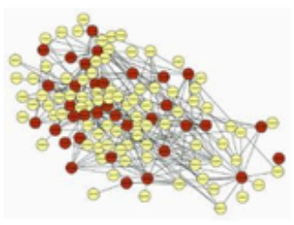
Web Images Maps Shopping More Search tools SafeSearch off

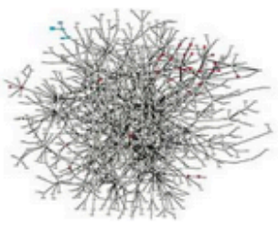



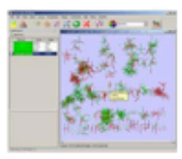
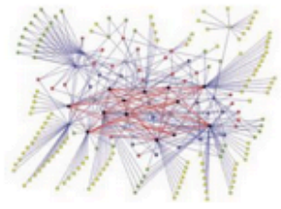
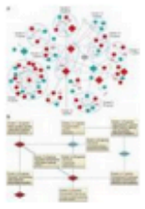




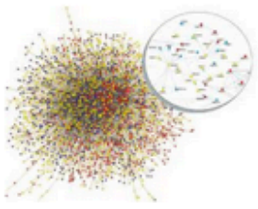
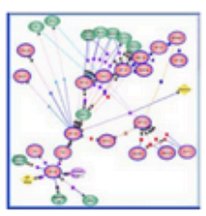
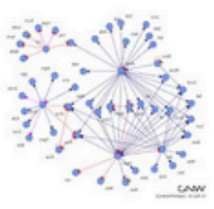





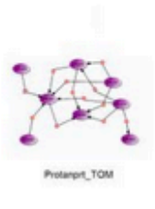




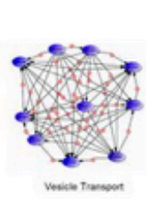



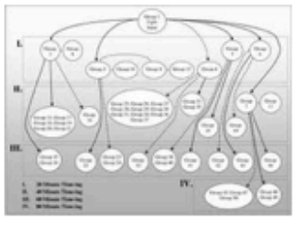
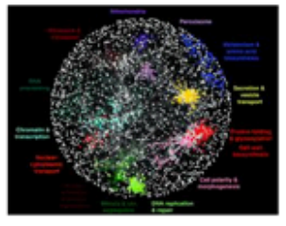
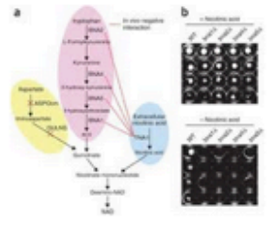




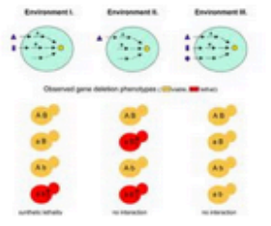






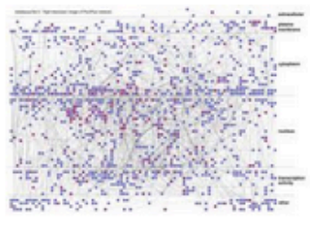
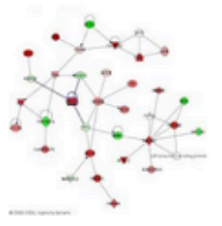
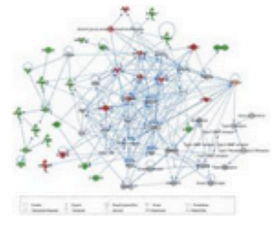
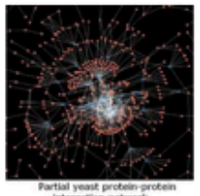




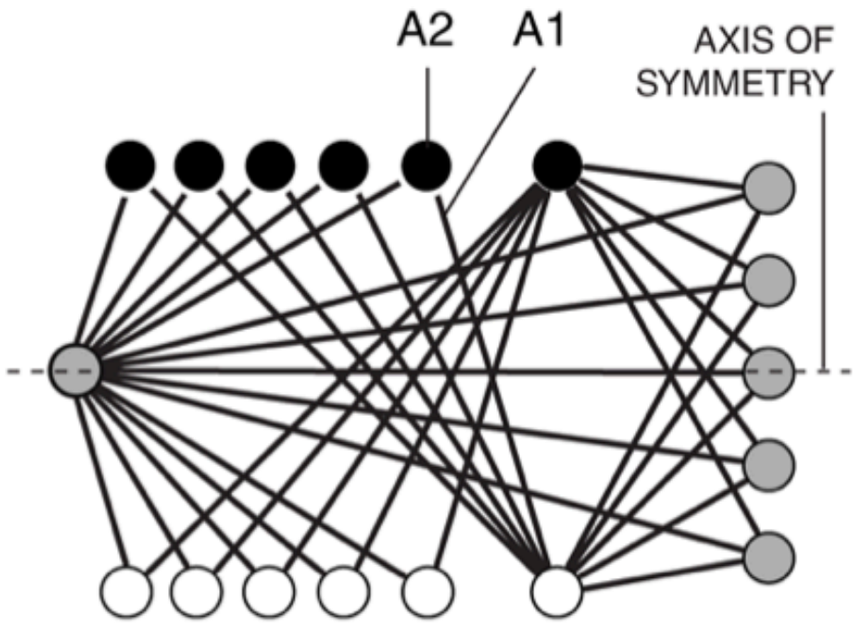






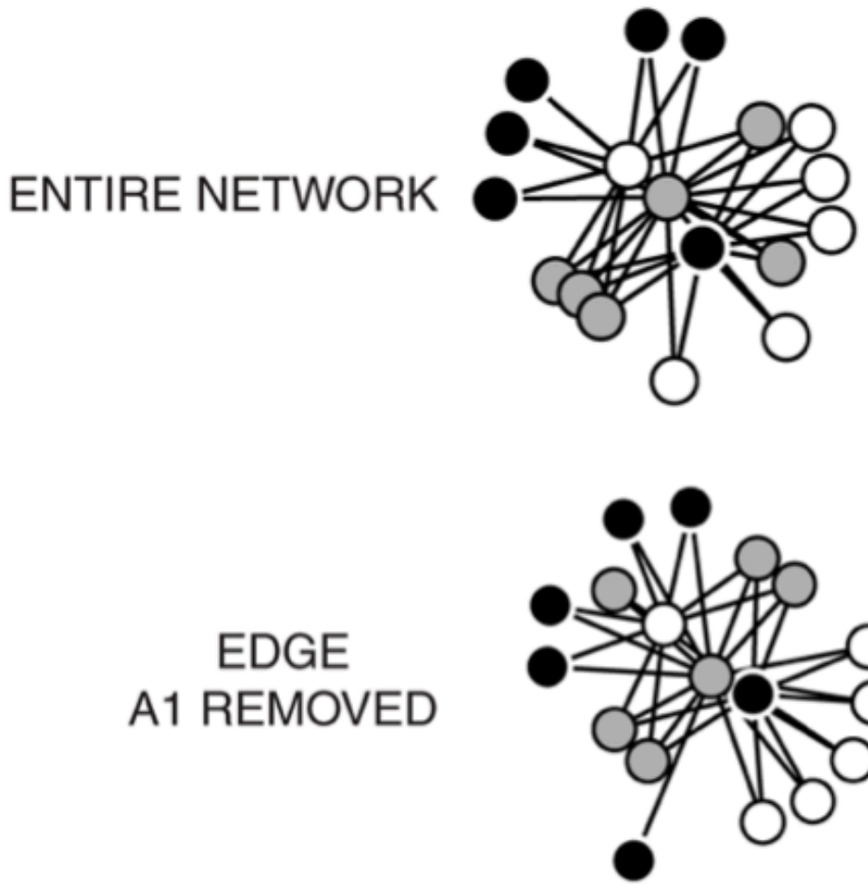


MOST LAYOUTS CANNOT BE COMPARED



MANUAL

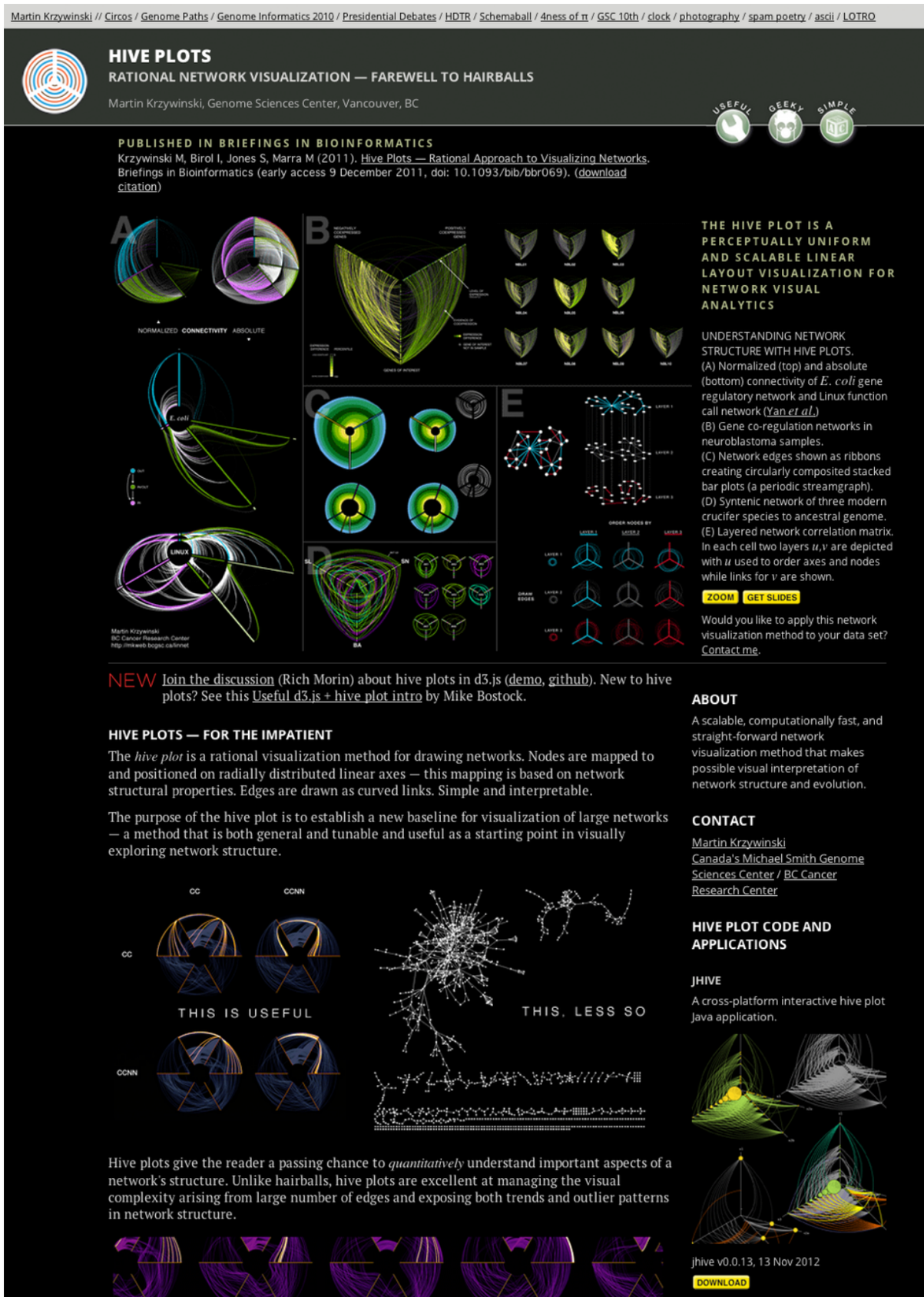


Krzywinski M, Birol I, Jones S, Marra M (2011). Hive Plots — Rational Approach to Visualizing Networks. Briefings in Bioinformatics (early access 9 December 2011, doi: 10.1093/bib/bbr069).

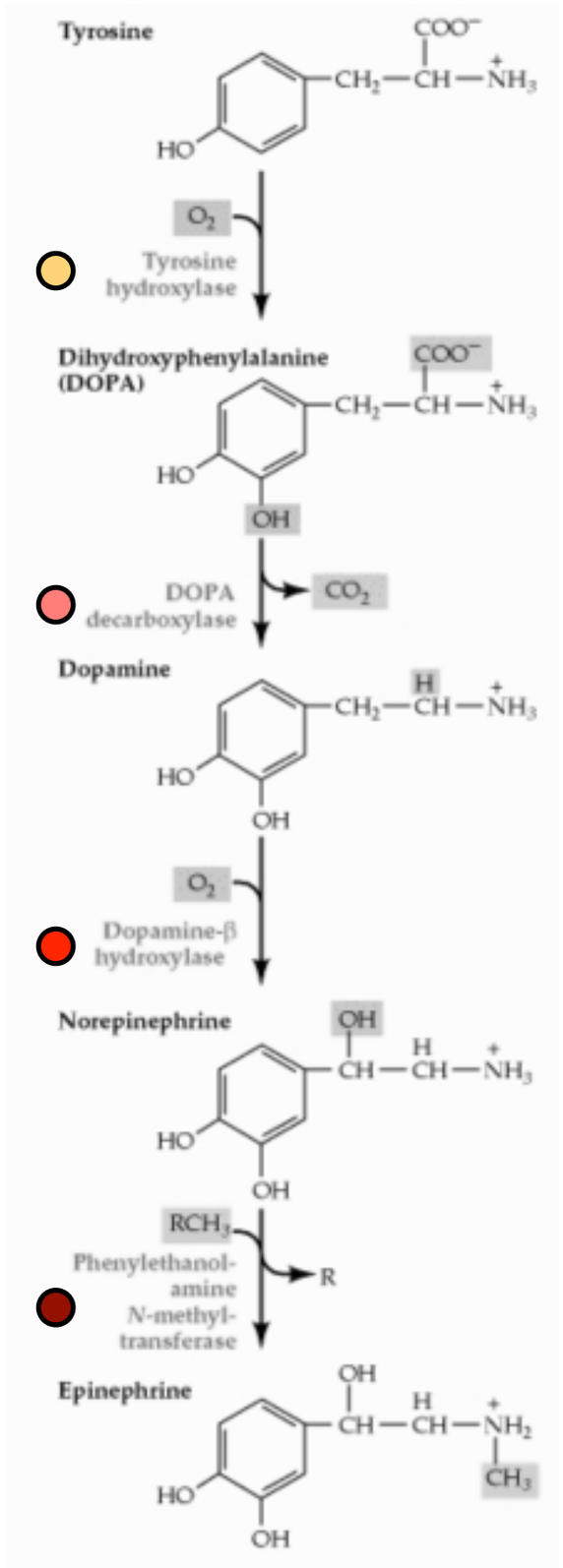
HIVE PLOTS — WWW.HIVEPLOT.COM

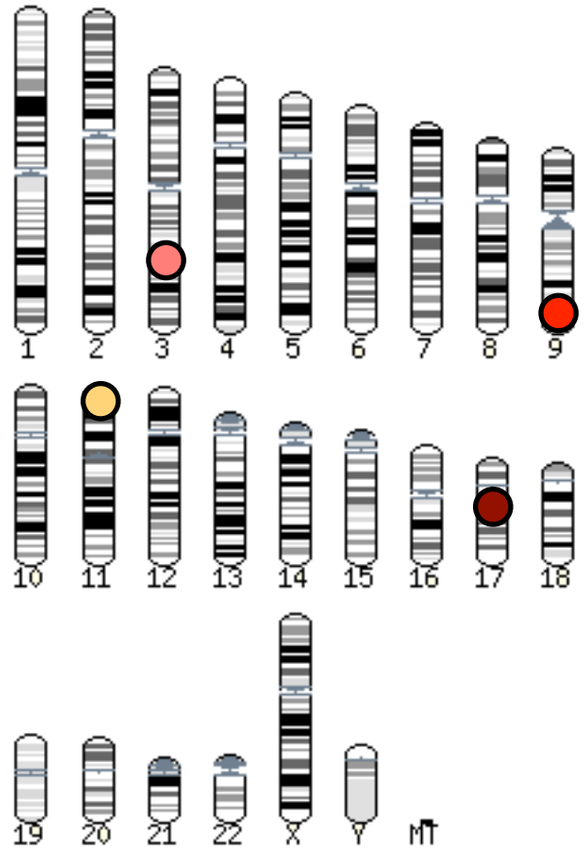
periodic parallel-coordinate
plots of topological properties

Krzywinski M, Birol I, Jones S, Marra M (2011). Hive Plots — Rational Approach to Visualizing Networks. Briefings in Bioinformatics (early access 9 December 2011, doi: 10.1093/bib/bbr069).



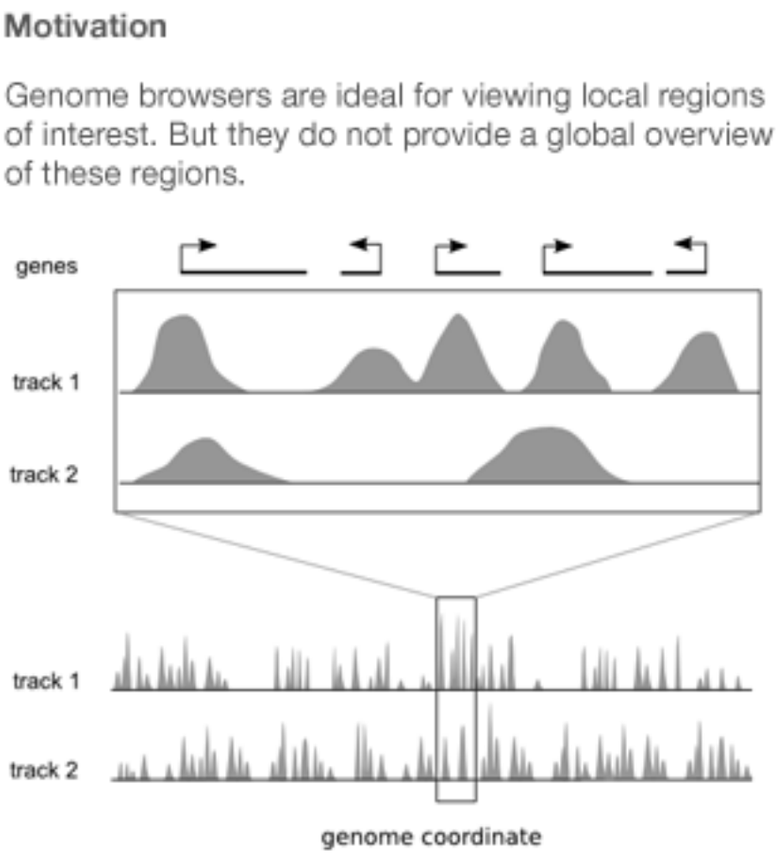
FUNCTION IS NOT RELATED TO GENOMIC POSITION

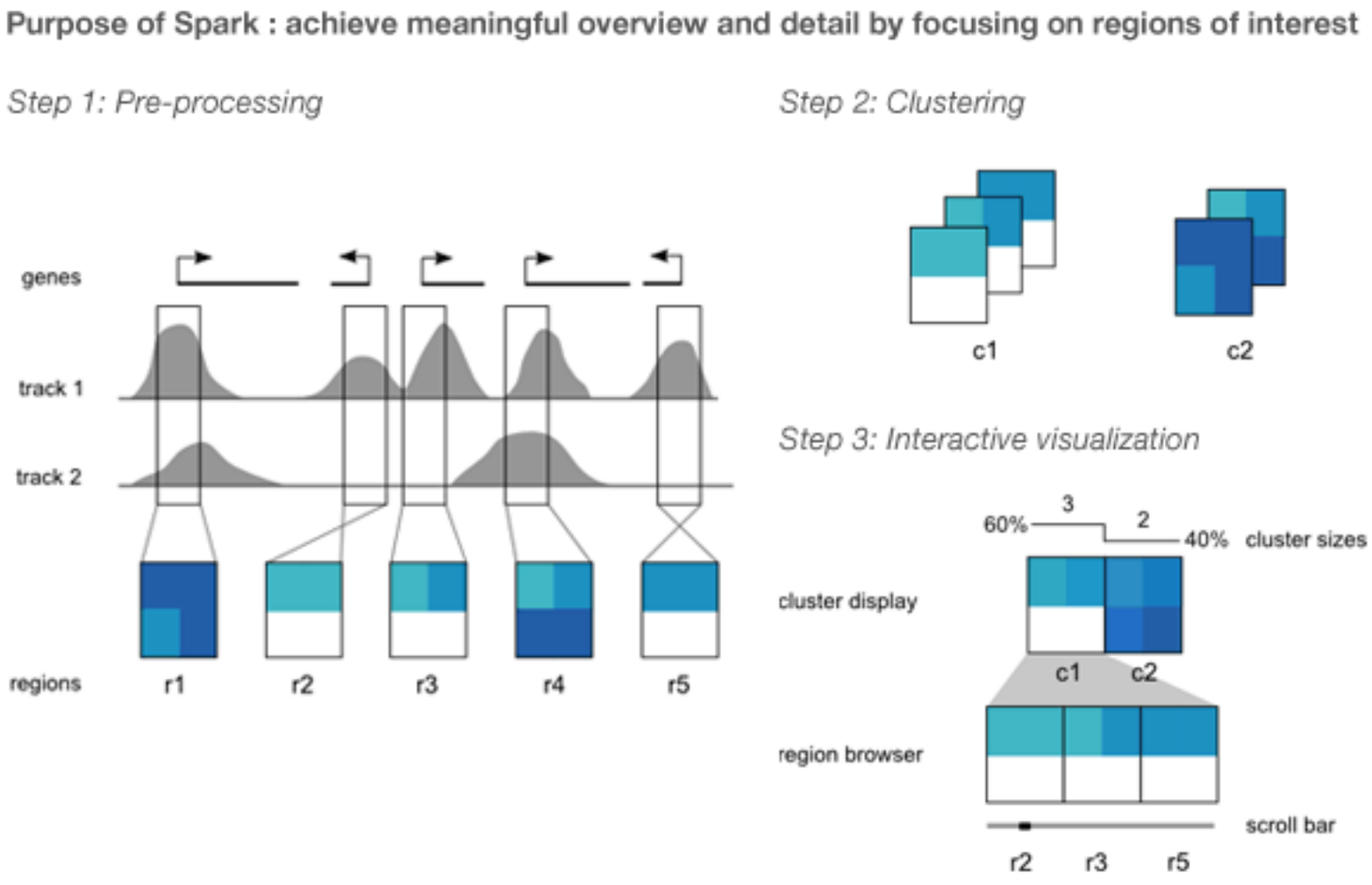




PHYSICAL COORDINATES ARE NATURAL, BUT LIMITING

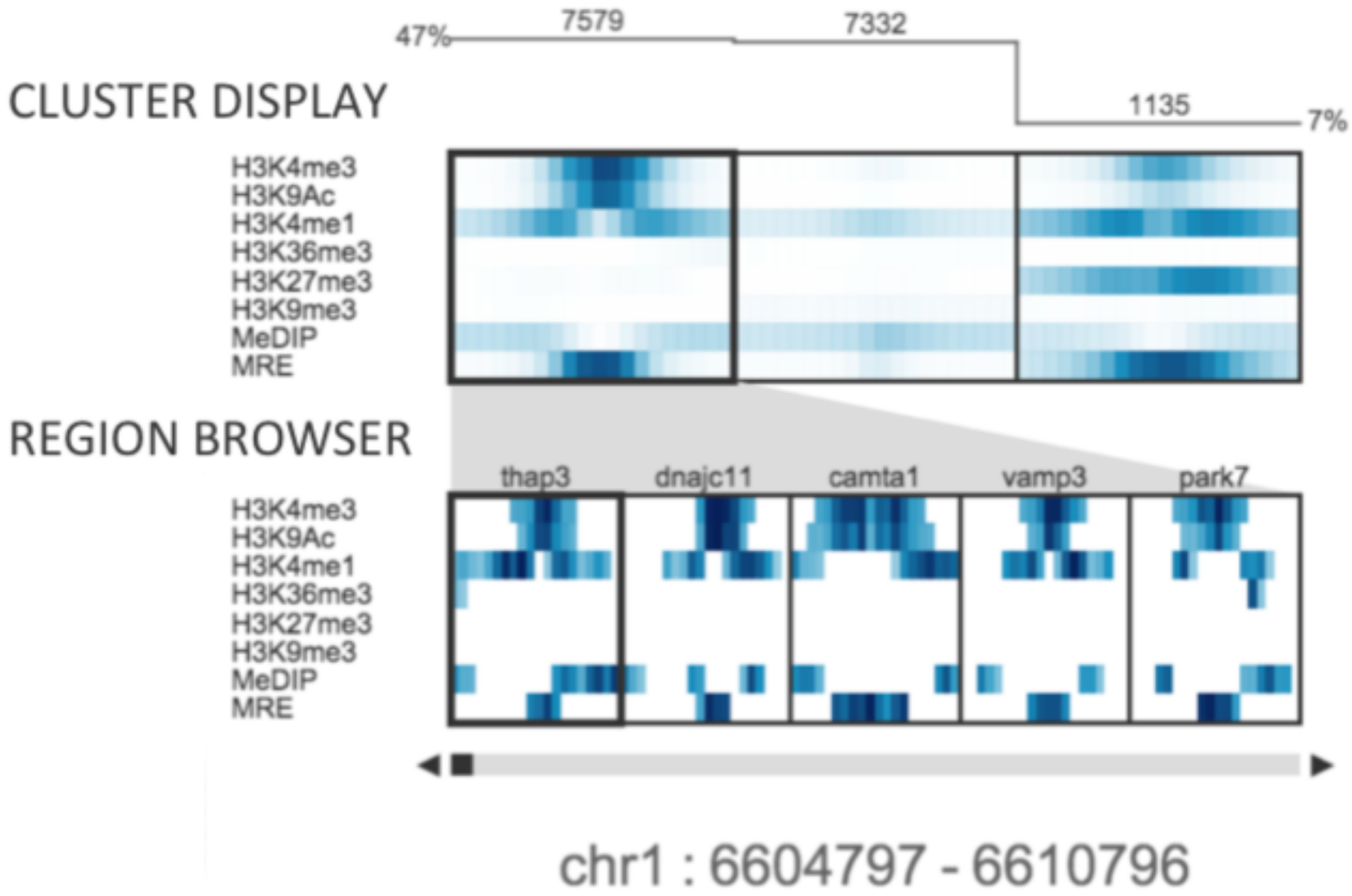
instead, use functional coordinates clustered by data profile





Nielsen et al. Spark: A navigational paradigm for genomic data exploration. *Genome Research* (2012) doi:10.1101/gr.140665.112.

SPARK



Nielsen et al. Spark: A navigational paradigm for genomic data exploration. Genome Research (2012) doi:10.1101/gr.140665.112.

SEQUENCE MOTIFS

```
HEM13   CCCATTGTTCTC
HEM13   TTTCTGGTTCTC
HEM13   TCAATTGTTTAG
ANB1    CTCATTGTTGTC
ANB1    TCCATTGTTCTC
ANB1    CCTATTGTTCTC
ANB1    TCCATTGTTCGT
ROX1    CCAATTGTTTTG

        YCHATTGTTCTC

A 002700000010
C 464100000505
G 000001800112
T 422087088261
```

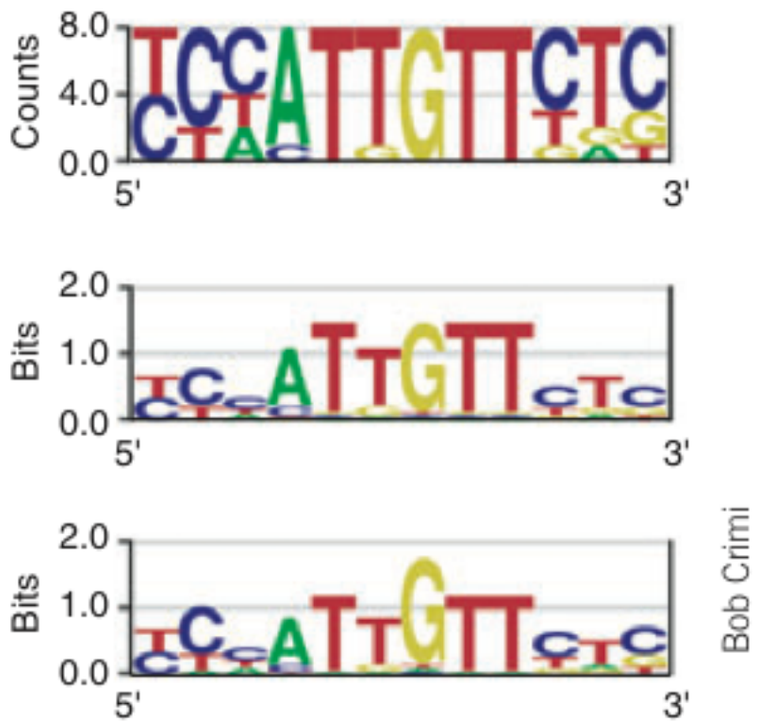
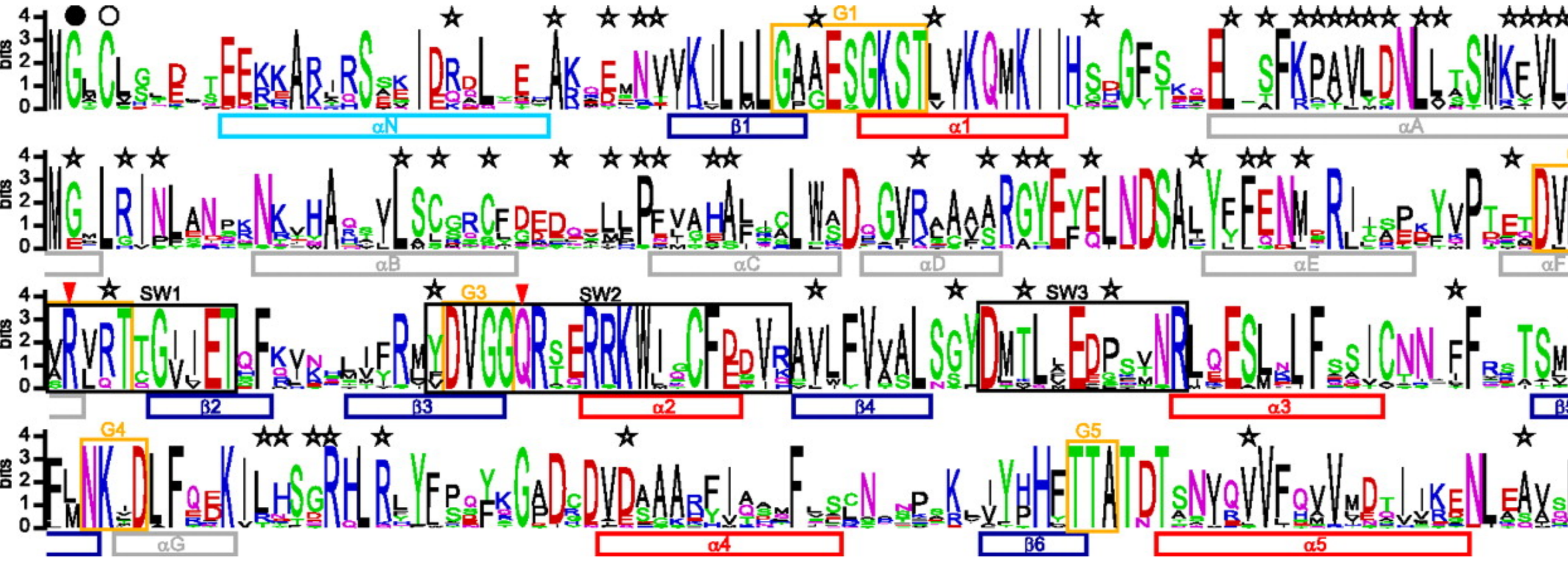


Figure 1 ROX1 binding sites and sequence motif. (**a**) Eight known genomic binding sites in three *S. cerevisiae* genes. (**b**) Degenerate consensus sequence. (**c**,**d**) Frequencies of nucleotides at each position. (**e**) Sequence logo showing the frequencies scaled relative to the information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

D'haeseleer P. *Nature Biotechnology* 24:423 (2006).

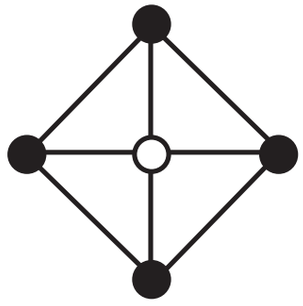
SEQUENCE LOGOS — VISUAL JARGON



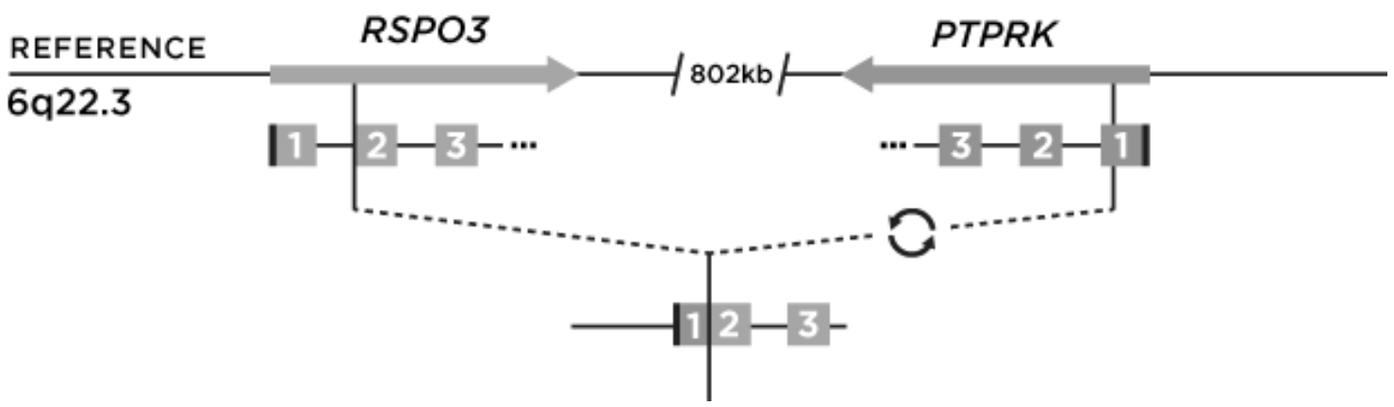
Oka Y, Saraiva LR, Kwan YY, Korsching SI (2009) The fifth class of Galpha proteins. Proc Natl Acad Sci U S A 106: 1484-1489.

EXPLORATION / COMMUNICATION

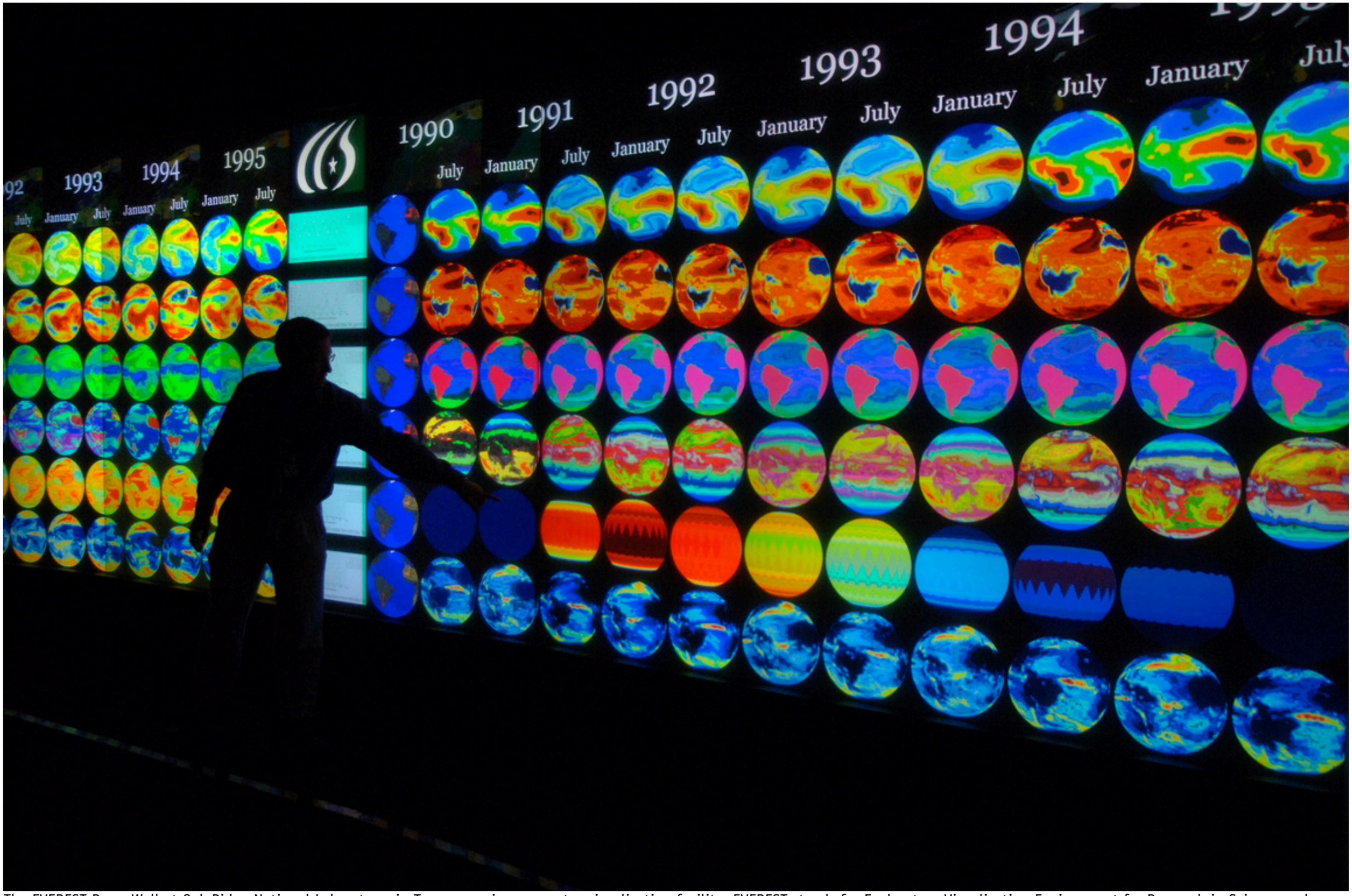
to explore data, use effective visual encodings



to communicate concepts and patterns, use effective design



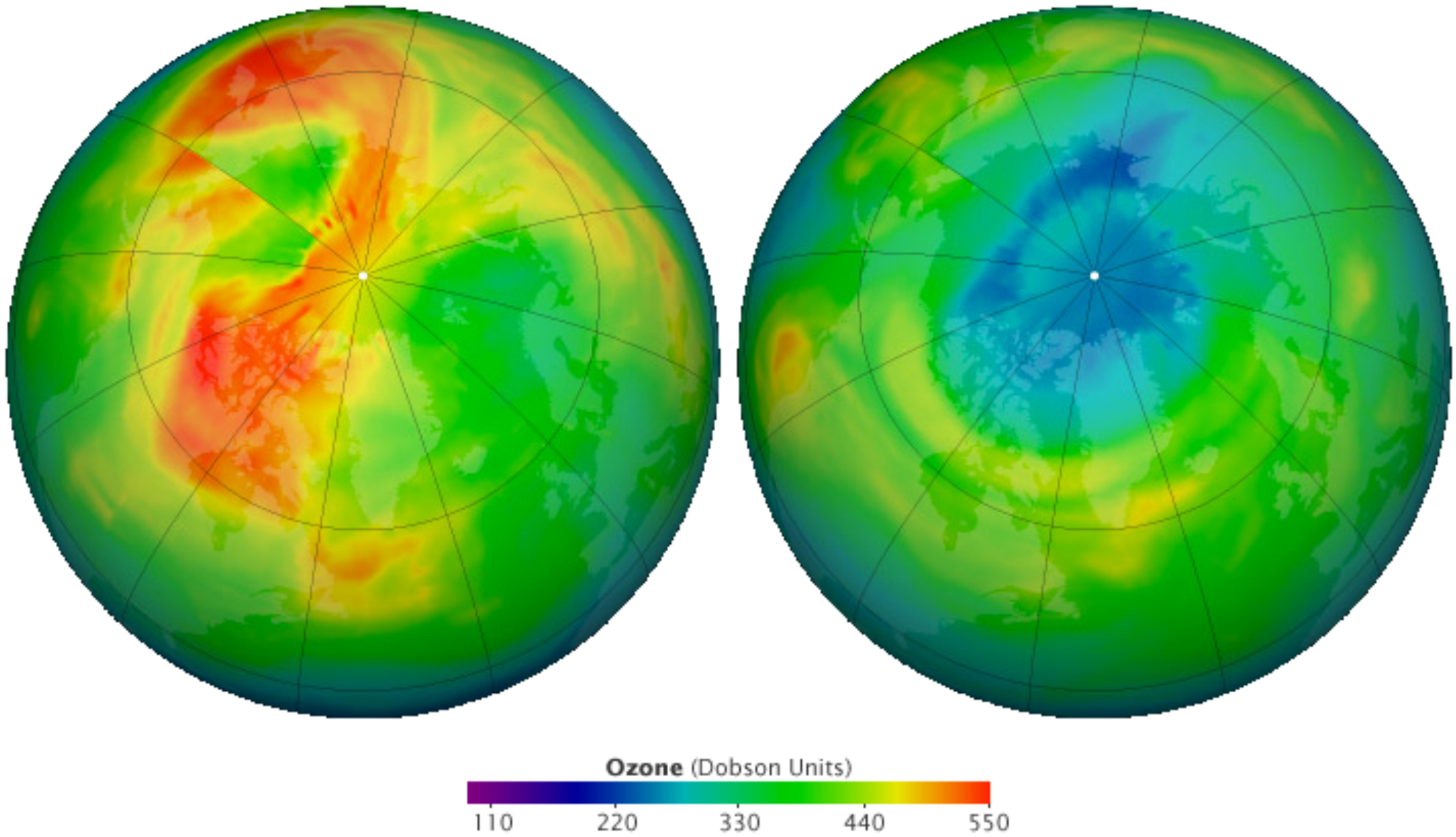
EXPLORING



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

CONSEQUENCES OF INAPPROPRIATE ENCODING

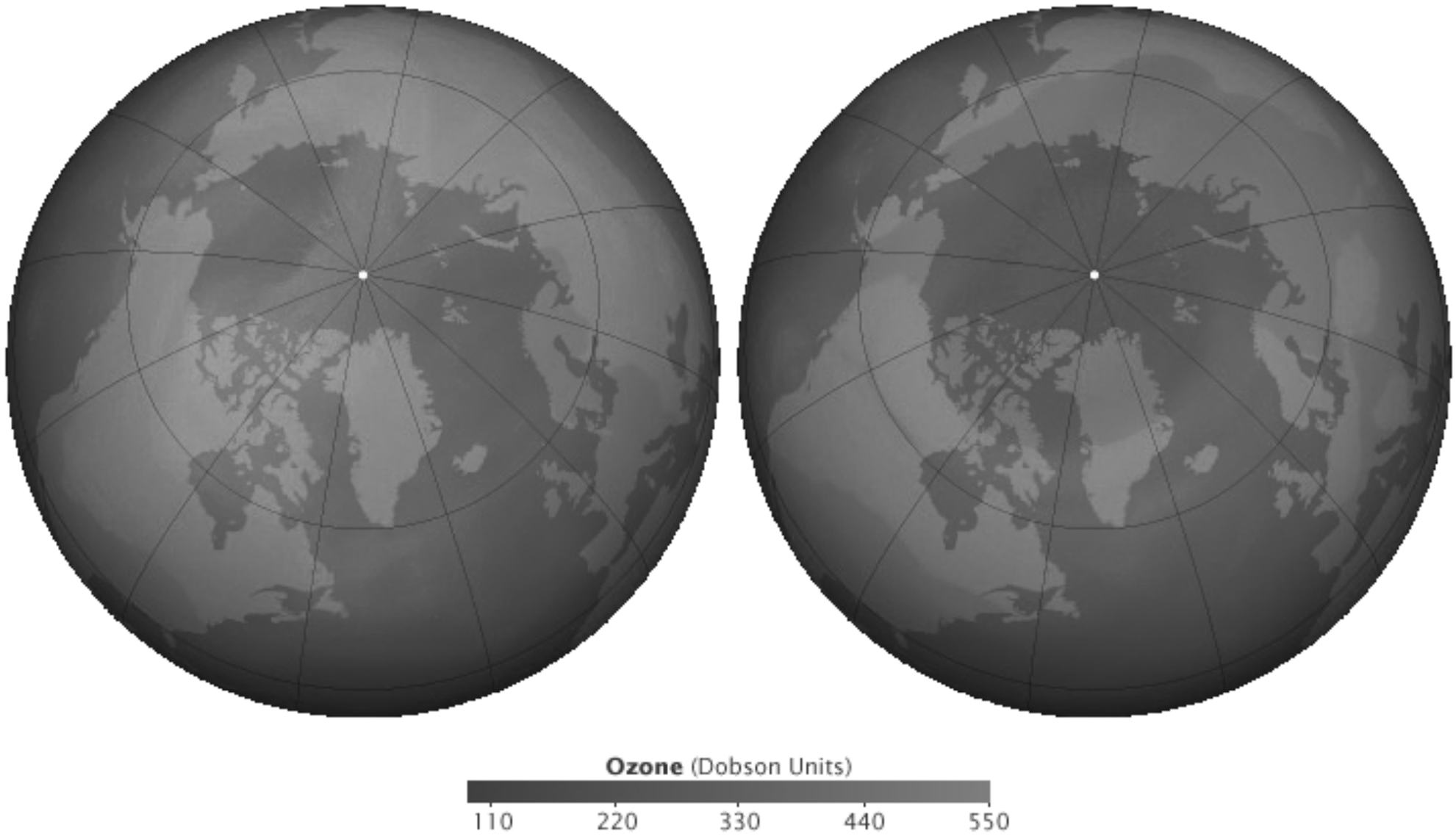
bad encoding doesn't mean the end of the world ... maybe



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

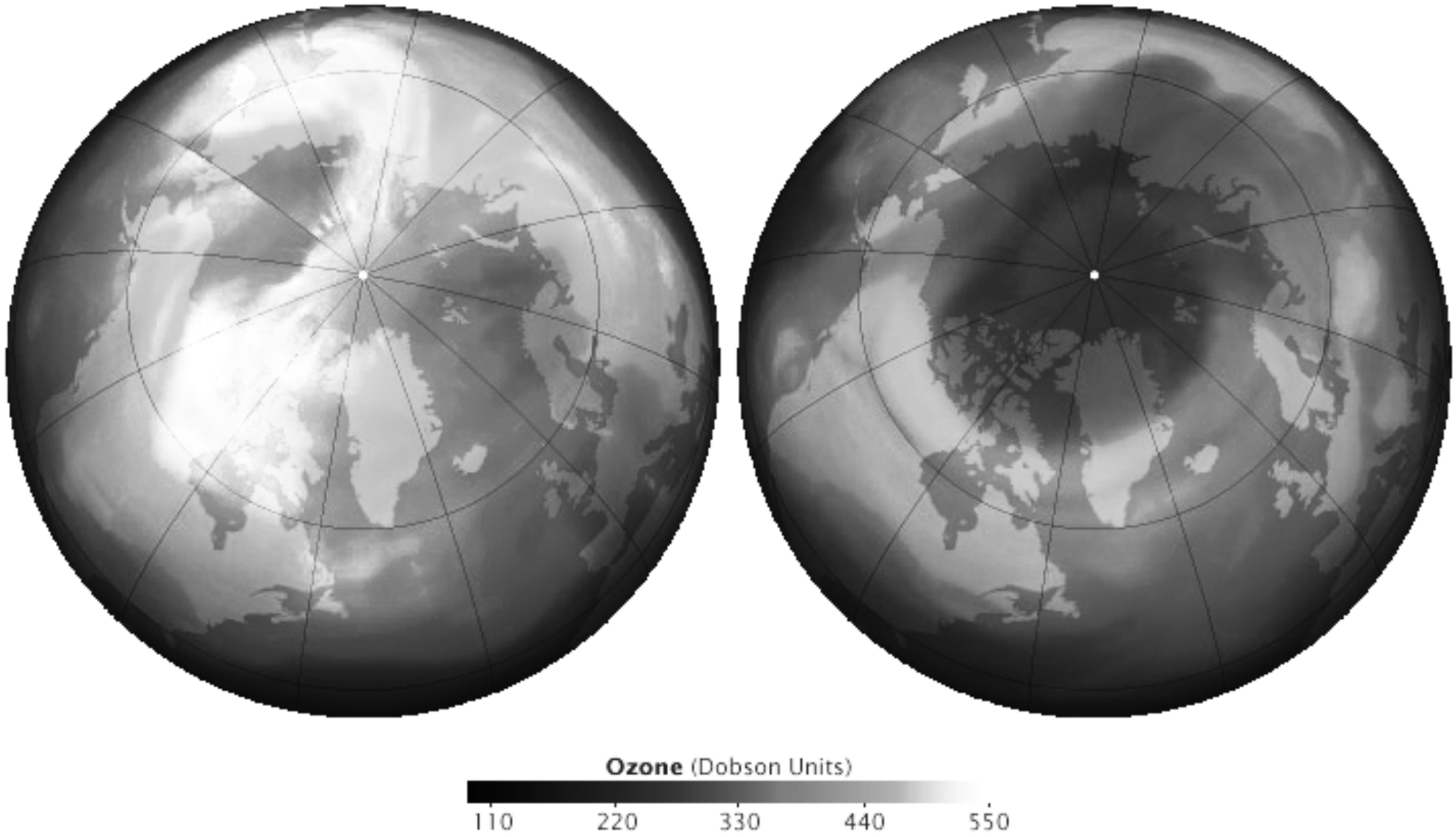
NYT did not use the figure – because information lost in b/w



Jonathan Corum (personal communication).

CONSEQUENCES OF INAPROPRIATE ENCODING

use tone instead of hue



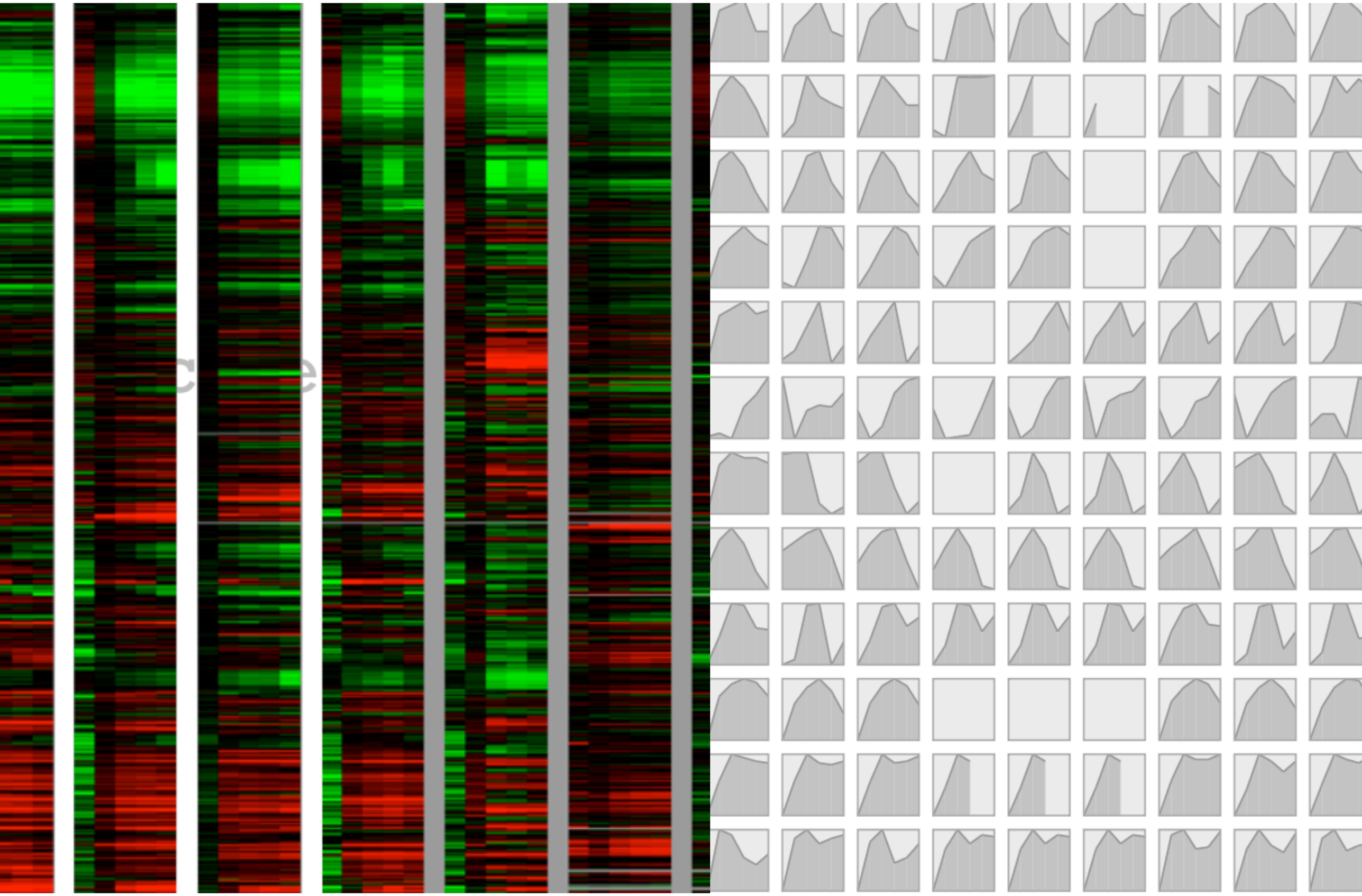
LUMINANCE EFFECT — THE LIER IN THE HEAT MAP

Same colour
looks different

Different colour
looks the same

* These rectangles
have the same colour
but look different

Wong B. Points of View. *Nature Methods* 7:573 (2010).



Mayer, M. et al. Pathline: A Tool For Comparative Functional Genomics.
Proc. EuroVis 29, 1043-1052 (2010)

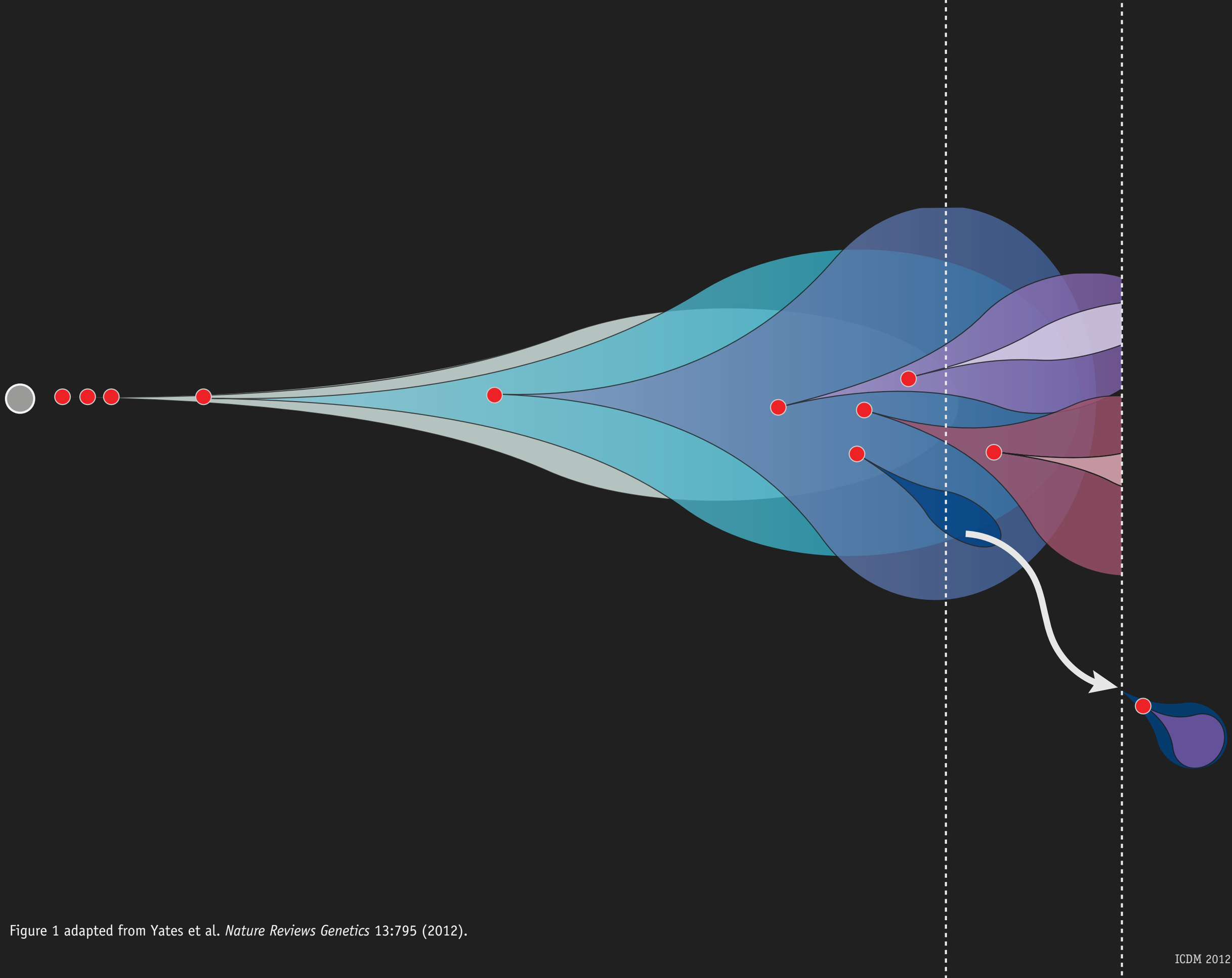
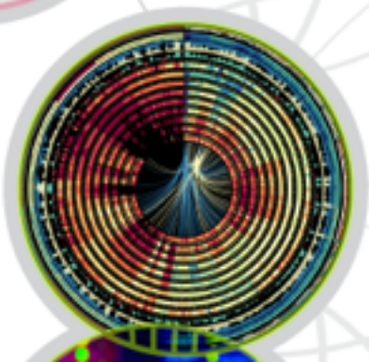


Figure 1 adapted from Yates et al. *Nature Reviews Genetics* 13:795 (2012).

GENOMICS

INNOVATION

CANADA'S MICHAEL SMITH
GENOME
SCIENCES
CENTRE



INFORMATICS

SEQUENCING

COMPUTING