Challenges of Modern Bioinformatics

István Albert

Bioinformatics Consulting Center

Pennsylvania State University

Background

- Physics PhD (2001) from Notre Dame, moved onto computer science first
- At Penn State since 2003, refocused on bioinformatics
- Since 2009 I teach courses on **Applied Bioinformatics** and **Bioinformatics Programming**
- Since 2010 Director of Bioinformatics Consulting Center at PSU
- I also love to program tools and web applications

What is bioinformatics?

- Computational analysis of genomic data

- **Genomics:** all information relating to the genetic sequence (DNA) of an organism

Beginnings: Human Genome Project

- Completed in 2000 at the cost of $3 billion

- Promised to bring about a revolution in the understanding of **biology in general** and **genomic medicine** in particular

It all seems so simple

- DNA is made up of simple elements: **A, T, G, C**

- What's good with tedious but well defined tasks?

DNA analysis + computer → match made in heaven

Published Online May 19 2011
Science 1 July 2011:
Vol. 333 no. 6038 pp. 53–58
DOI: 10.1126/science.1207018

RESEARCH ARTICLE

Widespread RNA and DNA Sequence Differences in the Human Transcriptome

Mingyao Li[1,*], Isabel X. Wang[2,*], Yun Li[3,4], Alan Bruzel[2], Allison L. Richards[5], Jonathan M. Toung[6], Vivian G. Cheung[2,7,8,†]

10,000 exonic sites where the RNA does not match the DNA,
All 12 possible categories of discordance have been observed

In total, we generated ~1.1 billion reads of 50 base pairs (bp) (~41 million reads and 2 Gb of

Next, we validated our findings experimentally by Sanger sequencing of both DNA and RNA

Proteomic evidence for RDD.

and gene density among chromosomes. RDD sites are significantly ($P < 10^{-10}$) enriched in genes

Comment on "Widespread RNA and DNA Sequence Differences in the Human Transcriptome"

Joseph K. Pickrell,[1]* Yoav Gilad,[1] Jonathan K. Pritchard[1,2]

1 year later

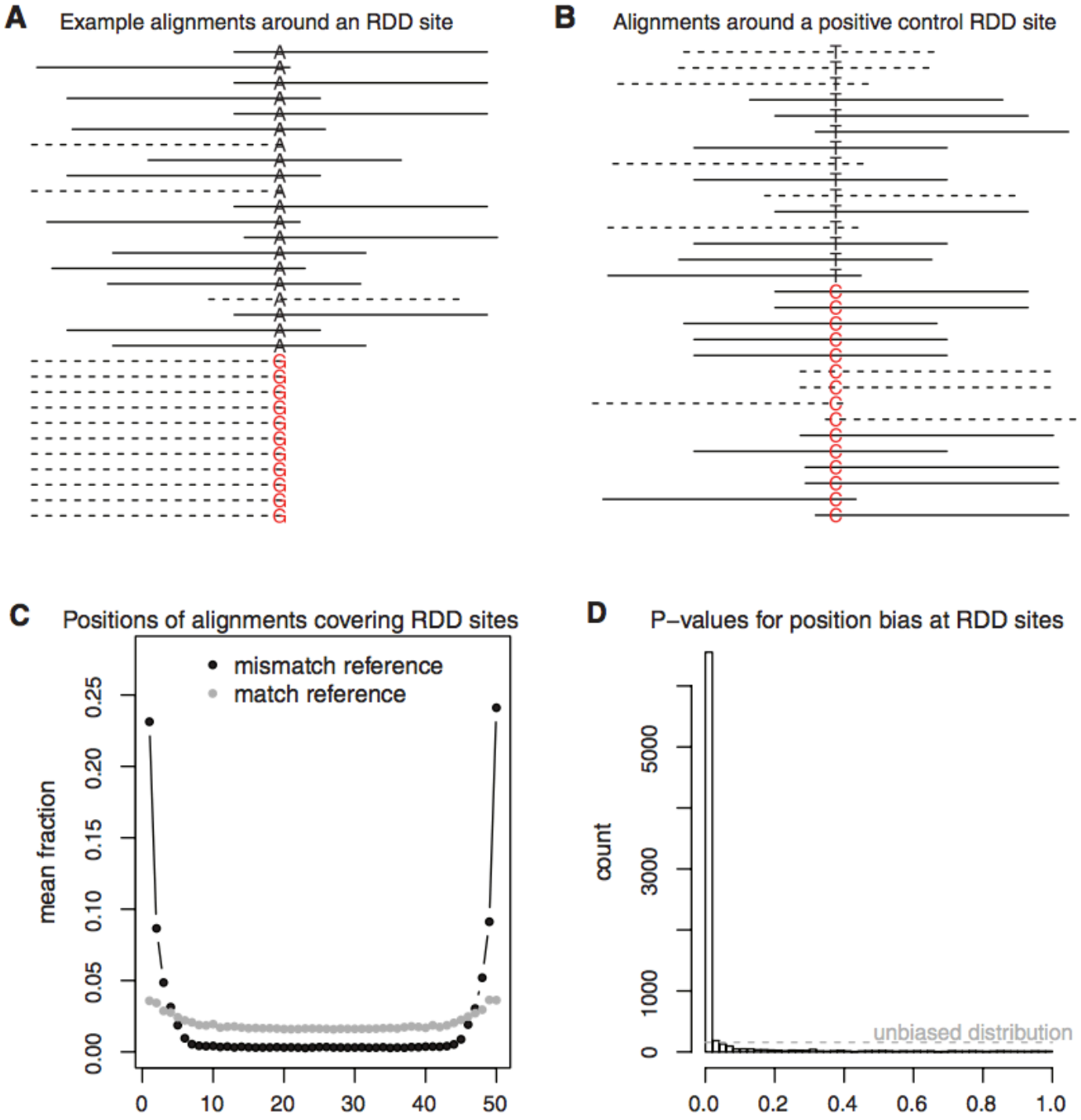
they attributed to previously unrecognized mechanisms of gene regulation. We found that at least 88% of these sequence mismatches can likely be explained by technical artifacts such as errors in mapping sequencing reads to a reference genome, sequencing errors, and genetic variation.

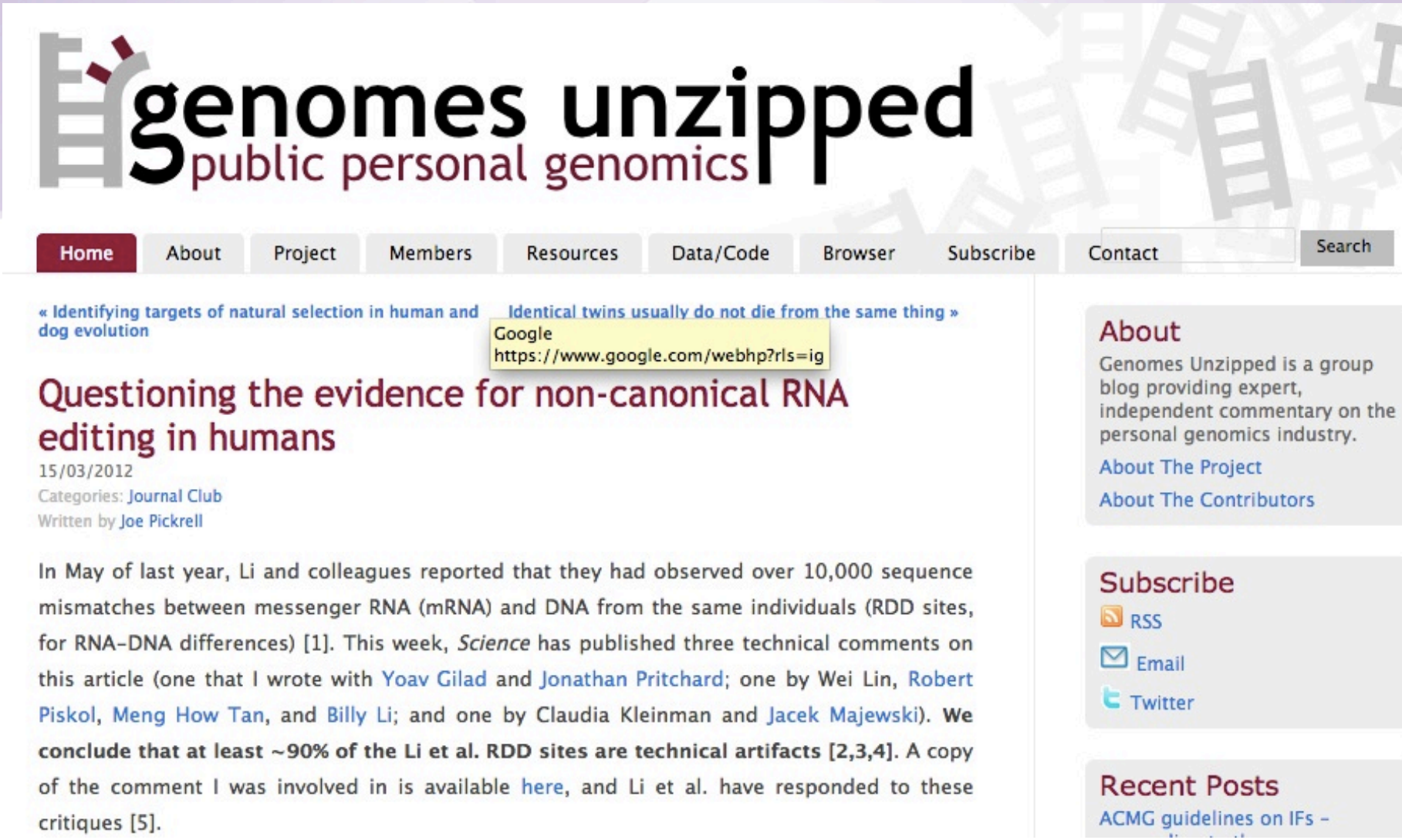
Comment on "Widespread RNA and DNA Sequence Differences in the Human Transcriptome"

Wei Lin,[1]* Robert Piskol,[2]* Meng How Tan,[2] Jin Billy Li[2]†

Critics say: at least 89% of the sites are false positives

12 possible mismatch types. Before accepting such a fundamental claim, a deeper analysis of the sequencing data is required to discern true differences between RNA and DNA from potential artifacts.


A
Example alignments around an RDD site
B
Alignments around a positive control RDD site
C
Positions of alignments covering RDD sites
mismatch reference
match reference
mean fraction
0.00
0.05
0.10
0.15
0.20
0.25
0
10
20
30
40
50
D
P−values for position bias at RDD sites
count
0
1000
3000
5000
unbiased distribution
0.0
0.2
0.4
0.6
0.8
1.0



Genomes Unzipped Blog

So perhaps it is not so simple after all

- The genomic patterns, variations and measurement errors make it surprisingly difficult to establish the standard by which we decide that a phenomena has been observed.

- On the same dataset the two “de facto” standards tools in SNP calling: GATK and SAMTOOLS produce results that are only about 80% concordant!

Back to history: rapid advances in technology

- High throughput sequencing instruments
- These make whole genome sequencing possible at a single institution
- Today the largest sequencing centers produce more sequence data in day than the combined sequence of all known organisms.

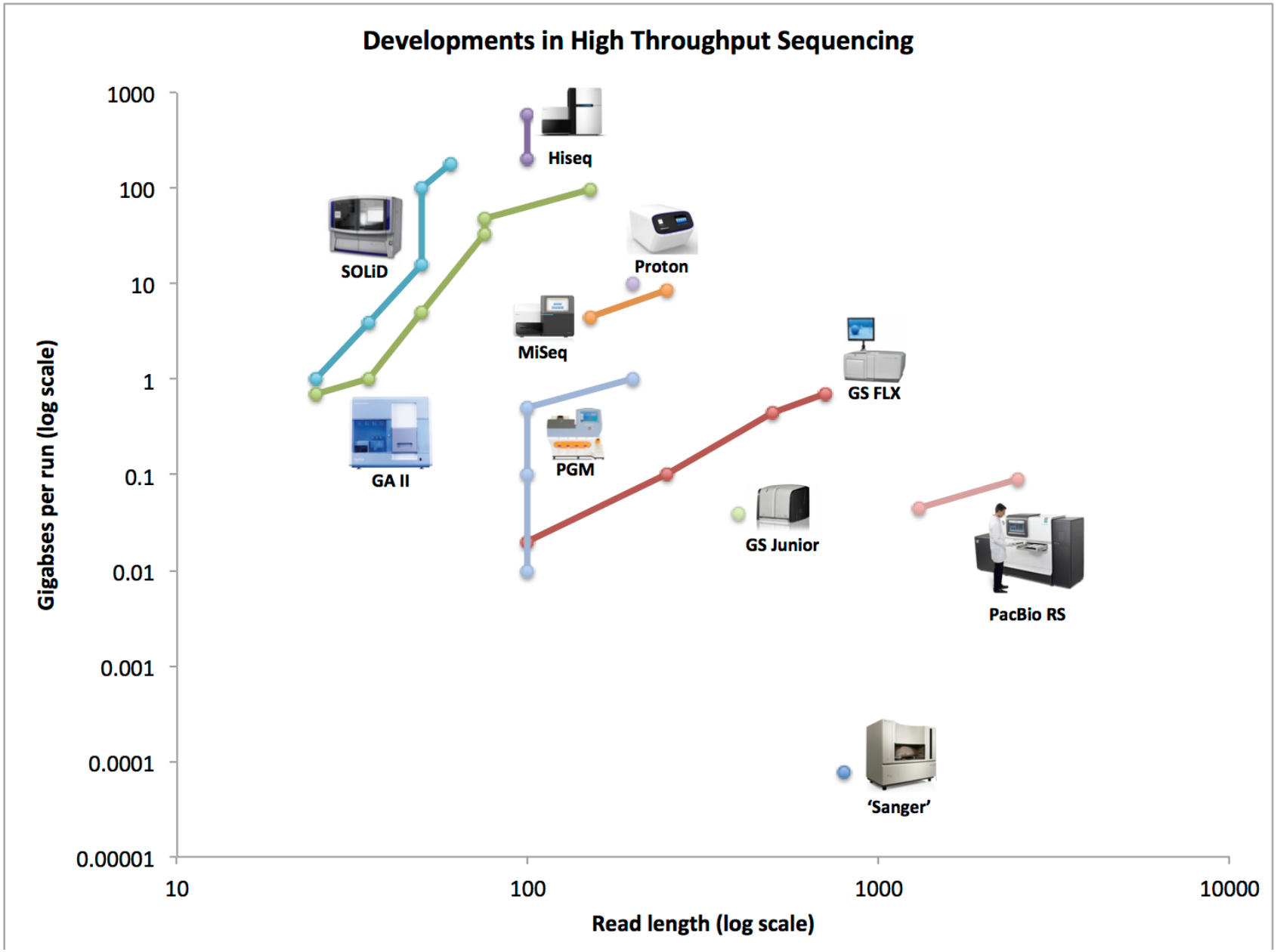
Rapid advances in sample preparation

- In the original approach the DNA was randomly sheared then these fragments are sequenced

- What if we one first isolate certain parts of the genome and sequence only those: Chip-Seq, RAD-Seq, Bisulfite sequencing, RNA-seq, 16S rRNA

- **Each of these techniques fundamentally alters the modes of interpretation for the data!**

From in between lines of code: http://flxlexblog.files.wordpress.com/



Field Guide to Next-Generation Sequencers

Molecular Ecology Resources (2011) 11, 759–769

Field guide to next-generation DNA sequencers.pdf (page 6 of 11)

Instrument	Run time[a]	Millions of reads/run	Bases/read[b]	Yield Mb/run	Reagent cost/run[c]	Reagent cost/Mb	Minimum unit cost (% run)[d]
3730xl (capillary)	2 h	0.000096	650	0.06	$96	$1500	$6 (1%)
Ion Torrent – '314'chip	2 h	0.10	100	>10	$500	<$50	~$750 (100%)
454 GS Jr. Titanium	10 h	0.10	400	50	$1100	$22	$1500 (100%)
Starlight*	†	~0.01	>1000	†	†	†	†
PacBio RS	0.5–2 h	0.01	860–1100	5–10	$110–900	$11–180	†
454 FLX Titanium	10 h	1	400	500	$6200	$12.4	$2000 (10%)
454 FLX+[e]	18–20 h	1	700	900	$6200	$7	$2000 (10%)
Ion Torrent – '316'chip*	2 h	1	>100	>100	$750	<$7.5	~$1000 (100%)
Helicos[f]	N/A	800	35	28 000	N/A	NA	$1100 (2%)
Ion Torrent – '318'chip*	2 h	4–8	>100	>1000	~$925	~$0.93	~$1200 (100%)
Illumina MiSeq*	26 h	3.4	150 + 150	1020	$750	$0.74	~$1000 (100%)
Illumina iScanSQ	8 days	250	100 + 100	50 000	$10 220	$0.20	$3000 (14%)
Illumina GAIIx	14 days	320	150 + 150	96 000	$11 524	$0.12	$3200 (14%)
SOLiD – 4	12 days	>840[g]	50 + 35	71 400	$8128	<$0.11	$2500 (12%)
Illumina HiSeq 1000	8 days	500	100 + 100	100 000	$10 220	$0.10	$3000 (12%)
Illumina HiSeq 2000	8 days	1000	100 + 100	200 000	$20 120[h]	$0.10	$3000 (6%)
SOLiD – 5500 (PI)*	8 days	>700[g]	75 + 35	77 000	$6101	<$0.08	$2000 (12%)
SOLiD – 5500xl (4hq)*	8 days	>1410[g]	75 + 35	155 100	$10 503[h]	<$0.07	$2000 (12%)
Illumina HiSeq 2000 – v3[i]*	10 days	≤3000	100 + 100	≤600 000	$23 470[h]	≥$0.04	~$3500 (6%)

2010: Human Genome at 10

Science 18 February 2011:
Vol. 331 no. 6019 pp. 861–862
DOI: 10.1126/science.1198039

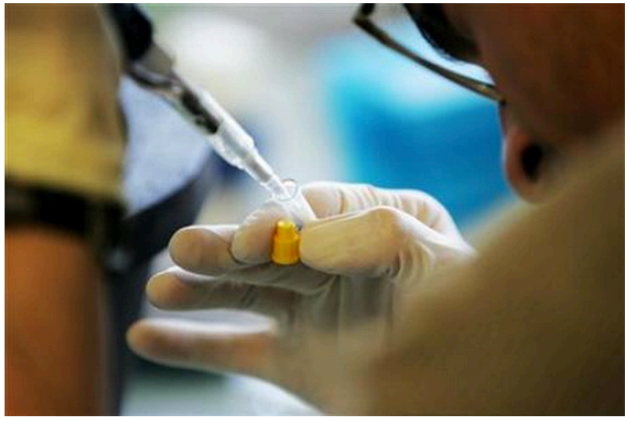
POLICY FORUM

GENOMICS

Deflating the Genomic Bubble

James P. Evans[1,*], Eric M. Meslin[2], Theresa M. Marteau[3], a

In cancer science, many "discoveries" don't hold up



By Sharon Begley
NEW YORK | Wed Mar 28, 2012 2:09pm EDT

(Reuters) - A former researcher at Amgen Inc has found that many basic studies on cancer -- a high proportion of them from university labs -- are unreliable, with grim consequences for producing new medicines in the future.

Related News

Analysis: Food security focus fuels new worries over crop chemicals
Tue, Mar 27 2012

Weight-loss surgery cut blood sugar more than drugs
Mon, Mar 26 2012

Monthly shots of Amgen drug slash cholesterol up to 66 percent
Sun, Mar 25 2012

Science 23 November 2012:
Vol. 338 no. 6110 pp. 1016–1017
DOI: 10.1126/science.338.6110.1016

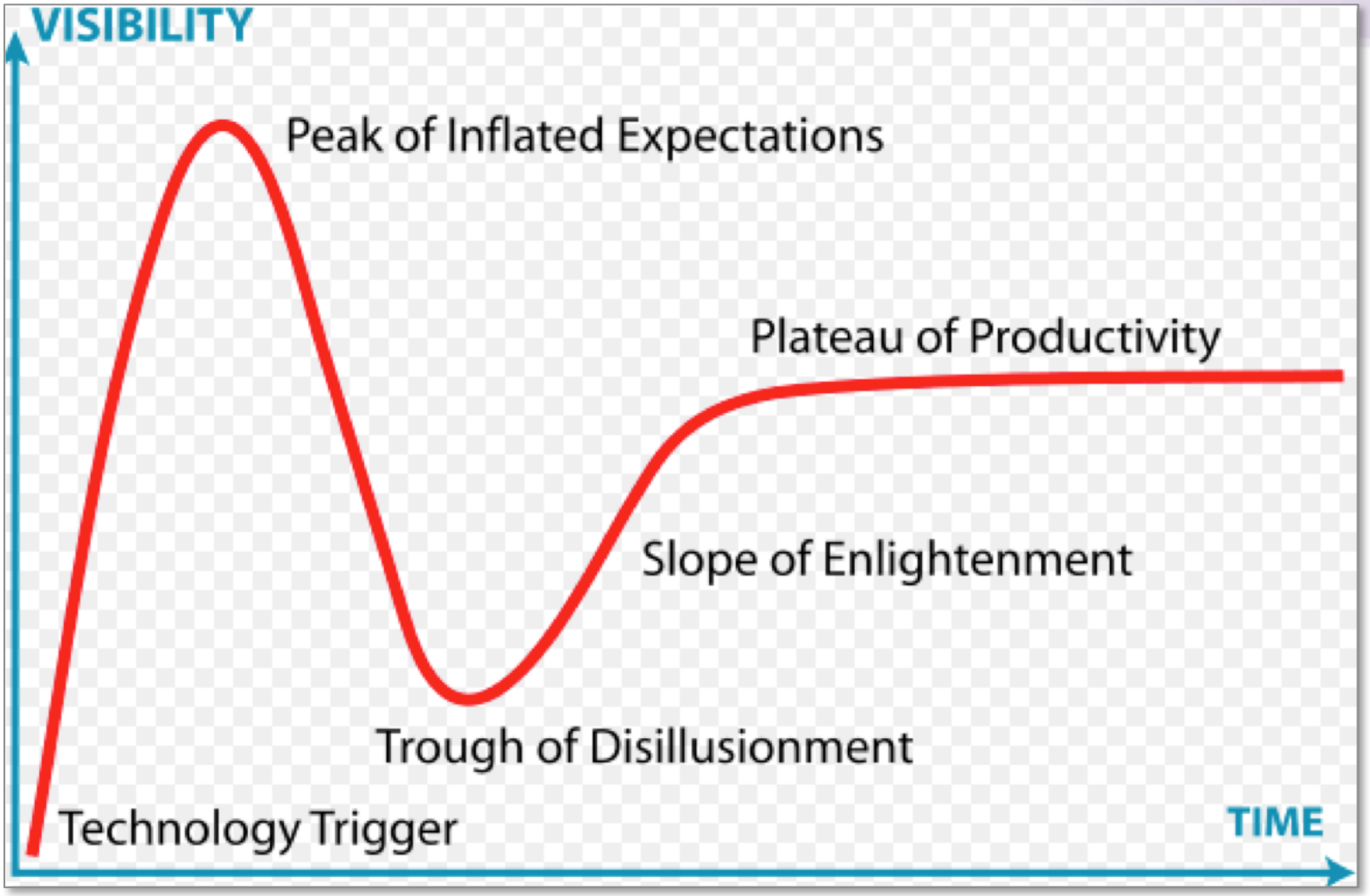
NEWS & ANALYSIS

HUMAN GENETICS

Genetic Influences on Disease Remain Hidden

Jocelyn Kaiser

Hype Cycle


VISIBILITY
Peak of Inflated Expectations
Plateau of Productivity
Slope of Enlightenment
Trough of Disillusionment
Technology Trigger
TIME

Two Sides of Bioinformatics

Descriptive

Properties, characteristics.
structure

Actionable

Diagnosis, predictions

Descriptive Bioinformatics

- Latest release of the ENCODE project 2012
- 30 simultaneous papers: Nature, Genome Research, Genome Biology
- **580** authors!

Main ENCODE findings
(some very contentious)

- **1.2**% of genome represents protein coding genes (20,678 protein coding genes with **6** spliced transcripts per locus)
- **62%** of genomic bases are present in long RNA molecules. **622,403** transcriptional start sites
- **8%** of the genome enriched for DNA binding, most locations with binding motifs (200Mb)
- First attempt to systematically test long range chromosomal interactions

1000 genomes project

- Another rousing success!
- **455** authors!

NATURE | ARTICLE OPEN

日本語要約

An integrated map of genetic variation from 1,092 human genomes

The 1000 Genomes Project Consortium

Affiliations | Contributions | Corresponding author

Nature **491**, 56–65 (01 November 2012) | doi:10.1038/nature11632

Received 04 July 2012 | Accepted 01 October 2012 | Published online 31 October 2012

Main Findings

- **3.6** million SNPs per individual
- **350,000** small insertions and deletions
- **717** large deletions

A few shocking observations (these are all healthy individuals):

- **2500** non-synonymous variants at conserved positions
- **20** – **40** damaging mutations
- **150** complete loss of function (LOF) mutations, many homozygous!

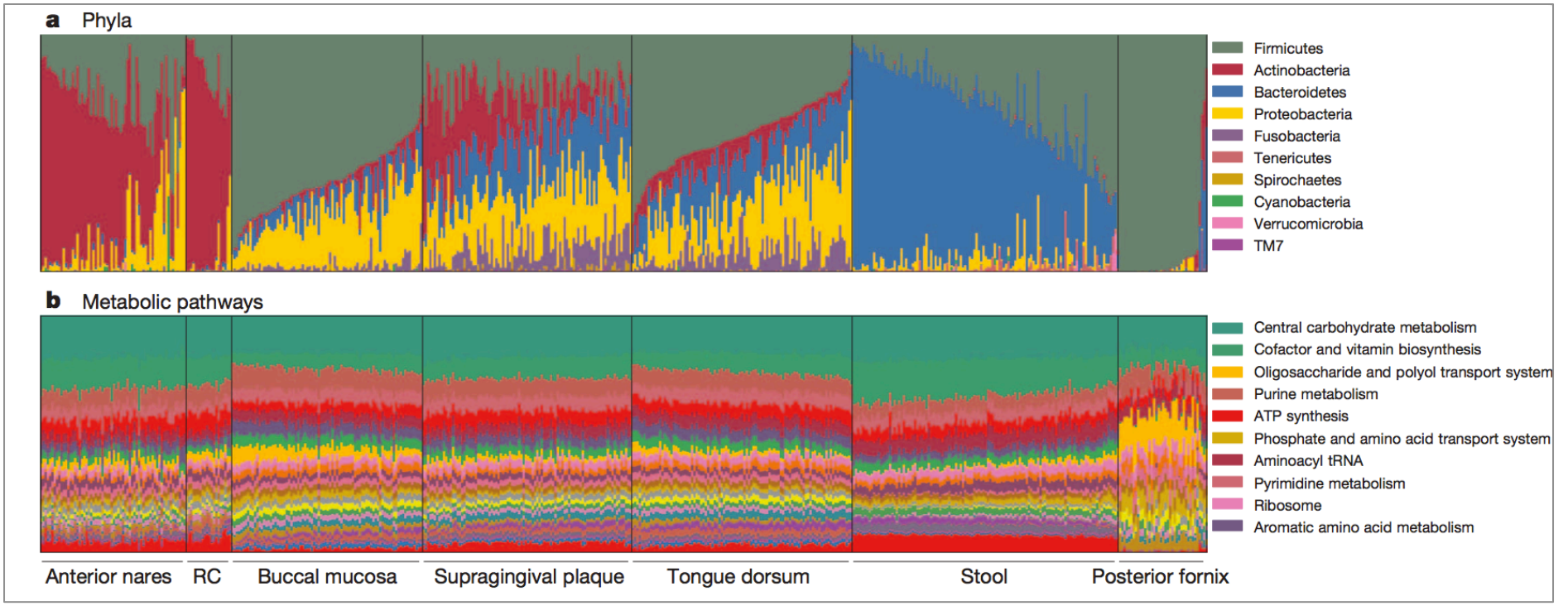
Structure, function and diversity of the healthy human microbiome

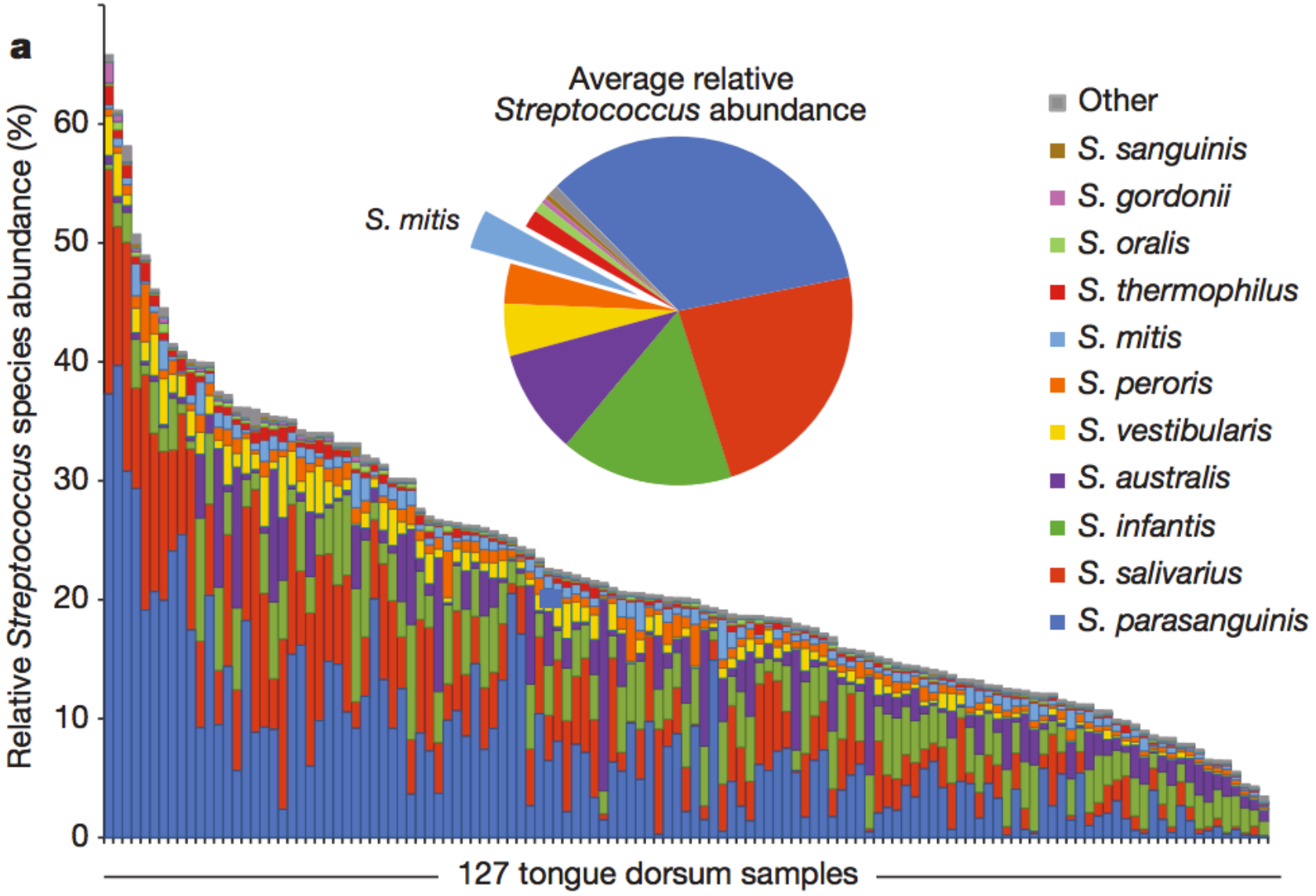
The Human Microbiome Project Consortium (247 authors)

Affiliations | **Contributions** | **Corresponding author**

Nature **486**, 207–214 (14 June 2012) | doi:10.1038/nature11234

Received 02 November 2011 | Accepted 16 May 2012 | Published online 13 June 2012




a
Relative Streptococcus species abundance (%)
0
10
20
30
40
50
60
127 tongue dorsum samples
Average relative Streptococcus abundance
S. mitis
Other
S. sanguinis
S. gordonii
S. oralis
S. thermophilus
S. mitis
S. peroris
S. vestibularis
S. australis
S. infantis
S. salivarius
S. parasanguinis

Two Sides of Bioinformatics

Descriptive

Properties, characteristics. structure

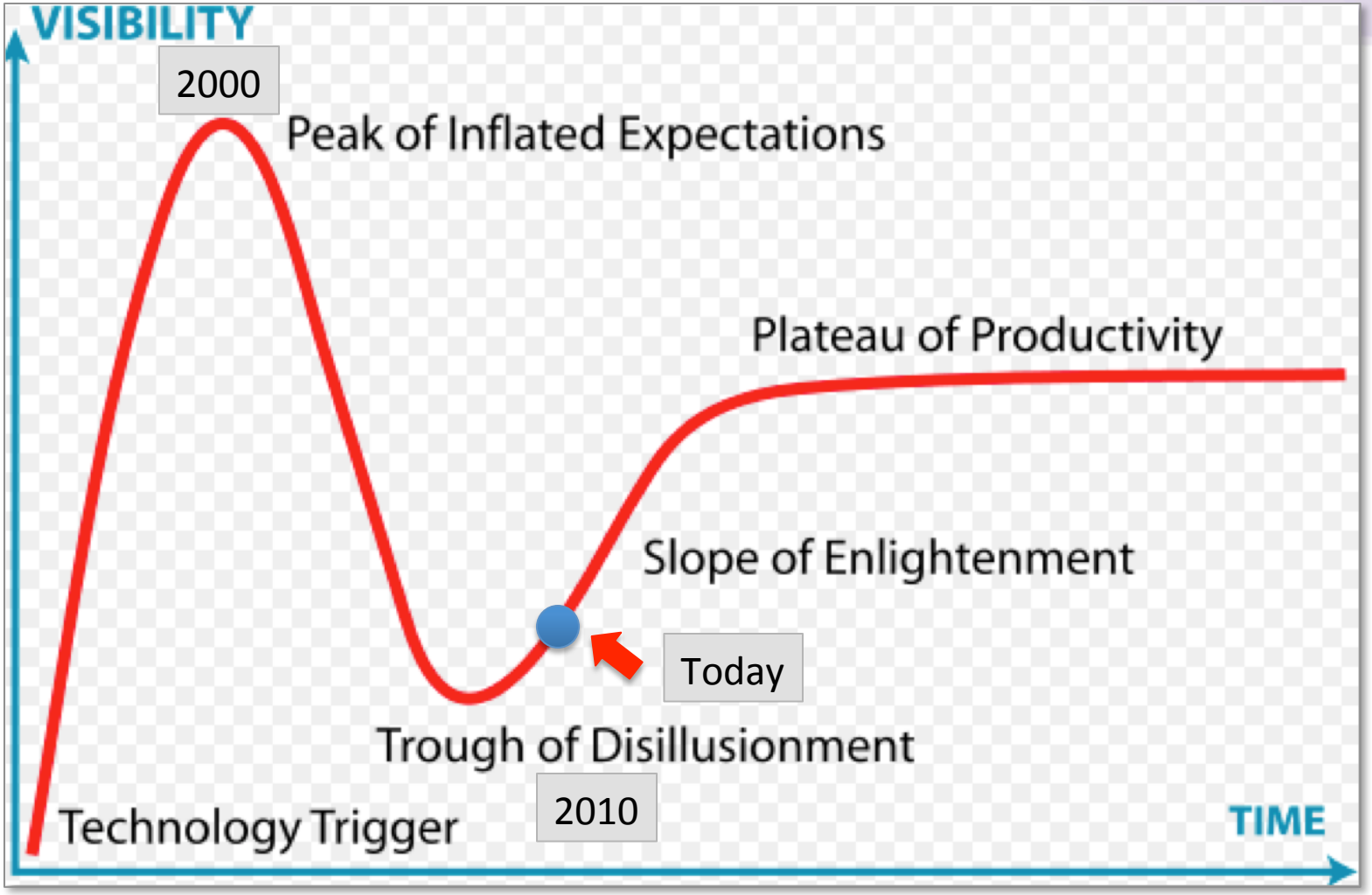
Substantial progress!

Actionable

Diagnosis, predictions

Surprisingly little progress!

Hype Cycle


VISIBILITY
2000
Peak of Inflated Expectations
Plateau of Productivity
Slope of Enlightenment
Today
Trough of Disillusionment
2010
Technology Trigger
TIME

Why are the advances in

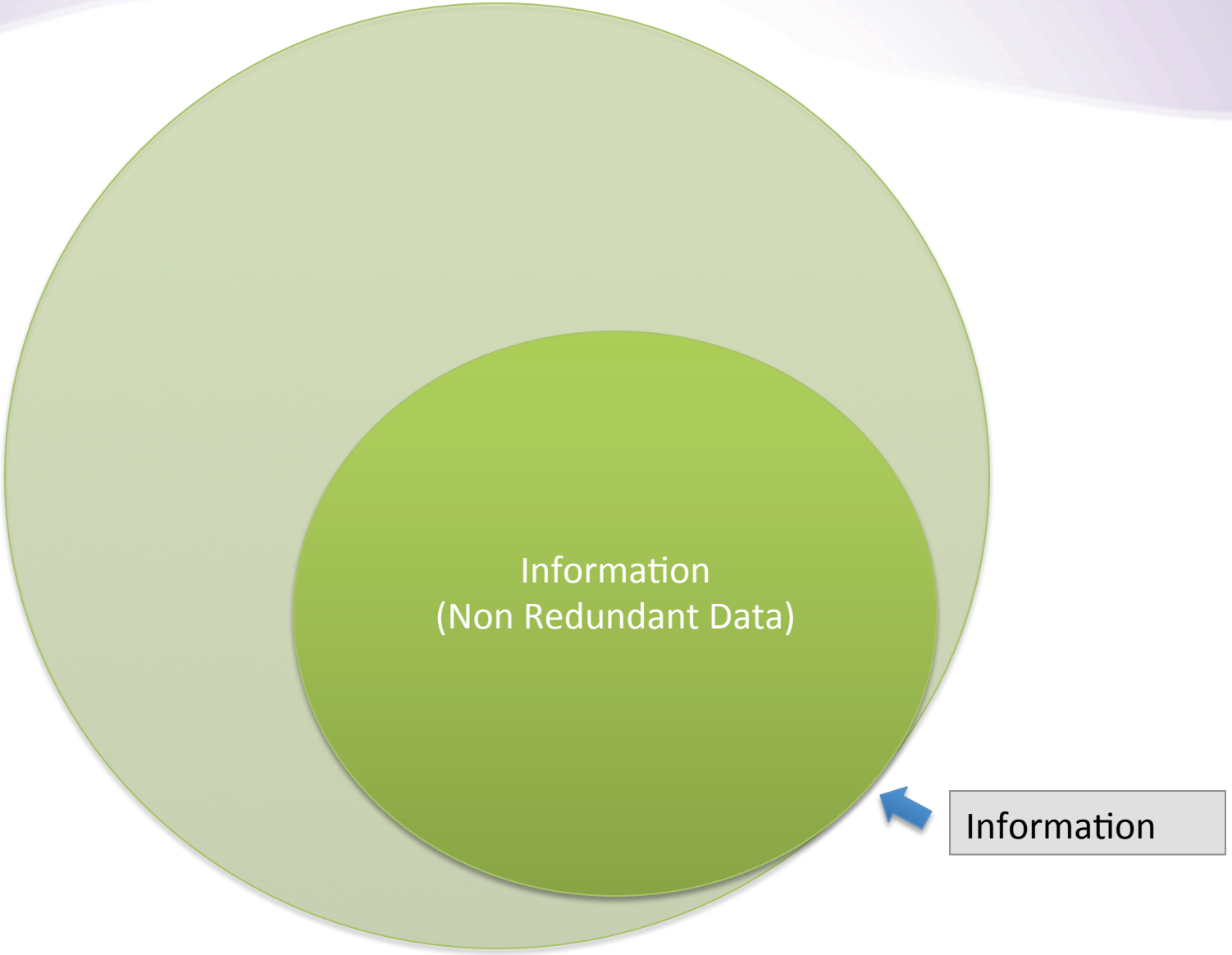
Actionable Bioinformatics

so slow?

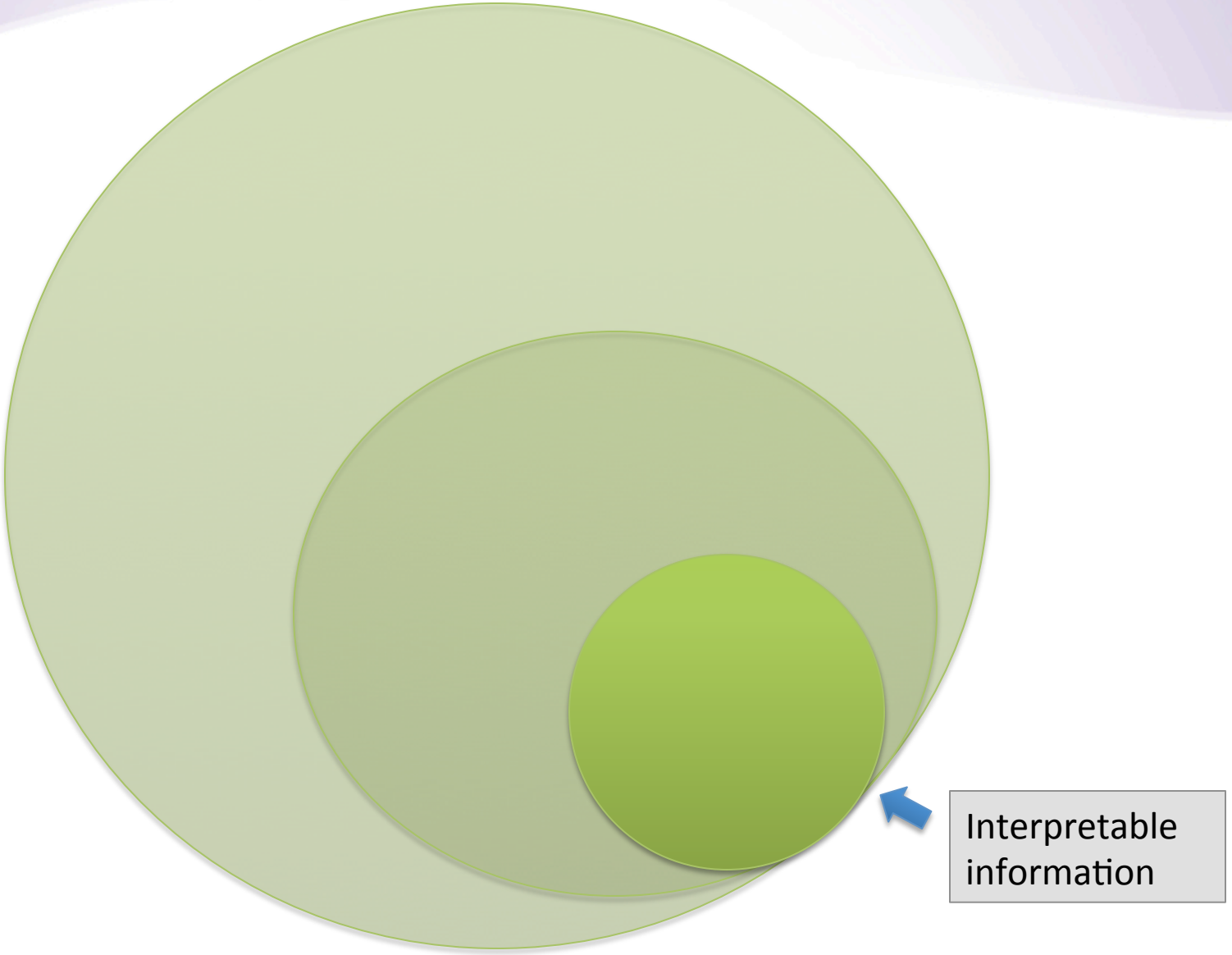
BIG DATA

Big Data ≠ Useful Information

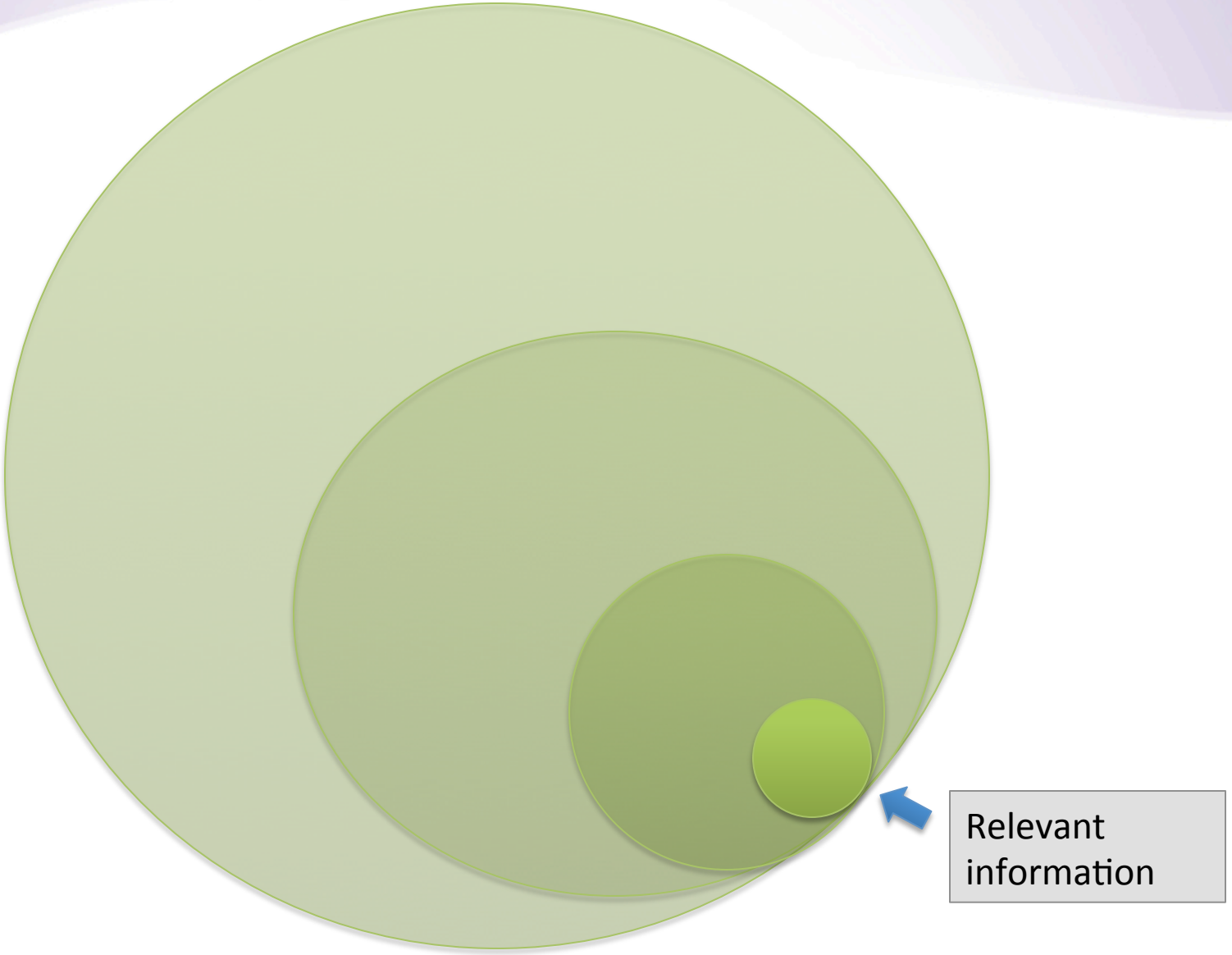
Big Data ≠ Useful Information



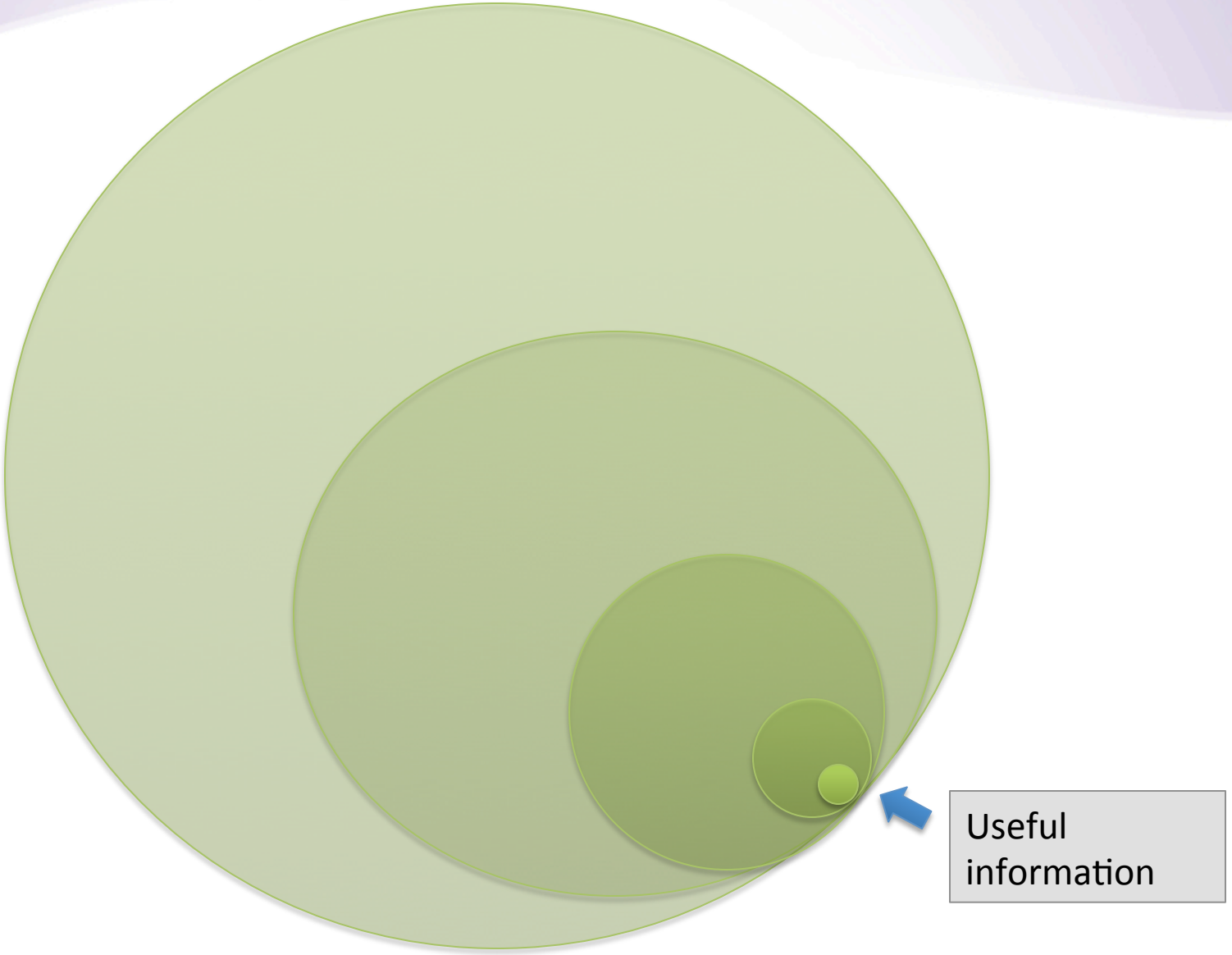
Big Data ≠ Useful Information



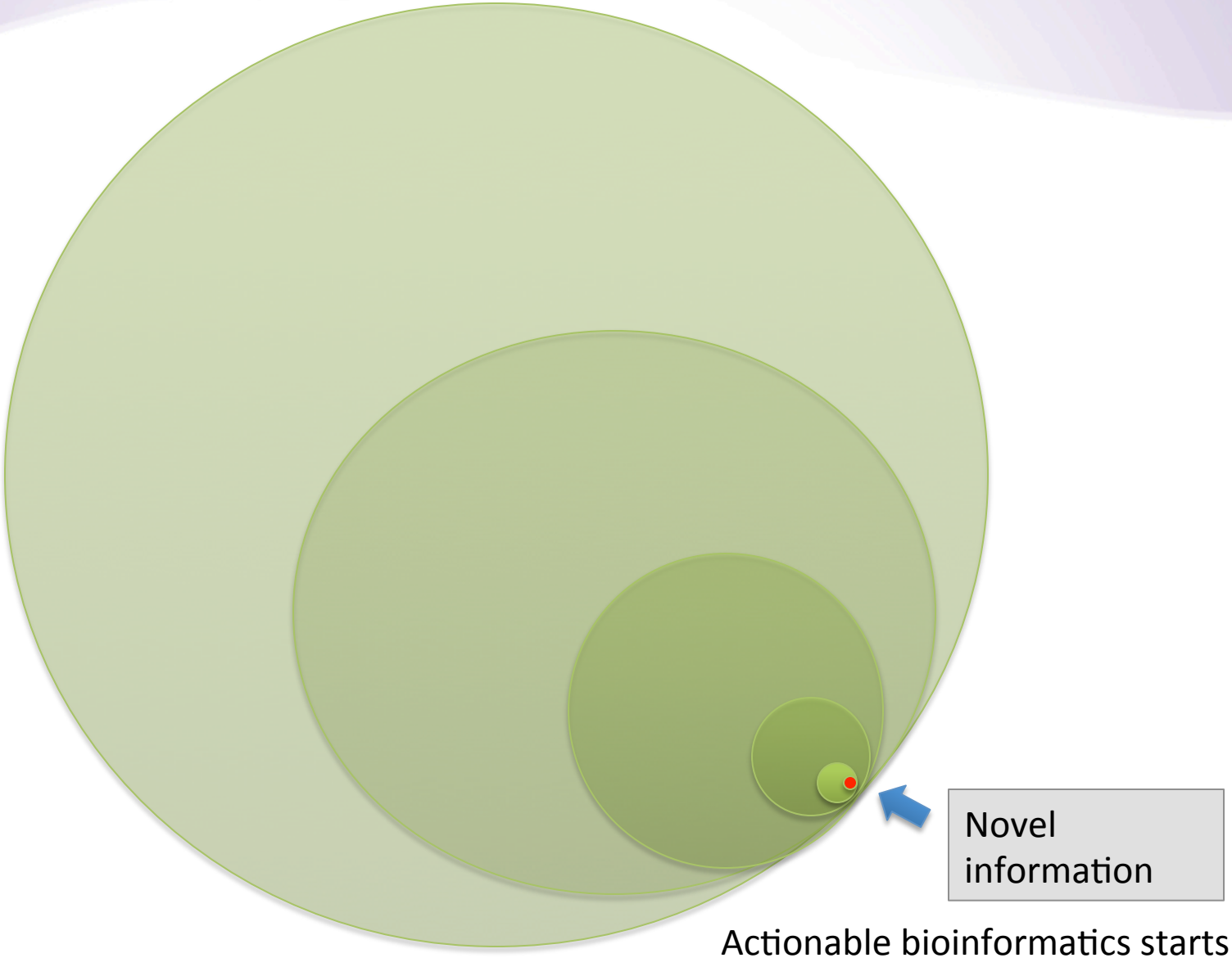
Big Data ≠ Useful Information



Big Data ≠ Useful Information

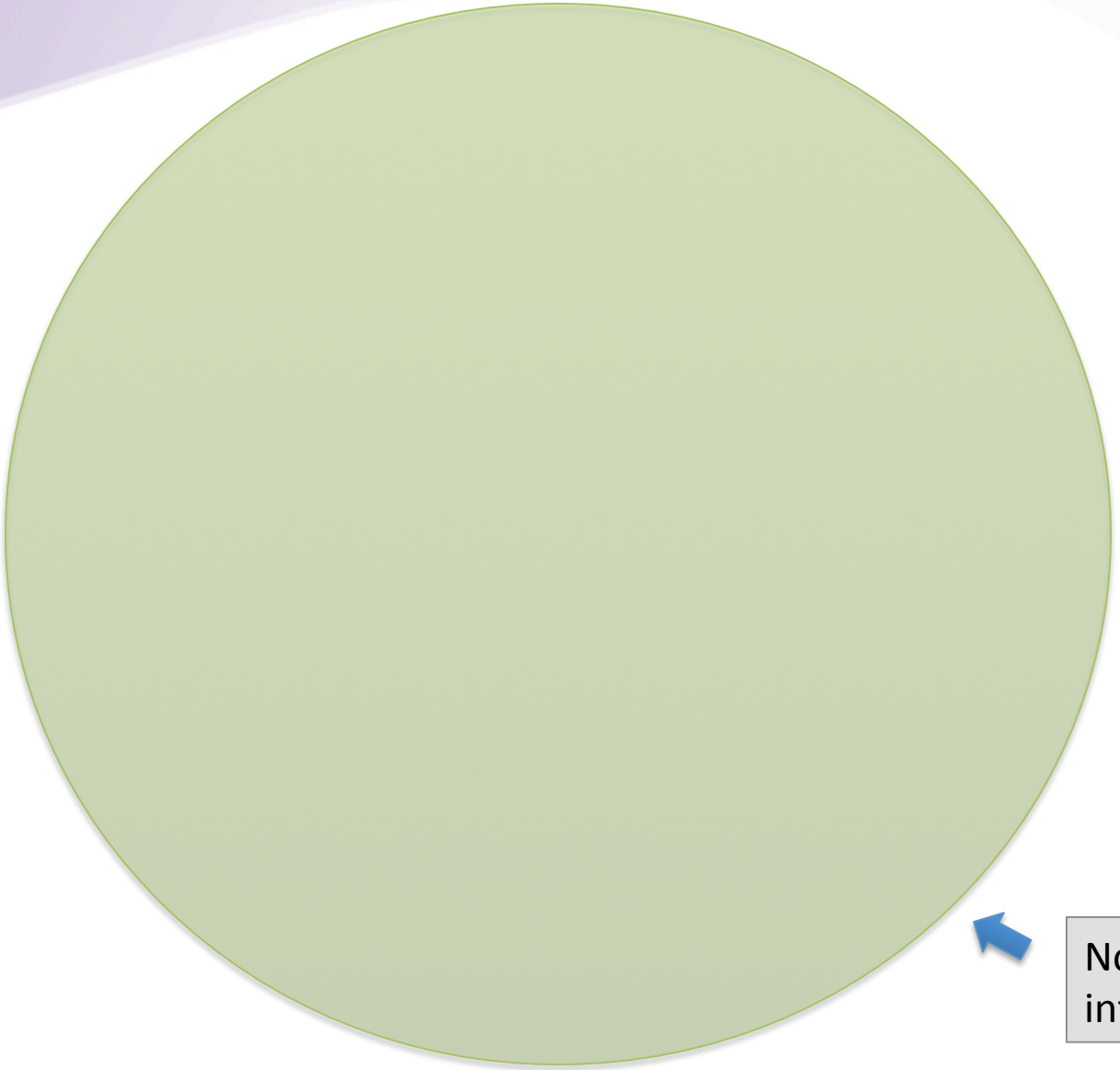


Big Data ≠ Useful Information



Actionable bioinformatics starts here

Big Data ≠ Useful Information



Let's make it realistic.

Big circle = data from the human genome covered at 10 fold coverage
(R = 6)

Suppose that a disease is caused by a single SNP that happens to be covered with 10 measurements.

Then the radius of the small circle will be

r ~ R/sqrt(10^9)= 0.0002

Novel information

Why does the data grow so large?

1. **Technological Limitations** – it is easier to generate more **reads** than longer ones. Insufficient financial incentive from customers.

2. **Representational Limitations** → the empirical data standards are often inefficient and seemingly impossible to change/adapt.

3. **Lack of education** on the researchers' side leads to a incorrect approach

Big data → descriptive bioinformatics

- Big Data lends itself to making general observations about the large scale characteristics of a genome

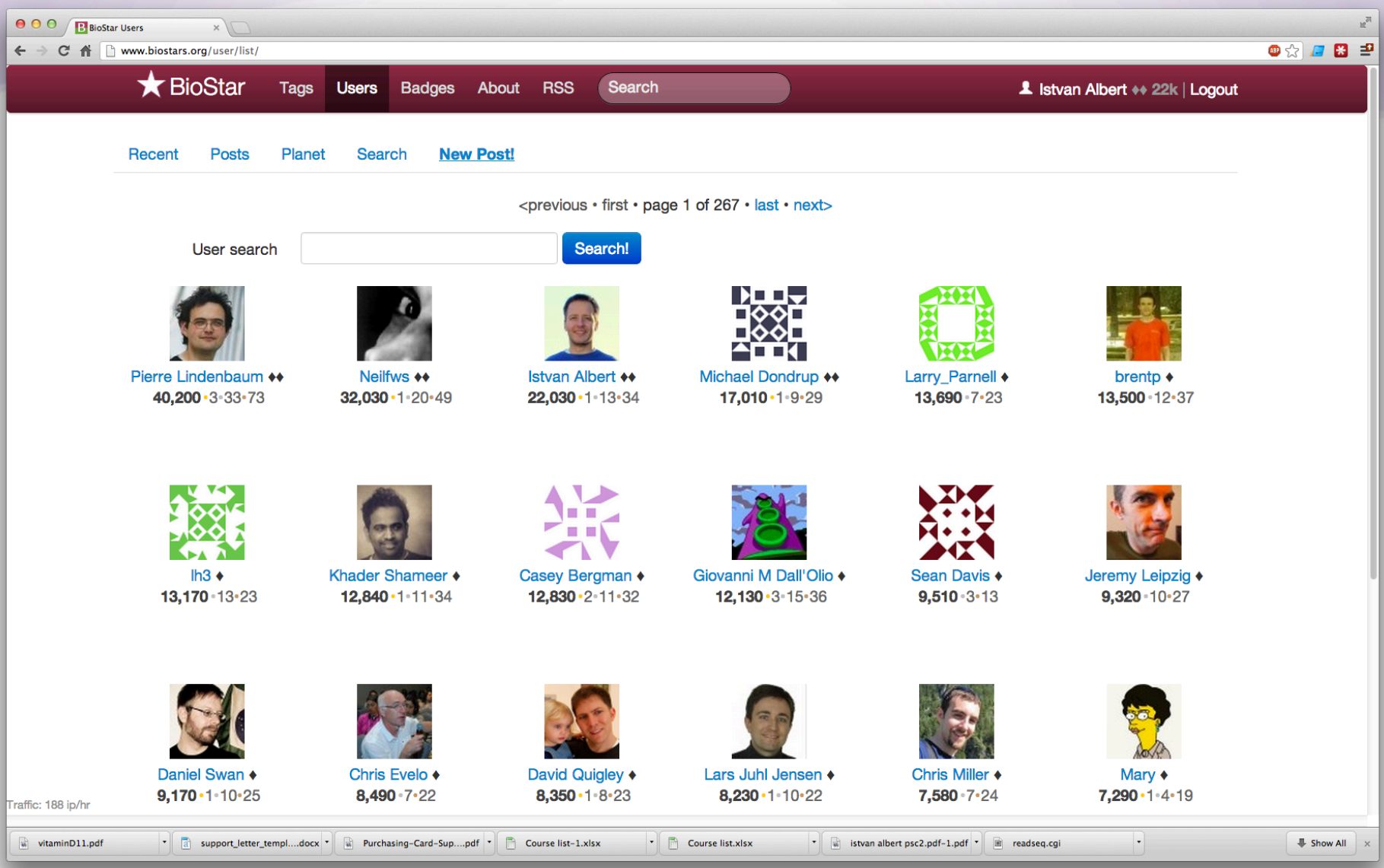
- It makes it exceedingly difficult to pinpoint one particular characteristic – and this is unlikely to change!

Education is the key!

- Making an analysis easy by providing a button that one can click is not the right approach!
- A typical analysis involves making hundreds of decisions – most of which need to be correct!
- One needs to understand the process at a deeper level.

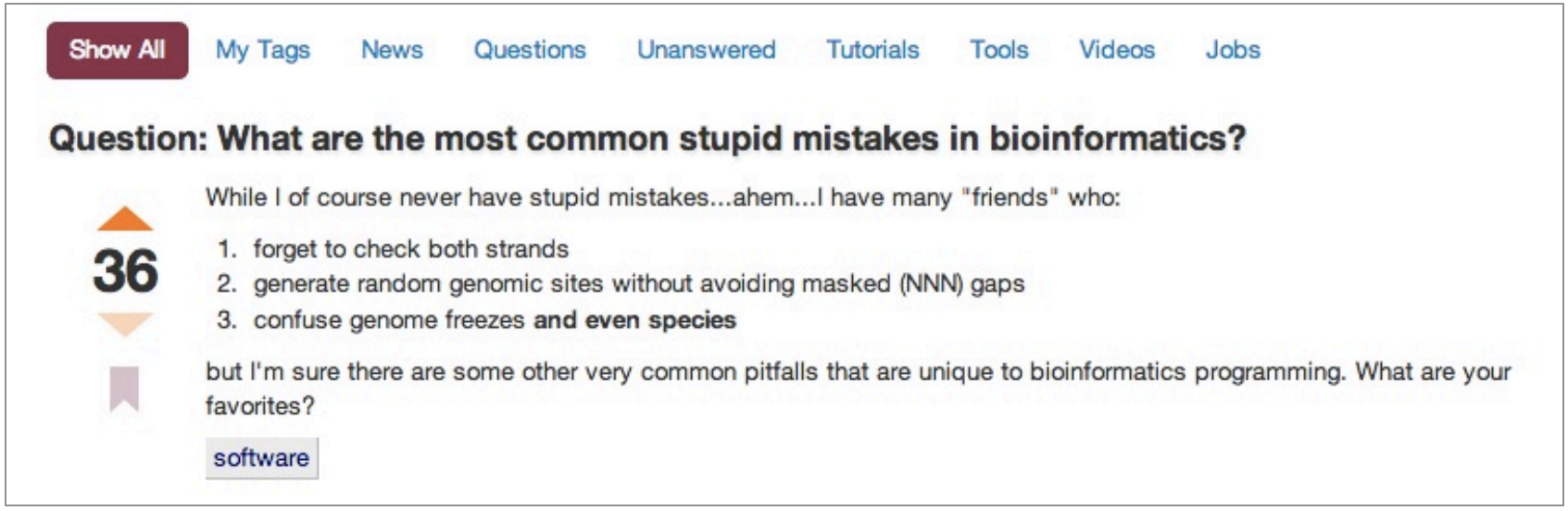
http://www.biostars.org

BioStar Users



BioStar a window into the challenges of bioinformatics analysis

- Eye opening as well as enlightening
- Plus entertaining



How to make a difference

- The right training can provide **computational competence** within six months/one year period.
- Treating bioinformatics as validation of a hypothesis not as discovery.
- The real challenge is to combine existing information, validate results, visualize data.