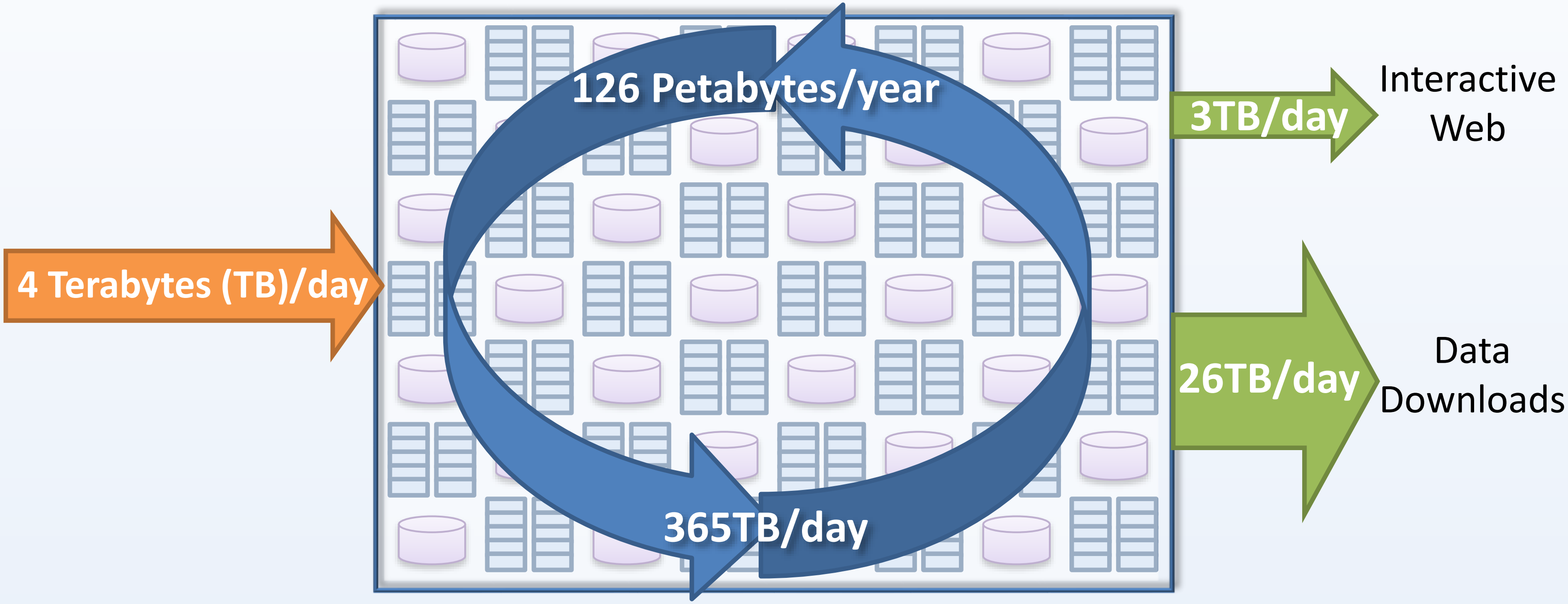
What is Big Data?

- Information Retrieval *not* Big Data
- Computing across TB’s of data (Higgs boson) *is* Big Data
- Computing across credit card data to detect possible fraud is Big Data

TARGET & Diapers == *Predictive Analytics*

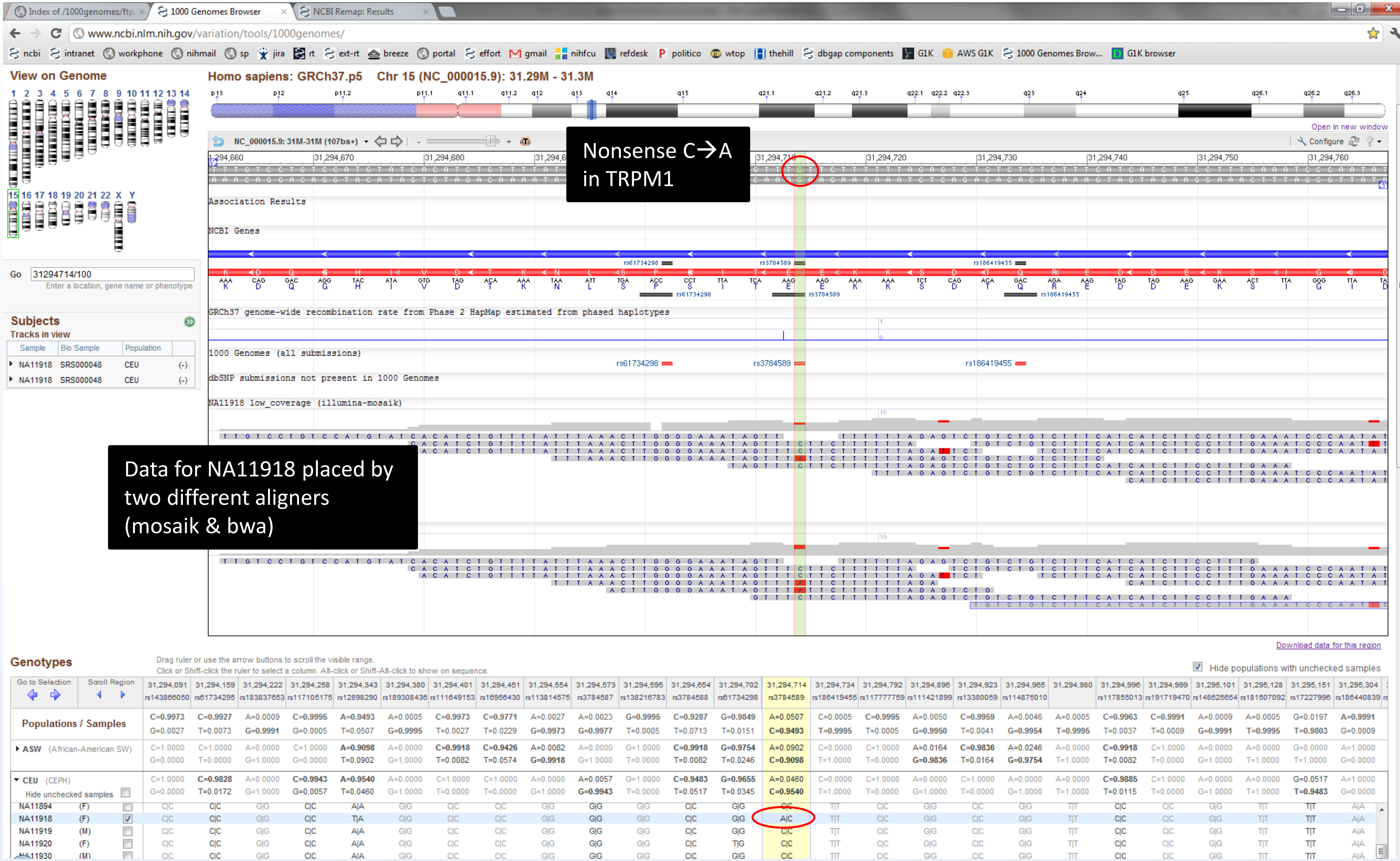
http://www.nytimes.com/2012/02/19/magazine/shopping-habits.html?smid=pl-share

Daily Data Processing at NCBI



2012 data…

Stationary night blindness due to premature termination in TRPM1


Index of /1000genomes/ftp
1000 Genomes Browser
NCBI Remap: Results
www.ncbi.nlm.nih.gov/variation/tools/1000genomes/
ncbi
intranet
workphone
nihmail
sp
jira
rt
ext-rt
breeze
portal
effort
gmail
nihfcu
refdesk
politico
wtop
thehill
dbgap components
G1K
AWS G1K
1000 Genomes Brow...
G1K browser
View on Genome
1 2 3 4 5 6 7 8 9 10 11 12 13 14
15 16 17 18 19 20 21 22 X Y
Go
31294714/100
Enter a location, gene name or phenotype
Homo sapiens: GRCh37.p5
Chr 15 (NC_000015.9): 31.29M - 31.3M
p13 p12 p11.2 p11.1 q11.1 q11.2 q12 q13 q14 q15 q21.1 q21.2 q21.3 q22.1 q22.2 q22.3 q23 q24 q25 q26.1 q26.2 q26.3
Open in new window
NC_000015.9: 31M-31M (107bs+)
Configure
1,294,660
31,294,670
31,294,680
31,294,6
31,294,71
31,294,720
31,294,730
31,294,740
31,294,750
31,294,760
T T T G T C C T G T C C A T G T A T C A C A T C T G T T T T A T
A A A C A G G A C A G G T A C A T A G T G T A G A C A A A A T A
G T T T C T T C T T T T T A G A G T C T G T C T G T C T T T C A T C A T C T T C C T T T G A A A T C C C A A T A T
C A A A G A A G A A A A A A T C T C A G A C A G A C A G A A A G T A G T A G A A G G A A A C T T T A G G G T T A T A
Association Results
NCBI Genes
rs61734298
rs3784589
rs186419455
GRCh37 genome-wide recombination rate from Phase 2 HapMap estimated from phased haplotypes
1000 Genomes (all submissions)
rs61734298
rs3784589
rs186419455
dbSNP submissions not present in 1000 Genomes
NA11918 low_coverage (illumina-mosaik)
Subjects
Tracks in view
Sample
Bio Sample
Population
Download data for this region
☑ Hide populations with unchecked samples

Sample	Bio Sample	Population	
NA11918	SRS000048	CEU	(-)
NA11918	SRS000048	CEU	(-)

Genotypes

Drag ruler or use the arrow buttons to scroll the visible range.
Click or Shift-click the ruler to select a column. Alt-click or Shift-Alt-click to show on sequence.

Go to Selection / Scroll Region		31,294,091 rs143866050	31,294,159 rs61734295	31,294,222 rs183837653	31,294,258 rs117105175	31,294,343 rs12898290	31,294,380 rs189308436	31,294,401 rs111649153	31,294,451 rs16956430	31,294,554 rs113814575	31,294,573 rs3784587	31,294,595 rs138216783	31,294,654 rs3784588	31,294,702 rs61734298
Populations / Samples		C=0.9973 G=0.0027	C=0.9927 T=0.0073	A=0.0009 G=0.9991	C=0.9995 G=0.0005	A=0.9493 T=0.0507	A=0.0005 G=0.9995	C=0.9973 T=0.0027	C=0.9771 T=0.0229	A=0.0027 G=0.9973	A=0.0023 G=0.9977	G=0.9995 T=0.0005	C=0.9287 T=0.0713	G=0.9849 T=0.0151
ASW (African-American SW)		C=1.0000 G=0.0000	C=1.0000 T=0.0000	A=0.0000 G=1.0000	C=1.0000 G=0.0000	A=0.9098 T=0.0902	A=0.0000 G=1.0000	C=0.9918 T=0.0082	C=0.9426 T=0.0574	A=0.0082 G=0.9918	A=0.0000 G=1.0000	G=1.0000 T=0.0000	C=0.9918 T=0.0082	G=0.9754 T=0.0246
CEU (CEPH) Hide unchecked samples ☐		C=1.0000 G=0.0000	C=0.9828 T=0.0172	A=0.0000 G=1.0000	C=0.9943 G=0.0057	A=0.9540 T=0.0460	A=0.0000 G=1.0000	C=1.0000 T=0.0000	C=1.0000 T=0.0000	A=0.0000 G=1.0000	A=0.0057 G=0.9943	G=1.0000 T=0.0000	C=0.9483 T=0.0517	G=0.9655 T=0.0345
NA11894 (F)	☐	C|C	C|C	G|G	C|C	A|A	G|G	C|C	C|C	G|G	G|G	G|G	C|C	G|G
NA11918 (F)	☑	C|C	C|C	G|G	C|C	T|A	G|G	C|C	C|C	G|G	G|G	G|G	C|C	G|G
NA11919 (M)	☐	C|C	C|C	G|G	C|C	A|A	G|G	C|C	C|C	G|G	G|G	G|G	C|C	G|G
NA11920 (F)	☐	C|C	C|C	G|G	C|C	A|A	G|G	C|C	C|C	G|G	G|G	G|G	C|C	T|G
NA11930 (M)	☐	C|C	C|C	G|G	C|C	A|A	G|G	C|C	C|C	G|G	G|G	G|G	C|C	G|G

Populations / Samples	31,294,714 rs3784589	31,294,734 rs186419455	31,294,792 rs117777759	31,294,896 rs111421899	31,294,923 rs13380059	31,294,965 rs114875010	31,294,980	31,294,996 rs117855013	31,294,999 rs191719470	31,295,101 rs148625654	31,295,128 rs181507092	31,295,151 rs17227996	31,295,304 rs186440839
	A=0.0507 C=0.9493	C=0.0005 T=0.9995	C=0.9995 T=0.0005	A=0.0050 G=0.9950	C=0.9959 T=0.0041	A=0.0046 G=0.9954	A=0.0005 T=0.9995	C=0.9963 T=0.0037	C=0.9991 T=0.0009	A=0.0009 G=0.9991	A=0.0005 T=0.9995	G=0.0197 T=0.9803	A=0.9991 G=0.0009
ASW (African-American SW)	A=0.0902 C=0.9098	C=0.0000 T=1.0000	C=1.0000 T=0.0000	A=0.0164 G=0.9836	C=0.9836 T=0.0164	A=0.0246 G=0.9754	A=0.0000 T=1.0000	C=0.9918 T=0.0082	C=1.0000 T=0.0000	A=0.0000 G=1.0000	A=0.0000 T=1.0000	G=0.0000 T=1.0000	A=1.0000 G=0.0000
CEU (CEPH)	A=0.0460 C=0.9540	C=0.0000 T=1.0000	C=1.0000 T=0.0000	A=0.0000 G=1.0000	C=1.0000 T=0.0000	A=0.0000 G=1.0000	A=0.0000 T=1.0000	C=0.9885 T=0.0115	C=1.0000 T=0.0000	A=0.0000 G=1.0000	A=0.0000 T=1.0000	G=0.0517 T=0.9483	A=1.0000 G=0.0000
NA11894 (F)	C|C	T|T	C|C	G|G	C|C	G|G	T|T	C|C	C|C	G|G	T|T	T|T	A|A
NA11918 (F)	A|C	T|T	C|C	G|G	C|C	G|G	T|T	C|C	C|C	G|G	T|T	T|T	A|A
NA11919 (M)	C|C	T|T	C|C	G|G	C|C	G|G	T|T	C|C	C|C	G|G	T|T	T|T	A|A
NA11920 (F)	C|C	T|T	C|C	G|G	C|C	G|G	T|T	C|C	C|C	G|G	T|T	T|T	A|A
NA11930 (M)	C|C	T|T	C|C	G|G	C|C	G|G	T|T	C|C	C|C	G|G	T|T	T|T	A|A

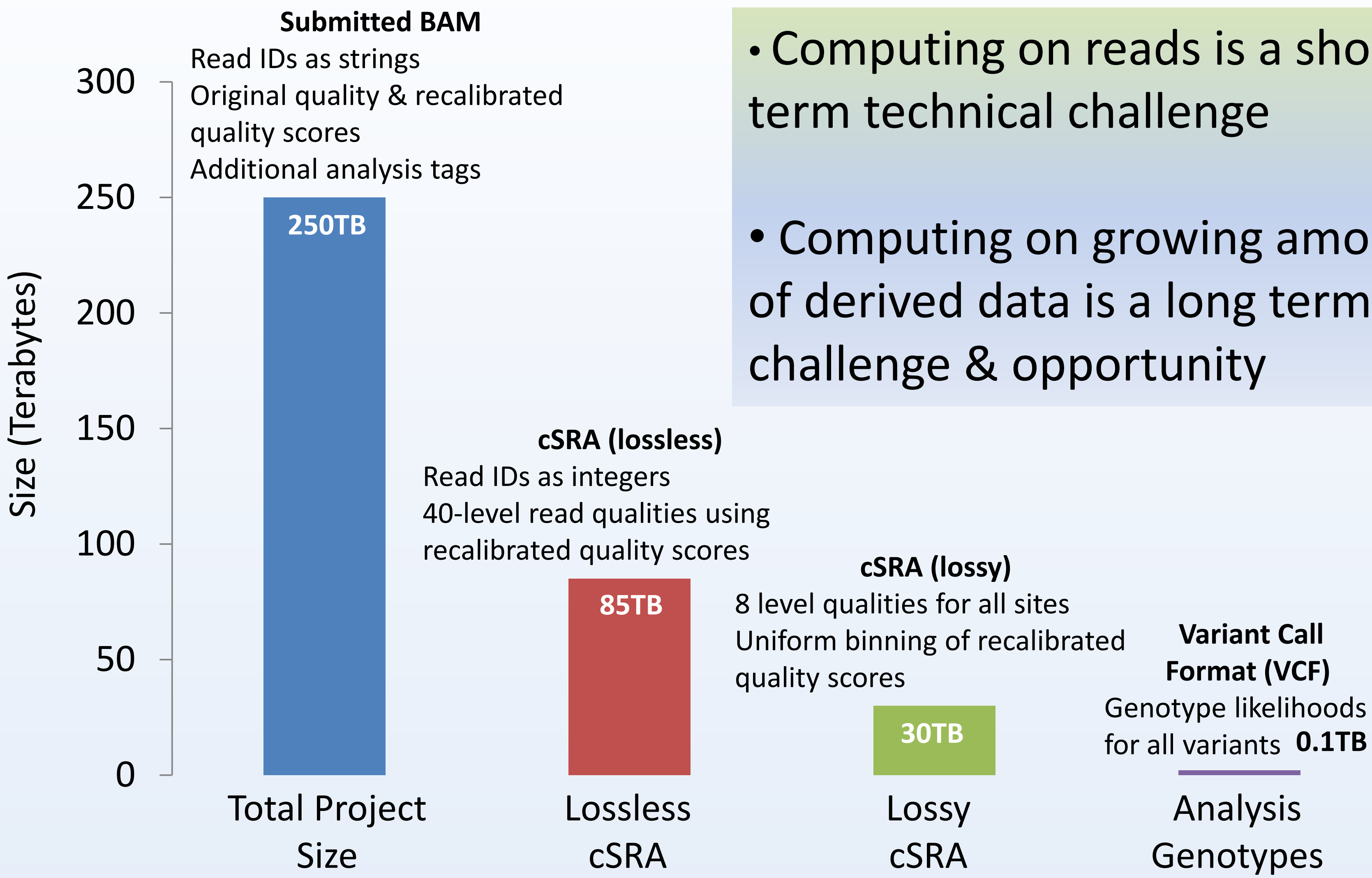
Nonsense C→A in TRPM1

Data for NA11918 placed by two different aligners (mosaik & bwa)

All individual genotypes For rs3784589

What is the Big Data Problem in Biology?
Example: Reducing the 1000 Genome Dataset

- Computing on reads is a short term technical challenge
- Computing on growing amounts of derived data is a long term challenge & opportunity



Submitted BAM
Read IDs as strings
Original quality & recalibrated quality scores
Additional analysis tags

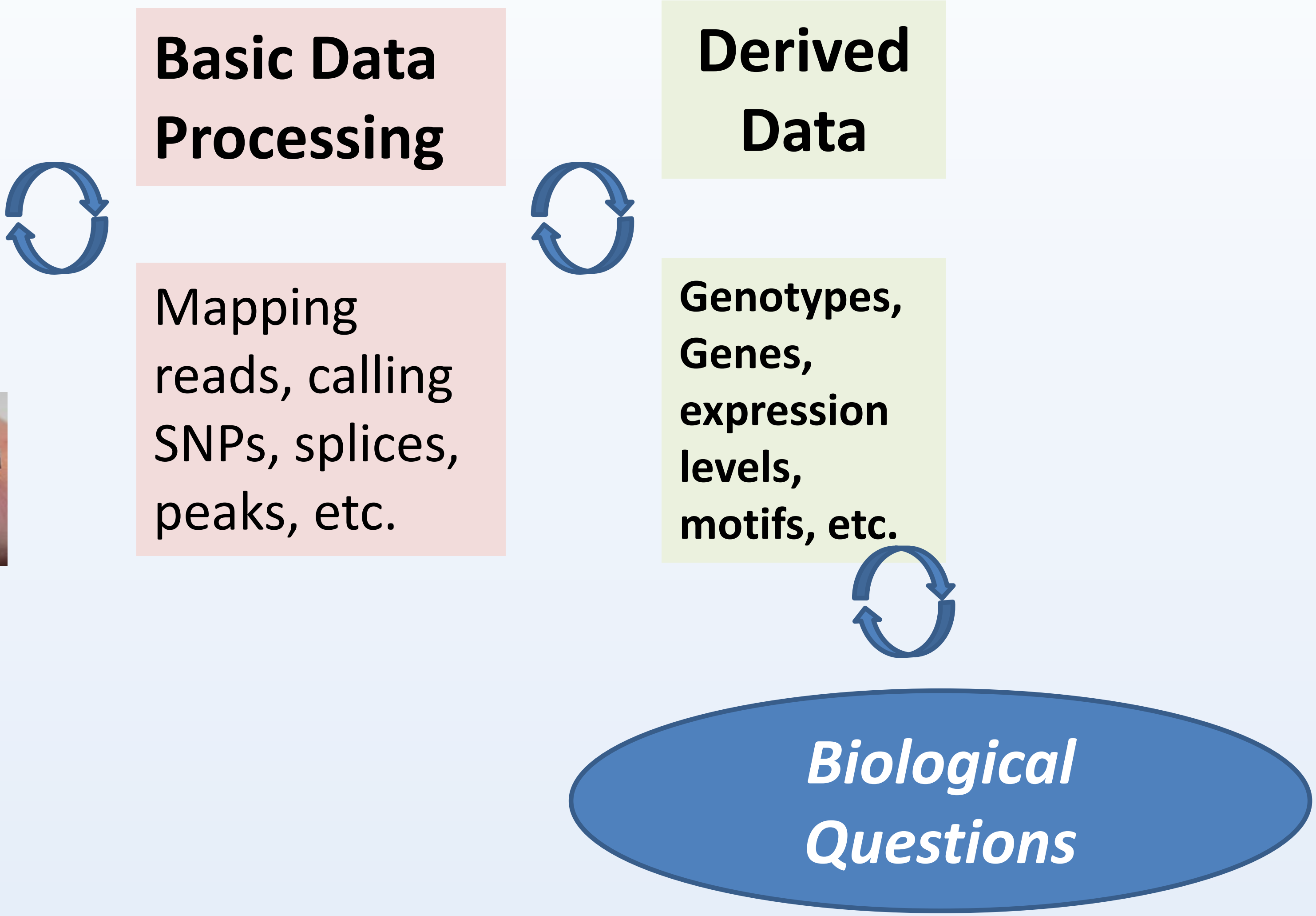
cSRA (lossless)
Read IDs as integers
40-level read qualities using recalibrated quality scores

cSRA (lossy)
8 level qualities for all sites
Uniform binning of recalibrated quality scores

Variant Call Format (VCF)
Genotype likelihoods for all variants

NextGen Churn

Etc….



Basic Data Processing

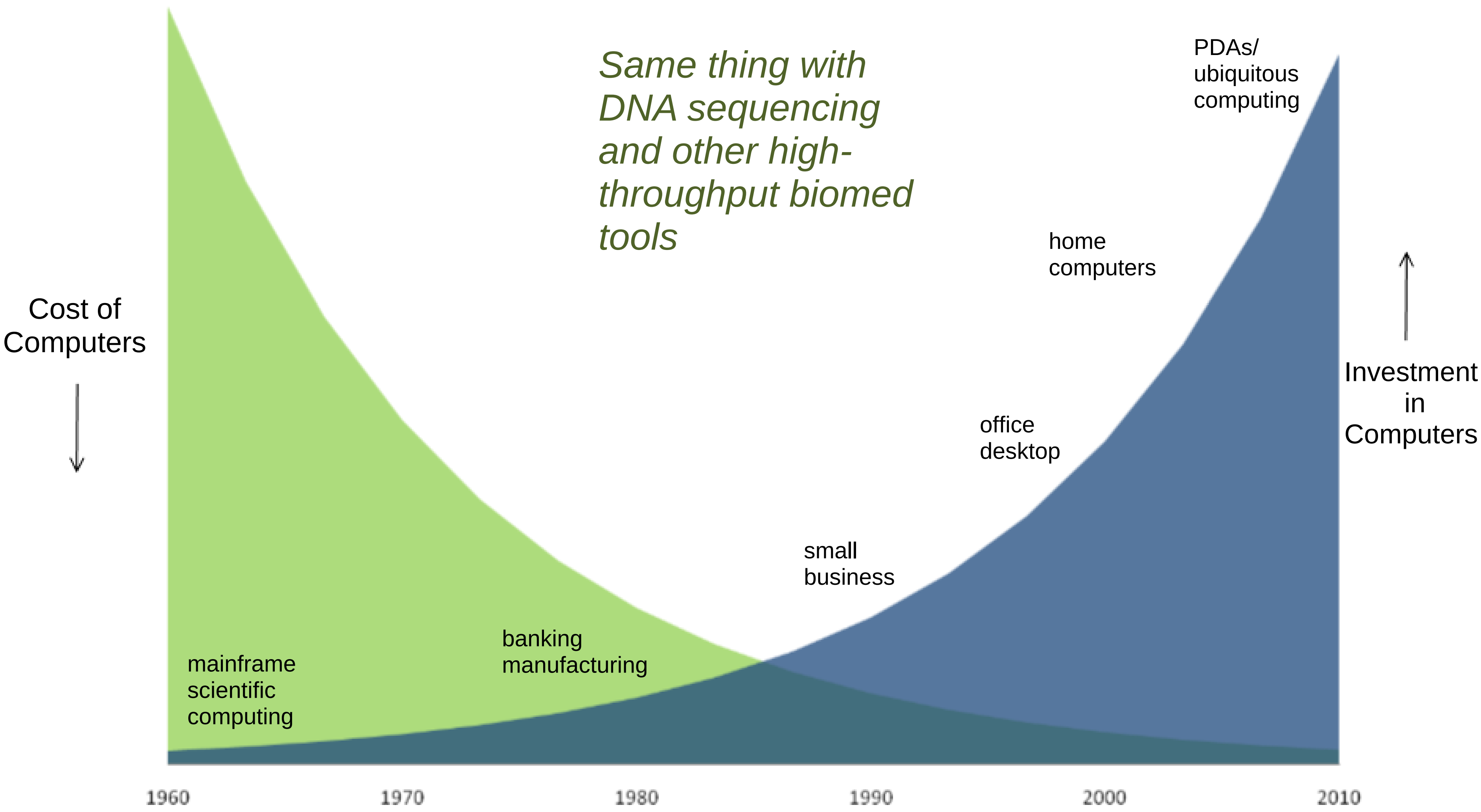
Mapping reads, calling SNPs, splices, peaks, etc.

Derived Data

Genotypes, Genes, expression levels, motifs, etc.

Biological Questions

As cost of computers decreased, overall investments increased...



Same thing with DNA sequencing and other high-throughput biomed tools

Economics of NextGen

Grantees want to maximize impact of budget

Sequencing will continue to increase if ***total cost*** *continues to drop:*

sample prep + raw sequencing + IT

Big Data:

(In

Ro

Ema

htt

What mess?

When

BIG DATA meets Small Signal

the signal has to be **teased out** of the data

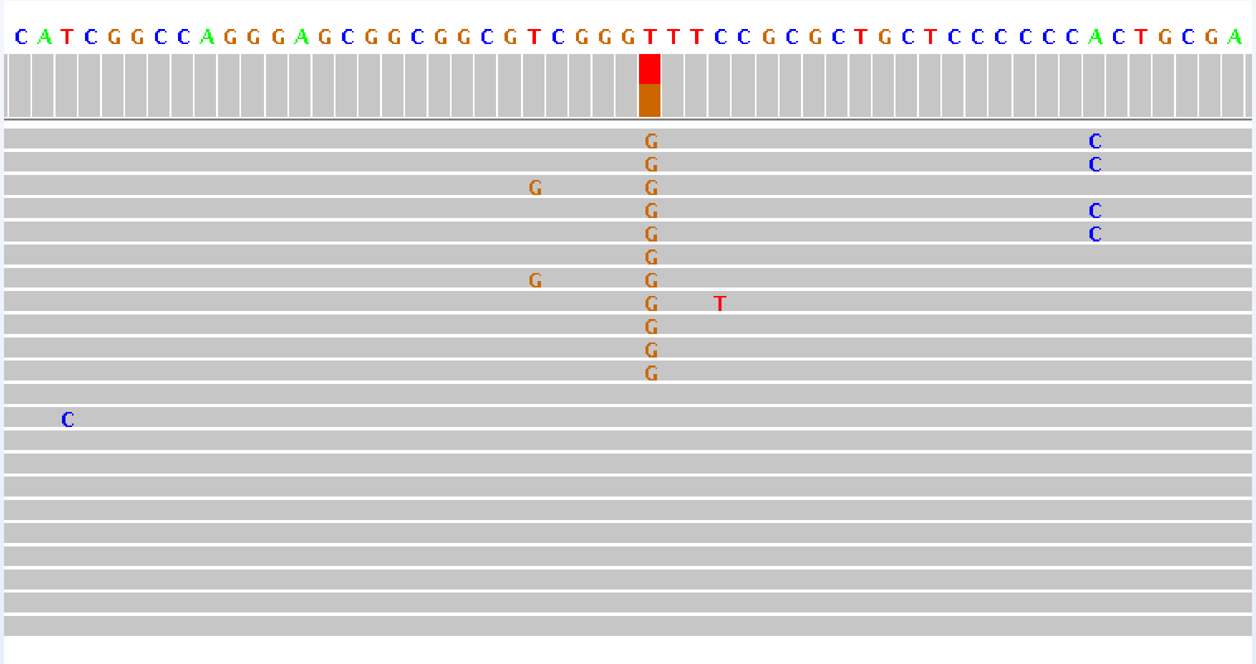
- There are many ways to do this: a **small change** in the analysis details can cause a **large change** in the results.
- It is too easy to distort your findings, either by **fooling yourself** or **on purpose**.

Systematic errors are routinely observed and have been reported in many papers

- Identification and correction of systematic error in high-throughput sequence data by Frazer Meacham, Dario Boffelli, Joseph Dhahbi, David IK Martin, Meromit Singer and Lior Pachter

  BMC Bioinformatics. 2011 Nov 21;12:451

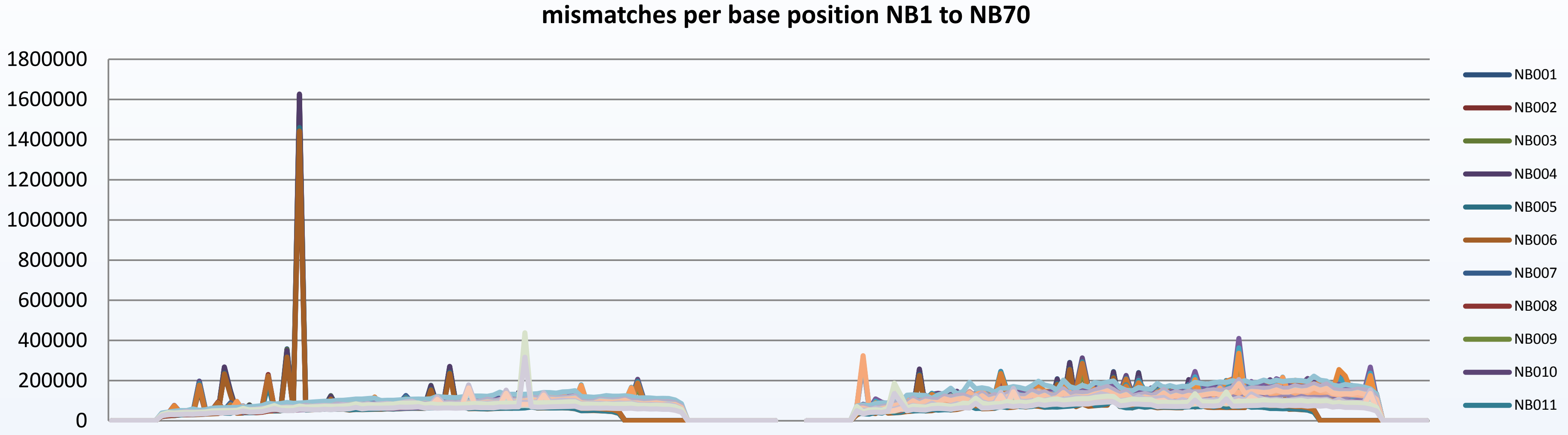
This paper shows the existence of systematic errors, even at high coverage, often strand dependent but not always



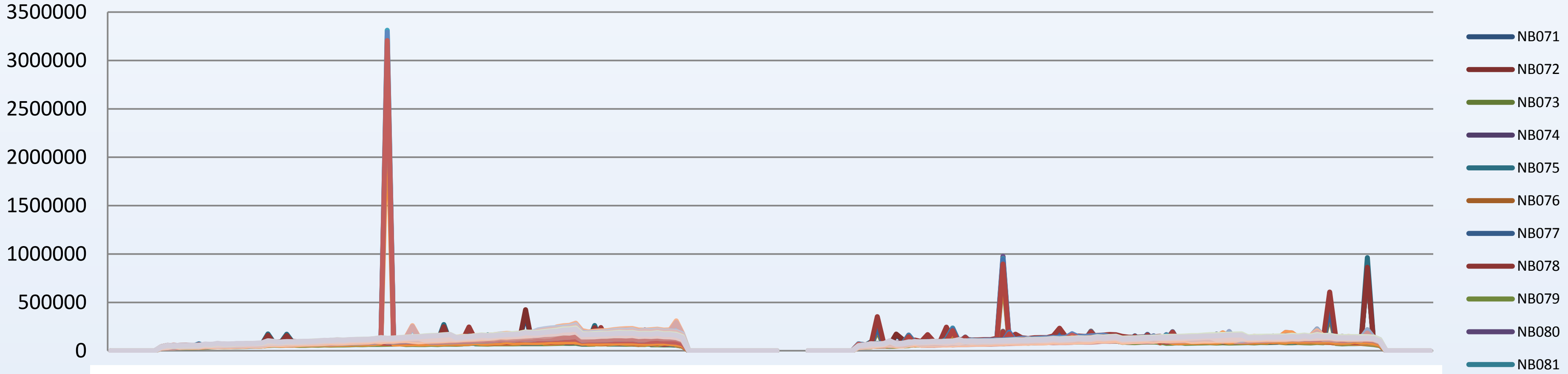
Strand +

Same fragment Strand -

Mismatch profiles are dominated by ‘spikes’ occurring at particular sequencing cycles (sequencing batch effect). Furthermore, each problematic cycle has a specific limited profile of mismatch types, adding to the bias



This effect on Illumina HiSeq is sequencing lane dependent

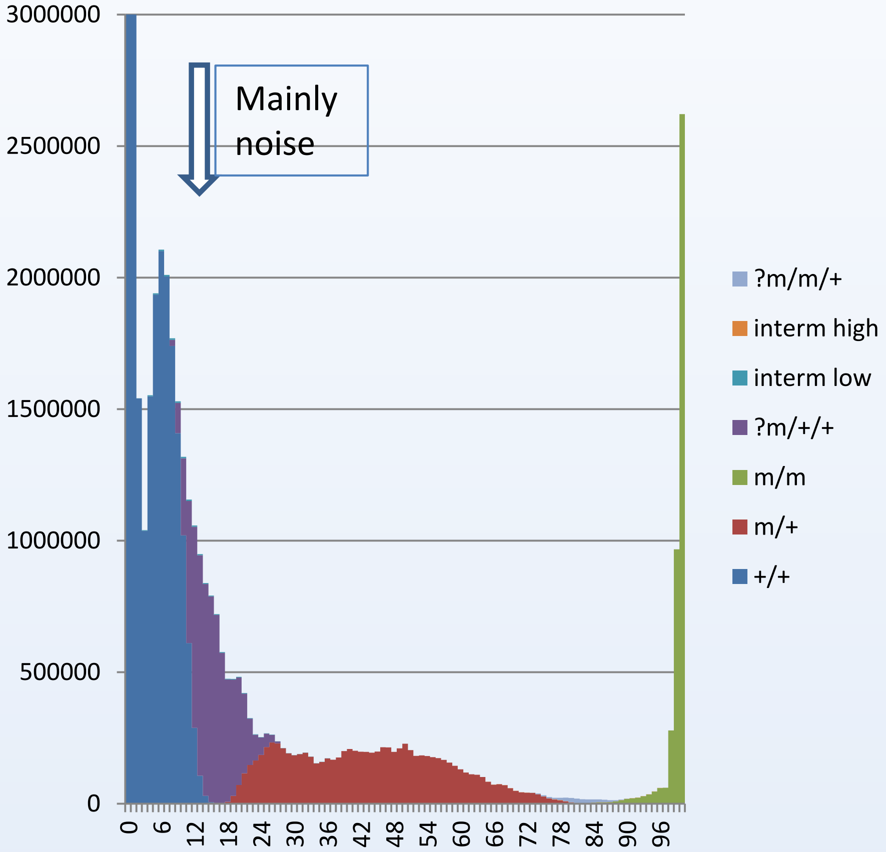


100 + 100 bp paired end reads, Illumina, Read 1 -------→ ----------------Read 2 -----------------------→

SEQC neuroblastoma study, L Shi, Fischer et al

Systematic errors generate noise in the low to intermediate allele fraction (1 to 30%), making identification of true SNPs hard in that area, even at high coverage

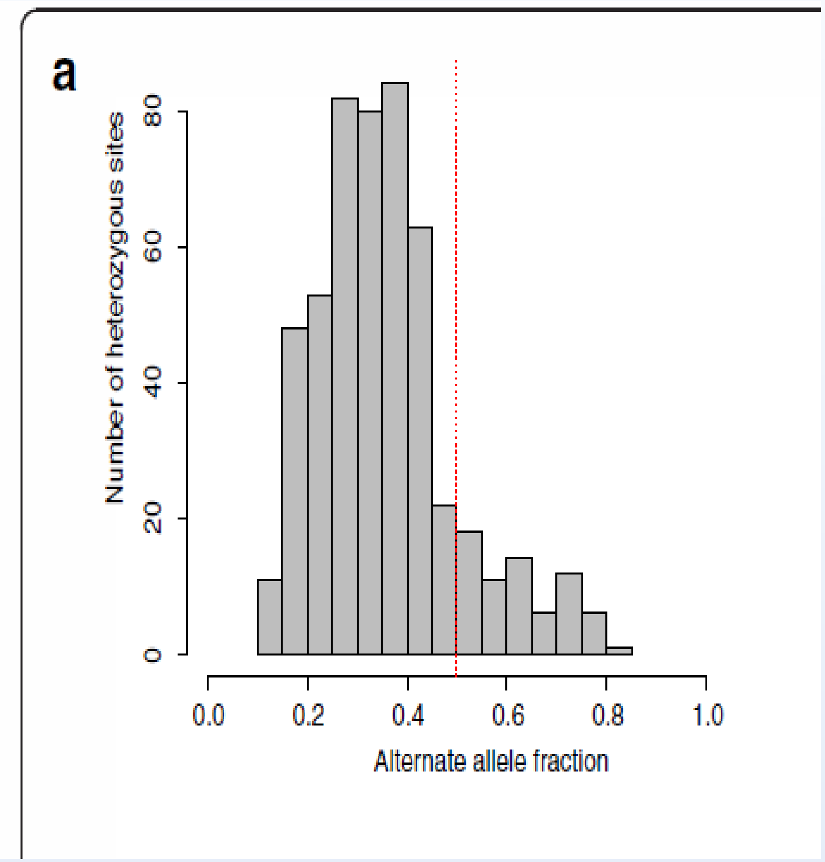
Histogram of candidate variant allele fraction before filtering: massive presence of SNP candidates with low to intermediate variant allele fraction values



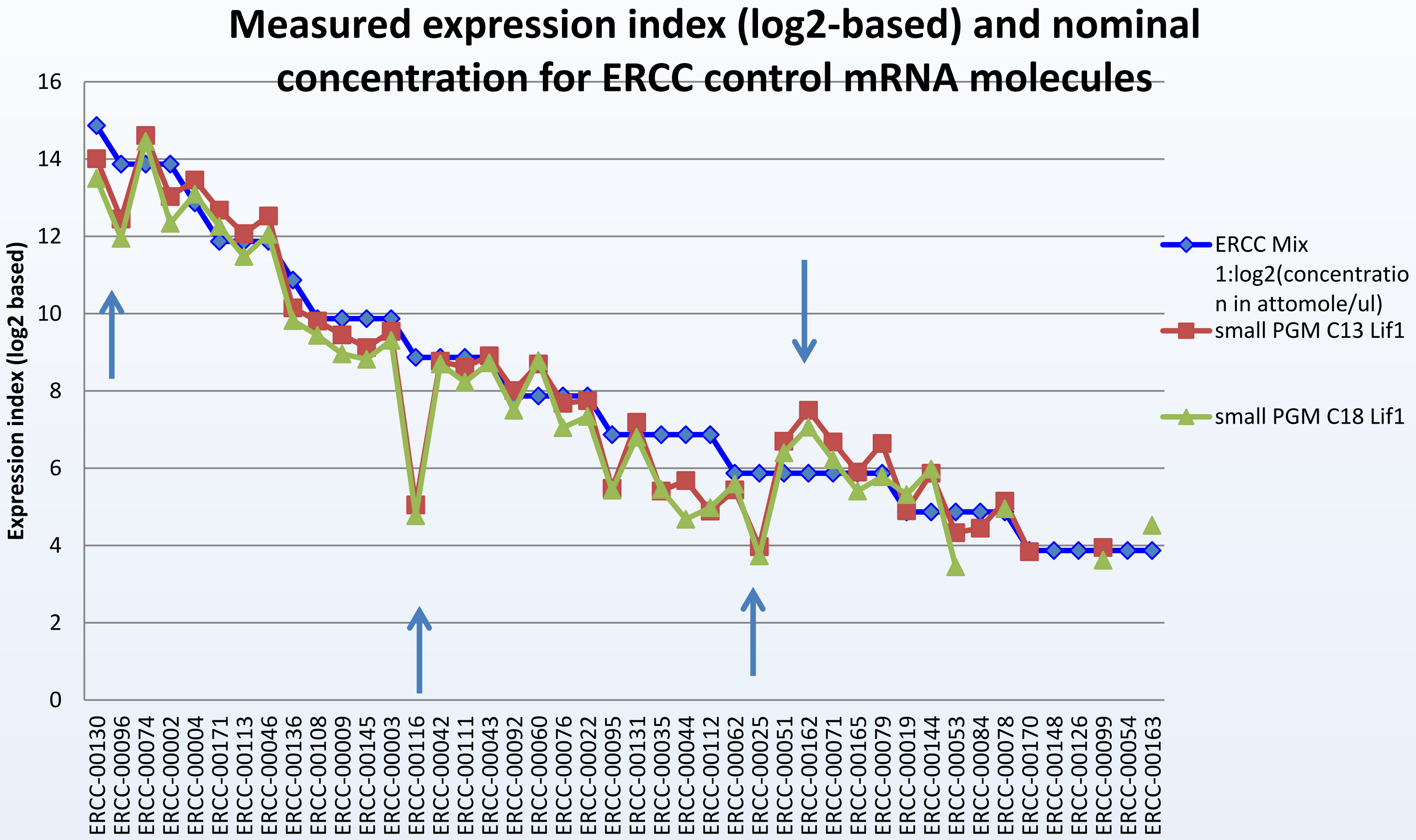
Pacific biosciences sequencing technology for genotyping and variation discovery in human data

Mauricio O Carneiro, Carsten Russ, Michael G Ross, Stacey B Gabriel, Chad Nusbaum and Mark A DePristo

***BMC Genomics* 2012, 13:375**



RNA-Seq often differs from 'true' concentration by a factor 2 or more. This affects all platforms, in particular Illumina and PGM/Proton/Solid



The two red and green runs should be superimposed on the blue nominal concentration.
Yet some specific ERCC molecules are vastly different, e.g. ERCC116 is measured 16 fold below nominal.

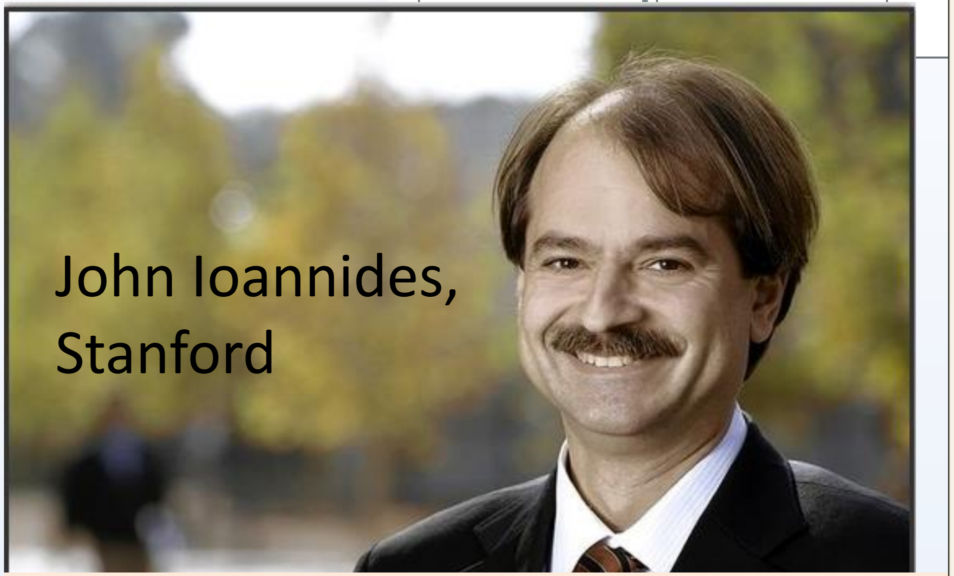
Data by Setterquist, LifeTech, QC for PGM

Friday, May 4, 2012 As of 10:24 PM

THE WALL STREET JOURNAL. | HEALTH

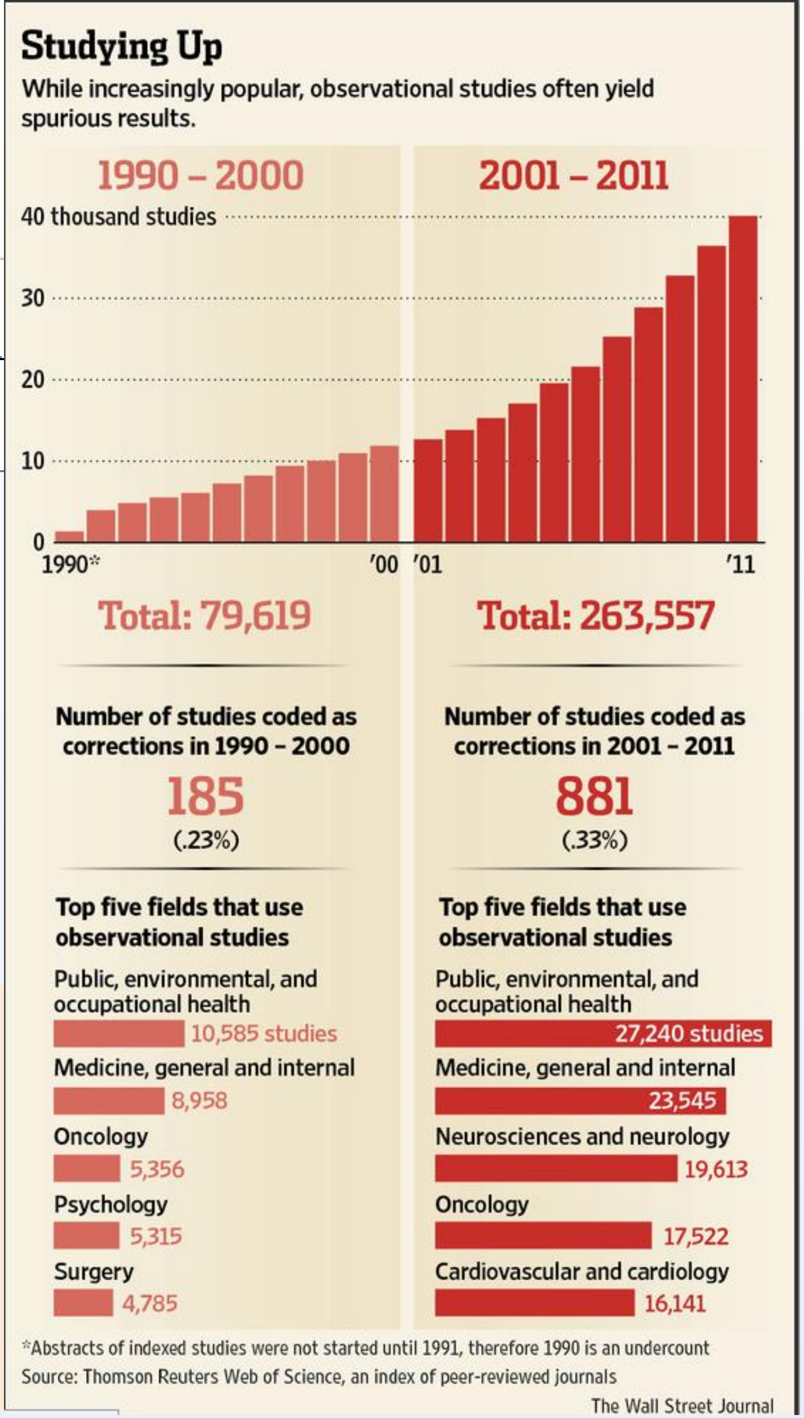
HEALTH INDUSTRY | Updated May 4, 2012, 10:24

Analytical Trend Trou



John Ioannides, Stanford

“That partly explains why observational studies in general can be replicated only 20% of the time, versus 80% for large, well-designed randomly controlled trials, says Dr. Ioannidis. Dr. Young, meanwhile, pegs the replication rate for observational data at an even lower 5% to 10%.”



Studying Up

While increasingly popular, observational studies often yield spurious results.

1990 – 2000	2001 – 2011
Total: 79,619	Total: 263,557
Number of studies coded as corrections in 1990 – 2000	Number of studies coded as corrections in 2001 – 2011
185	881
(.23%)	(.33%)

Top five fields that use observational studies (1990 – 2000)

Public, environmental, and occupational health	10,585 studies
Medicine, general and internal	8,958
Oncology	5,356
Psychology	5,315
Surgery	4,785

Top five fields that use observational studies (2001 – 2011)

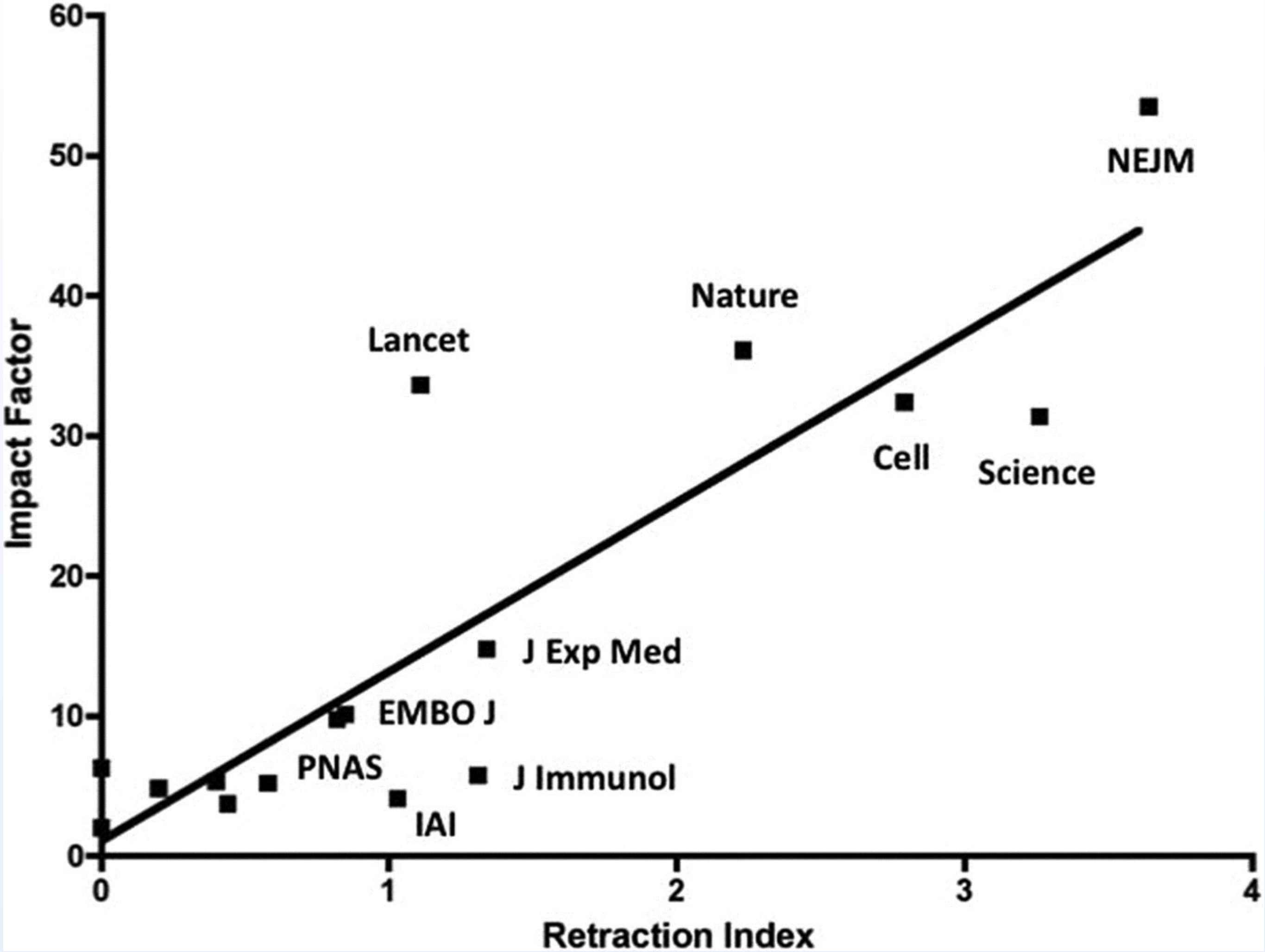
Public, environmental, and occupational health	27,240 studies
Medicine, general and internal	23,545
Neurosciences and neurology	19,613
Oncology	17,522
Cardiovascular and cardiology	16,141

*Abstracts of indexed studies were not started until 1991, therefore 1990 is an undercount

Source: Thomson Reuters Web of Science, an index of peer-reviewed journals

The Wall Street Journal

Correlation between impact factor and retraction index.



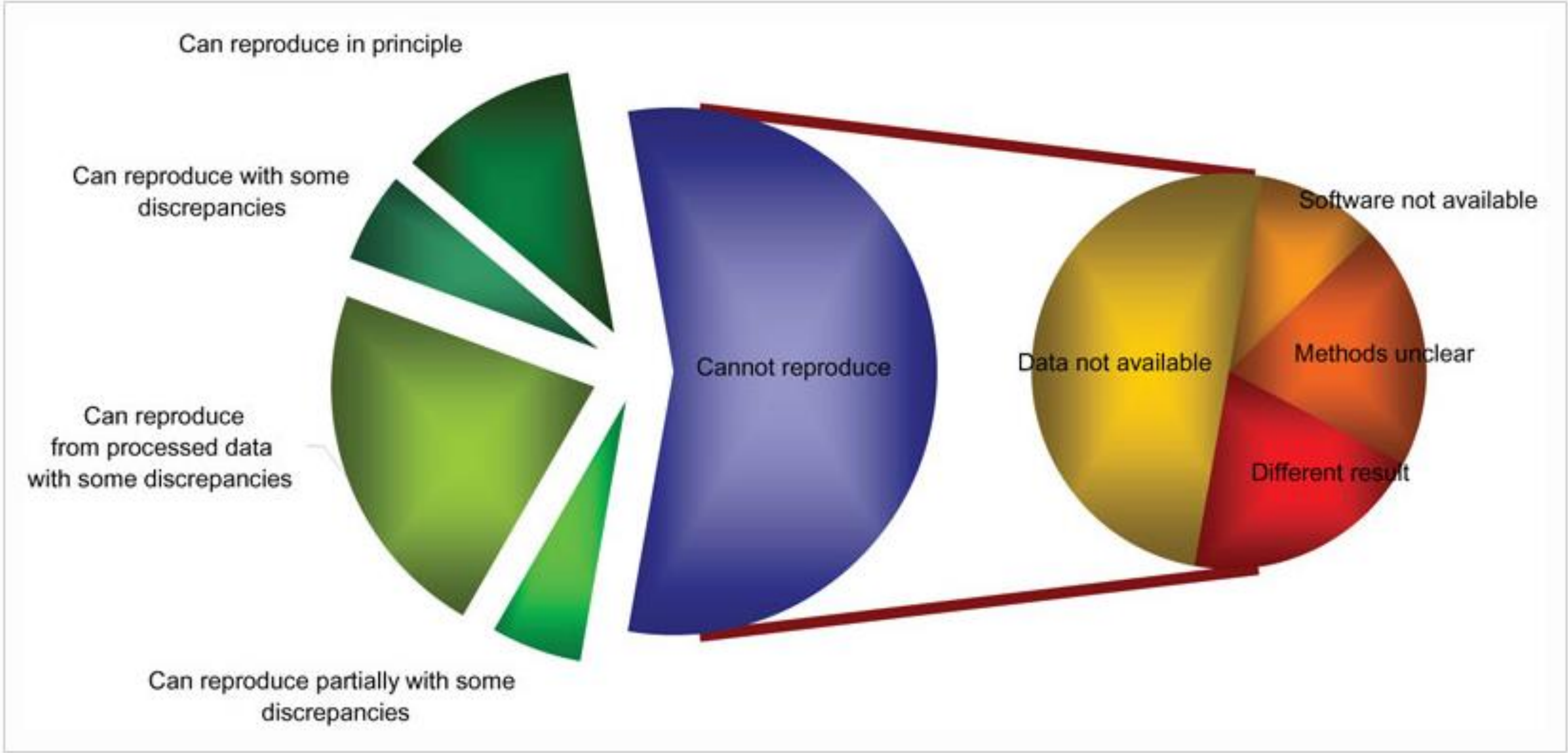
Fang F C , Casadevall A Infect. Immun. 2011;79:3855-3859

Infection and Immunity

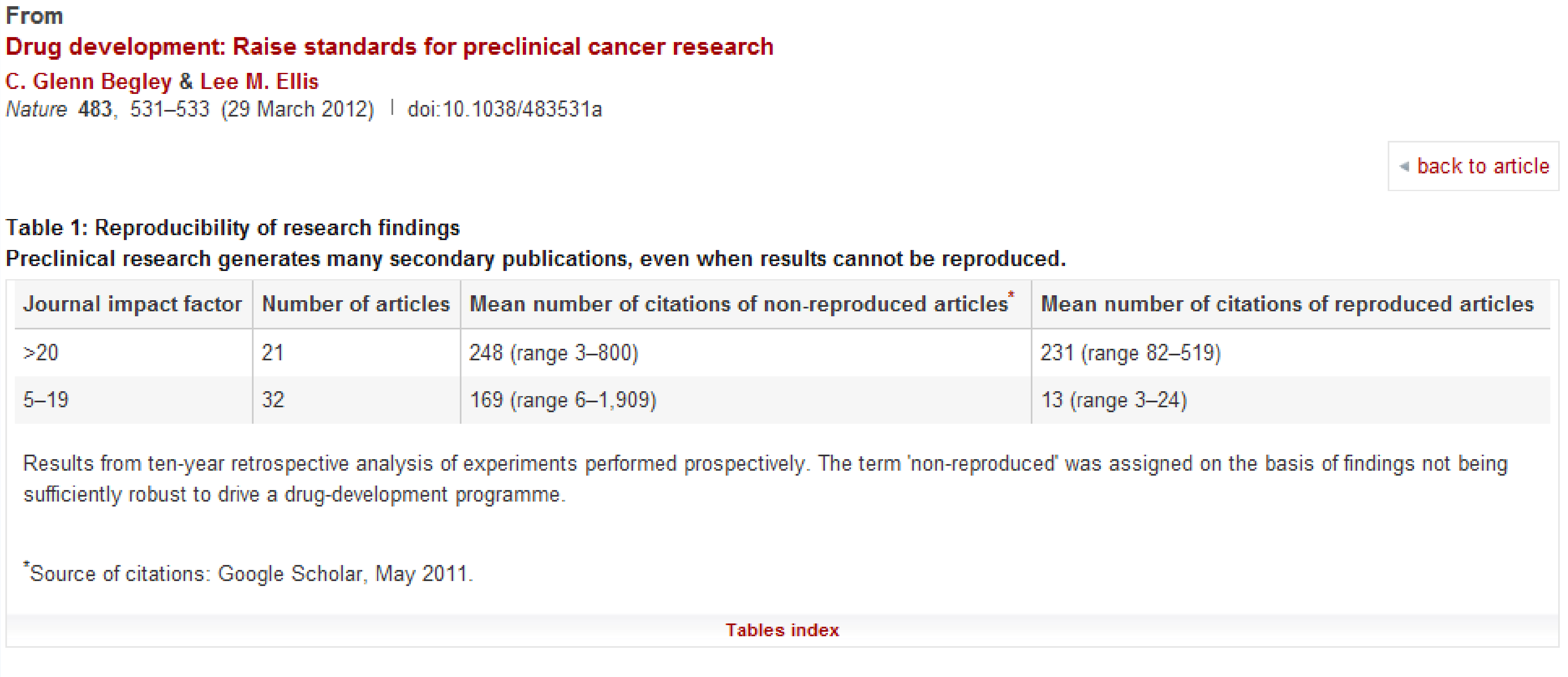
Journals.ASM.org | Copyright © American Society for Microbiology. All Rights Reserved.

Repeatability of published microarray gene expression analyses

Ioannidis et al. *Nature Genetics* **41**, 149 - 155 (2009)



“Preclinical research generates many secondary publications, even when results cannot be reproduced”



From

Drug development: Raise standards for preclinical cancer research

C. Glenn Begley & Lee M. Ellis

Nature **483**, 531–533 (29 March 2012) | doi:10.1038/483531a

Table 1: Reproducibility of research findings

Preclinical research generates many secondary publications, even when results cannot be reproduced.

Journal impact factor	Number of articles	Mean number of citations of non-reproduced articles*	Mean number of citations of reproduced articles
>20	21	248 (range 3–800)	231 (range 82–519)
5–19	32	169 (range 6–1,909)	13 (range 3–24)

Results from ten-year retrospective analysis of experiments performed prospectively. The term 'non-reproduced' was assigned on the basis of findings not being sufficiently robust to drive a drug-development programme.

*Source of citations: Google Scholar, May 2011.

- Biomarkers – most highly cited studies overestimated effect sizes (*JAMA* 2011; 305(21): 2200-2210)
- Faculty & Trainee survey at MD Anderson – 50% had experienced at least one case of irreproducibility (PLOS One 2013 May 15; 8(5))
- More first-in-class small molecule drugs approved between 1999-2008 identified by "classical" methods than genomics approaches (J. Biomol Screen 2013 Dec; 18(10): 1143-55)
- "Koch's Postulates" for assigning causality between genetic variants & disease phenotypes (Cell 2013 Sep 26; 155(1) : 21-6)

Lab Mistakes Hobble Cancer Studies But Scientists Slow to Take Remedies

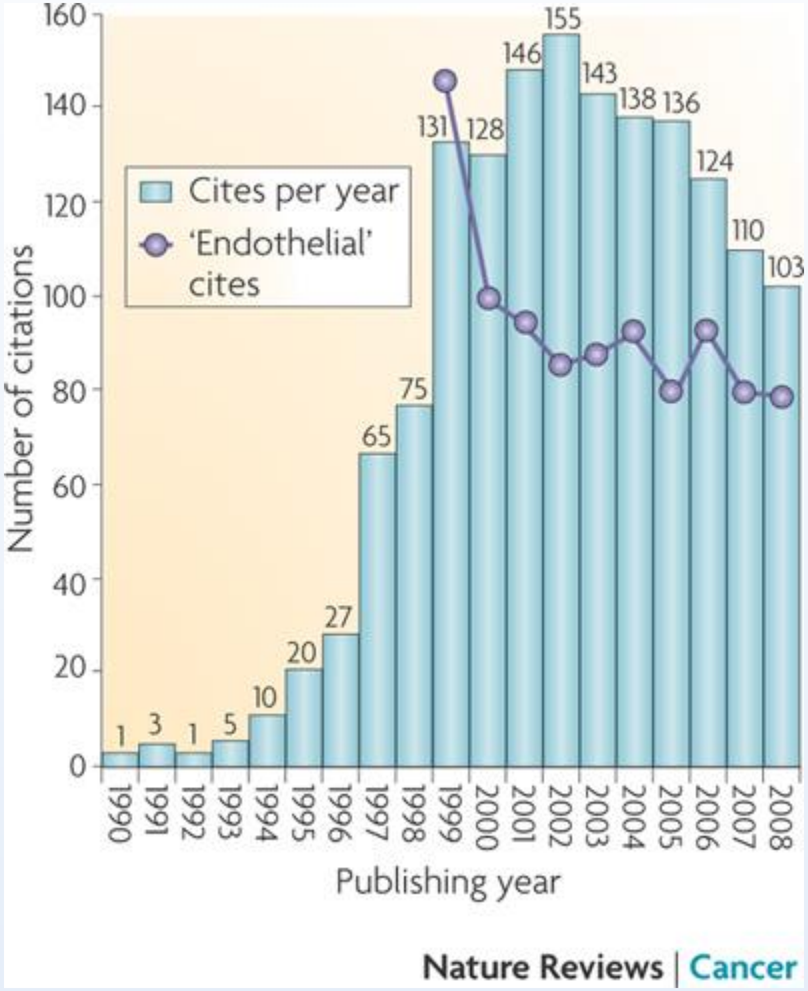
WSJ April 20, 2012 AMY DOCKSER MARCUS

"Cancer experts seeking to solve the problem have found that a fifth to a third or more of cancer cell lines tested were mistakenly identified—with researchers unwittingly studying the wrong cancers, slowing progress toward new treatments and wasting precious time and money."

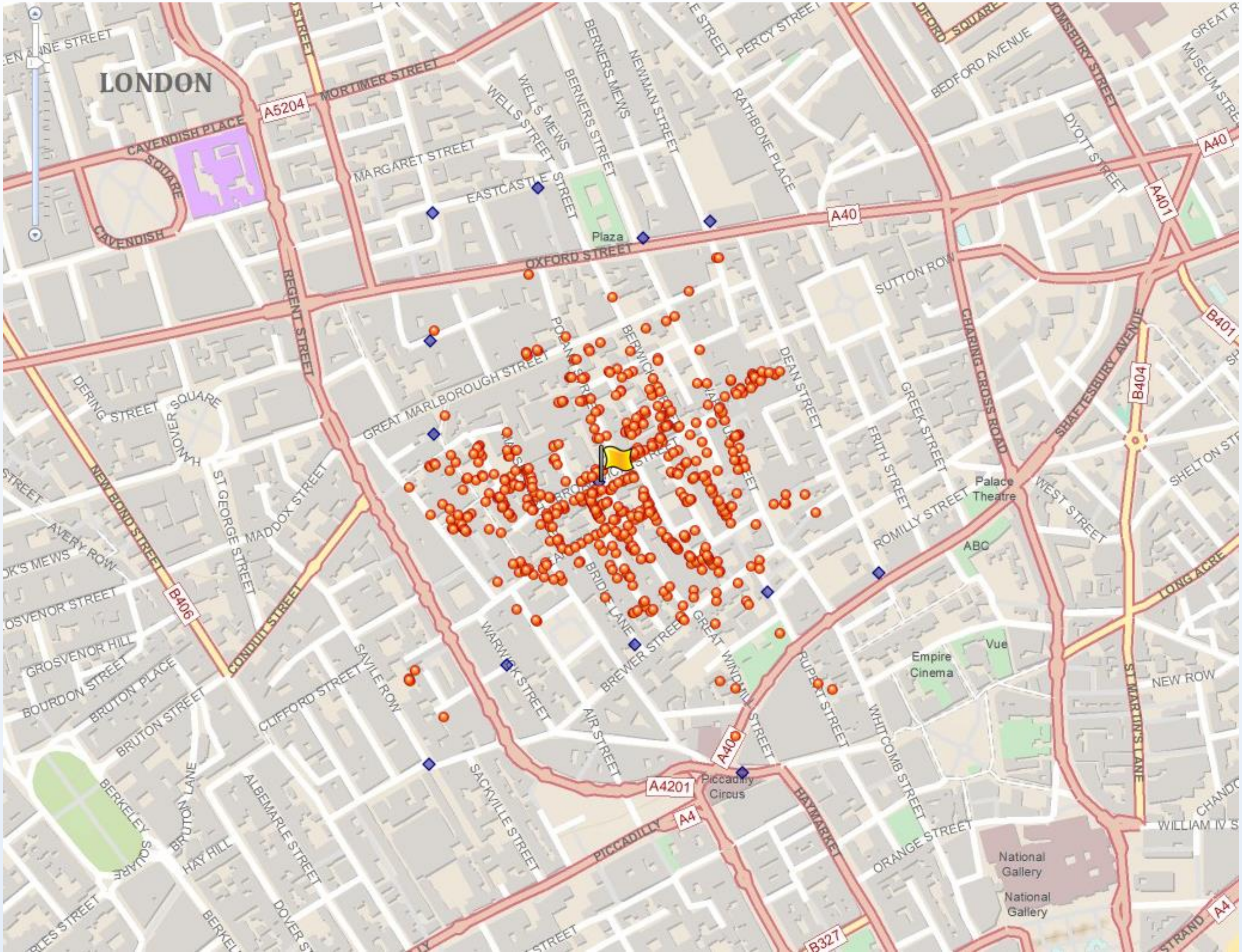
"…Dr. Masters, in a study of scientific papers published between 2000 and 2004, found nearly a 1,000 citations of the same contaminated cancer lines revealed in Dr. Gartler's 1966 findings, which have since been replicated many times using more advanced techniques."

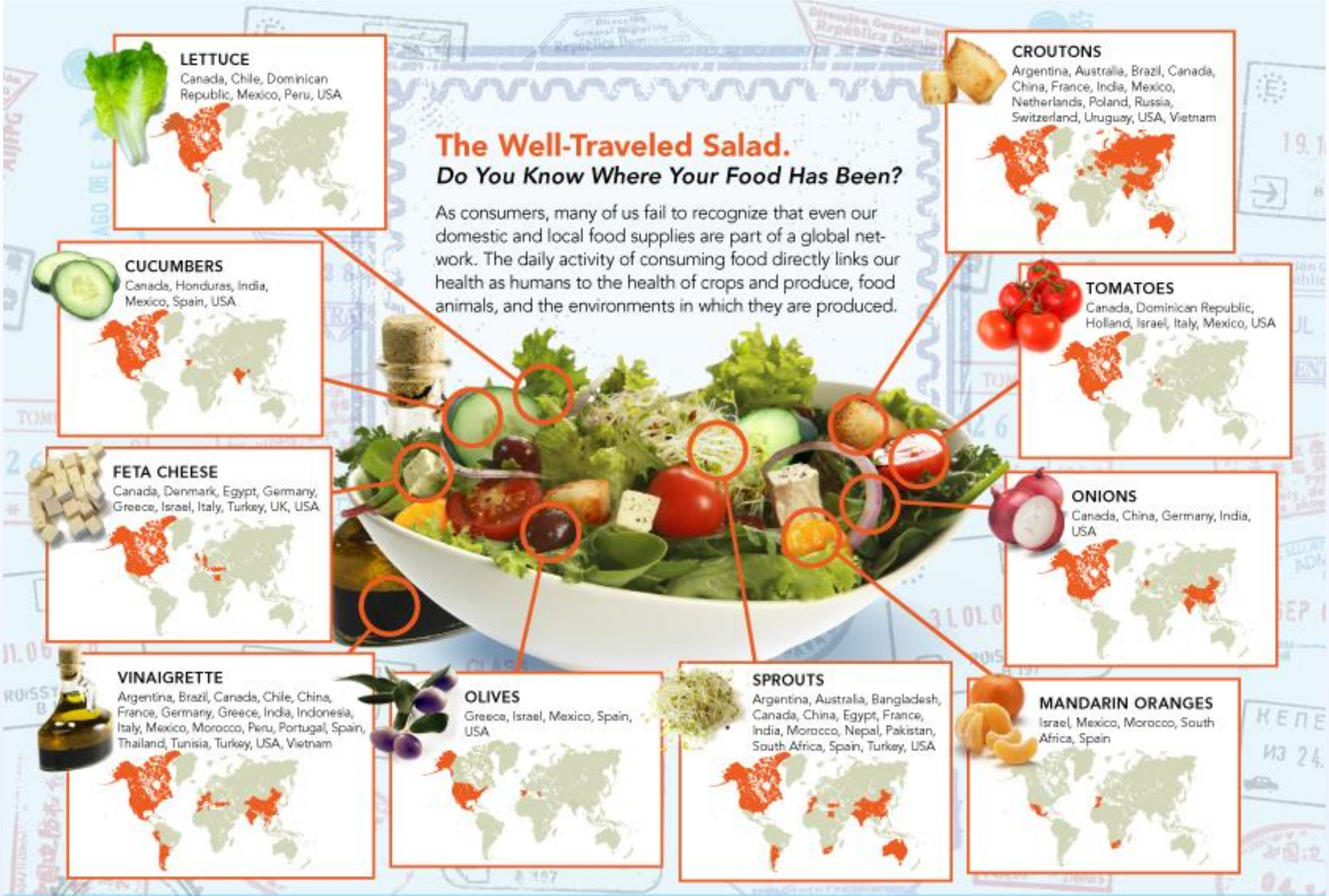
Citations of T24 bladder cancer cells referred to as normal endothelial cells.

Nature Reviews Cancer **10**, (June 2010) | doi:10.1038/nrc2852



Dr. John Snow - location of pumps and cholera deaths, London, England, 1854





The Well-Traveled Salad.

Do You Know Where Your Food Has Been?

As consumers, many of us fail to recognize that even our domestic and local food supplies are part of a global network. The daily activity of consuming food directly links our health as humans to the health of crops and produce, food animals, and the environments in which they are produced.

LETTUCE
Canada, Chile, Dominican Republic, Mexico, Peru, USA

CUCUMBERS
Canada, Honduras, India, Mexico, Spain, USA

FETA CHEESE
Canada, Denmark, Egypt, Germany, Greece, Israel, Italy, Turkey, UK, USA

VINAIGRETTE
Argentina, Brazil, Canada, Chile, China, France, Germany, Greece, India, Indonesia, Italy, Mexico, Morocco, Peru, Portugal, Spain, Thailand, Tunisia, Turkey, USA, Vietnam

OLIVES
Greece, Israel, Mexico, Spain, USA

SPROUTS
Argentina, Australia, Bangladesh, Canada, China, Egypt, France, India, Morocco, Nepal, Pakistan, South Africa, Spain, Turkey, USA

MANDARIN ORANGES
Israel, Mexico, Morocco, South Africa, Spain

ONIONS
Canada, China, Germany, India, USA

TOMATOES
Canada, Dominican Republic, Holland, Israel, Italy, Mexico, USA

CROUTONS
Argentina, Australia, Brazil, Canada, China, France, India, Mexico, Netherlands, Poland, Russia, Switzerland, Uruguay, USA, Vietnam

A "One Health" approach to food safety—bringing together expertise and resources from the clinical, veterinary, wildlife health, and ecology communities—has the potential to reveal the sources, pathways, and factors driving the outbreaks of foodborne illness and possibly prevent them from occurring in the first place.

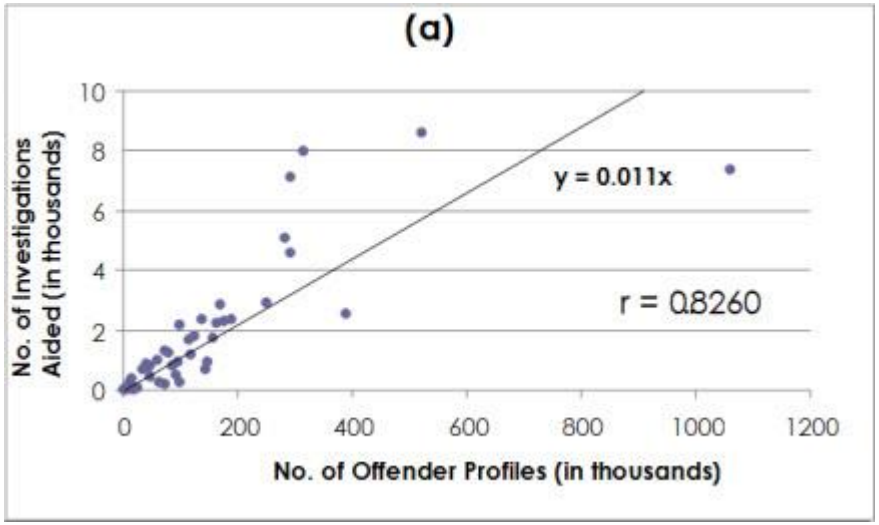
***NOTE:** Countries are listed in alphabetical order and not by volume of export.*

INSTITUTE OF MEDICINE
OF THE NATIONAL ACADEMIES
www.iom.edu

Foodborne Illness

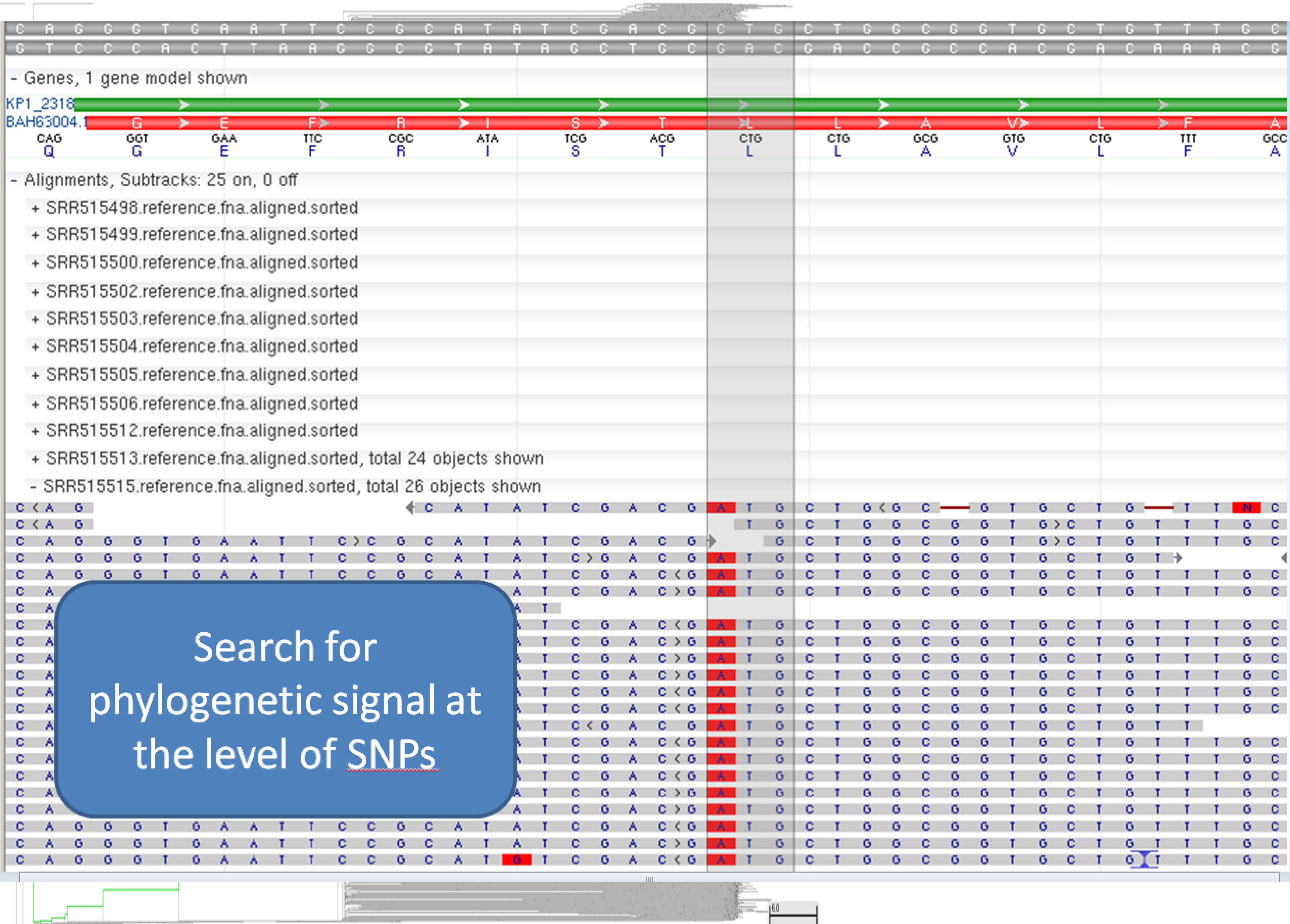
- ~48 million US cases annually (CDC)
  - 128,000 hospitalized
  - 3,000 deaths
- Trends show little evidence of progress

DNA Forensics…


(a)
No. of Investigations Aided (in thousands)
0
2
4
6
8
10
0
200
400
600
800
1000
1200
No. of Offender Profiles (in thousands)
y = 0.011x
r = 0.8260

Functional prediction can be developed and refined more slowly from this base.

A Pathogen Genome Is The Fingerprint



```
C A G G G T G A A T T C C G C A T A T C G A C G C T G C T G G C G G T G C T G T T T G C
G T C C C A C T T A A G G C G T A T A G C T G C G A C G A C C G C C A C G A C A A A C G
```

- Genes, 1 gene model shown

KP1_2318

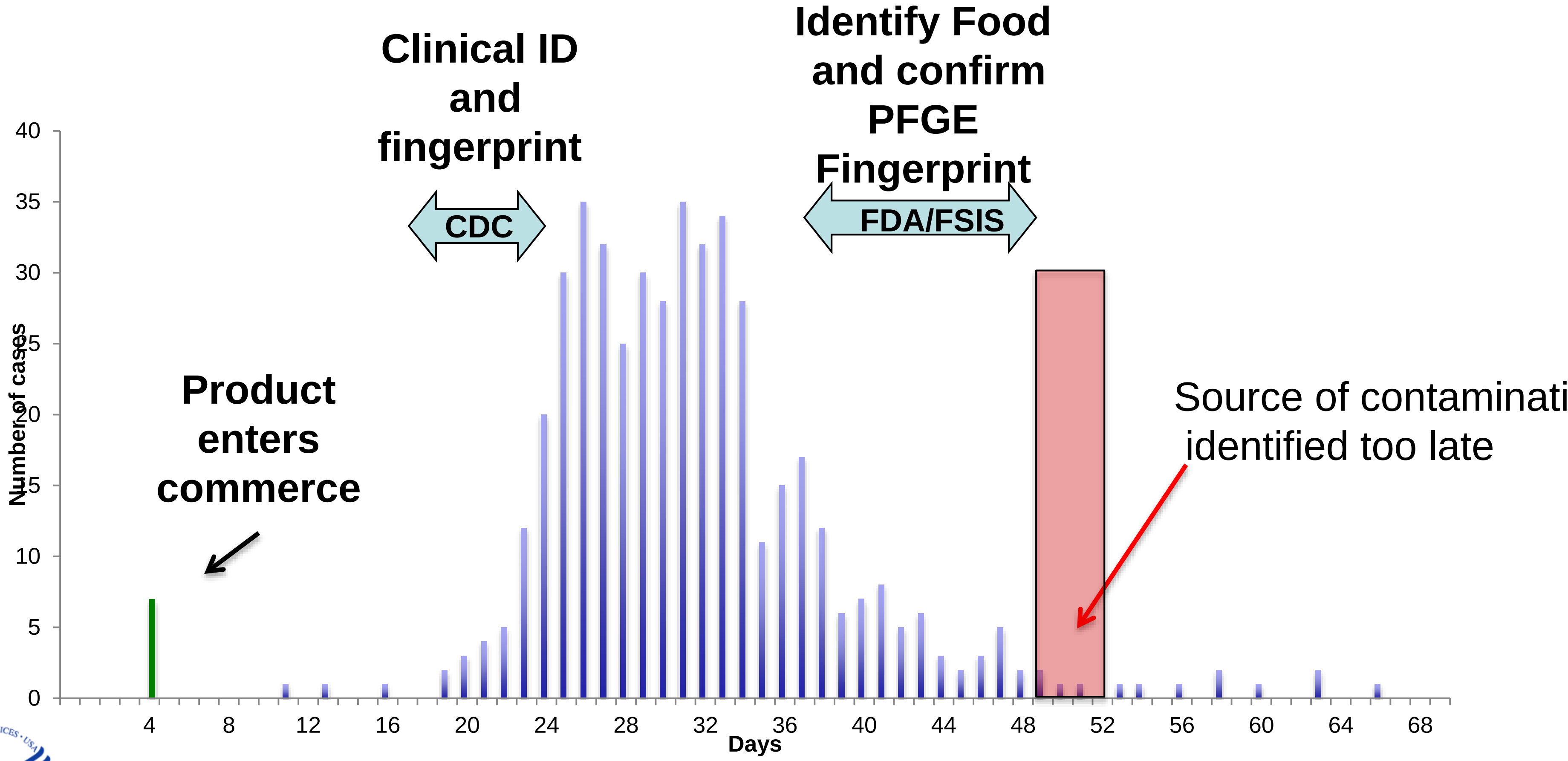
BAH63004.1

```
CAG GGT GAA TTC CGC ATA TCG ACG CTG CTG GCG GTG CTG TTT GCC
 Q   G   E   F   R   I   S   T   L   L   A   V   L   F   A
```

- Alignments, Subtracks: 25 on, 0 off
  + SRR515498.reference.fna.aligned.sorted
  + SRR515499.reference.fna.aligned.sorted
  + SRR515500.reference.fna.aligned.sorted
  + SRR515502.reference.fna.aligned.sorted
  + SRR515503.reference.fna.aligned.sorted
  + SRR515504.reference.fna.aligned.sorted
  + SRR515505.reference.fna.aligned.sorted
  + SRR515506.reference.fna.aligned.sorted
  + SRR515512.reference.fna.aligned.sorted
  + SRR515513.reference.fna.aligned.sorted, total 24 objects shown
  - SRR515515.reference.fna.aligned.sorted, total 26 objects shown

Search for phylogenetic signal at the level of SNPs

Why do we need WGS? To Shift the Paradigm from a "low resolution" Public Health Approach to A Real-Time "high resolution" Approach



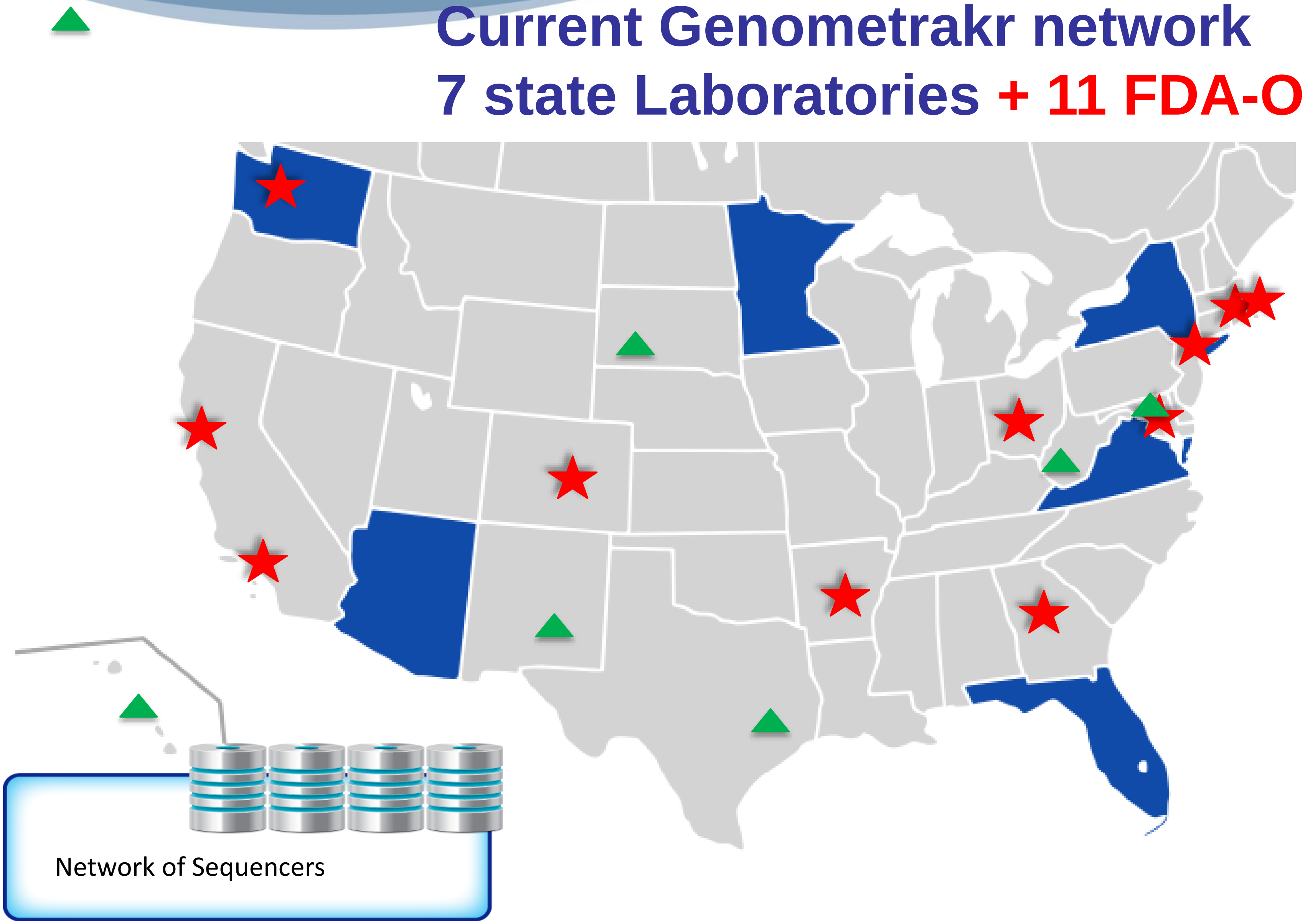
Product enters commerce

Clinical ID and fingerprint

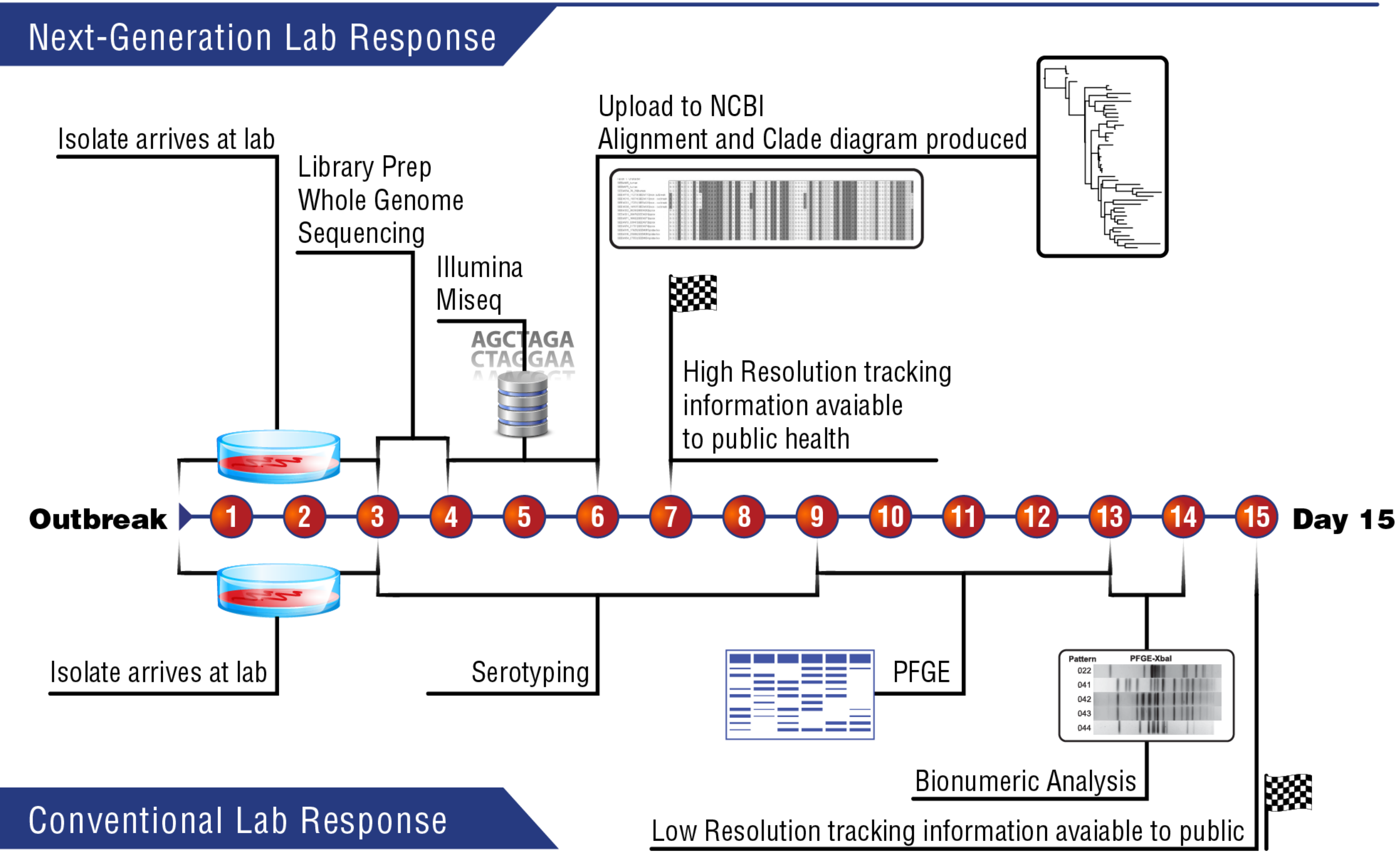
Identify Food and confirm PFGE Fingerprint

Source of contaminati identified too late

Current Genometrakr network
7 state Laboratories + 11 FDA-ORA



Next-Generation Lab Response vs. Conventional Lab Response



Next-Generation Lab Response

Isolate arrives at lab

Library Prep
Whole Genome
Sequencing

Illumina
Miseq

Upload to NCBI
Alignment and Clade diagram produced

High Resolution tracking
information avaiable
to public health

Isolate arrives at lab

Serotyping

PFGE

Bionumeric Analysis

Low Resolution tracking information avaiable to public

Conventional Lab Response

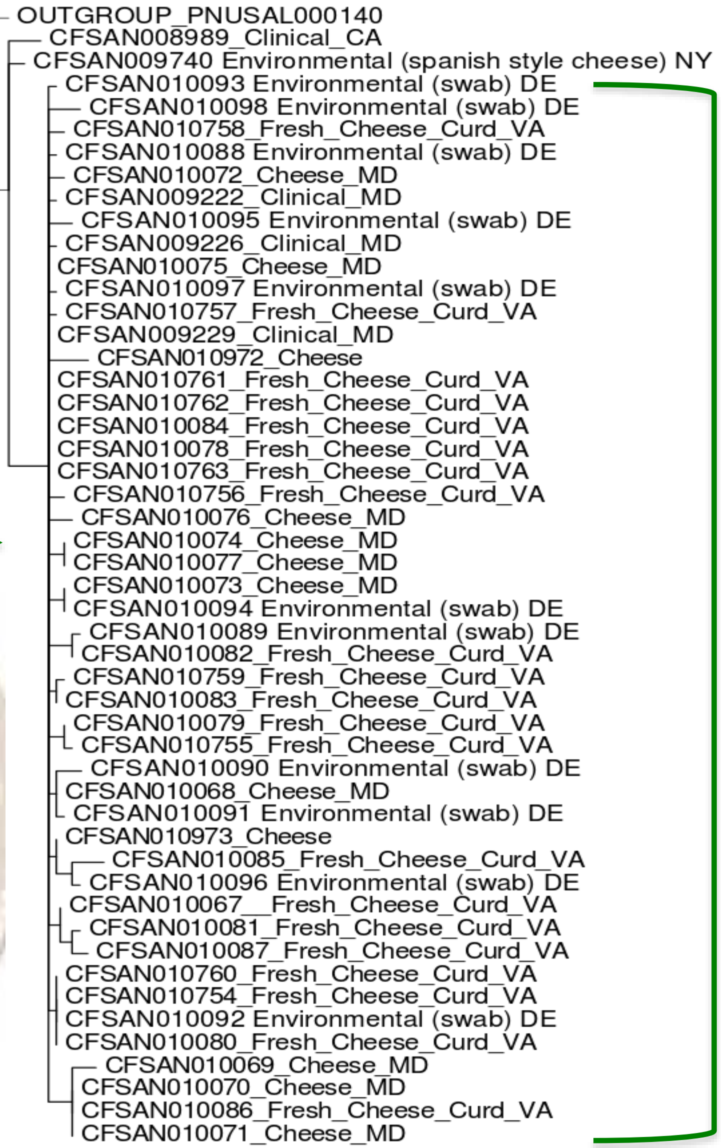
09/16/2013

WGS In its First Official Regulatory Action

First use of Genometrakr network and WGS approach to support regulatory action and positive public health outcome in real-time

***Listeria monocytogenes* project with CDC, FDA, NCBI, USDA March 2014**

> 6 SNPs

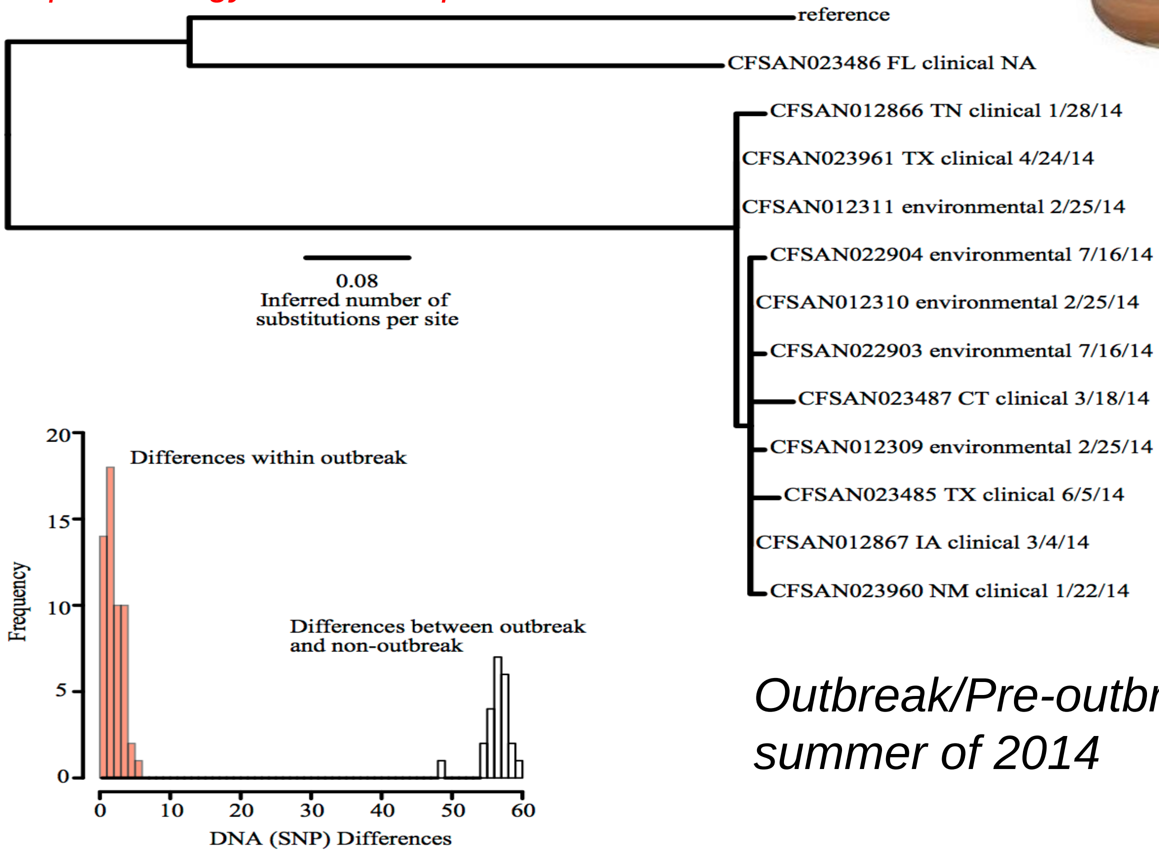

OUTGROUP_PNUSAL000140
CFSAN008989_Clinical_CA
CFSAN009740 Environmental (spanish style cheese) NY
CFSAN010093 Environmental (swab) DE
CFSAN010098 Environmental (swab) DE
CFSAN010758_Fresh_Cheese_Curd_VA
CFSAN010088 Environmental (swab) DE
CFSAN010072_Cheese_MD
CFSAN009222_Clinical_MD
CFSAN010095 Environmental (swab) DE
CFSAN009226_Clinical_MD
CFSAN010075_Cheese_MD
CFSAN010097 Environmental (swab) DE
CFSAN010757_Fresh_Cheese_Curd_VA
CFSAN009229_Clinical_MD
CFSAN010972_Cheese
CFSAN010761_Fresh_Cheese_Curd_VA
CFSAN010762_Fresh_Cheese_Curd_VA
CFSAN010084_Fresh_Cheese_Curd_VA
CFSAN010078_Fresh_Cheese_Curd_VA
CFSAN010763_Fresh_Cheese_Curd_VA
CFSAN010756_Fresh_Cheese_Curd_VA
CFSAN010076_Cheese_MD
CFSAN010074_Cheese_MD
CFSAN010077_Cheese_MD
CFSAN010073_Cheese_MD
CFSAN010094 Environmental (swab) DE
CFSAN010089 Environmental (swab) DE
CFSAN010082_Fresh_Cheese_Curd_VA
CFSAN010759_Fresh_Cheese_Curd_VA
CFSAN010083_Fresh_Cheese_Curd_VA
CFSAN010079_Fresh_Cheese_Curd_VA
CFSAN010755_Fresh_Cheese_Curd_VA
CFSAN010090 Environmental (swab) DE
CFSAN010068_Cheese_MD
CFSAN010091 Environmental (swab) DE
CFSAN010973_Cheese
CFSAN010085_Fresh_Cheese_Curd_VA
CFSAN010096 Environmental (swab) DE
CFSAN010067__Fresh_Cheese_Curd_VA
CFSAN010081_Fresh_Cheese_Curd_VA
CFSAN010087_Fresh_Cheese_Curd_VA
CFSAN010760_Fresh_Cheese_Curd_VA
CFSAN010754_Fresh_Cheese_Curd_VA
CFSAN010092 Environmental (swab) DE
CFSAN010080_Fresh_Cheese_Curd_VA
CFSAN010069_Cheese_MD
CFSAN010070_Cheese_MD
CFSAN010086_Fresh_Cheese_Curd_VA
CFSAN010071_Cheese_MD
0.2

Isolates from cheese facility, distributed product, and patients who consumed product

And it gets even better….

An increased degree of certainty that comes with matching strains of pathogens through whole genome sequencing allowed for detection of this Salmonella contamination event in nut butter across several states with low level contamination and a widely distributed product. In this case, WGS identifies the link and preempts an outbreak even w/o availability of food - it informs the epidemiology and our inspectors.

Photo courtesy of the MaraNatha website



Outbreak/Pre-outbraek summer of 2014

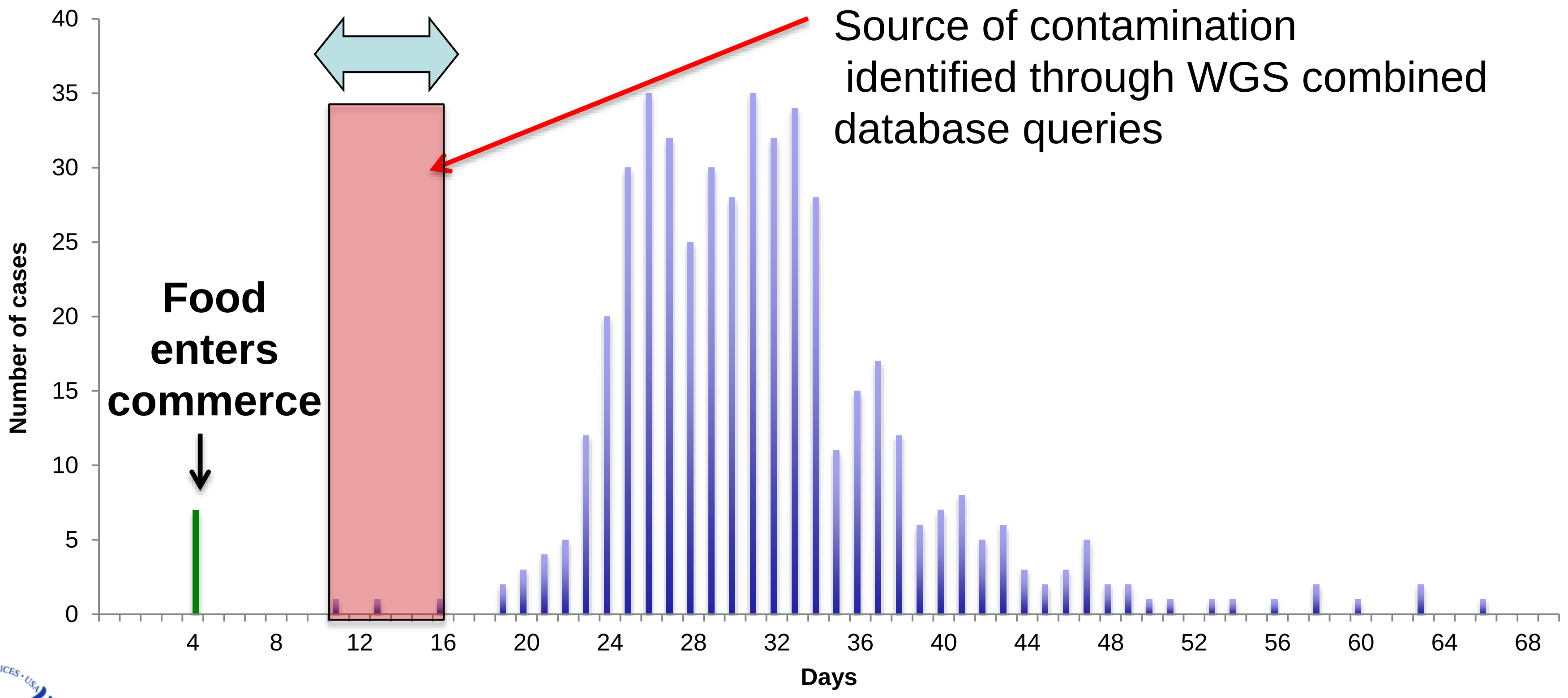
Health and Economic Impact of Active WGS-based Surveillance

Comparison of Nspired and Sunland contamination events.

- ❑ **Similar facilities – broad domestic distribution.**
- ❑ **Sunland 42 cases and 10 hospitalizations with as many as 1,260 illnesses unreported (Fall 2012)**
- ❑ **Nspired – 4 confirmed cases, 1 hospitalization (Summer 2014)**
- ❑ **WGS informed investigation prevented significant illness and hospitalizations**
  - **– lower illness rate and treatment cost ($3000-$9000) + fraction of longterm and chronic onset complications associated with Salmonella infection (ie, Reiter’s syndrome, GBS)**

The New Microbiology Approach to Public Health

Clinical ID WGS in real-time and in parallel food and environmental WGS FDA, CDC, FSIS, States



National Digital Immune system: Big Data?

- WGS + metadata
- PulseNet - ~50,000 isolates per year + Hospital based infections (AMR strains) 50-500K per year (??) + Environmental samples
- Ecology & population genetics of pathogens…