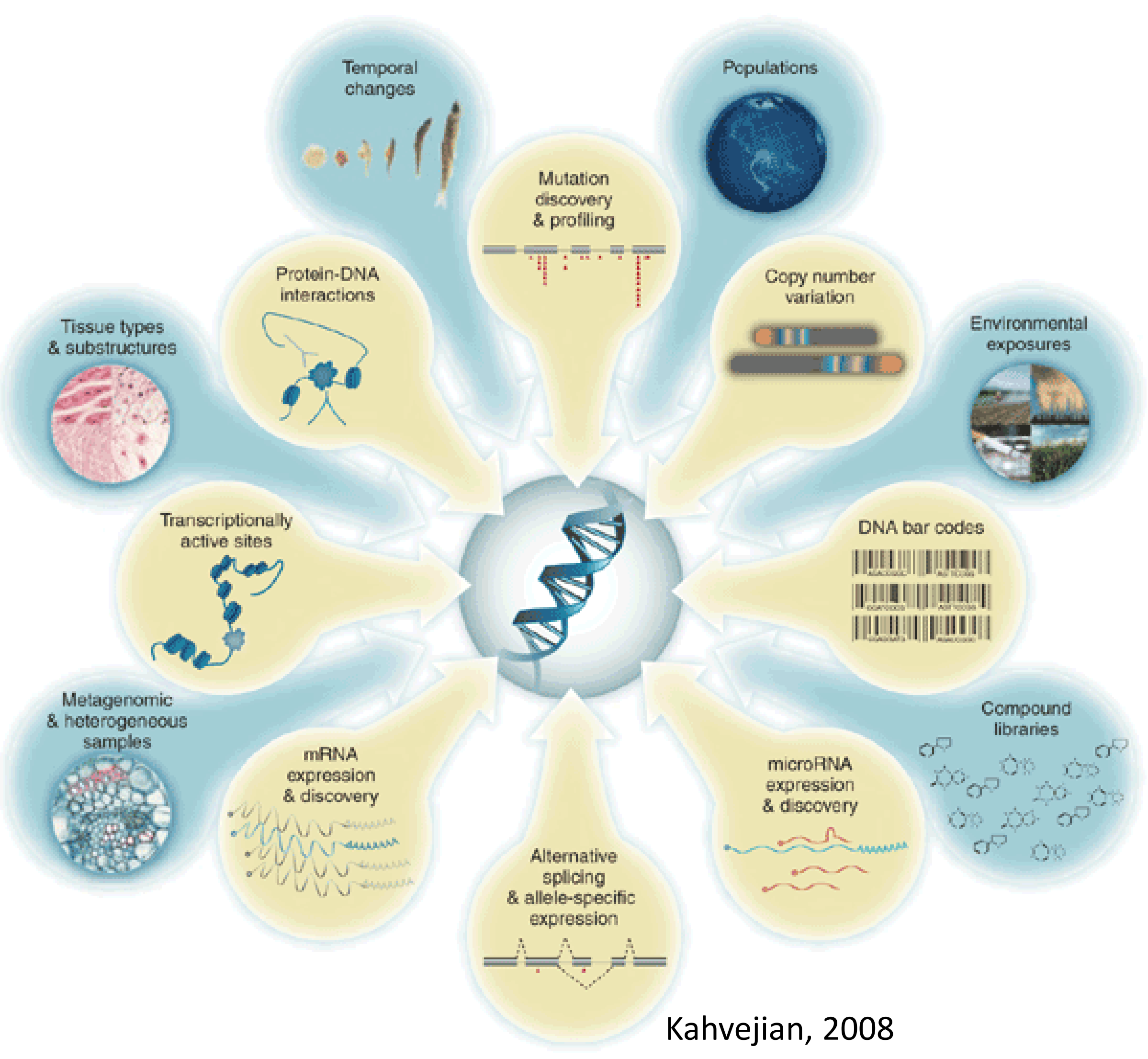
Use of Statistical Techniques in High-throughput Sequencing

Christopher E. Mason
Assistant Professor
The Institute for Computational Biomedicine,
Department of Physiology and Biophysics at
Weill Cornell Medical College, and the
Tri-I Program (Cornell, MSKCC, Rockefeller) on Comp. Bio. & Medicine
February 23rd , 2011

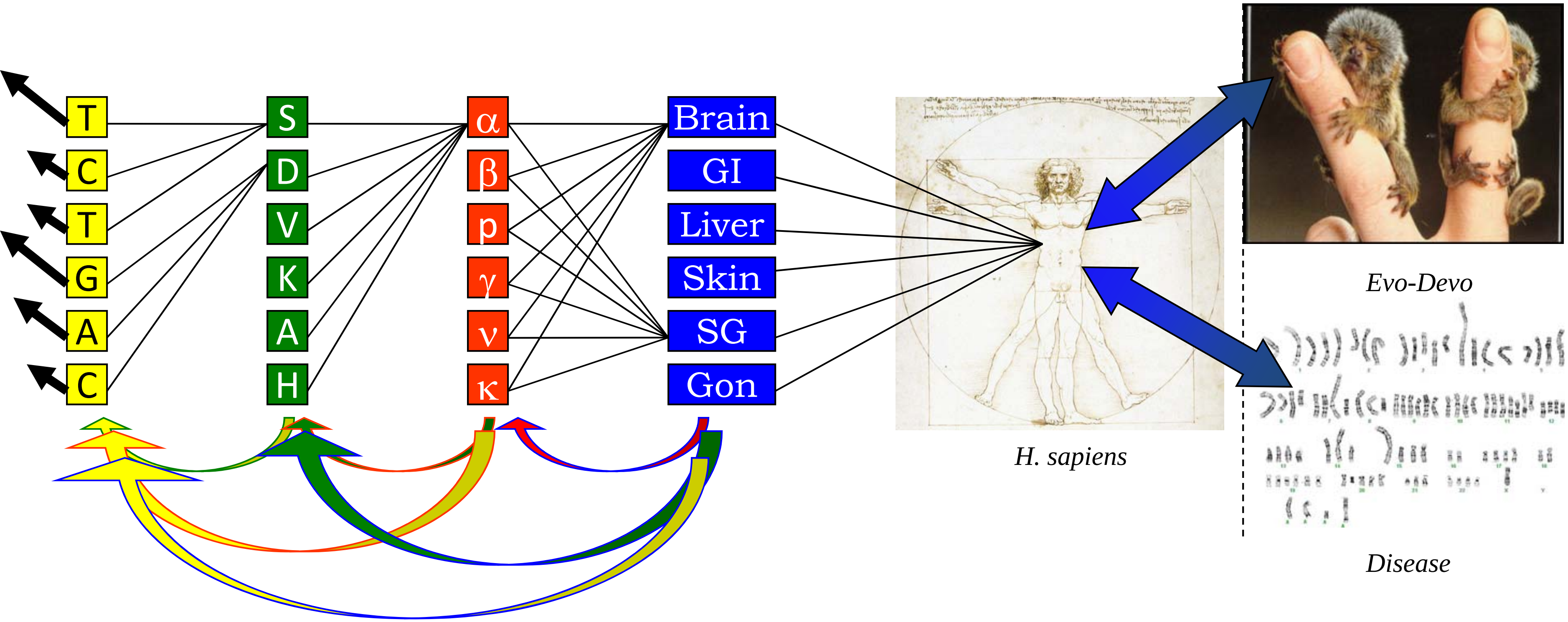
Since DNA defines the biochemical recipe for the genesis of organisms, sequencing allows us to create molecular portraits of development and disease at single-base resolution.



Kahvejian, 2008

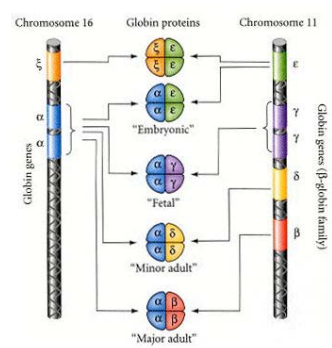

PHASE TWO : INTERPRETATION
I THINK I FOUND A CORNER PIECE.
3 BILLION PIECES
GENOME

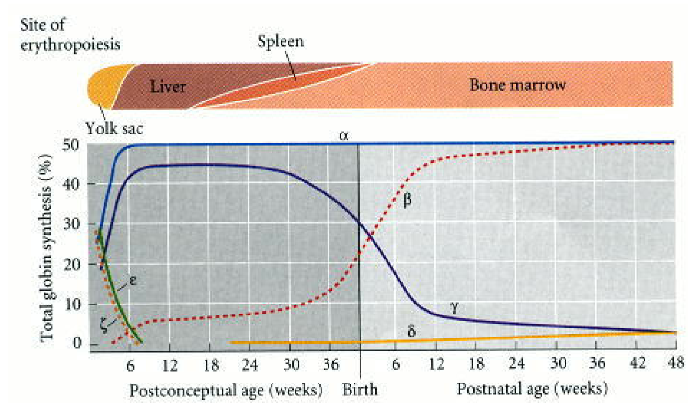
Understanding the genome's mutation, selection, and/or drift



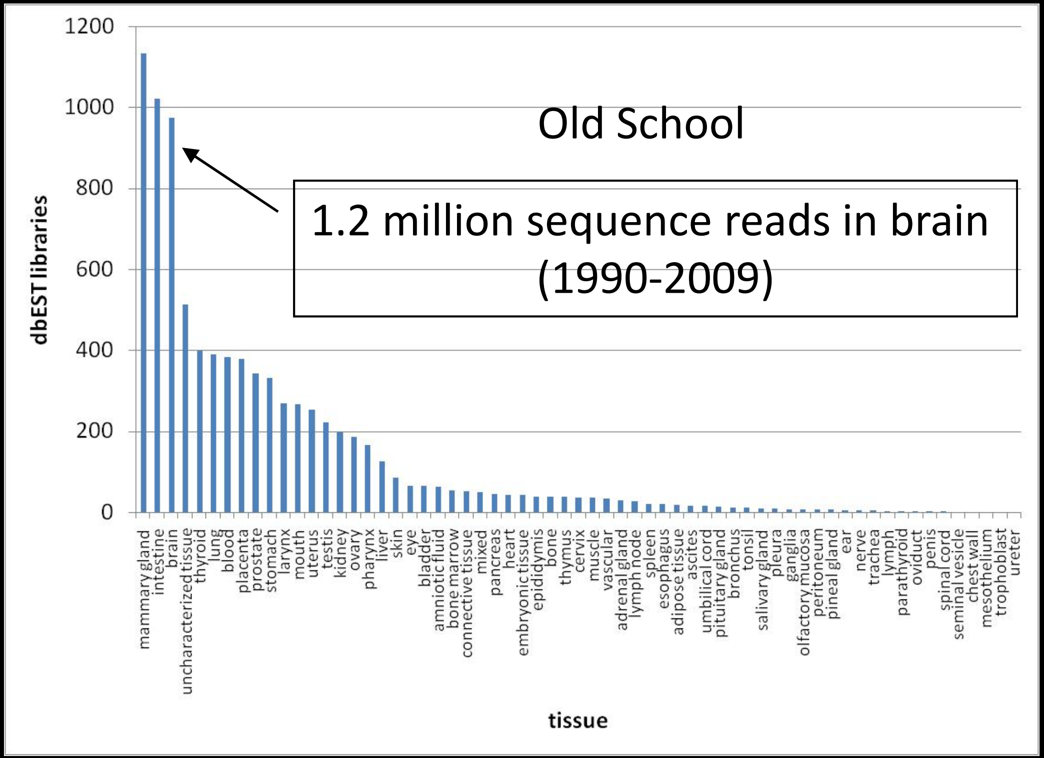
How can we understand these interactions?

Integrated, spatiotemporal molecular profiling.





What erudition do we have now on the functional elements?



- ➢Currently limited amount of EST info at NCBI
- ➢EST data is expensive, time-consuming (cloning), and exhibits 3’ bias.
- ➢Much EST and cDNA data is for whole brains, and few libraries exist with region-specific data.

New School:
One run of a NGS machine = billions of sequence reads in days

Description/Discussion of the Various Technologies

- The goal of the Archon X prize in Genomics is to enable a $1,000 genome,
- Currently at $5,000-$50,000
- Certain platforms are better suited for certain tasks:
  - Counting applications (ChIP-Seq, RNA-Seq) need more reads
  - *De novo* assembly work needs longer reads
  - Whole genome re-sequencing requires lower errors rate and high processivity

But, there are many options:

Platform	Library/ template preparation	NGS chemistry	Read length (bases)	Run time (days)	Gb per run	Machine cost (US$)	Pros	Cons	Biological applications	Refs
Roche/454's GS FLX Titanium	Frag, MP/ emPCR	PS	330*	0.35	0.45	500,000	Longer reads improve mapping in repetitive regions; fast run times	High reagent cost; high error rates in homo-polymer repeats	Bacterial and insect genome *de novo* assemblies; medium scale (<3 Mb) exome capture; 16S in metagenomics	D. Muzny, pers. comm.
Illumina/ Solexa's GA_{II}	Frag, MP/ solid-phase	RTs	75 or 100	$4^{‡}$, $9^{§}$	$18^{‡}$, $35^{§}$	540,000	Currently the most widely used platform in the field	Low multiplexing capability of samples	Variant discovery by whole-genome resequencing or whole-exome capture; gene discovery in metagenomics	D. Muzny, pers. comm.
Life/APG's SOLiD 3	Frag, MP/ emPCR	Cleavable probe SBL	50	$7^{‡}$, $14^{§}$	$30^{‡}$, $50^{§}$	595,000	Two-base encoding provides inherent error correction	Long run times	Variant discovery by whole-genome resequencing or whole-exome capture; gene discovery in metagenomics	D. Muzny, pers. comm.
Polonator G.007	MP only/ emPCR	Non-cleavable probe SBL	26	$5^{§}$	$12^{§}$	170,000	Least expensive platform; open source to adapt alternative NGS chemistries	Users are required to maintain and quality control reagents; shortest NGS read lengths	Bacterial genome resequencing for variant discovery	J. Edwards, pers. comm.
Helicos BioSciences HeliScope	Frag, MP/ single molecule	RTs	32*	$8^{‡}$	$37^{‡}$	999,000	Non-bias representation of templates for genome and seq-based applications	High error rates compared with other reversible terminator chemistries	Seq-based methods	91
Pacific Biosciences (target release: 2010)	Frag only/ single molecule	Real-time	964*	N/A	N/A	N/A	Has the greatest potential for reads exceeding 1 kb	Highest error rates compared with other NGS chemistries	Full-length transcriptome sequencing; complements other resequencing efforts in discovering large structural variants and haplotype blocks	S. Turner, pers. comm.

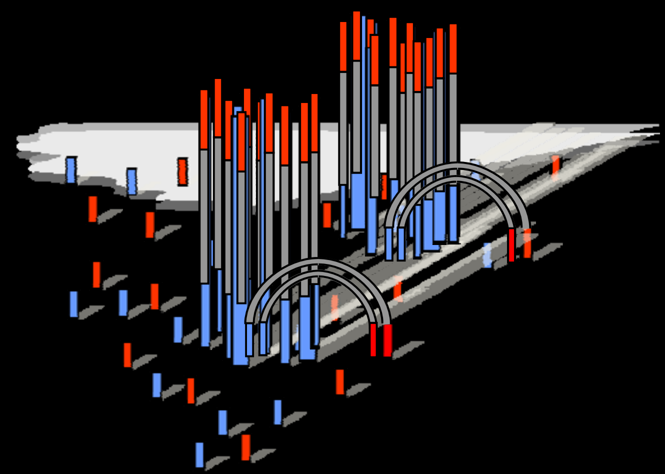
Michael Metzker, 2010

Illumina SBS Technology

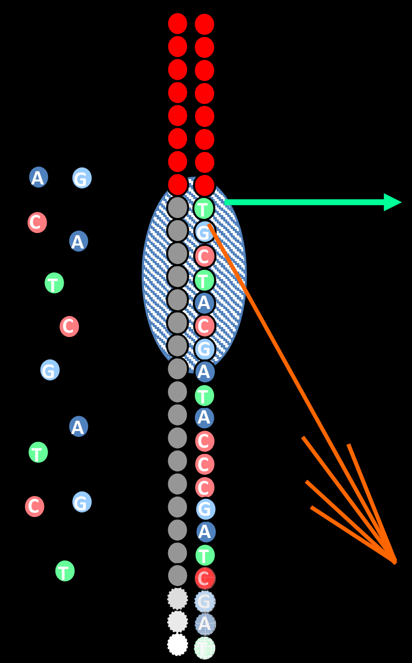
Reversible Terminator Chemistry Foundation

DNA (0.1-1.0 ug)

Sample preparation



Cluster growth



Sequencing

1 2 3 4 5 6 7 8 9

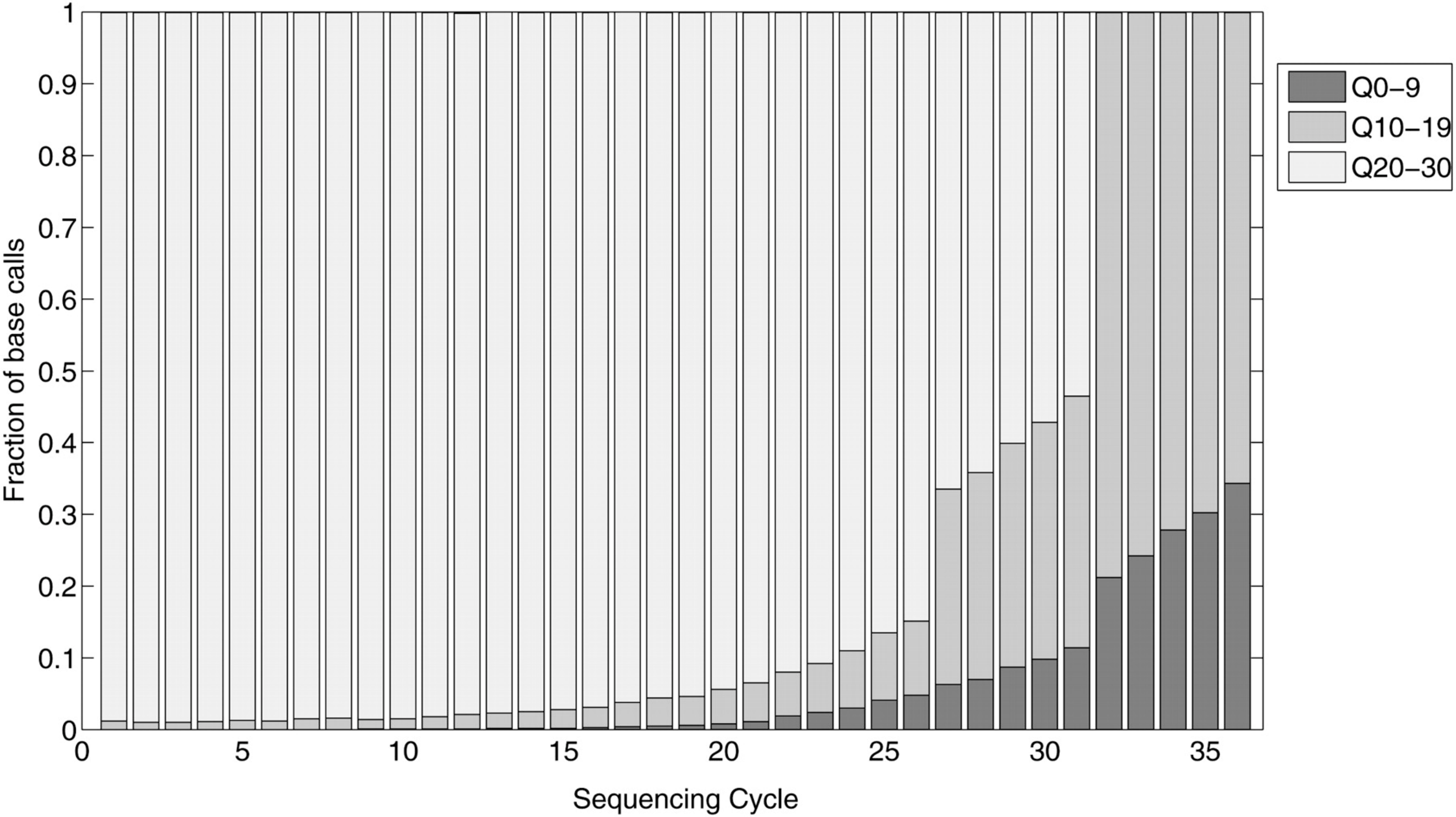
Image acquisition

T G C T A C G A T …

Base calling

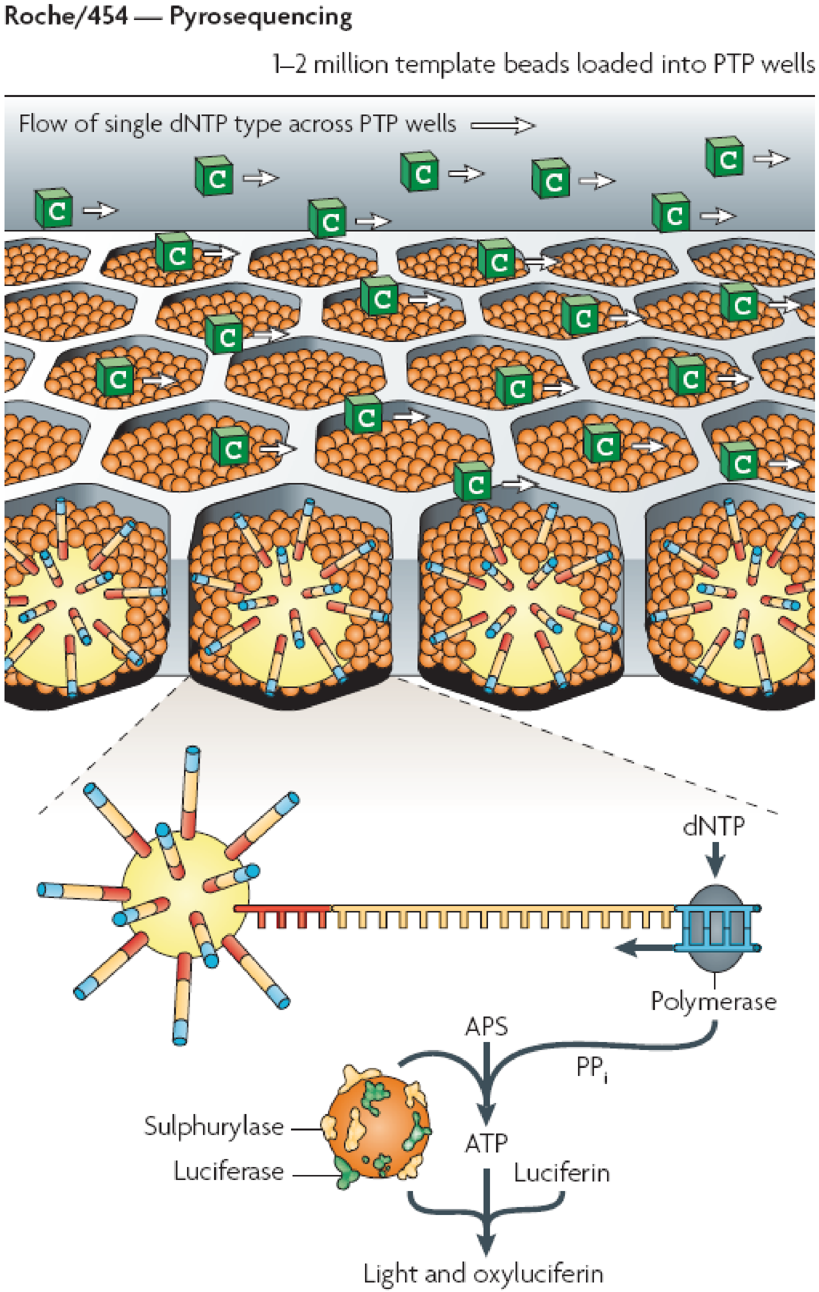
http://www.illumina.com/technology/sequencing_technology.ilmn

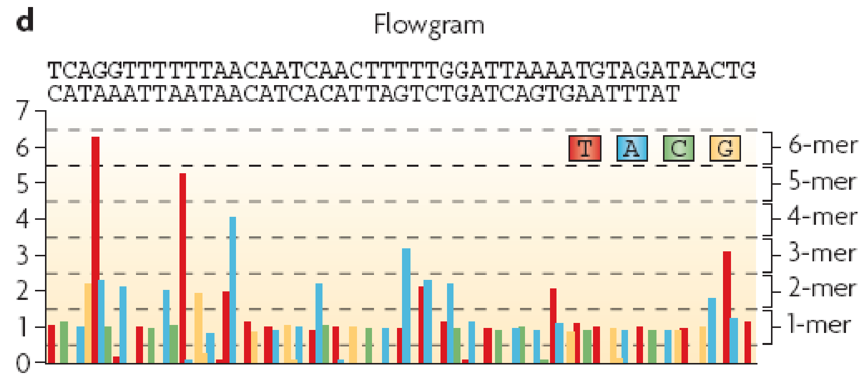
Quality Scores vary

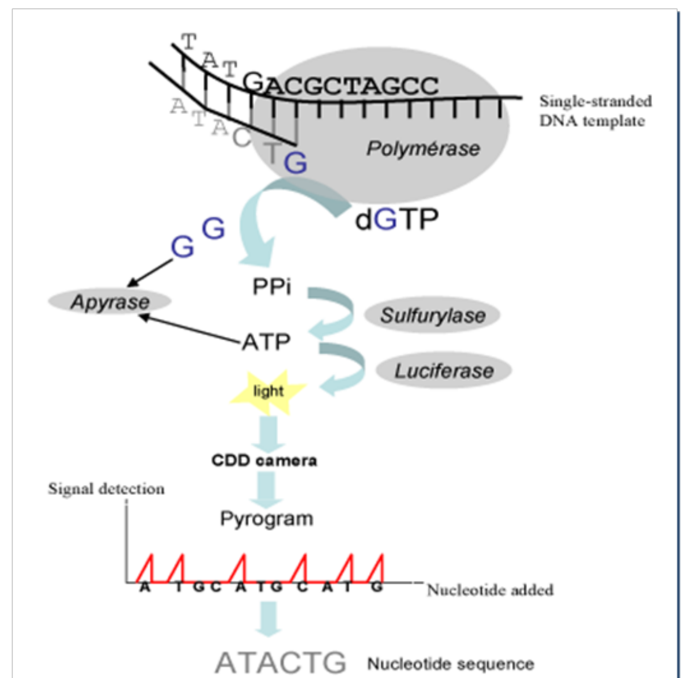


Bansal et al, 2010

Pyrosequencing

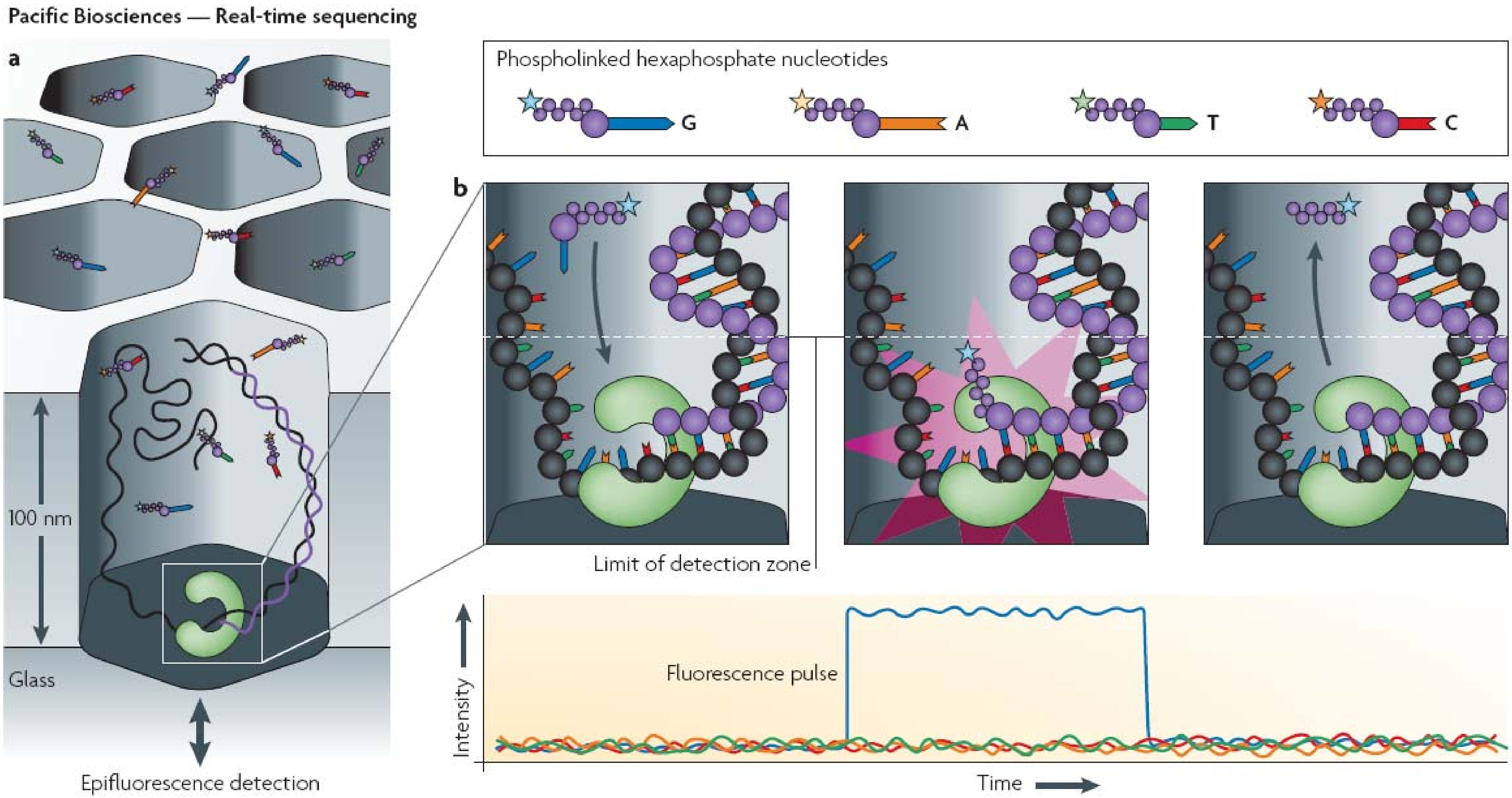






Metzker, 2010

Single Molecule Real-Time

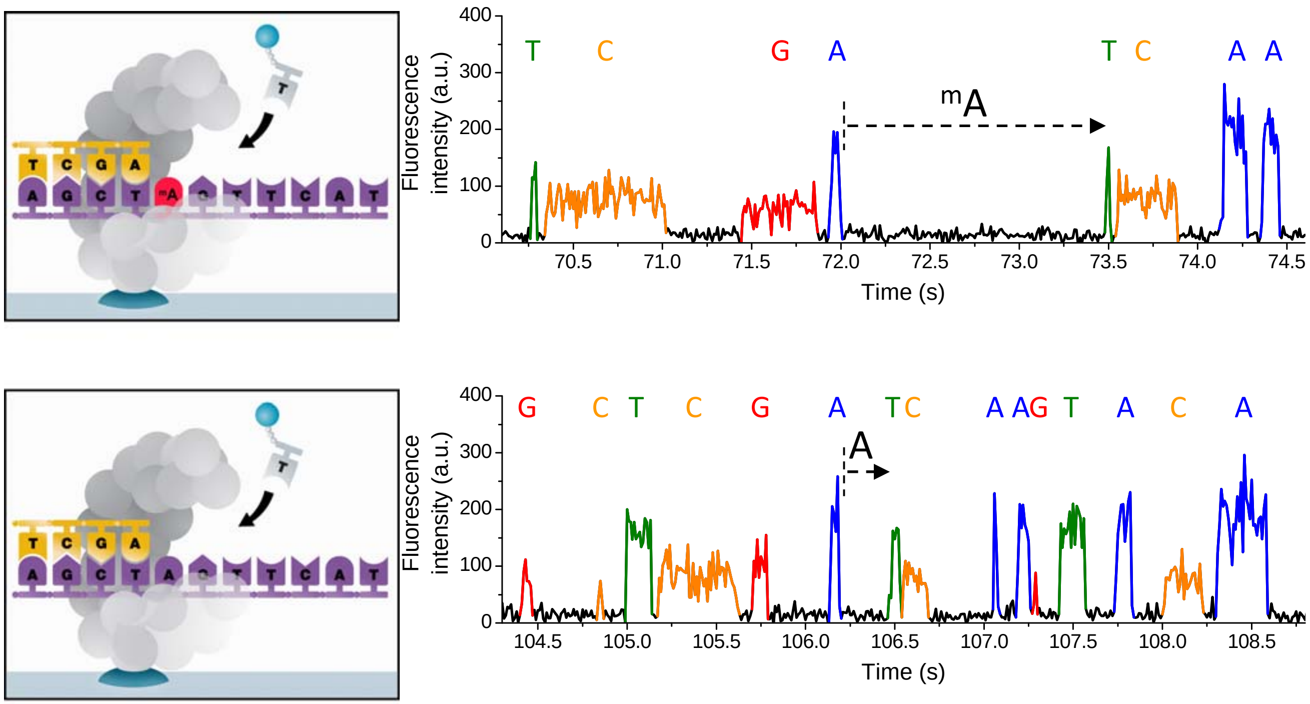


Metzker, 2010

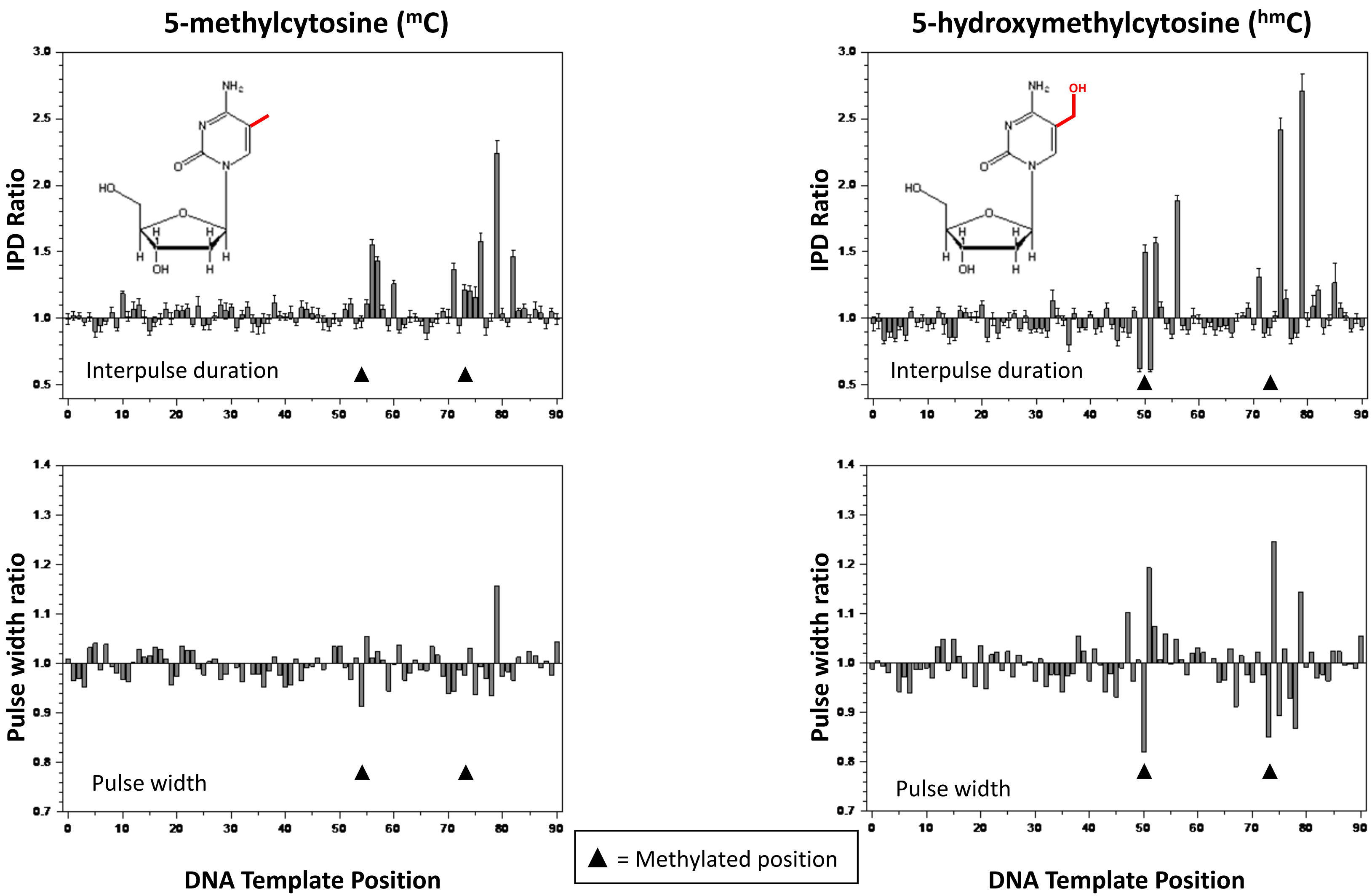
Direct Detection of Methylation

Approach: Kinetic detection of methylated bases during SMRT DNA sequencing

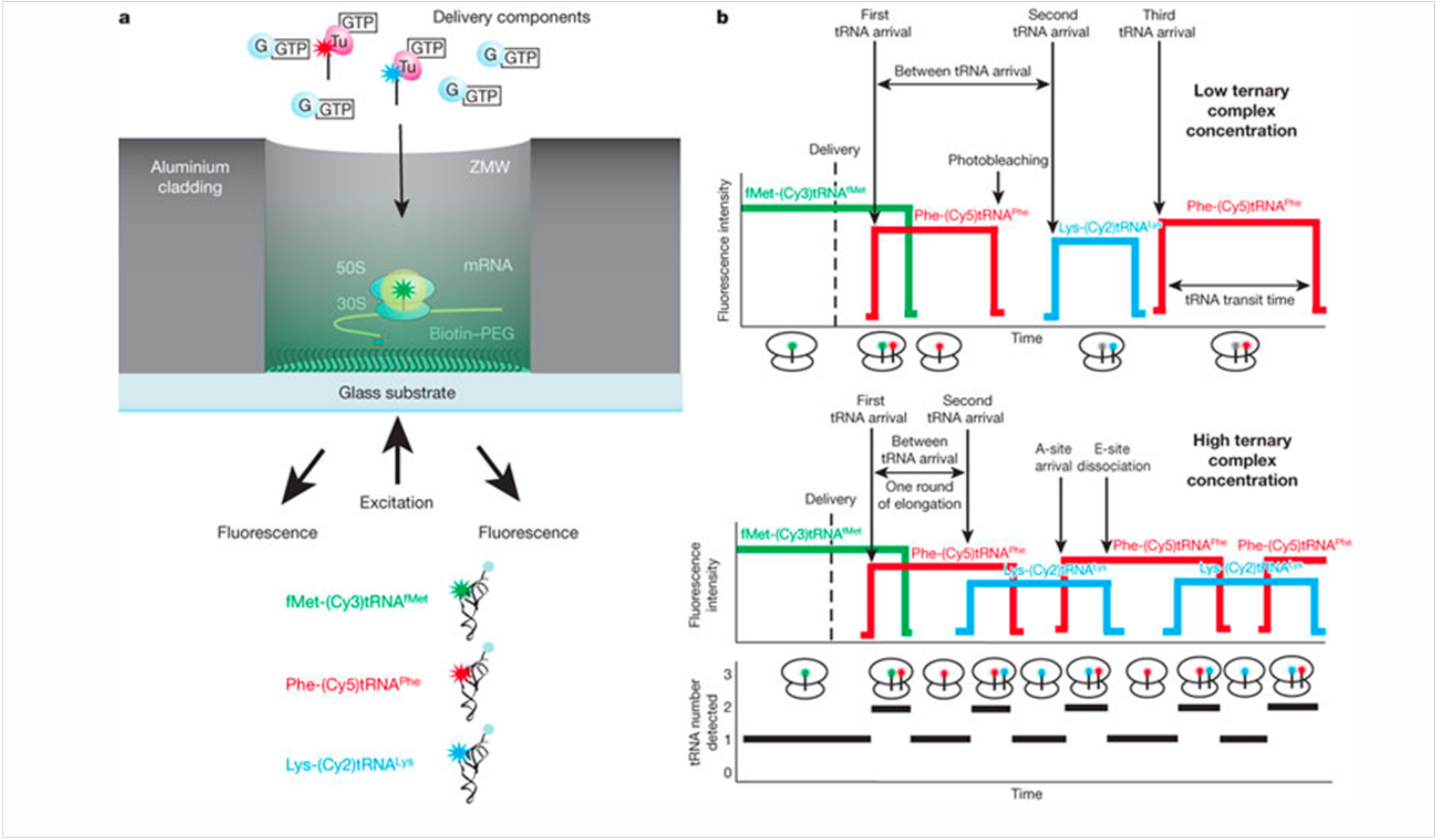
Example: N^6-methyladenosine (^mA)



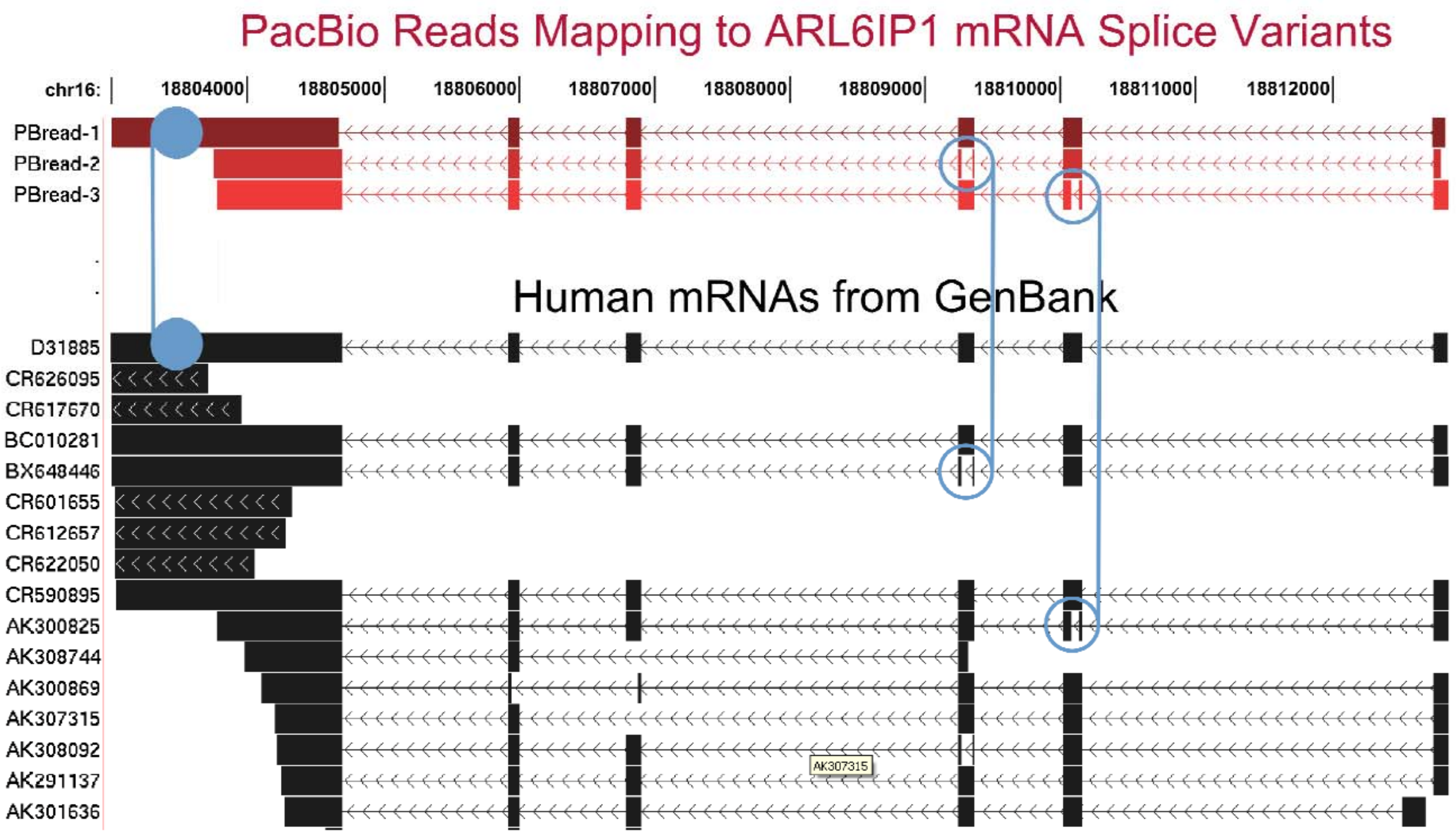
detect other base modifications



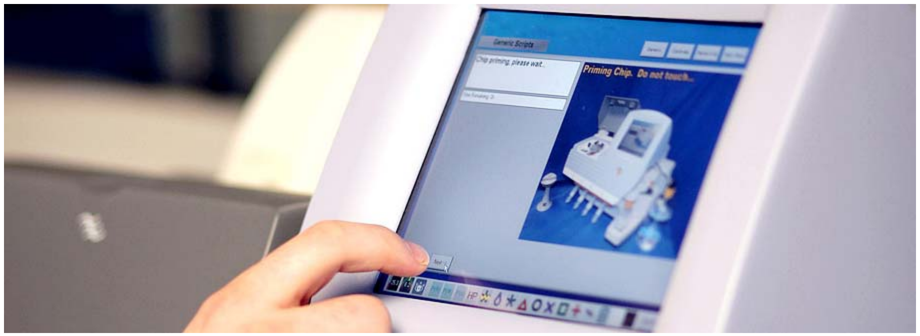
Watch Translation in Real-Time

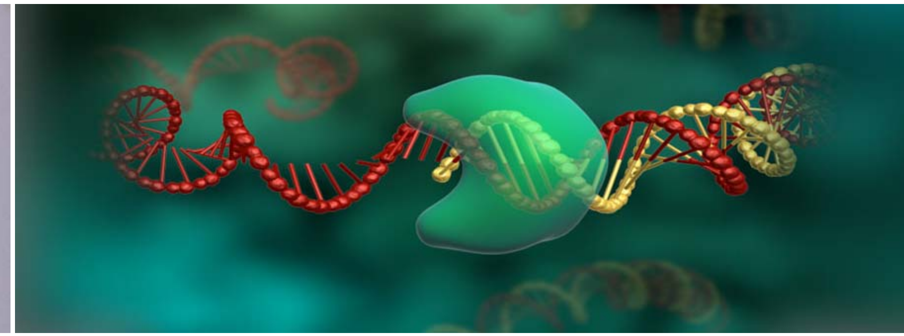


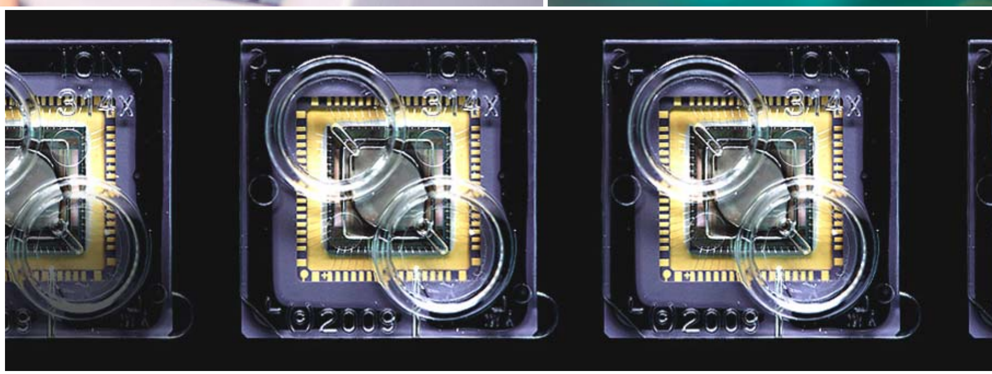
Full-length cDNA sequencing



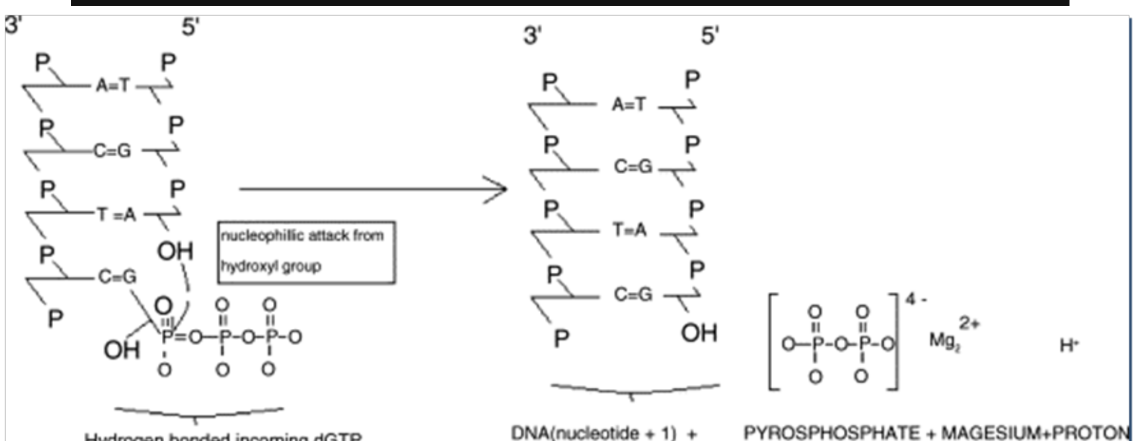
"Post-Light,"
Semi-Conductor Sequencing







Essentially,
7 million
very small
pH meters

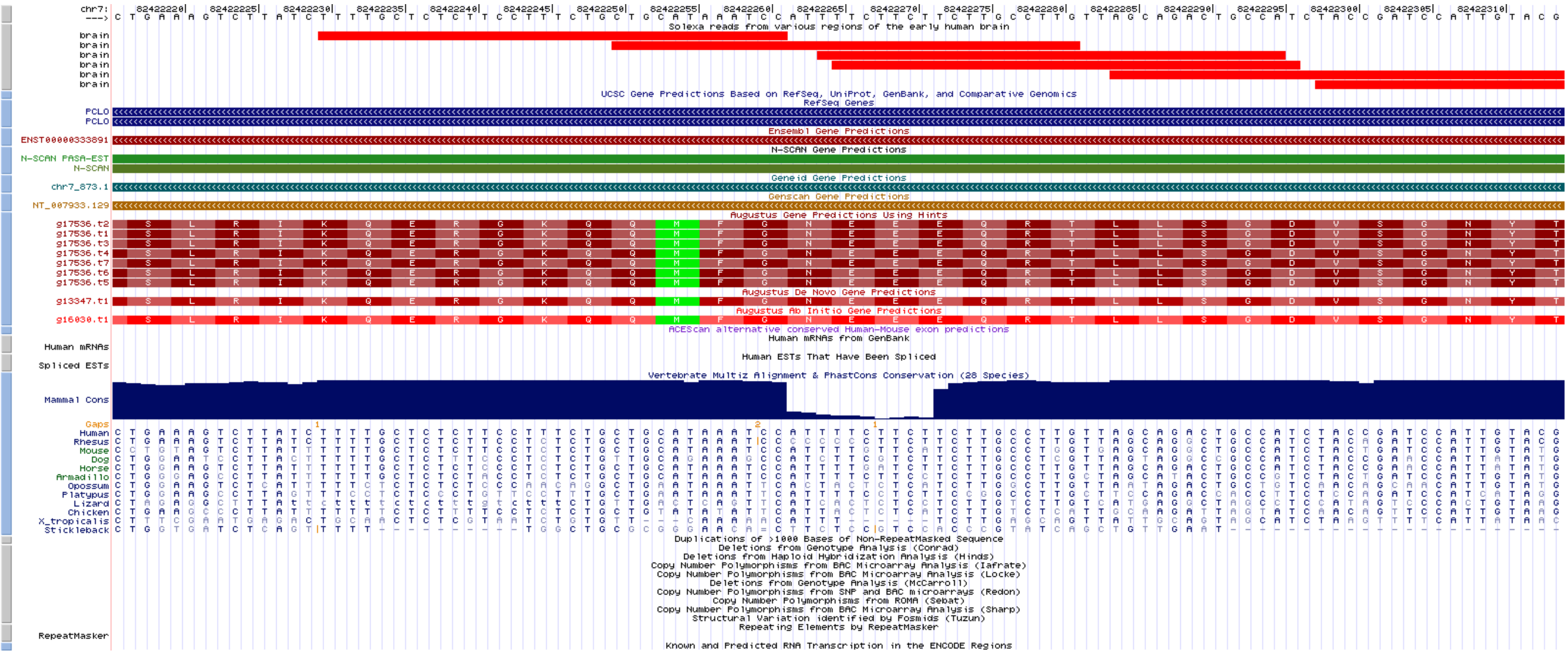


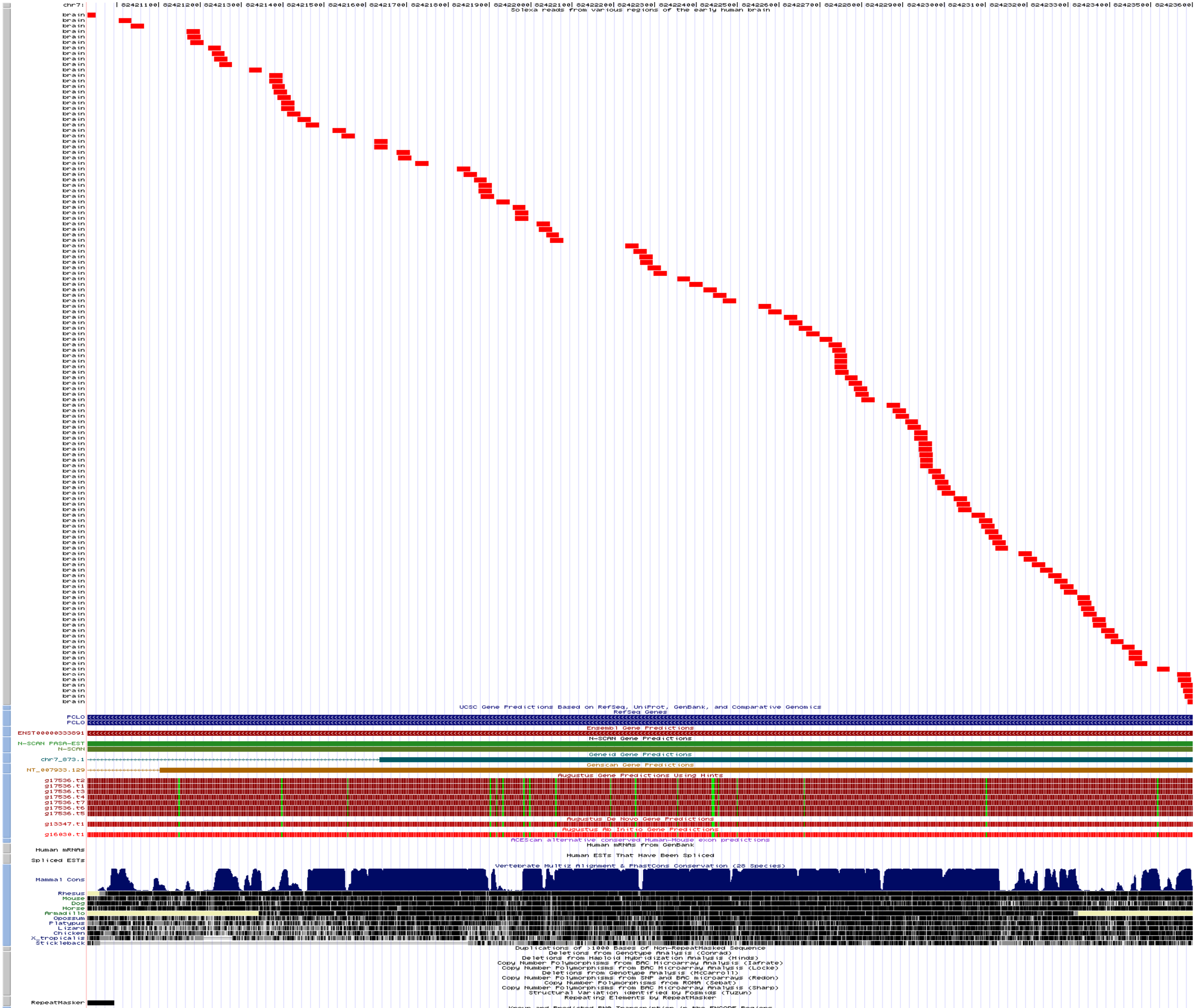
Purushothaman *et al*, 2005
IonTorrent, Inc.

Each Platform has various sources of noise, and thus Error

- De-Phasing
  - Lagging strand dephasing from incomplete extension
  - Leading strand dephasing from over-extension
- Dark Nucleotides
- Polymerase errors (10^{-5} to 10^{-7})
- Platform-specific errors
  - Illumina more likely to have error after 'G'
  - PCR-based methods miss GC- and AT-rich regions

Alignment to the genome




chr7:
82421100 82421200 82421300 82421400 82421500 82421600 82421700 82421800 82421900 82422000 82422100 82422200 82422300 82422400 82422500 82422600 82422700 82422800 82422900 82423000 82423100 82423200 82423300 82423400 82423500 82423600
Solexa reads from various regions of the early human brain
brain
UCSC Gene Predictions Based on RefSeq, UniProt, GenBank, and Comparative Genomics
RefSeq Genes
PCLO
PCLO
ENST00000333891
Ensembl Gene Predictions
N-SCAN Gene Predictions
N-SCAN PASA-EST
N-SCAN
chr7_073.1
Geneid Gene Predictions
NT_007933.129
Genscan Gene Predictions
Augustus Gene Predictions Using Hints
g17536.t2
g17536.t1
g17536.t3
g17536.t4
g17536.t7
g17536.t6
g17536.t5
Augustus De Novo Gene Predictions
g13347.t1
Augustus Ab Initio Gene Predictions
g16030.t1
ACEScan alternative conserved Human-Mouse exon predictions
Human mRNAs from GenBank
Human mRNAs
Human ESTs That Have Been Spliced
Spliced ESTs
Vertebrate Multiz Alignment & PhastCons Conservation (28 Species)
Mammal Cons
Rhesus
Mouse
Dog
Horse
Armadillo
Opossum
Platypus
Lizard
Chicken
X_tropicalis
Stickleback
Duplications of >1000 Bases of Non-RepeatMasked Sequence
Deletions from Genotype Analysis (Conrad)
Deletions from Haploid Hybridization Analysis (Hinds)
Copy Number Polymorphisms from BAC Microarray Analysis (Iafrate)
Copy Number Polymorphisms from BAC Microarray Analysis (Locke)
Deletions from Genotype Analysis (McCarroll)
Copy Number Polymorphisms from SNP and BAC microarrays (Redon)
Copy Number Polymorphisms from ROMA (Sebat)
Copy Number Polymorphisms from BAC Microarray Analysis (Sharp)
Structural Variation identified by Fosmids (Tuzun)
Repeating Elements by RepeatMasker
RepeatMasker
Known and Predicted RNA Transcription in the ENCODE Regions

Analyzing High-Resolution Data

- Bayesian Methods
- Hidden Markov Models
- Permutation Testing
- Circular Binary Segmentation
- Seed-seeking
- Least Squares Regression
- Democratic Voting

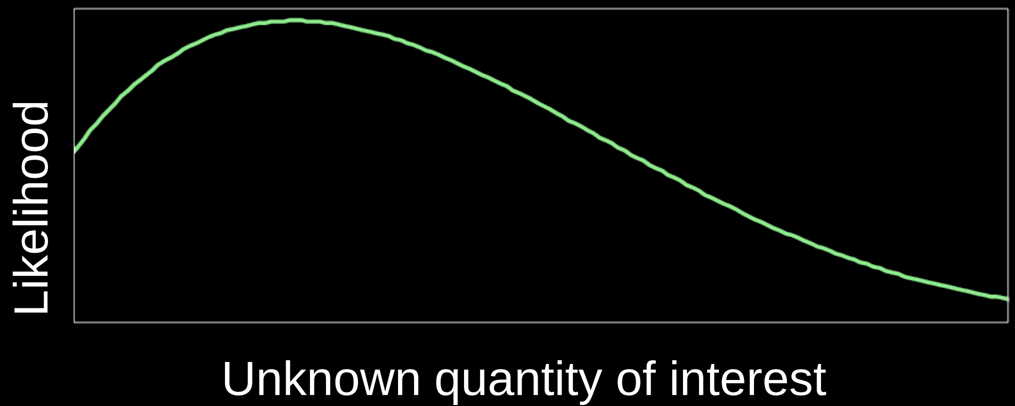
Prior

- The prior function Pr (H) gives the probability of different possible values of the quantity of interest *before* the data are considered – that is, it represents the state of knowledge *prior* to the data.
- Prior may be broad or flat if we have few data (non-informative prior), or peaked if we have more information (informative prior).

(Population size, length at sexual maturity, haplotype frequency, model parameter)

Likelihood

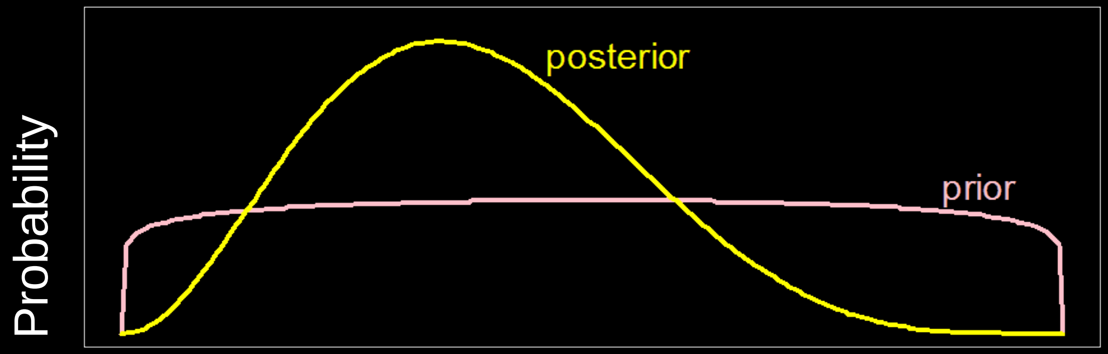
- The likelihood function Pr (data | H) gives the probability of obtaining the data, given different possible values of the unknown quantity of interest (the "hypothesis" H).
- The likelihood is calculated using a statistical model, which represents the process that produced the data. The likelihood function connects the parameters of the model to the data.



(Population size, length at sexual maturity, haplotype frequency, model parameter)

Posterior

- The posterior function Pr (H | data) gives the probability of different possible values of the quantity of interest *after* the data are considered – that is, it represents the state of knowledge *posterior* to the data.
- The posterior is a combination of the prior (what we knew before) and the likelihood (what the data told us).
- The difference between the prior and the posterior indicates how much we learned from the data.



Unknown quantity of interest

(Population size, length at sexual maturity, haplotype frequency, model parameter)

Paradigm for Bayesian inference

posterior likelihood x prior

new state of knowledge	information from new data	x	current state of knowledge

Thus, in Bayesian reasoning, new data update the current state of knowledge through Bayes' Theorem. The result is a new state of knowledge represented by the posterior.

Bayes' Theorem

(H) The Hypothesis (the unknown) and x = data

$$p\,\textbf{(H | data)} = \frac{p\,\textbf{(data | H)}\;\; p\,\textbf{(H)}}{p\,\textbf{(data)}}$$

posterior = likelihood x **prior**

2 fundamental Bayesian concepts:

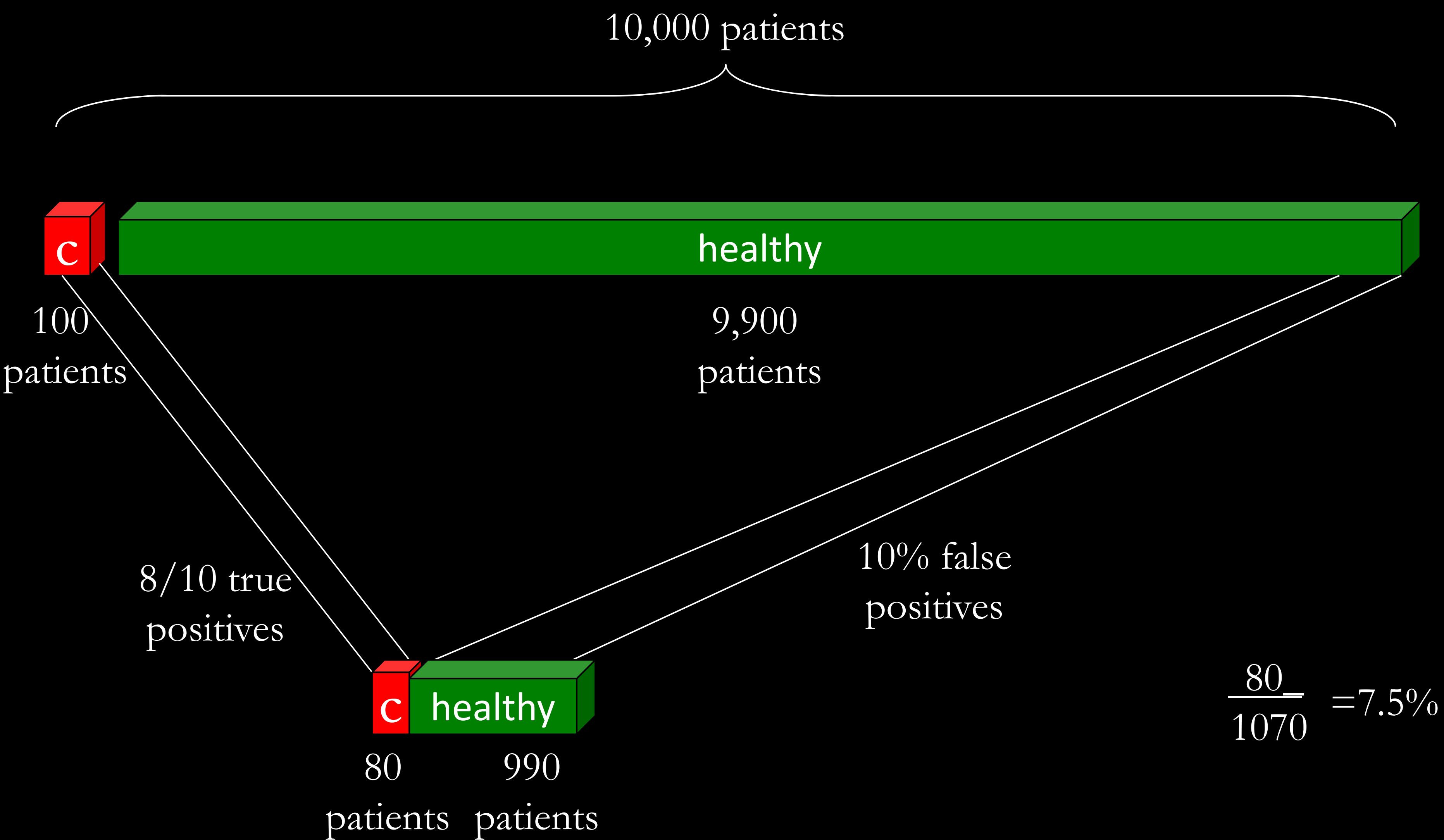
1. Things that are unknown are represented by probability distributions.

2. Things that are known (data) are used to improve the knowledge of unknowns through Bayes' Theorem.

Bayesian view of cancer

- 1% of women at 40 yrs have breast cancer
- Mammography diagnoses 8/10 correctly (true positive rate, false negative rate)
- 10% of mammographies are false positives

- If you get a positive result, what are your odds of having breast cancer?

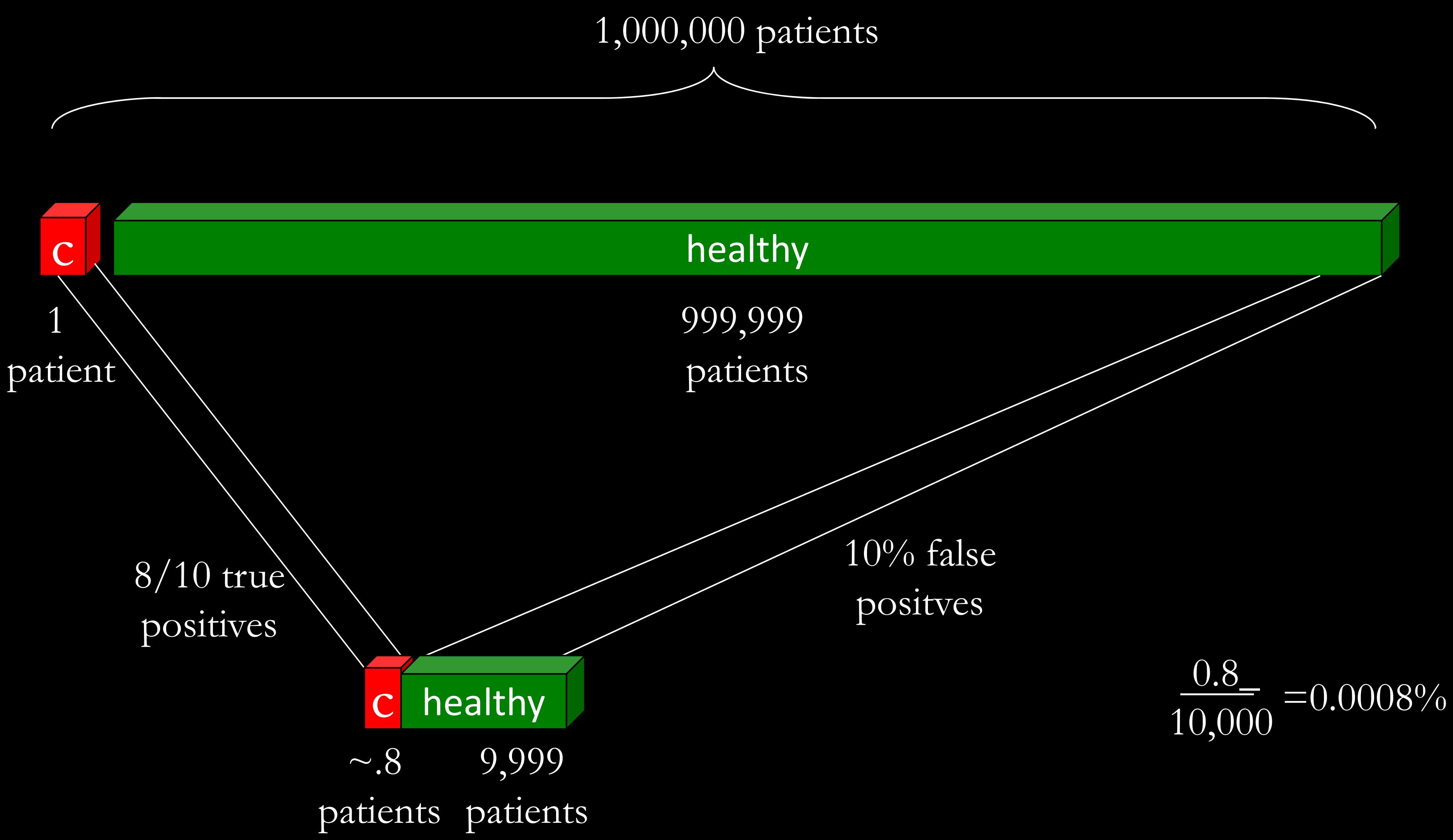

10,000 patients
c
healthy
100
patients
9,900
patients
8/10 true
positives
10% false
positives
c
healthy
80
patients
990
patients
80
1070
=7.5%

Bayes Theorem depends on the prior probability (pr(A)):

$$\Pr(A|B) = \frac{\Pr(B|A)\ \Pr(A)}{\Pr(B)}.$$

A = have cancer
B = positive result

$$\frac{.8*.01}{.8*.01 + .1*.99} = 7.5\%$$


1,000,000 patients
c
healthy
1
patient
999,999
patients
8/10 true
positives
10% false
positves
c
healthy
~.8
patients
9,999
patients
0.8
10,000
=0.0008%

Q. What is the Bayesian Conspiracy?

A. The Bayesian Conspiracy is a multinational, interdisciplinary, and shadowy group of scientists that controls publication, grants, tenure, and the illicit traffic in grad students. The best way to be accepted into the Bayesian Conspiracy is to join the Campus Crusade for Bayes in high school or college, and gradually work your way up to the inner circles. It is rumored that at the upper levels of the Bayesian Conspiracy exist nine silent figures known only as the Bayes Council.

http://yudkowsky.net/rational/bayes

Applications of Bayes to DNA-Seq

GATK single sample genotype likelihoods

Bayesian model

$$\overbrace{L(G \mid D)}^{\text{Likelihood for the genotype}} = \overbrace{P(G)}^{\text{Prior for the genotype}}\overbrace{P(D \mid G)}^{\text{Likelihood of the data given the genotype}} = \overbrace{\prod_{b \in \{good\_bases\}} P(b \mid G)}^{\text{Independent base model}}$$

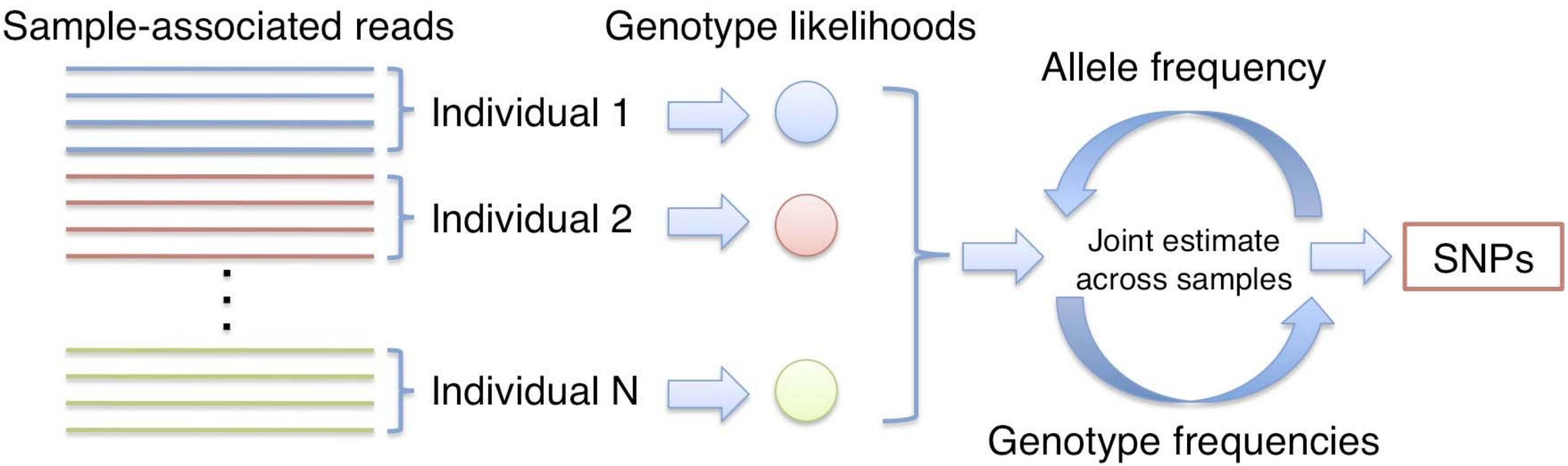
- Priors applied during multi-sample calculation; P(G) = 1
- Likelihood of data computed using pileup of bases and associated quality scores at given locus
- Only “good bases” are included: those satisfying minimum base quality, mapping read quality, pair mapping quality, NQS
- P(b | G) uses a platform-specific confusion matrix
- L(G|D) computed for all 10 genotypes

Each Platform is slightly different, and so intrinic errors are different

Figure Redacted
Unpublished Proprietary Data

Thierry-Mieg, SeQC Consortium

The Broad Unified Genotyper SNP caller multiple-sample allele frequency and genotype estimates



- This approach allows us to combine weak single sample calls to discover variation among samples with high confidence

See http://www.broadinstitute.org/gsa/wiki/index.php/Unified_genotyper for more information

SNIP-Seq SNP calling

For each potential variant site in the sequenced region:

Set the base quality value for each base call to the Illumina quality value

For k = 1, 2,…

a. Sample a genotype for each individual from the posterior distribution using a heterozygote prior of 0.001.

b. Recalibrate the quality score for each base call using genotypes for all individuals.

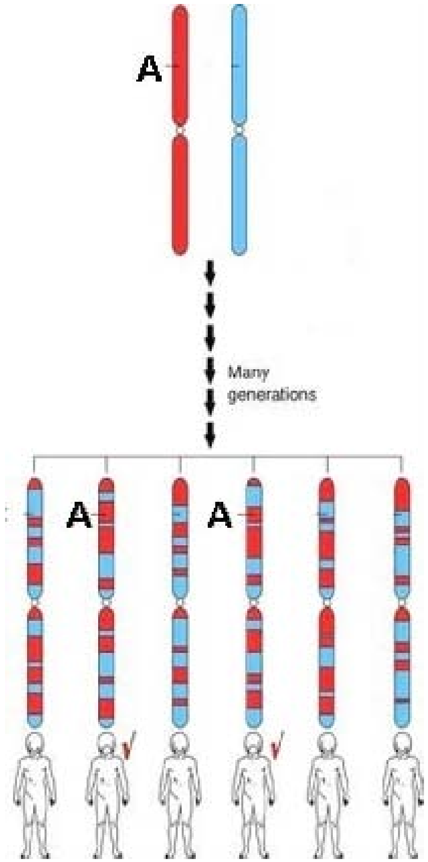
If the genotype of any individual is different from the reference, identify position as a SNP.

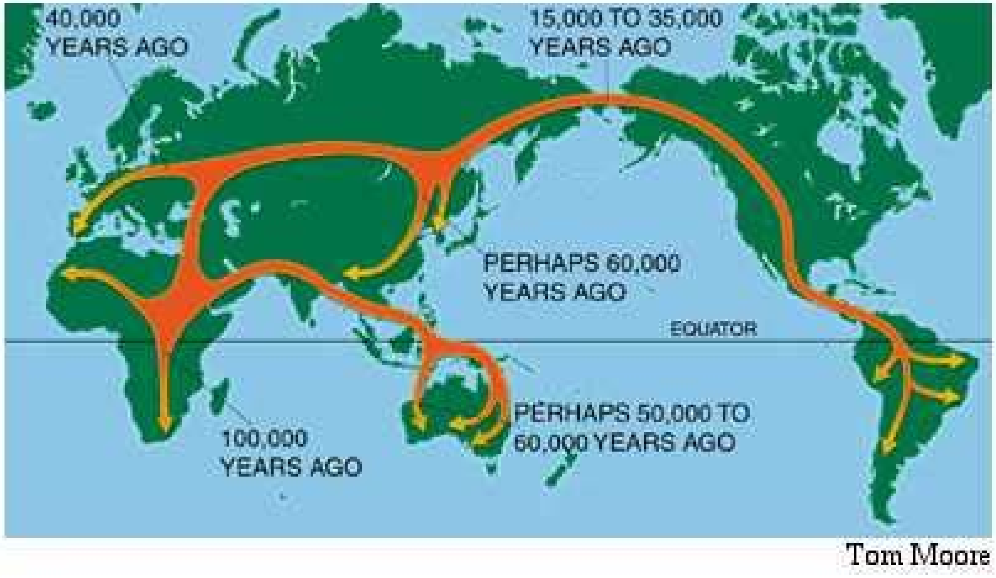
Sample a genotype for each individual from the posterior distribution computed using a heterozygote prior of 0.2.

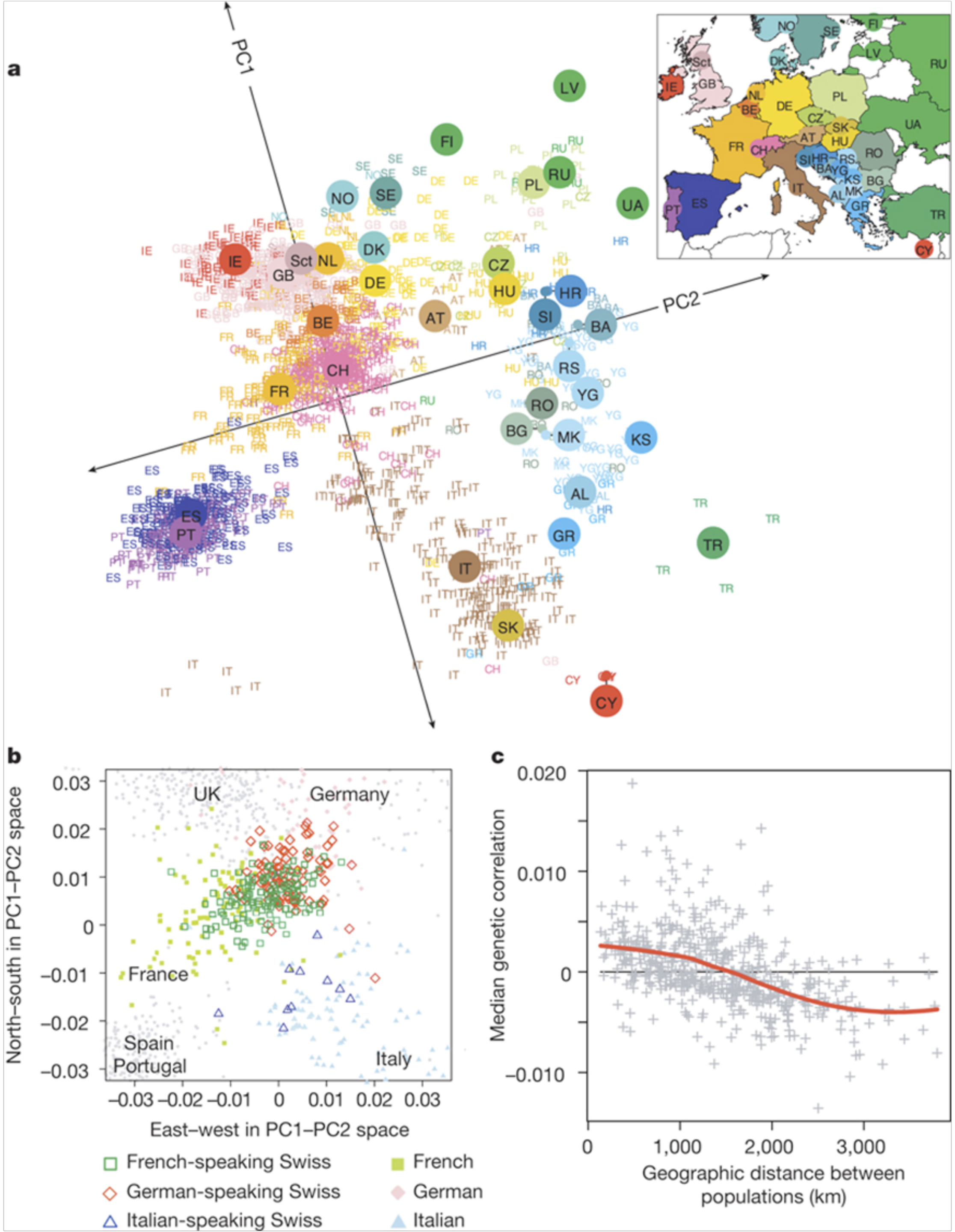
Bansal et al, 2010

Population-based models can overcome systematic errors, but…

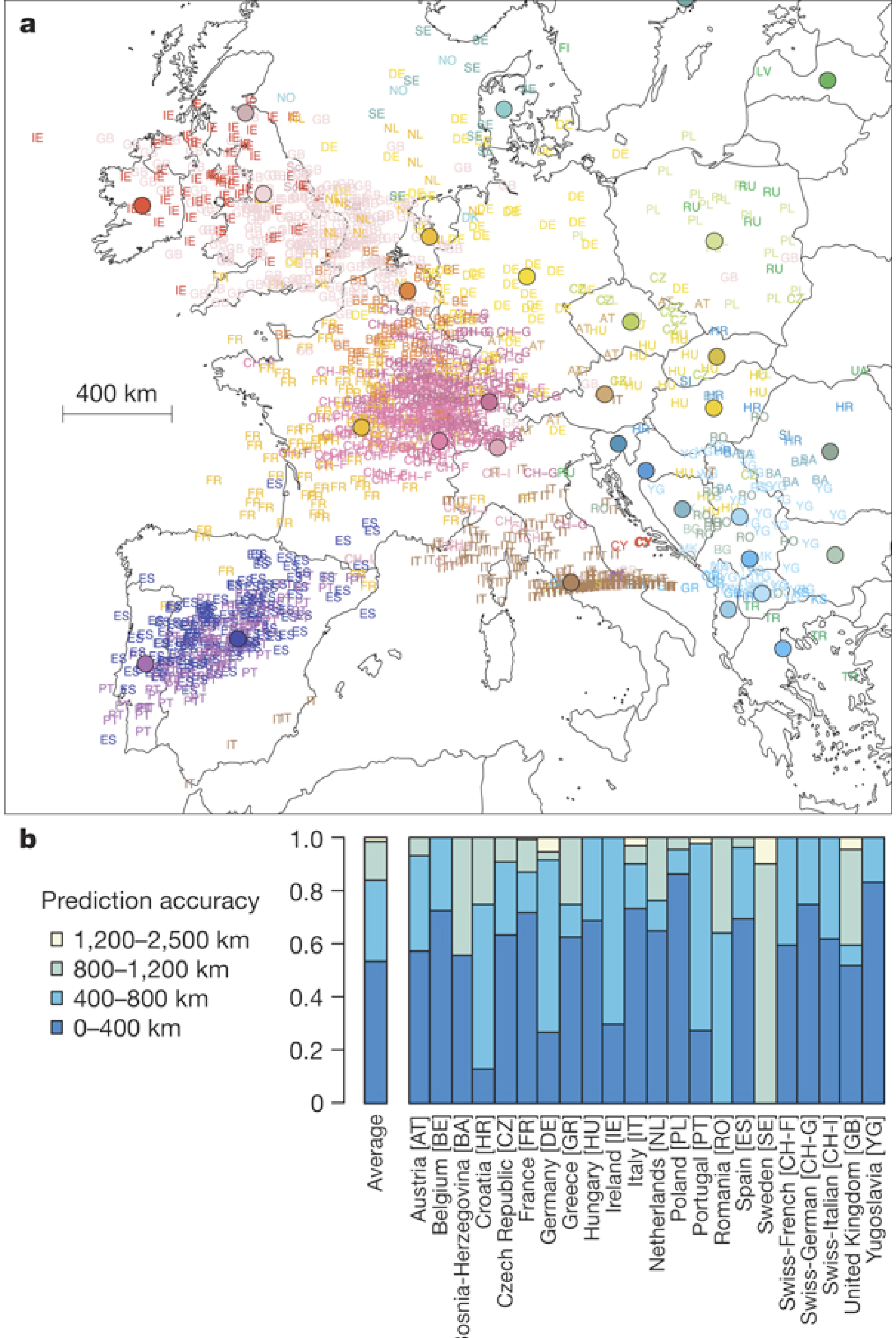
Population Stratification is from the migration patterns of haplotypes throughout human history







Genes mirror geography within Europe
Novembre et al, 2008



Genes mirror geography within Europe
Novembre et al, 2008

Applications of Bayes to RNA-Seq

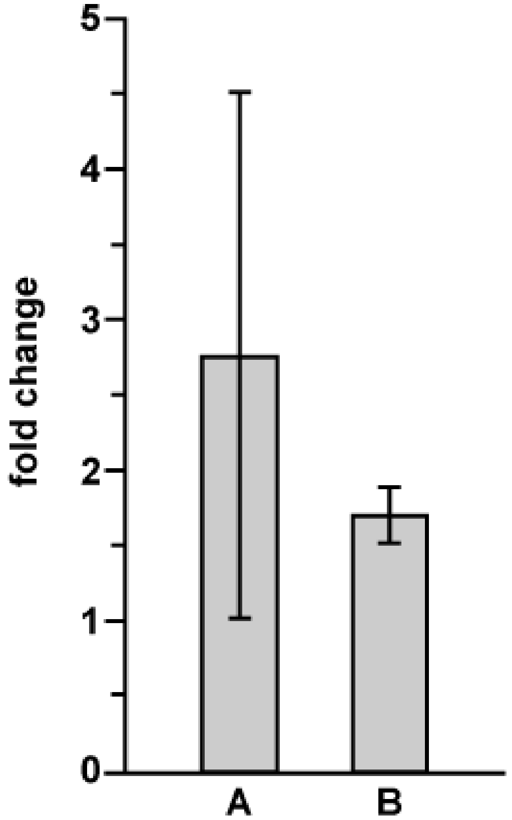
Microarrays vs. RNA-Seq

RNA-Seq: An assessment of technical reproducibility and comparison with gene expression arrays

Marioni and Mason *et. al*, Genome Research, 2008

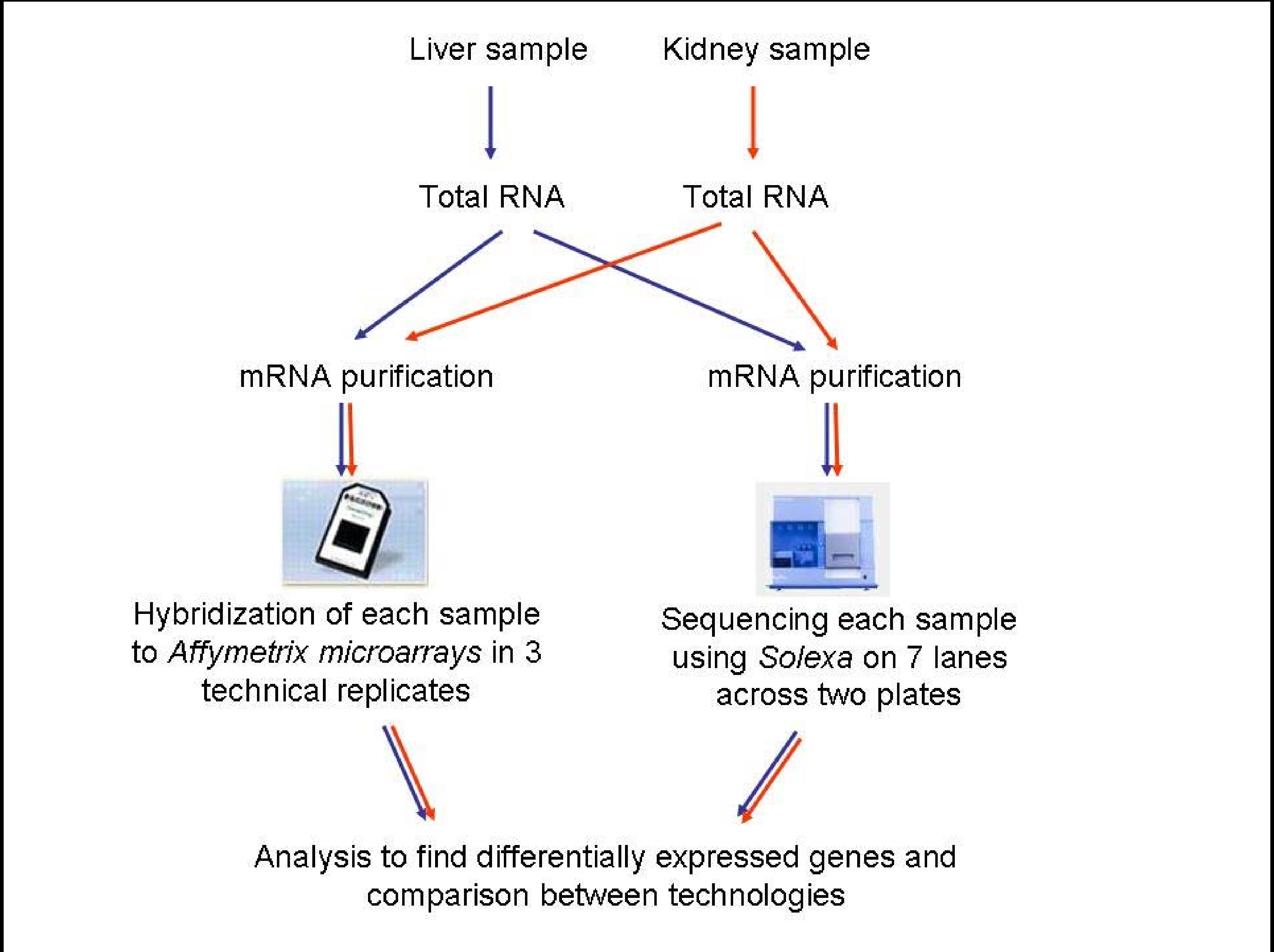
Data Analysis: What genes are differentially expressed?

- Early days—fold change cutoffs (e.g., 2x difference or better)
- not a very satisfying approach:
  - -doesn’t take into account variance
  - -misses any small changes



Here, “A” has a fold change >2.5, but varies greatly between replicate experiments. “B” has a fold change of only 1.75, but changes reliably each time the experiment is performed.

Experimental Design: Liver vs. Kidney



Metric for RNA-Seq Expression

RPKM:

Reads per Kilobase per Million Reads

Normalizes for (1) gene size and (2) sequencing depth

(~0.1-1 transcript/cell)

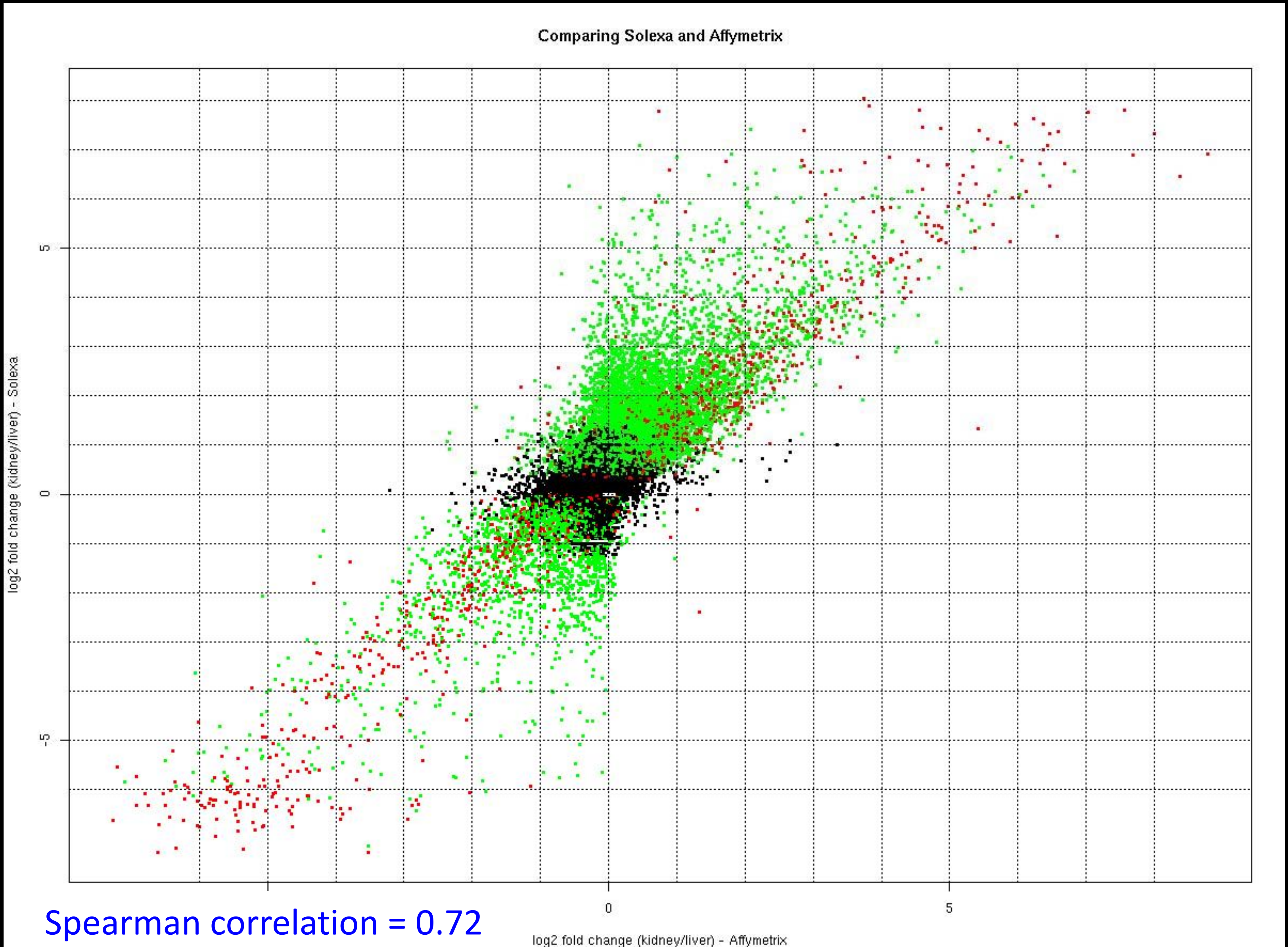
$$\mathrm{RPKM} = \frac{N\ reads}{1\ gene} \times \frac{1\ \mathrm{gene}}{B\ bp} \times \frac{1000\ \mathrm{bp}}{1Kb} \times \frac{1\ Million\ reads}{Y\ total\ reads}$$

Y = (exons, introns, intergenic reads)

FPKM=fragments-PKM
is for paired-end data

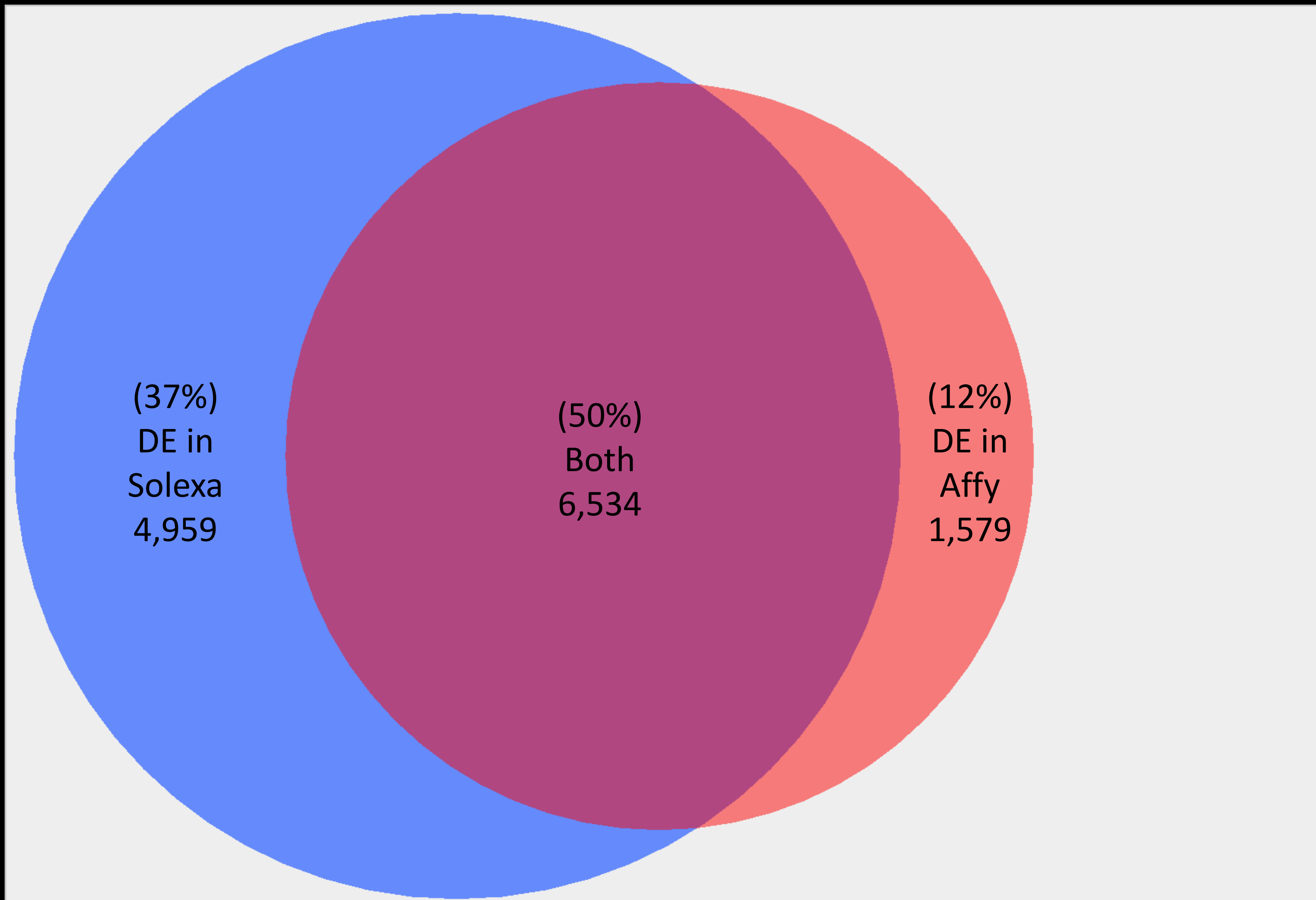
Mortazavi, Williams, *et al.*
Nature Methods, 2008

Comparing GA and Affy arrays

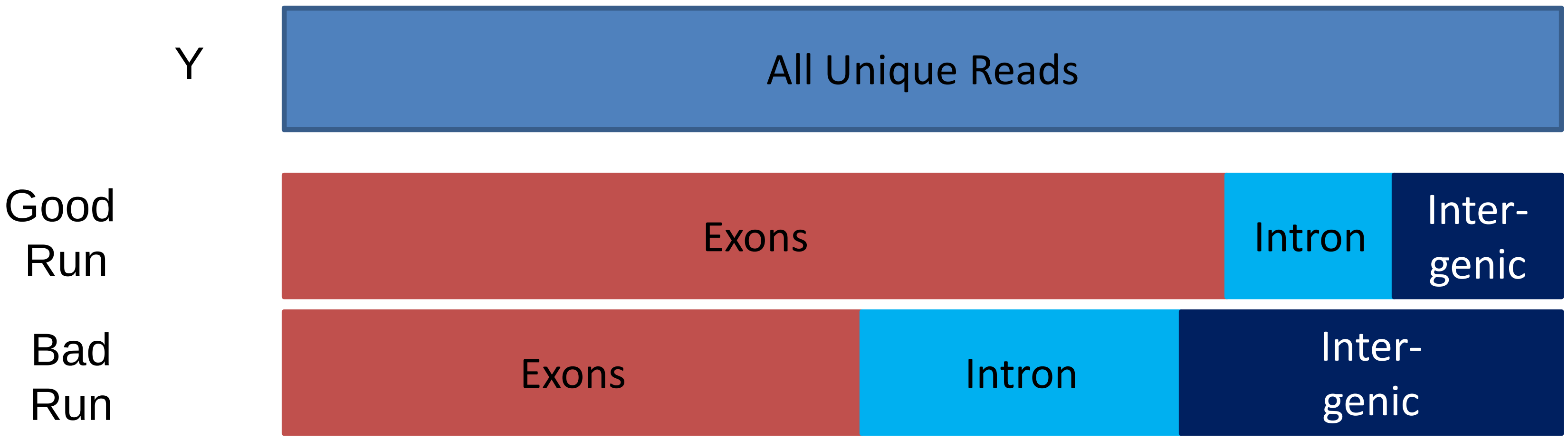


Spearman correlation = 0.72

13,072 Differentially Expressed (DE) Genes



Bias is introduced if these ratios are not kept:



Good Run RPKM = 128.4

Bad Run RPKM = 72.7

Mortazavi, Williams, *et al.*
Nature Methods, 2008

Normalization is needed

Define Y_{gk} as the observed count for gene g in library k summarized from the raw reads, μ_{gk} as the true and unknown expression level (number of transcripts), L_g as the length of gene g and N_k as total number of reads for library k. We can model the expected value of Y_{gk} as:

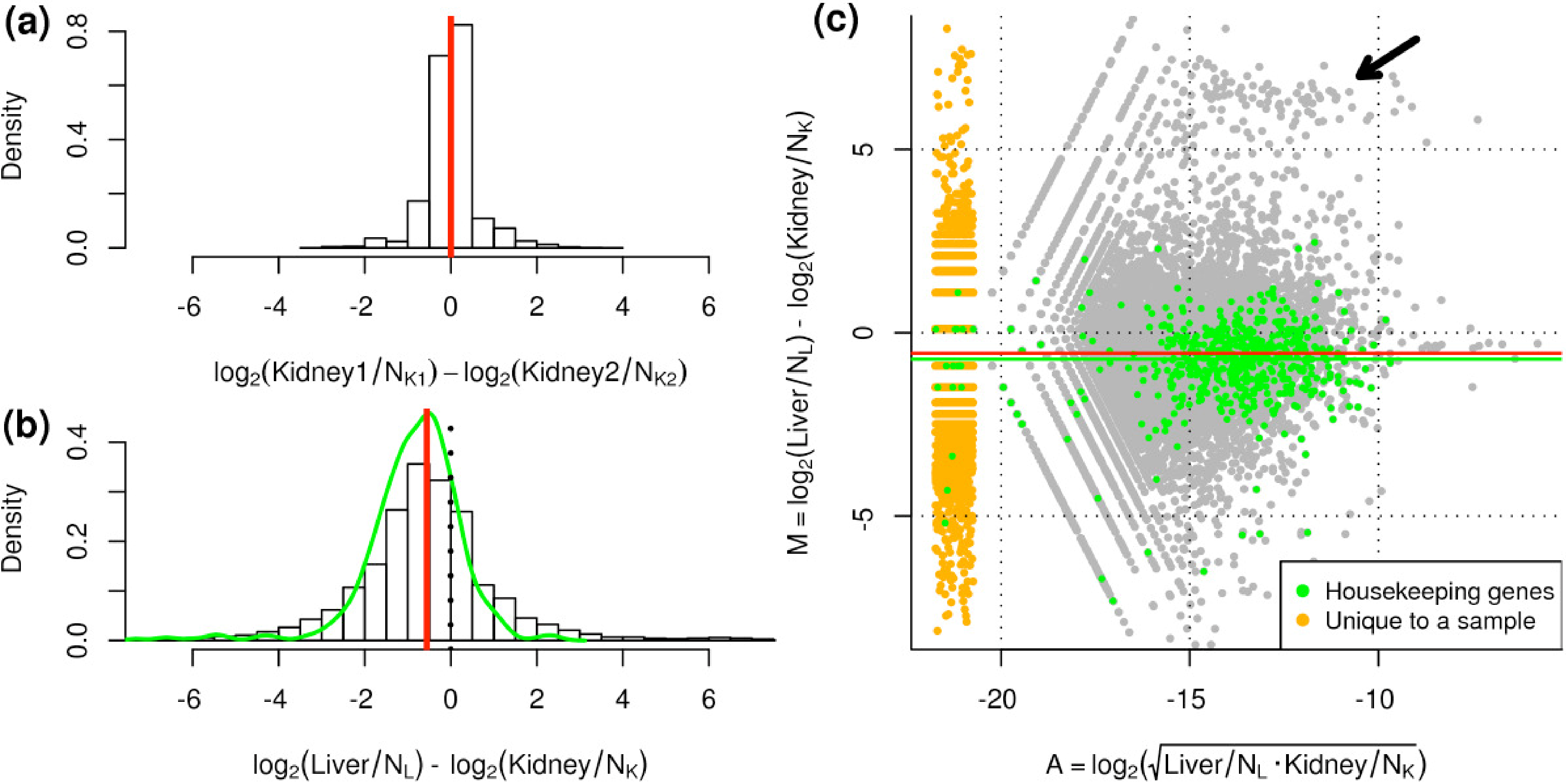
$$E[Y_{gk}] = \frac{\mu_{gk} L_g}{S_k} N_k$$

$$\text{where } S_k = \sum_{g=1}^{G} \mu_{gk} L_g;$$

S_k represents the total RNA output of a sample. The problem underlying the analysis of RNA-seq data is that while N_k is known, S_k is unknown and can vary drastically from sample to sample, depending on the RNA composition.

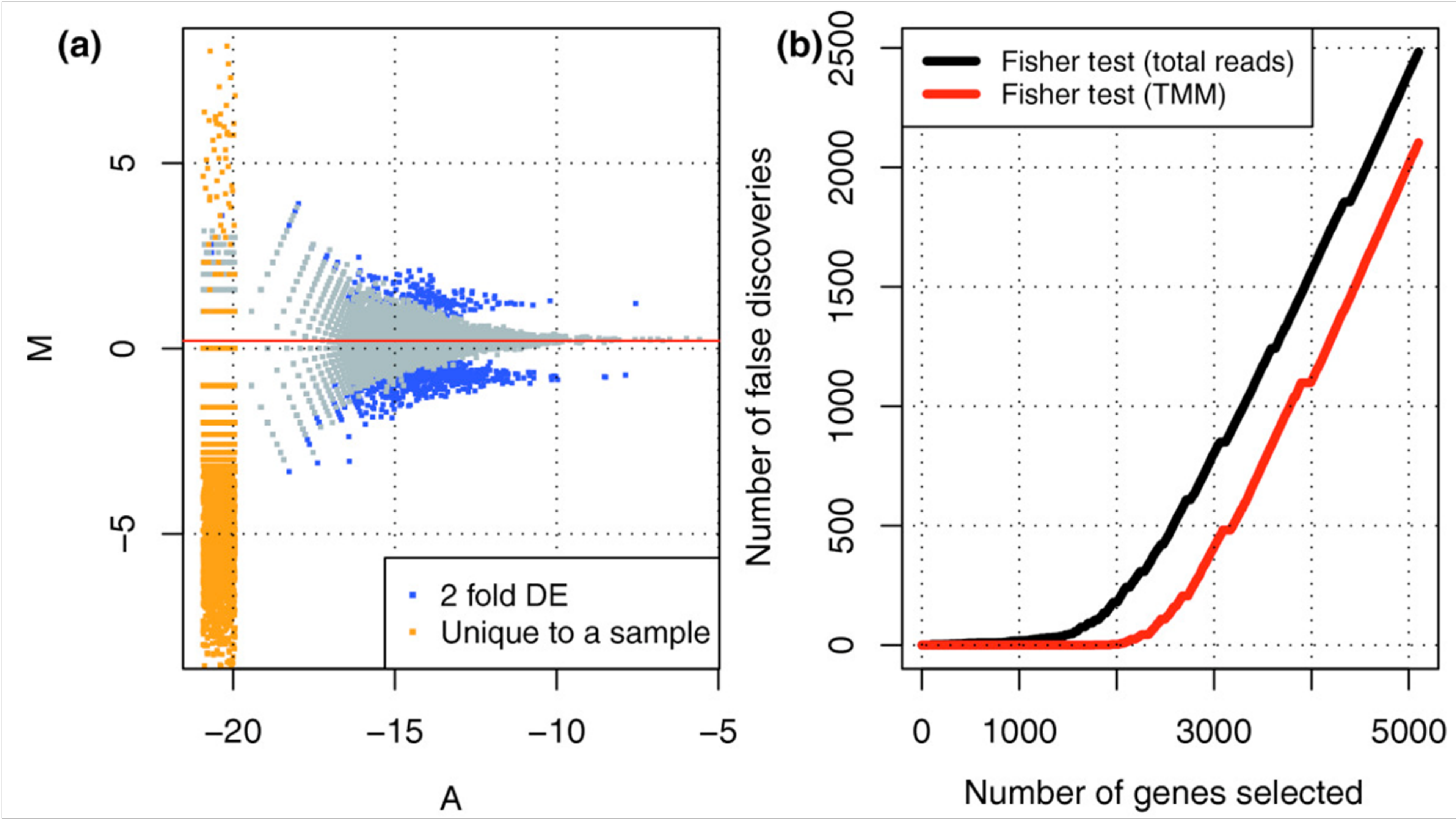
$$M_g = \log_2 \frac{Y_{gk} / N_k}{Y_{gk'} / N_{k'}}$$

Normalization is needed



Robinson and Oshlack Genome Biology 2010 11:R25

Normalization is needed



Robinson and Oshlack Genome Biology 2010 11:R25

MISO / Probabilistic analysis and design of RNA-Seq experiments for identifying mRNA isoform regulation

About MISO

About MISO

MISO (Mixture of Isoforms) is a probabilistic framework that quantitates the expression level of alternatively spliced genes from RNA-Seq data, and identifies differentially regulated isoforms or exons across samples. By modeling the generative process by which reads are produced from isoforms in RNA-Seq, the MISO model uses Bayesian inference to compute the probability that a read originated from a particular isoform.

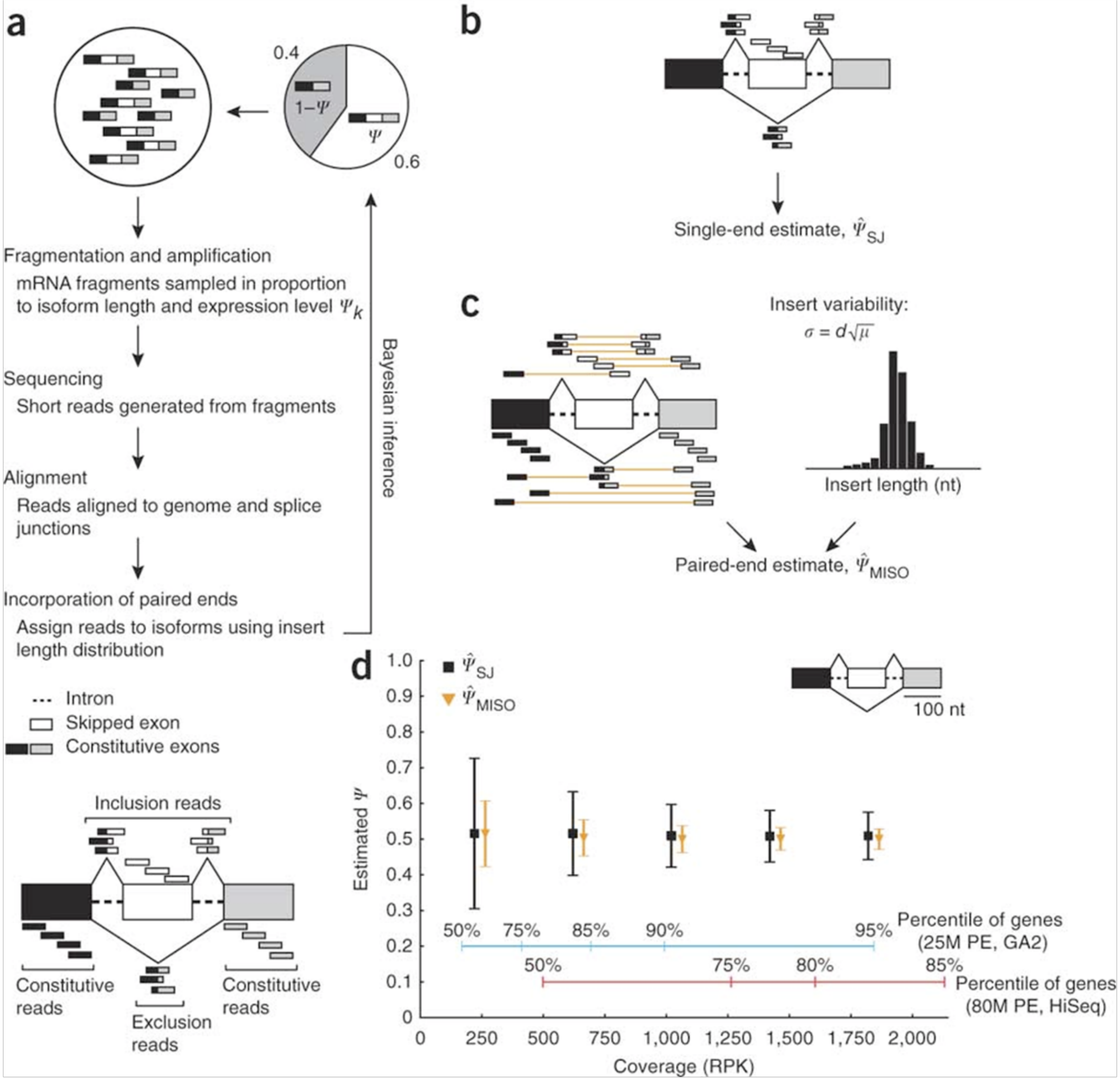
MISO uses the inferred assignment of reads to isoforms to quantitate the abundances of the underlying set of alternative mRNA isoforms. Confidence intervals over estimates can be obtained, which quantify the reliability of the estimates.

MIT
Depts. of Biology and Biological Engineering
Dept. of Brain and Cognitive Sciences

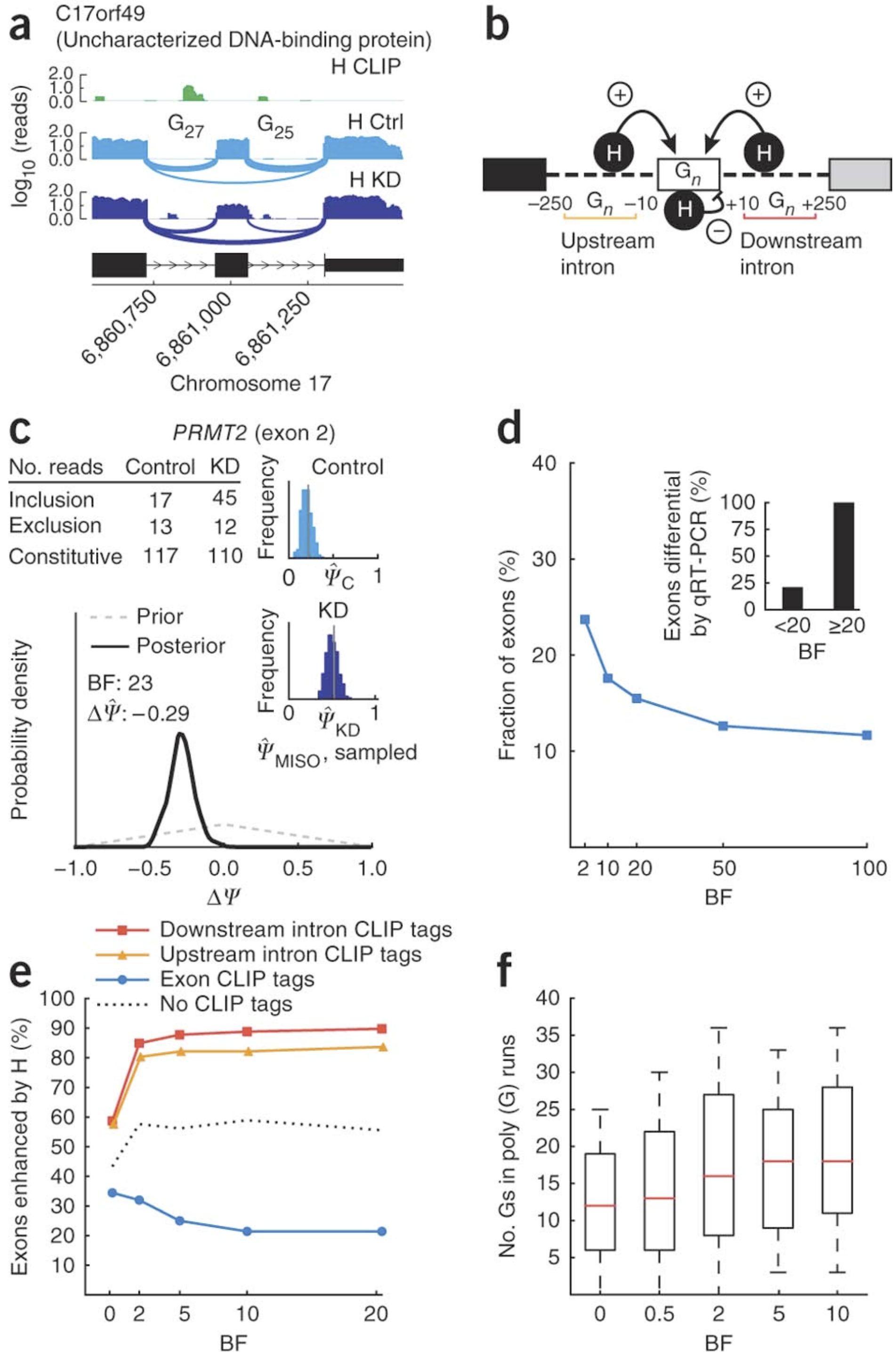
Harvard University
Dept. of Statistics
FAS Center for Systems Biology

Contact

Department of Biology MIT
31 Ames Street, 68-271A
Cambridge, MA 02139-4307



Katz et al, 2010



Katz et al, 2010

Coverage Requirements: How many lanes/plates/wells?

Depends on:

1. Read Length
2. Size of Transcriptome
3. Complexity of Transcriptome
4. Complexity of Tissue
5. Biological Variance
6. Errors (random and systematic)

Plateau of Information Starts @ ~500Mb

Number of lanes compared	Differentially expressed genes	Overlap with genes called from the array	Correlation of fold changes between Solexa and the array
One vs One	5670	4208	0.67
Two vs Two	7994	5340	0.70
Three vs Three	9482	5909	0.71
Four vs Four	10580	6278	0.72
Five vs Five	11493	6534	0.73

Liver

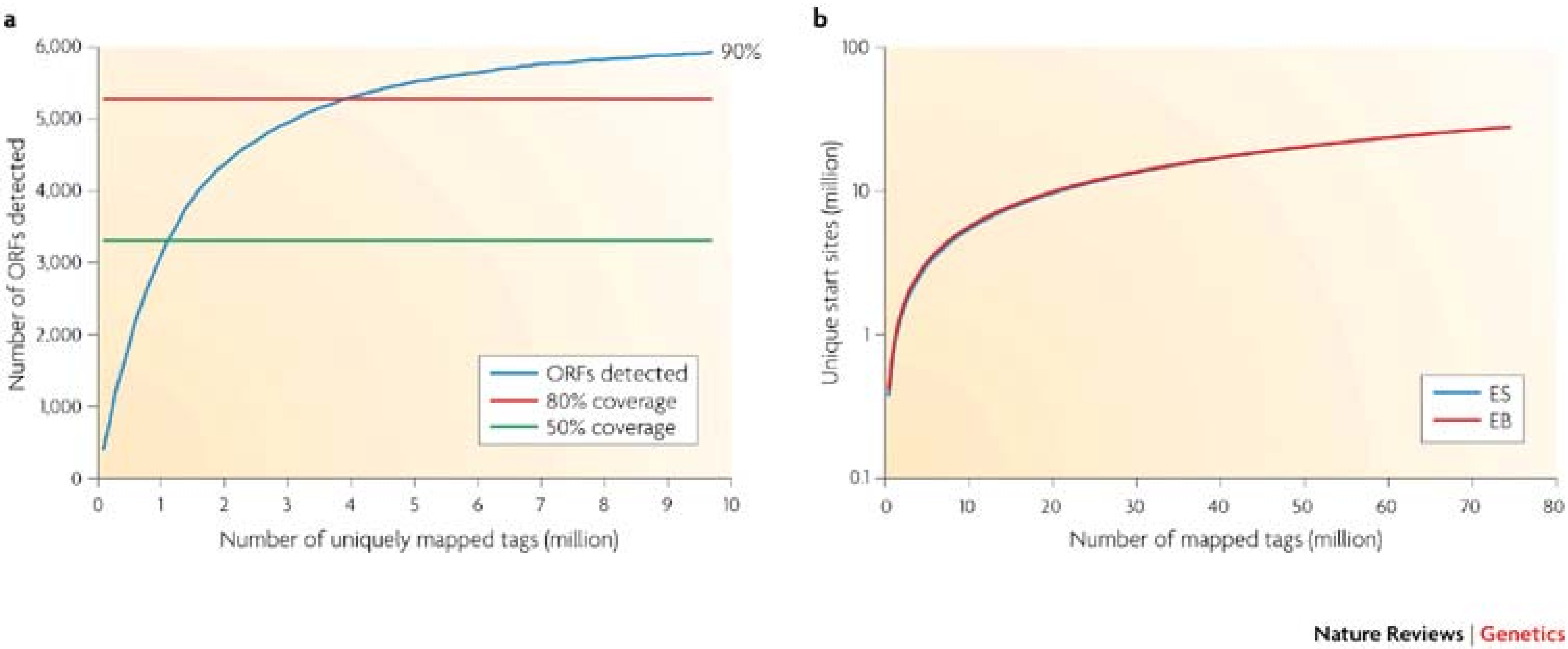
	No genes	Percentage
Five Lanes	20080	100
Four Lanes	19695	97.9
Three Lanes	19170	95.5
Two Lanes	18390	91.6
One Lane	16973	84.5

Kidney

	No genes	Percentage
Five Lanes	20921	100
Four Lanes	20552	98.2
Three Lanes	20064	96.0
Two Lanes	19355	92.5
One Lane	18080	86.4

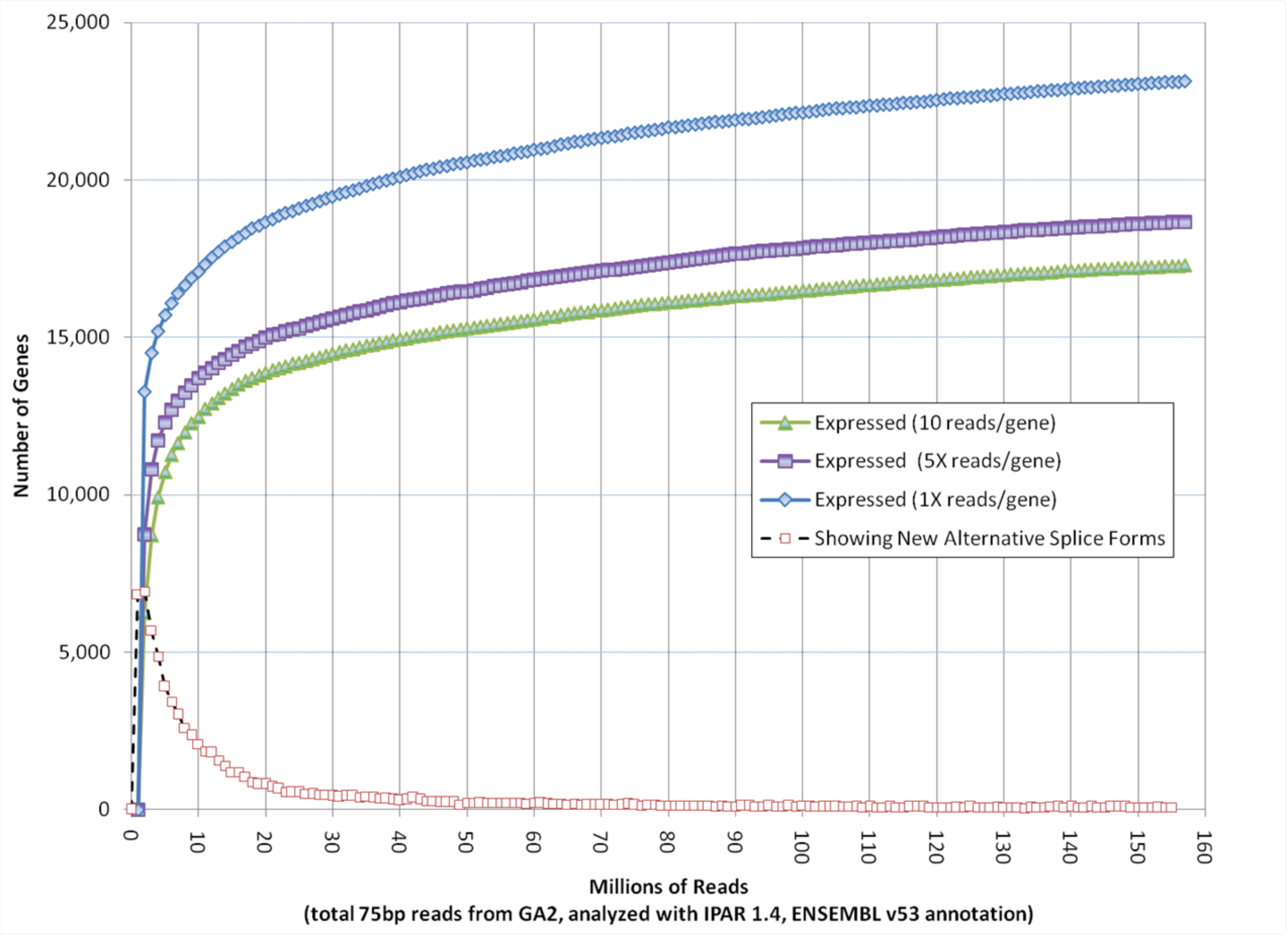
Marioni,Mason et al, Genome Research, 2008

Coverage Requirements



Wang, Gerstein, and Snyder, 2008

No current visible end of gene discovery



How many replicates do I need?

Calculation of the number of replicates depends on:

1. An estimate of σ^2 obtained from previous experiments.
2. The size of the difference (δ) to be detected.
3. The assurance with which it is desired to detect the difference (i.e., Power of the test = 1-β).
4. The level of significance to be used in the actual experiment (i.e., Type I error).
5. The test required, whether a one-tail or two-tail test.

To determine the number of replicates use the following formula :

$$\#reps = 2\left(Z_{\alpha/2} + Z_{\beta}\right)\left(\frac{\sigma}{\delta}\right)^2$$

where: *$Z\alpha/2$* is associated with the Type I error (two-tailed)
$Z\beta$ is associated with the Type II error
δ is the true difference to be detected, and
σ is the known variance obtained from previous experiments

Bayes in Chip-seq too!

Research article

Highly accessed Open Access

BayesPeak: Bayesian analysis of ChIP-seq data

Christiana Spyrou[1,3], **Rory Stark**[3], **Andy G Lynch**[4] and **Simon Tavaré**[2,4]

1 Statistical Laboratory, Centre for Mathematical Sciences, Wilberforce Road, Cambridge, UK
2 DAMTP, Centre for Mathematical Sciences, Wilberforce Road, Cambridge, UK
3 Cancer Research UK, Cambridge Research Institute, Li Ka Shing Centre, Robinson Way, Cambridge UK
4 Department of Oncology, University of Cambridge, Li Ka Shing Centre, Robinson Way, Cambridge, UK

author email corresponding author email

BMC Bioinformatics 2009, **10:**299 doi:10.1186/1471-2105-10-299

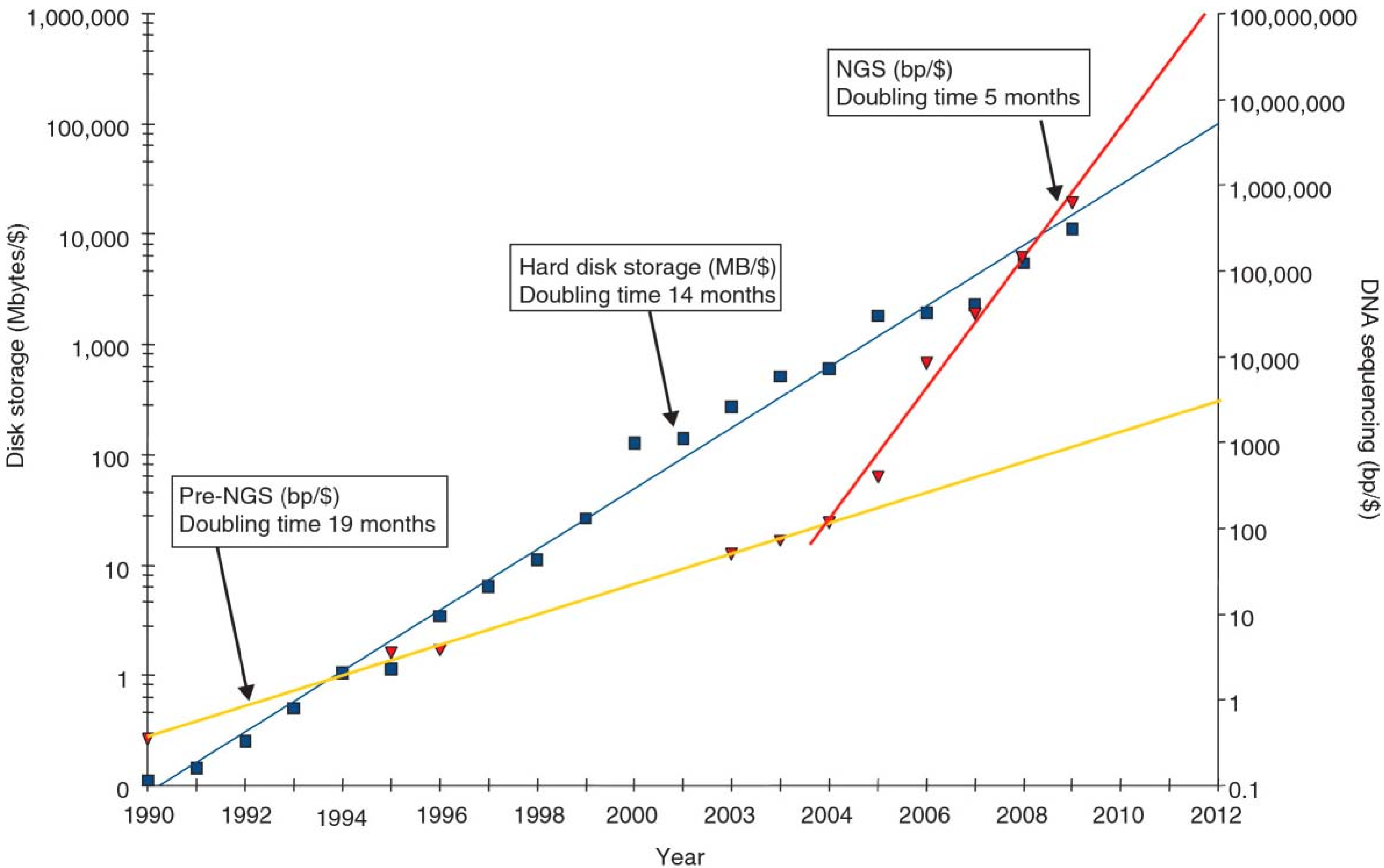
The electronic version of this article is the complete one and can be found online at: http://www.biomedcentral.com/1471-2105/10/299

Received: 8 May 2009
Accepted: 21 September 2009
Published: 21 September 2009

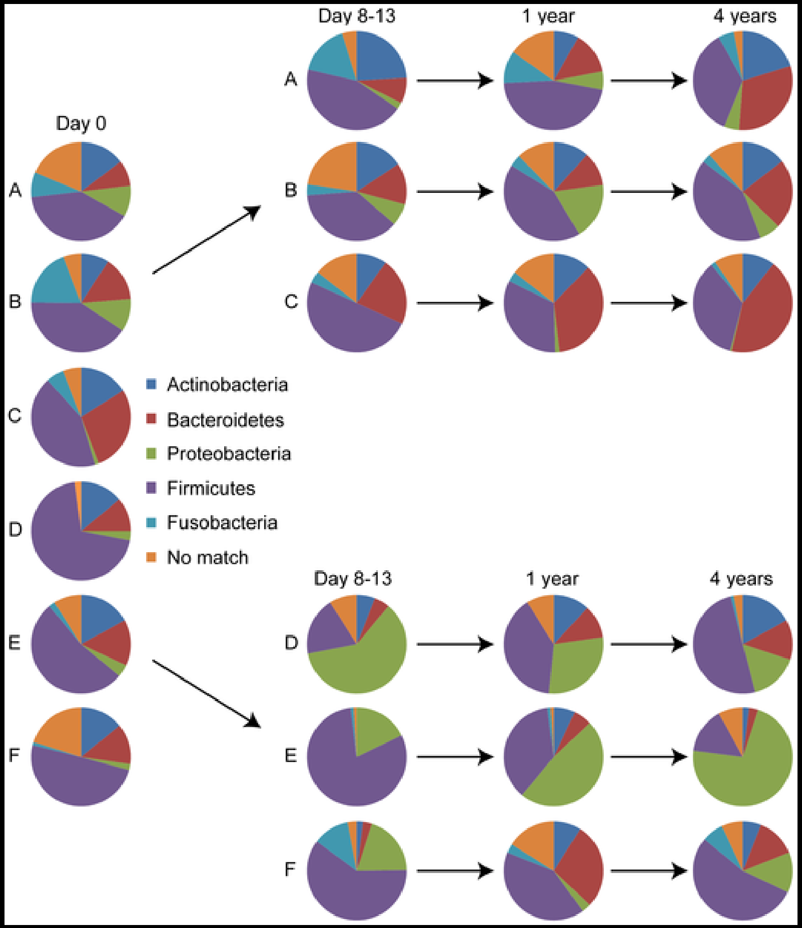
What's the problem with Bayesian statistics?
(according to non-Bayesians)

1. Priors can introduce subjective judgment into data analysis.
2. Priors affect the result. Different people can get different answers from the same data.
3. It's too hard. There are no simple point-and-click programs.

This requires a lot of space



Meta-genomic phenotypes can persist for years, and "passenger genomes" can be a phenotype, as well as their distributions.

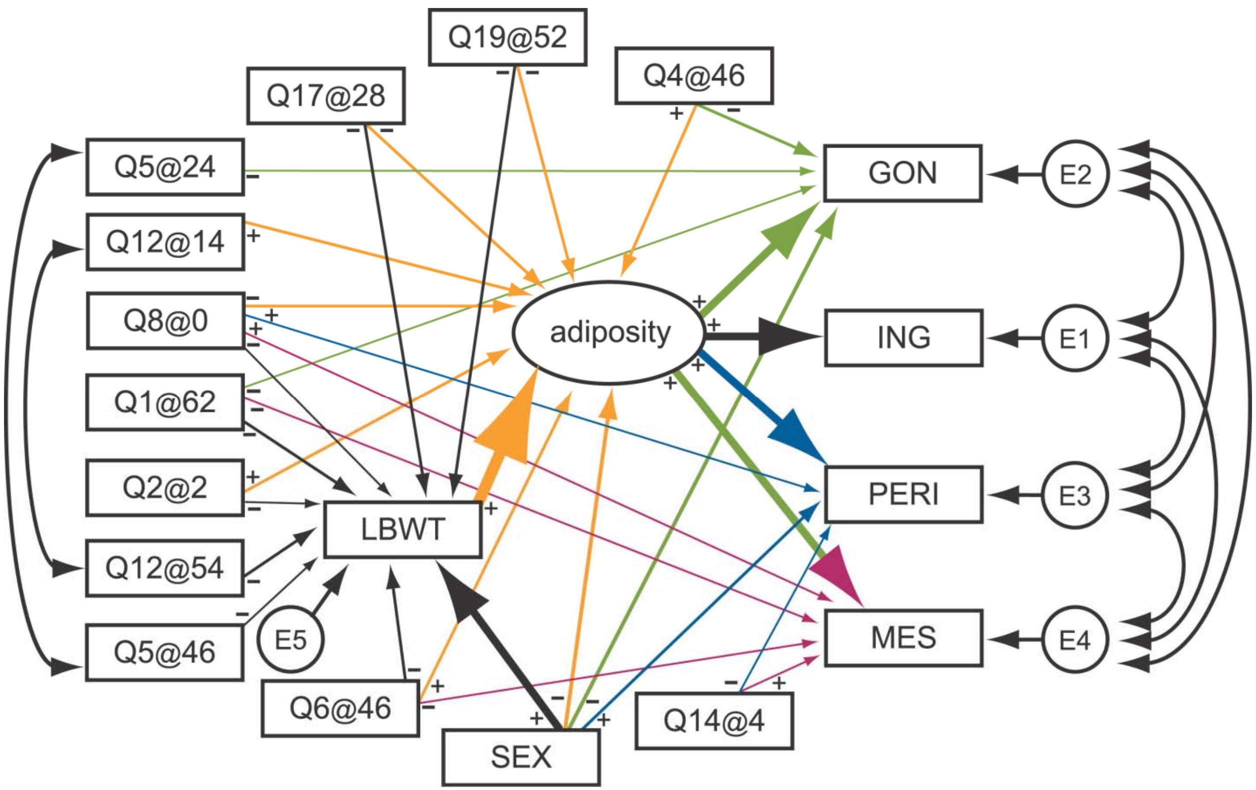


Normal throat

Throat + antibiotics

Jakobsson et al, 2010

Risk factors for diseases usually involve many genes and pathways



. Li et al, PLOS Genetics, 2007.

There are other factors than these!

nature International weekly journal of science

Search This journal

nature.com ▸ journal home ▸ current issue ▸ letter ▸ full text

NATURE | LETTER

Genome, epigenome and RNA sequences of monozygotic twins discordant for multiple sclerosis SHOW NO DIFFERENCE

Sergio E. Baranzini, Joann Mudge, Jennifer C. van Velkinburgh, Pouya Khankhanian, Irina Khrebtukova, Neil A. Miller, Lu Zhang, Andrew D. Farmer, Callum J. Bell, Ryan W. Kim, Gregory D. May, Jimmy E. Woodward, Stacy J. Caillier, Joseph P. McElroy, Refujia Gomez, Marcelo J. Pando, Leonda E. Clendenen, Elena E. Ganusova, Faye D. Schilkey, Thiruvarangan Ramaraj, Omar A. Khan, Jim J. Huntley, Shujun Luo, Pui-yan Kwok, Thomas D. Wu + *et al.*

Affiliations | Contributions | Corresponding authors

Nature **464**, 1351–1356 (29 April 2010) | doi:10.1038/nature08990

Received 25 July 2009 | Accepted 11 March 2010

Systems biology requires spatiotemporal monitoring of the genome, epigenome, transcriptome, proteome, metabolome, and the environment, to see the interactome