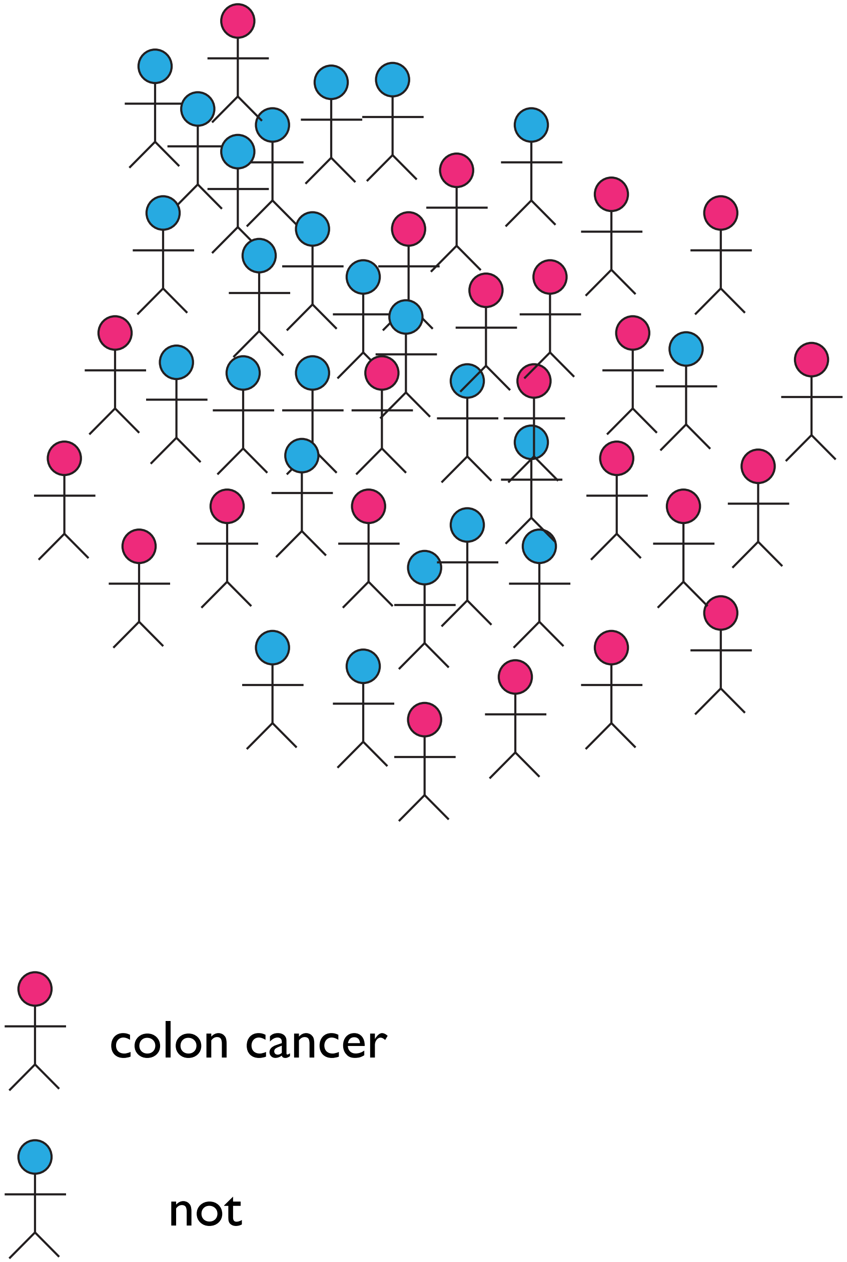
Computational and Statistical Methods for High-Throughput Genomics

Héctor Corrada Bravo
Dept. of Computer Science
Center for Bioinformatics and Computational Biology
University of Maryland

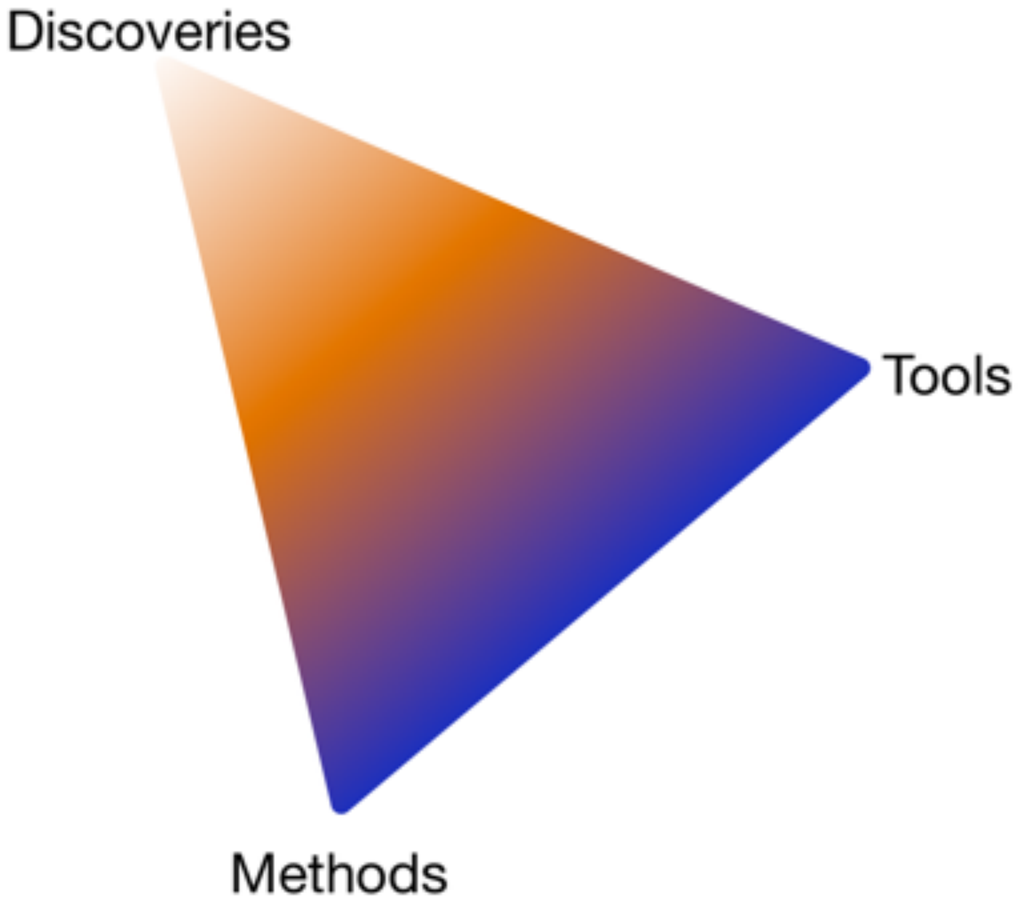
Computational Genomics



- Study the **molecular** basis of *variation* in development and disease

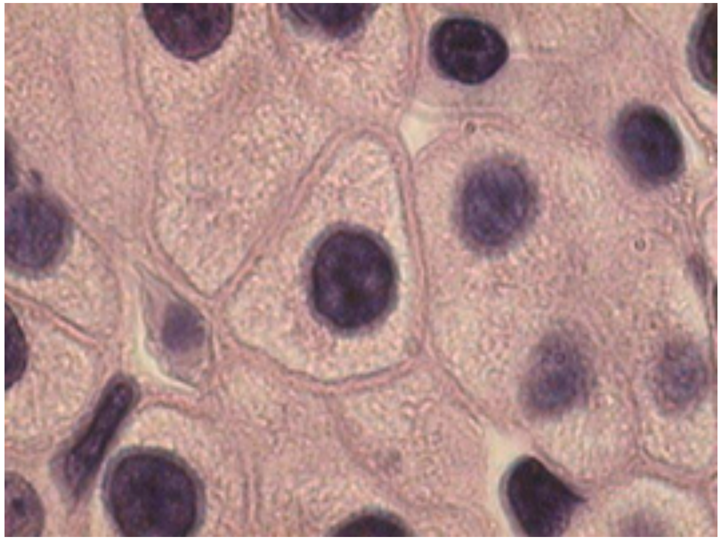
- Using **high-throughput** experimental methods
  - statistical learning
  - visualization
  - algorithms
  - data management

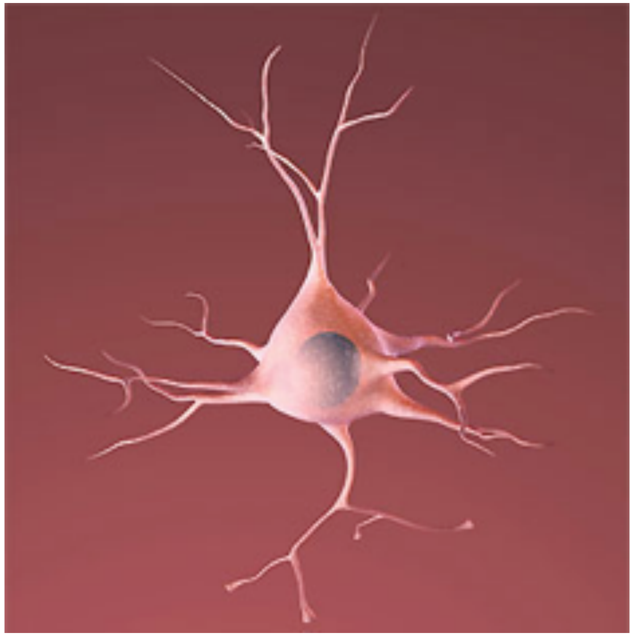
Computational Genomics



Differentiation

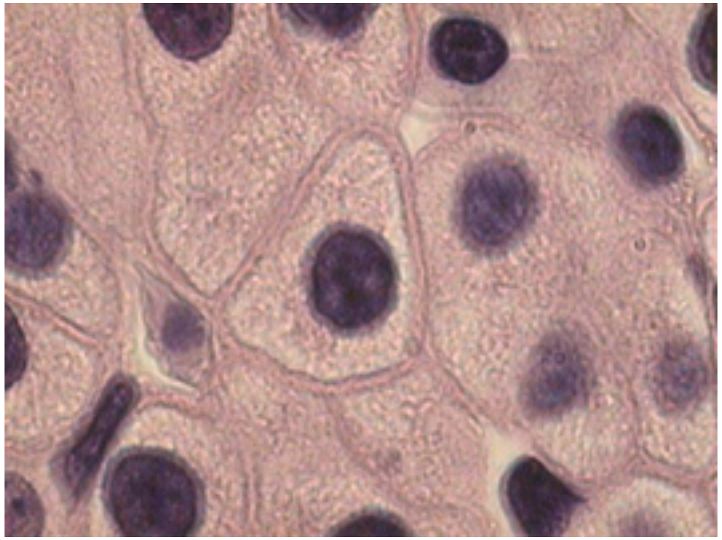
Different genes are **expressed** during different **stages** and in different **tissues**

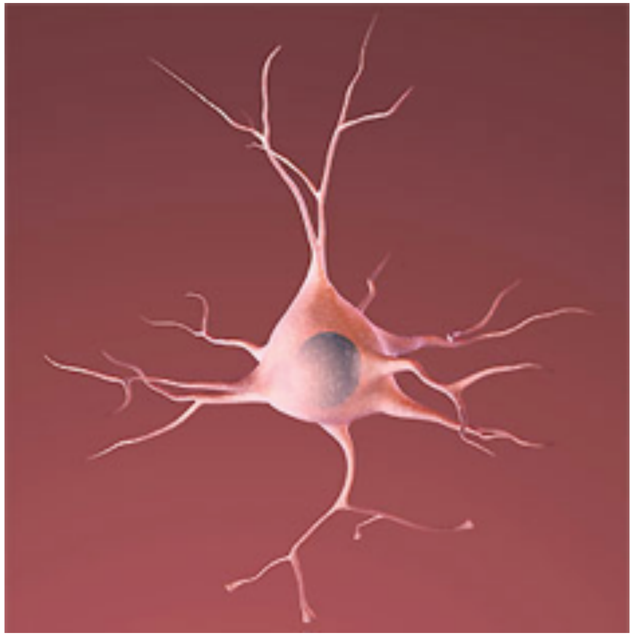




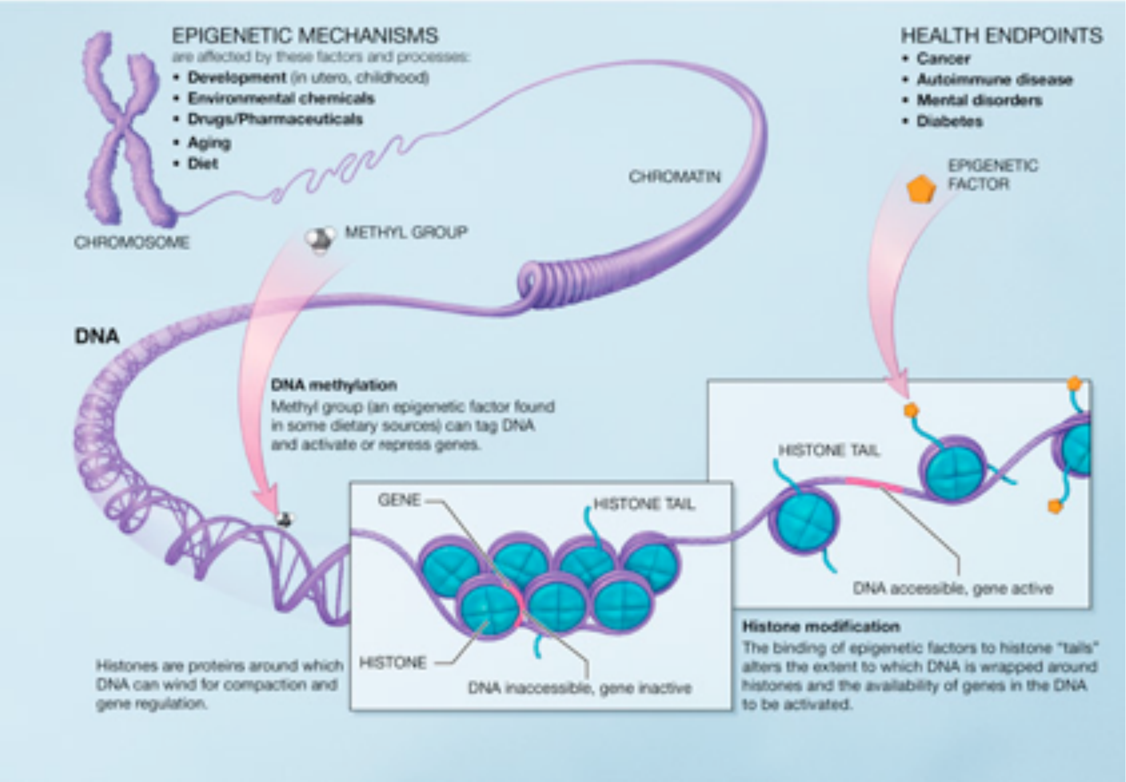
Differentiation and disease

Behavior of these genes in age and age-related disease





Computational Epigenetics



DNA methylation
Methyl group (an epigenetic factor found in some dietary sources) can tag DNA and activate or repress genes.

Histones are proteins around which DNA can wind for compaction and gene regulation.

Histone modification
The binding of epigenetic factors to histone "tails" alters the extent to which DNA is wrapped around histones and the availability of genes in the DNA to be activated.

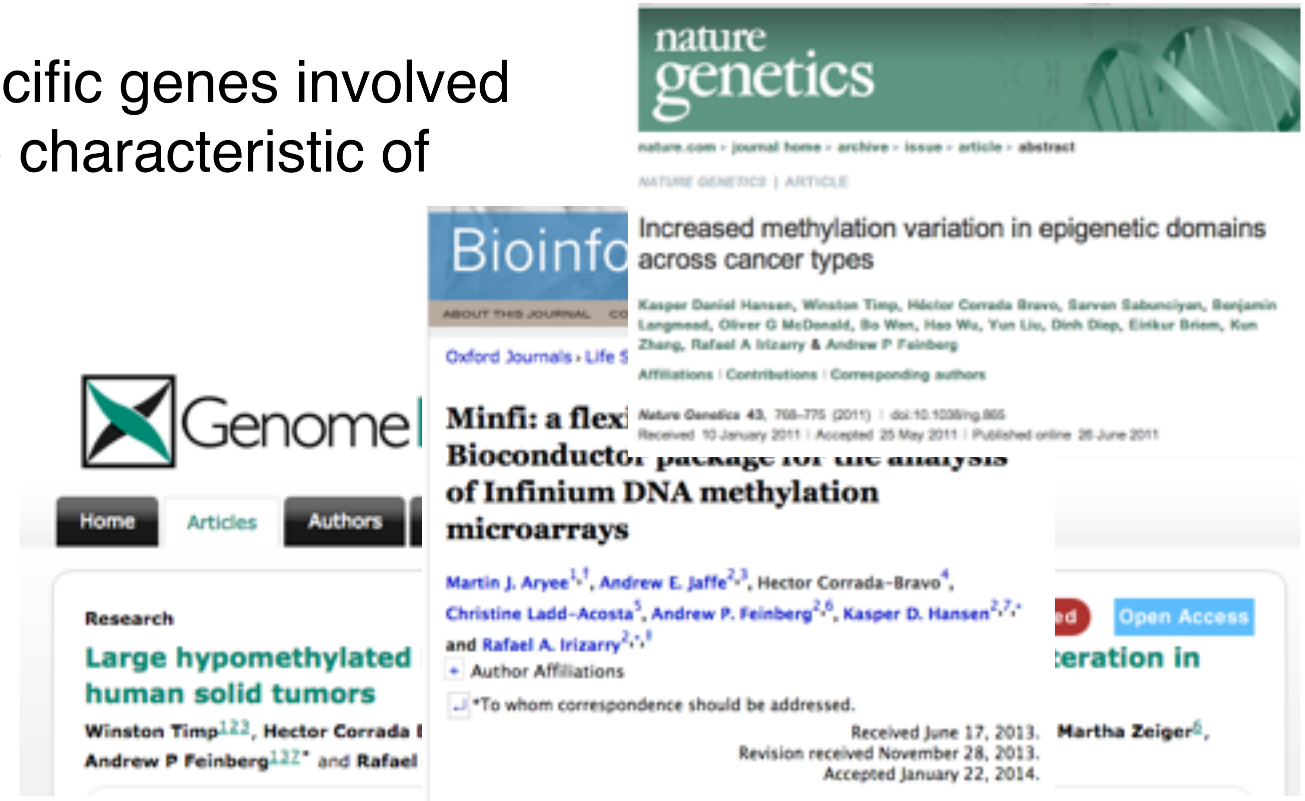
- DNA methylation
  - replicates after cell division
  - modified in **disease**

Computational Epigenetics

1. Novel computational analysis methods
2. First project to measure and characterize genome-wide in cancer

Big takeaway:

Hyper-variability of specific genes involved in *differentiation* is a stable characteristic of cancer



nature.com › journal home › archive › issue › article › abstract

NATURE GENETICS | ARTICLE

Increased methylation variation in epigenetic domains across cancer types

Kasper Daniel Hansen, Winston Timp, Héctor Corrada Bravo, Sarven Sabunciyan, Benjamin Langmead, Oliver G McDonald, Bo Wen, Hao Wu, Yun Liu, Dinh Diep, Eirikur Briem, Kun Zhang, Rafael A Irizarry & Andrew P Feinberg

Affiliations | Contributions | Corresponding authors

Nature Genetics 43, 768–775 (2011) | doi:10.1038/ng.865
Received 10 January 2011 | Accepted 25 May 2011 | Published online 26 June 2011

Oxford Journals › Life S

Minfi: a flexi Bioconductor package for the analysis of Infinium DNA methylation microarrays

Martin J. Aryee[1,*], Andrew E. Jaffe[2,3], Hector Corrada-Bravo[4], Christine Ladd-Acosta[5], Andrew P. Feinberg[2,6], Kasper D. Hansen[2,7,*] and Rafael A. Irizarry[2,*,†]

+ Author Affiliations

*To whom correspondence should be addressed.

Received June 17, 2013.
Revision received November 28, 2013.
Accepted January 22, 2014.

Research

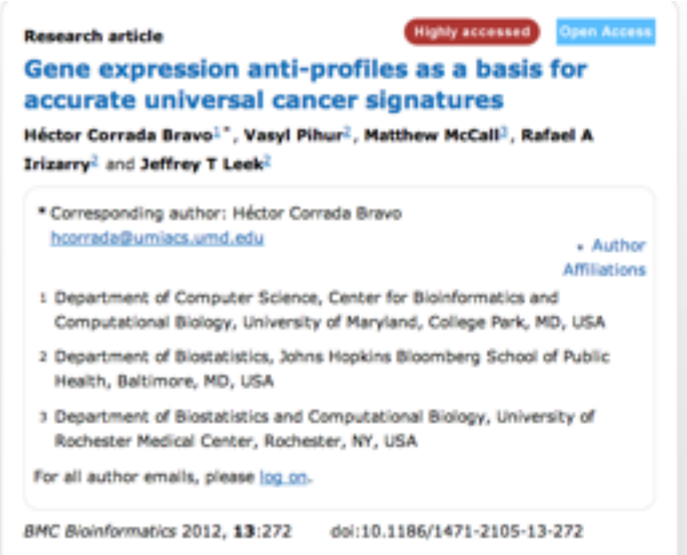
Large hypomethylated human solid tumors

teration in

Winston Timp[1,2,3], Hector Corrada Martha Zeiger[6], Andrew P Feinberg[1,2,7*] and Rafael

Anti-profiles

1. Understanding *hyper-variability*
2. How do we predict using *hyper-variability*?
   - Anomaly classification
   - Robustness of predictors



Research article

Highly accessed Open Access

Gene expression anti-profiles as a basis for accurate universal cancer signatures

Héctor Corrada Bravo[1]*, Vasyl Pihur[2], Matthew McCall[3], Rafael A Irizarry[2] and Jeffrey T Leek[2]

* Corresponding author: Héctor Corrada Bravo hcorrada@umiacs.umd.edu

+ Author Affiliations

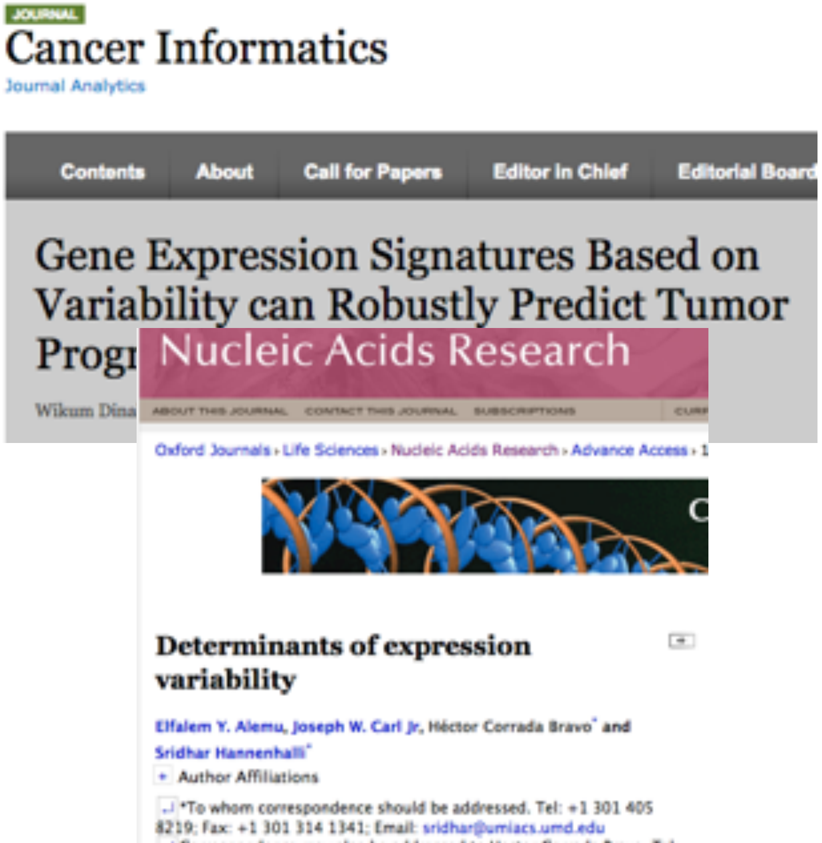
1 Department of Computer Science, Center for Bioinformatics and Computational Biology, University of Maryland, College Park, MD, USA

2 Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA

3 Department of Biostatistics and Computational Biology, University of Rochester Medical Center, Rochester, NY, USA

For all author emails, please log on.

BMC Bioinformatics 2012, **13**:272 doi:10.1186/1471-2105-13-272



JOURNAL

Cancer Informatics

Journal Analytics

Gene Expression Signatures Based on Variability can Robustly Predict Tumor Progr

Wikum Dina

Nucleic Acids Research

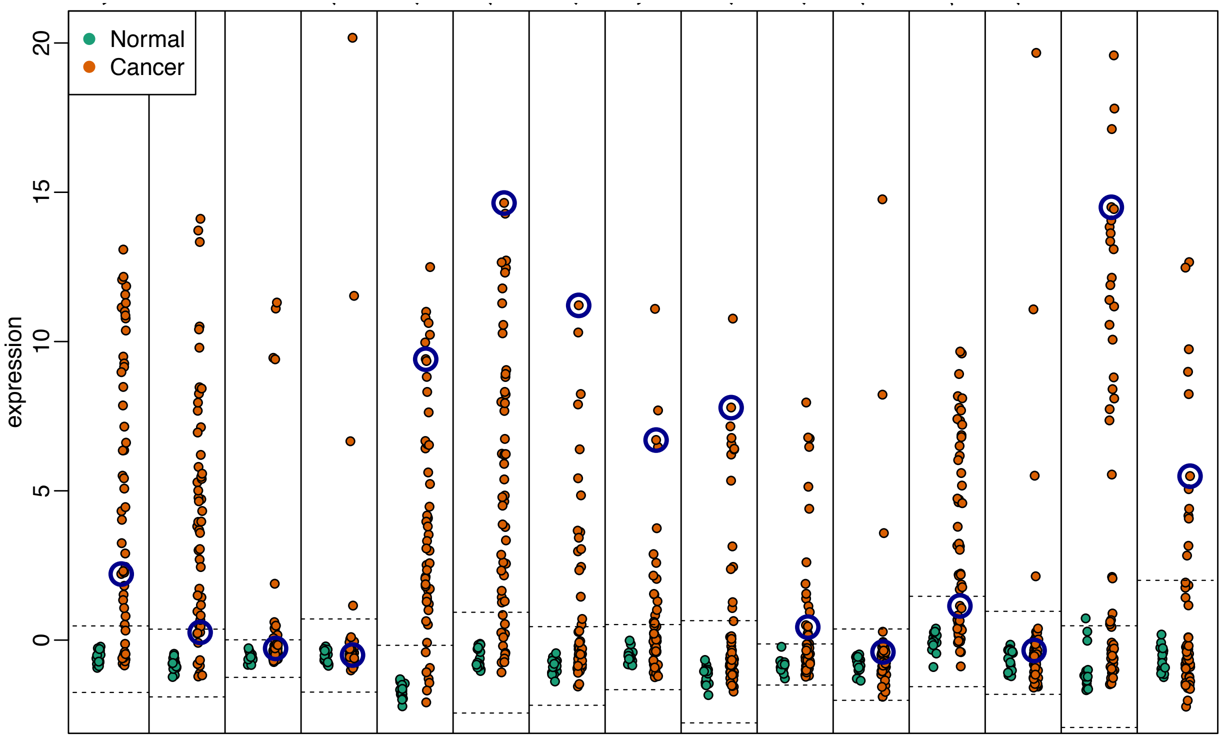
Oxford Journals › Life Sciences › Nucleic Acids Research › Advance Access › 1

Determinants of expression variability

Elfalem Y. Alemu, Joseph W. Carl Jr, Héctor Corrada Bravo* and Sridhar Hannenhalli*

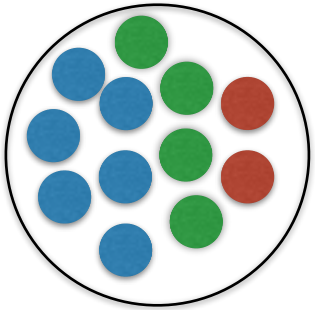
+ Author Affiliations

*To whom correspondence should be addressed. Tel: +1 301 405 8219; Fax: +1 301 314 1341; Email: sridhar@umiacs.umd.edu

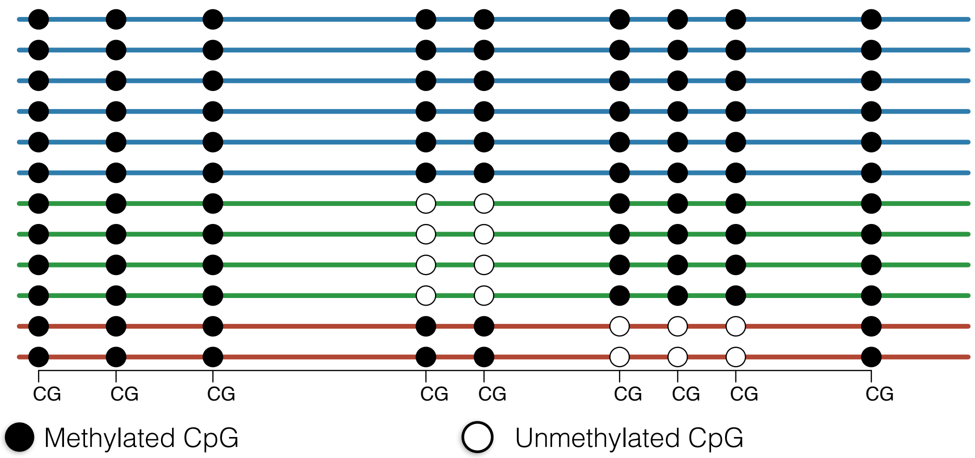


Epigenomic Heterogeneity

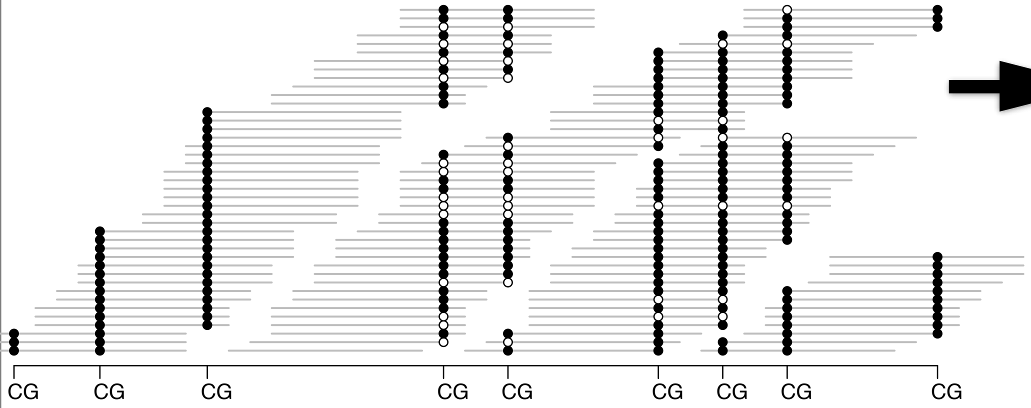
Heterogeneous
Celltype Composition



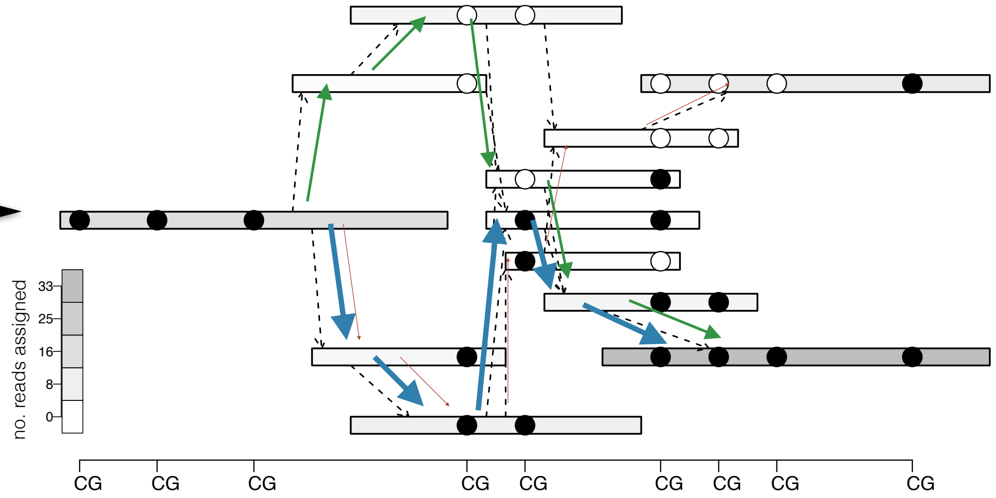
Celltype-Specific
Methylation Patterns



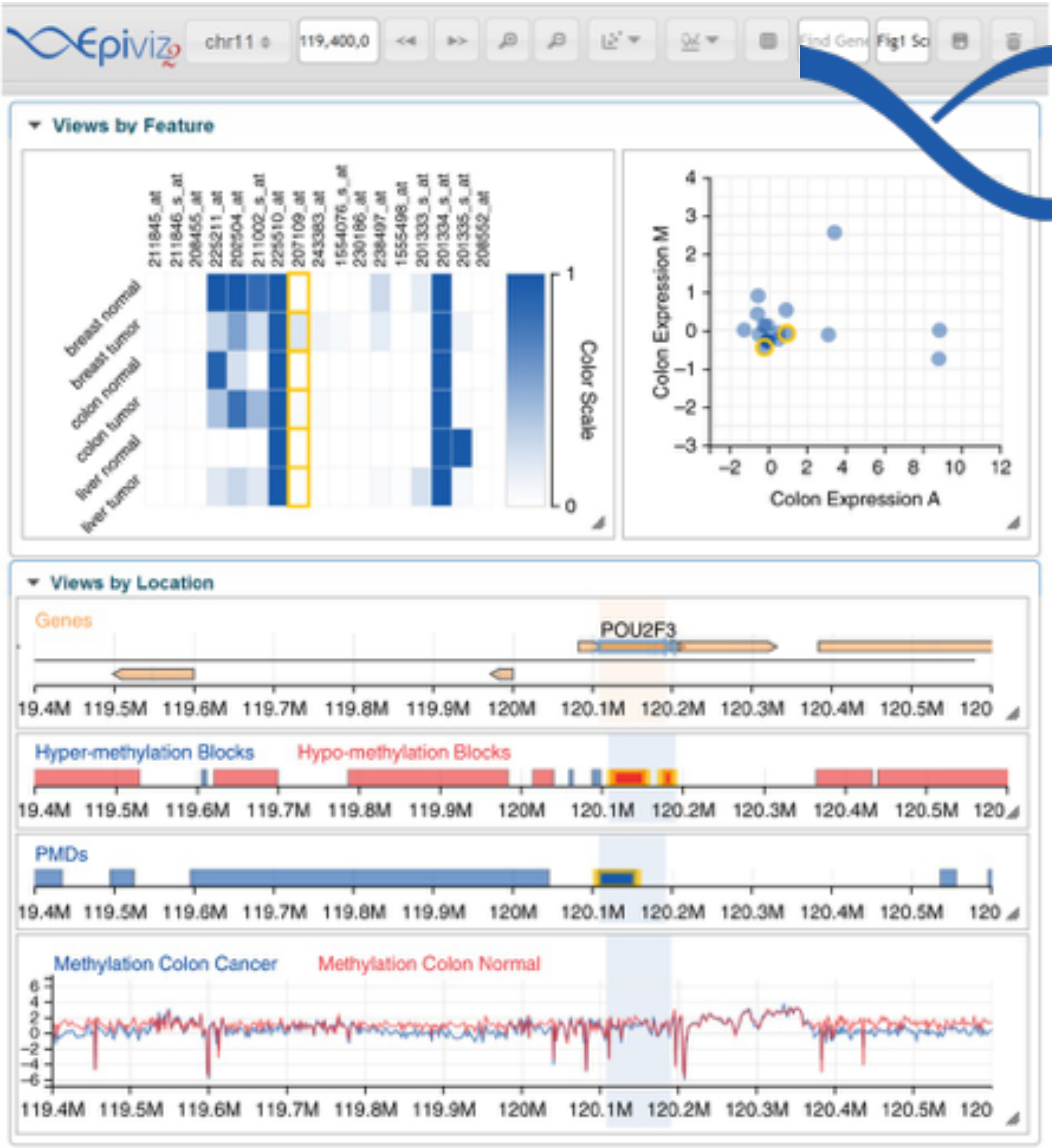
Aligned Reads



Region Graph

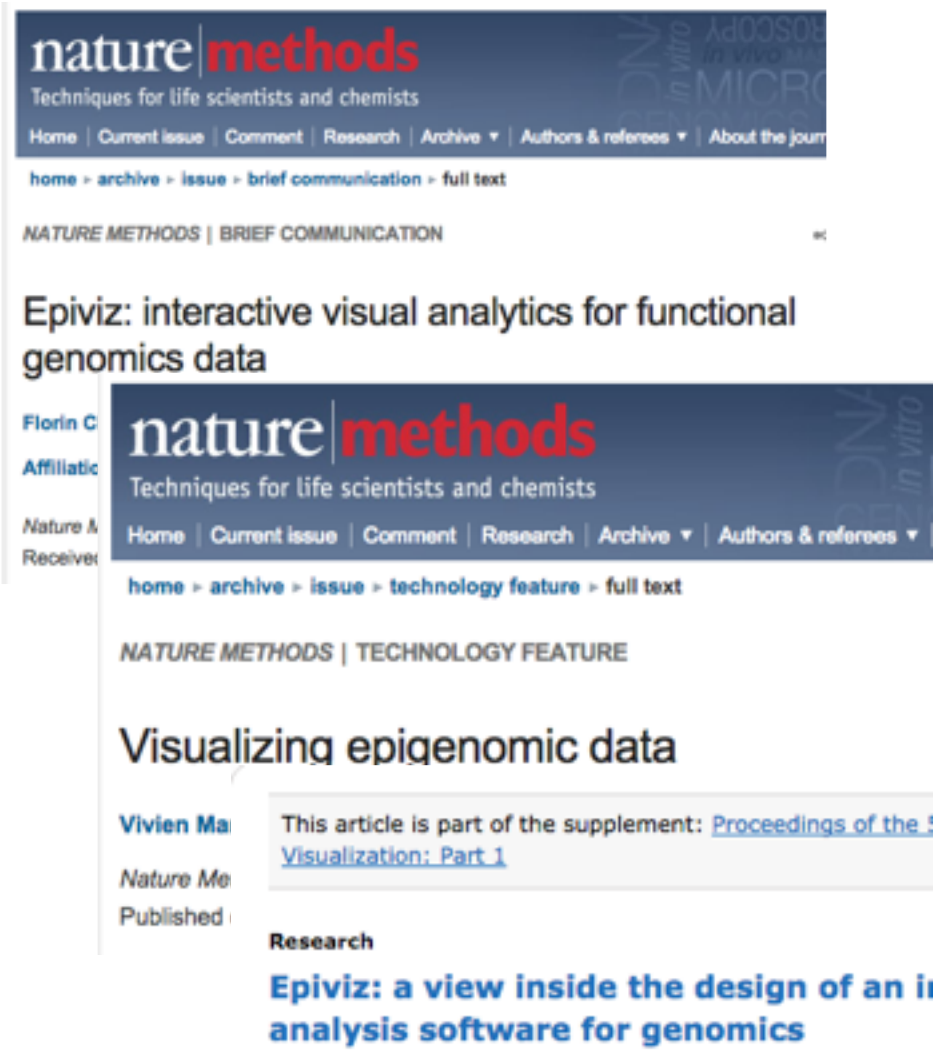


Interactive Visualization



Epiviz3

http://epiviz.cbcb.umd.edu



nature methods

Techniques for life scientists and chemists

Home | Current issue | Comment | Research | Archive ▾ | Authors & referees ▾ | About the journ

home ▸ archive ▸ issue ▸ brief communication ▸ full text

NATURE METHODS | BRIEF COMMUNICATION

Epiviz: interactive visual analytics for functional genomics data

Florin C

Affiliatic

Nature M

Received

nature methods

Techniques for life scientists and chemists

Home | Current issue | Comment | Research | Archive ▾ | Authors & referees ▾

home ▸ archive ▸ issue ▸ technology feature ▸ full text

NATURE METHODS | TECHNOLOGY FEATURE

Visualizing epigenomic data

Vivien Ma

Nature Me

Published

This article is part of the supplement: Proceedings of the
Visualization: Part 1

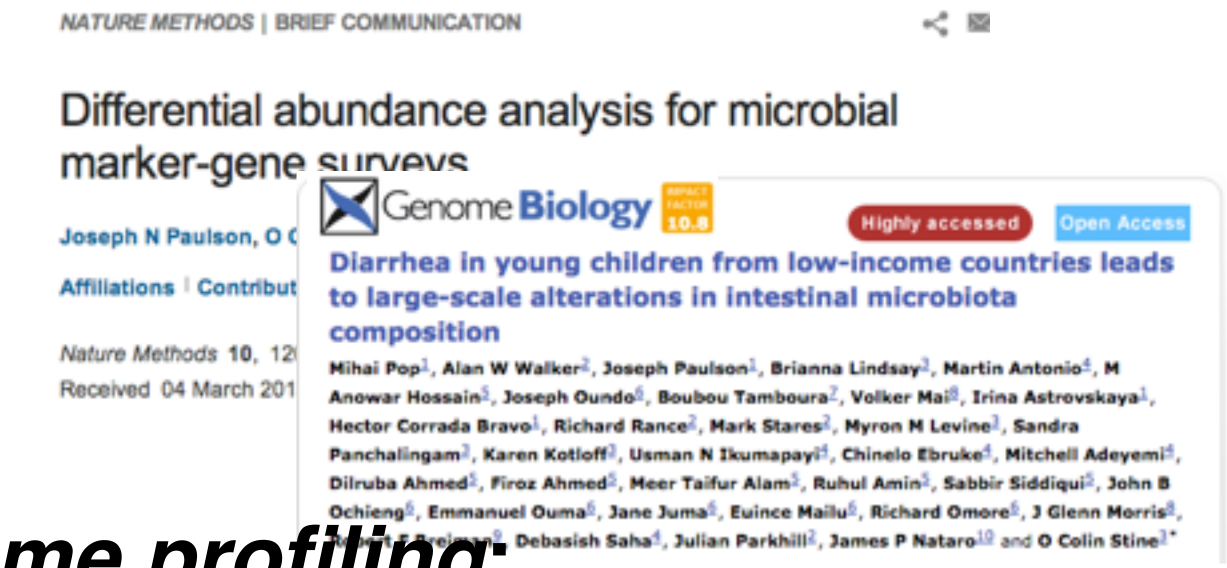
Research

Epiviz: a view inside the design of an i
analysis software for genomics

Host-Pathogen Systems

Microbiome: *environmental sequencing*

- Detecting microbes associated with disease
- Modeling and detecting differences in microbial community composition in health and disease



NATURE METHODS | BRIEF COMMUNICATION

Differential abundance analysis for microbial marker-gene surveys

Joseph N Paulson, O C

Affiliations | Contribut

Nature Methods 10, 12

Received 04 March 201

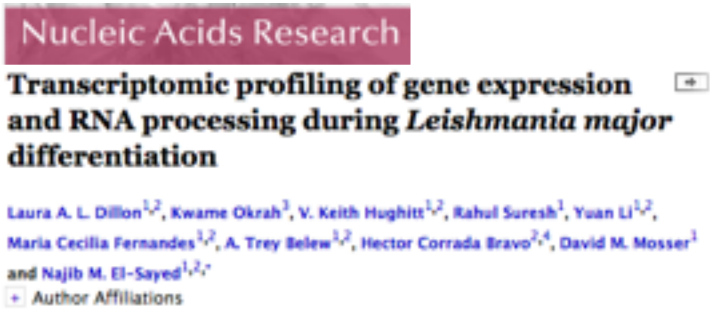
Genome Biology 10.8 Highly accessed Open Access

Diarrhea in young children from low-income countries leads to large-scale alterations in intestinal microbiota composition

Mihai Pop[1], Alan W Walker[2], Joseph Paulson[1], Brianna Lindsay[3], Martin Antonio[4], M Anowar Hossain[5], Joseph Oundo[6], Boubou Tamboura[7], Volker Mai[8], Irina Astrovskaya[1], Hector Corrada Bravo[1], Richard Rance[2], Mark Stares[2], Myron M Levine[3], Sandra Panchalingam[3], Karen Kotloff[3], Usman N Ikumapayi[4], Chinelo Ebruke[4], Mitchell Adeyemi[4], Dilruba Ahmed[5], Firoz Ahmed[5], Meer Taifur Alam[5], Ruhul Amin[5], Sabbir Siddiqui[5], John B Ochieng[6], Emmanuel Ouma[6], Jane Juma[6], Euince Mailu[6], Richard Omore[6], J Glenn Morris[8], Robert F Breiman[9], Debasish Saha[4], Julian Parkhill[2], James P Nataro[10] and O Colin Stine[3*]

***Joint transcriptome profiling*:**

Understanding host-pathogen interaction *throughout* course of infection



Nucleic Acids Research

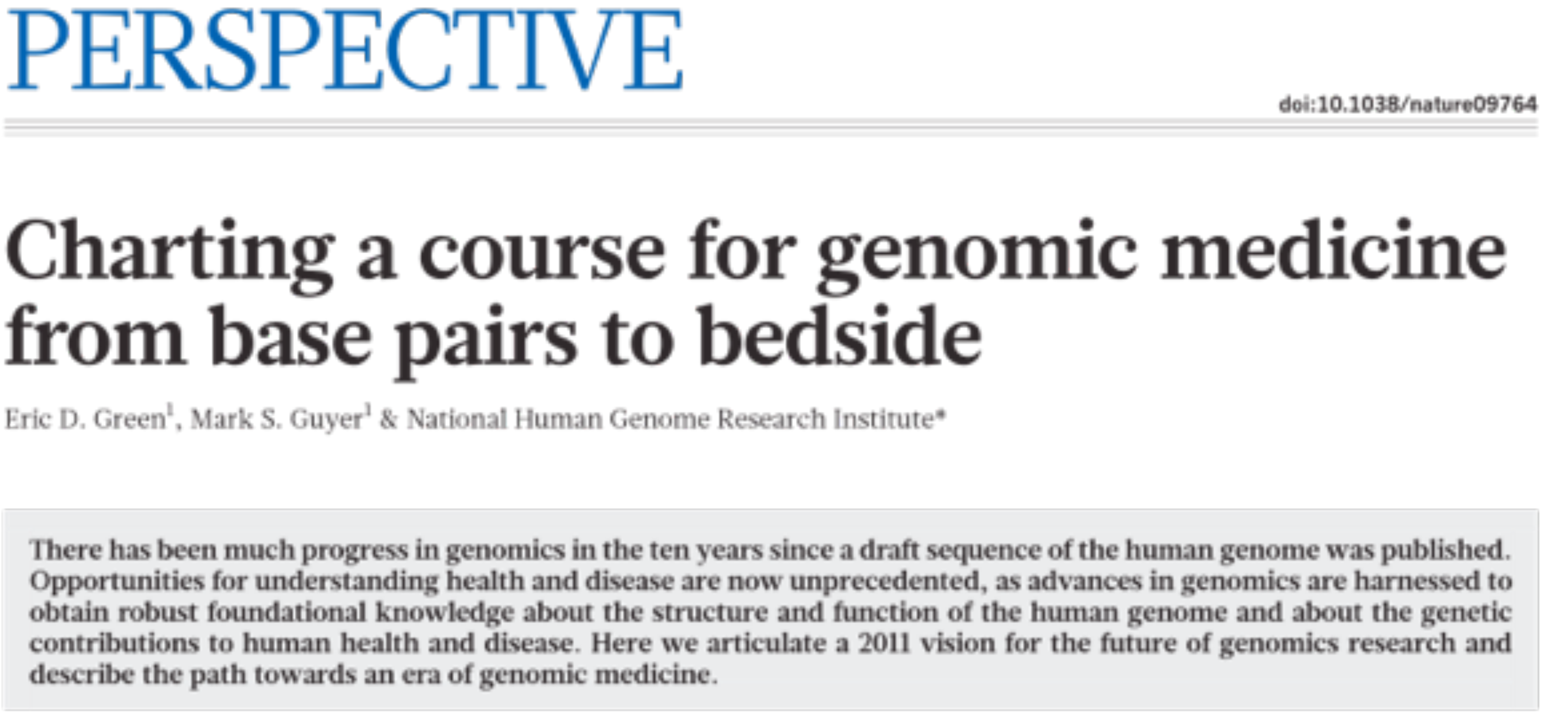
Transcriptomic profiling of gene expression and RNA processing during *Leishmania major* differentiation

Laura A. L. Dillon[1,2], Kwame Okrah[3], V. Keith Hughitt[1,2], Rahul Suresh[1], Yuan Li[1,2], Maria Cecilia Fernandes[1,2], A. Trey Belew[1,2], Hector Corrada Bravo[2,4], David M. Mosser[1] and Najib M. El-Sayed[1,2,*]

+ Author Affiliations

NHGRI strategic plan

- What does the NIH think genomics should be for the next 10 years?



PERSPECTIVE

doi:10.1038/nature09764

Charting a course for genomic medicine from base pairs to bedside

Eric D. Green[1], Mark S. Guyer[1] & National Human Genome Research Institute*

There has been much progress in genomics in the ten years since a draft sequence of the human genome was published. Opportunities for understanding health and disease are now unprecedented, as advances in genomics are harnessed to obtain robust foundational knowledge about the structure and function of the human genome and about the genetic contributions to human health and disease. Here we articulate a 2011 vision for the future of genomics research and describe the path towards an era of genomic medicine.

[Nature, Feb. 2011]

Where do we fit in?

- The major bottleneck in genome sequencing is no longer data generation—the computational challenges around data analysis, display and integration are now rate limiting. New approaches and methods are required to meet these challenges.
- **Data analysis**
  - Computational tools are quickly becoming inadequate for analysing the amount of genomic data that can now be generated, and this mismatch will worsen. Innovative approaches to analysis, involving close coupling with data production, are essential.
- **Data integration**
  - Genomics projects increasingly produce disparate data types (for example, molecular, phenotypic, environmental and clinical), so computational approaches must not only keep pace with the volume of genomic data, but also their complexity. New integrative methods for analysis and for building predictive models are needed.
- **Visualization**
  - In the past, visualizing genomic data involved indexing to the one-dimensional representation of a genome. New visualization tools will need to accommodate the multidimensional data from studies of molecular phenotypes in different cells and tissues, physiological states and developmental time. Such tools must also incorporate non-molecular data, such as phenotypes and environmental exposures. The new tools will need to accommodate the scale of the data to deliver information rapidly and efficiently.
- **Computational tools and infrastructure**
  - Generally applicable tools are needed in the form of robust, well-engineered software that meets the distinct needs of genomic and non-genomic scientists. Adequate computational infrastructure is also needed, including sufficient storage and processing capacity to accommodate and analyse large, complex data sets (including metadata) deposited in stable and accessible repositories, and to provide consolidated views of many data types, all within a framework that addresses privacy concerns. Ideally, multiple solutions should be developed105.

Where do we fit in?

- Meeting the computational challenges for genomics requires scientists with expertise in biology as well as in informatics, computer science, mathematics, statistics and/or engineering.

- A new generation of investigators who are proficient in two or more of these fields must be trained and supported.

Education and Training

- PhDs: Computer Science, Applied Statistics, Scientific Computation, Computational Biology
  - MIT/Broad, Harvard/Dana Farber Cancer Center, U. of Chicago, Genentech, Johns Hopkins Medicine, Dow Jones Data Science
- New UG courses in Bioinformatics and Data Science