Perspectives on the Priorities for Bioinformatics Education in the 21st Century

Oyekanmi Nash, PhD
Associate Professor & Director
Genetics, Genomics & Bioinformatics
National Biotechnology Development Agency, NABDA/FMST
Abuja. Nigeria
Oyekan.nash@gmail.com

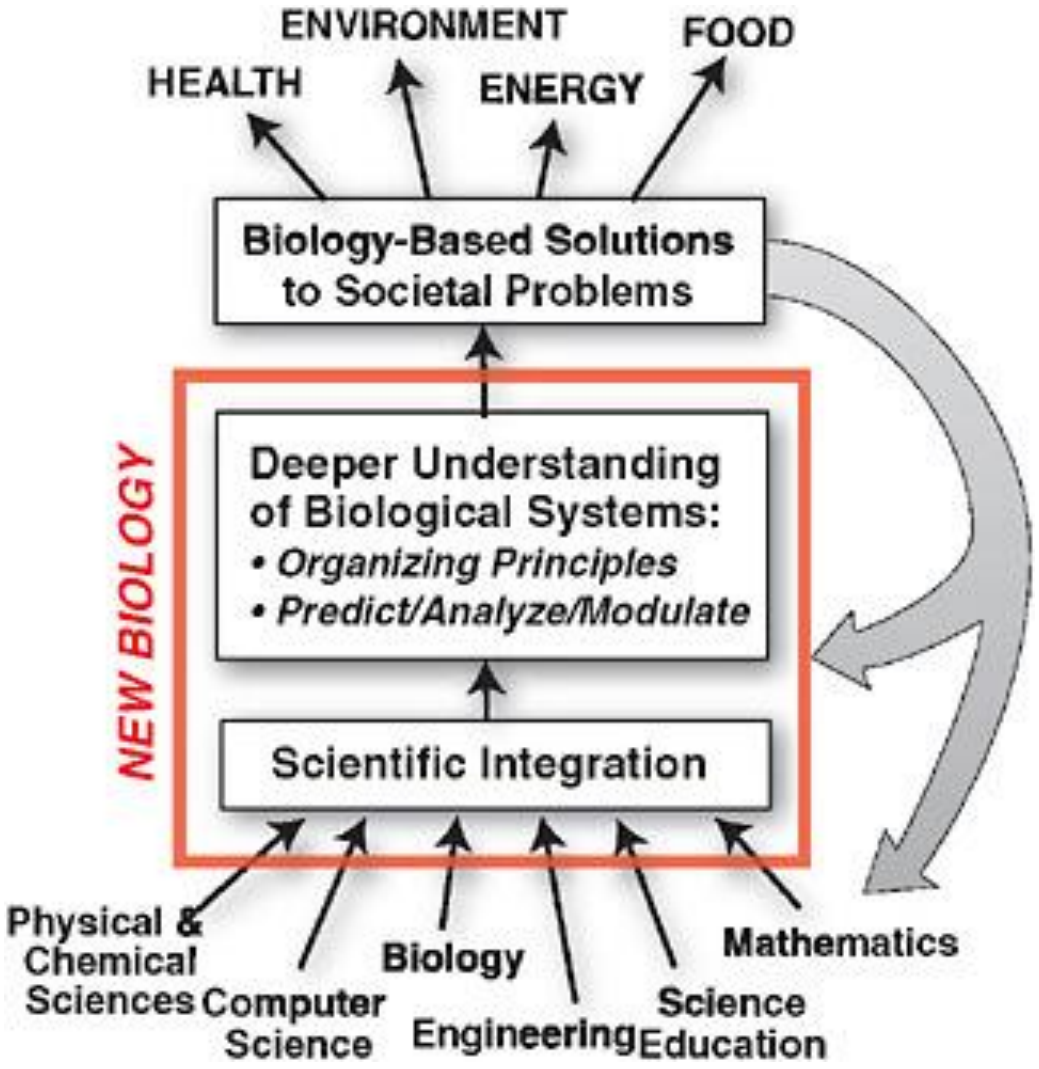
Co-Author: Raphael D. Isokpehi, PhD
Associate Professor (Microbiology & Bioinformatics)
Bethune-Cookman University, Daytona Beach, Florida, United States

Bioinformatics Curriculum Development Workshop
University of Botswana, Gaborone, Botswana
11-12 March, 2014

Outline

- 21st Century Biology
- Bioinformatics Definitions
- Scope of Bioinformatics
- Big Data Challenges in Biology
- Challenges in Data Science
- Addressing these challenges in Training Students to Master Bioinformatics
    - Visual Analytics
- Conclusion
- Acknowledgements

NEW BIOLOGY IN THE 21ST CENTURY



What is the New Biology?
SOURCE: Committee on a New Biology for the 21st Century.
http://books.nap.edu/openbook.php?record_id=12764&page=18

Definitions

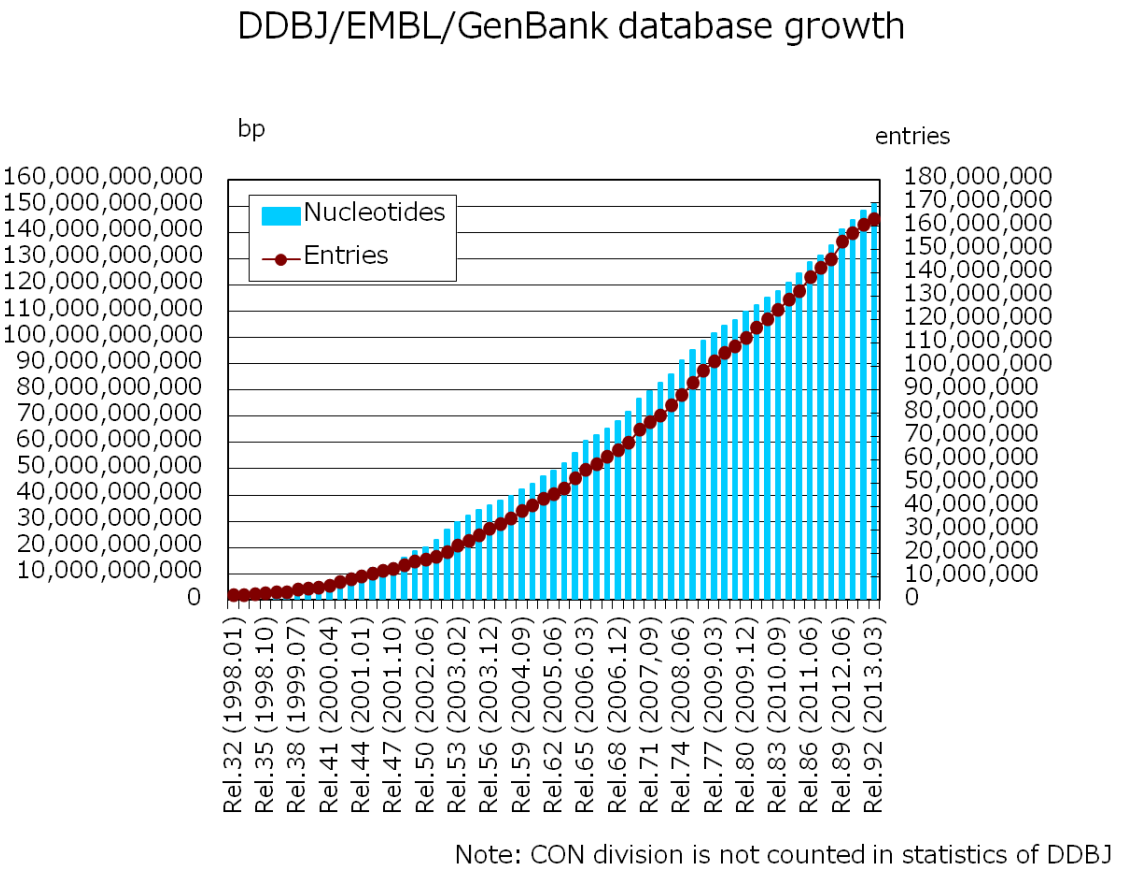
Bioinformatics:

- Research, development, or application of
  - computational tools and approaches for
  - expanding the use of biological, medical, or behavioral or health data including those to
  - acquire, store, organize, archive, analyze, or visualize such data.

.

Focus of Bioinformatics – Large Datasets

- Analyses in Bioinformatics predominantly focus on three types of large datasets available in molecular biology:
  - macromolecular structures,
  - genome sequences, and
  - the results of functional genomics experiments (e.g. expression data).



http://www.ddbj.nig.ac.jp/breakdown_stats/dbgrowth-e.html#dbgrowth-graph

What is bioinformatics? A proposed definition and overview of the field.
http://www.ncbi.nlm.nih.gov/pubmed/11552348

Focus of Bioinformatics - Techniques

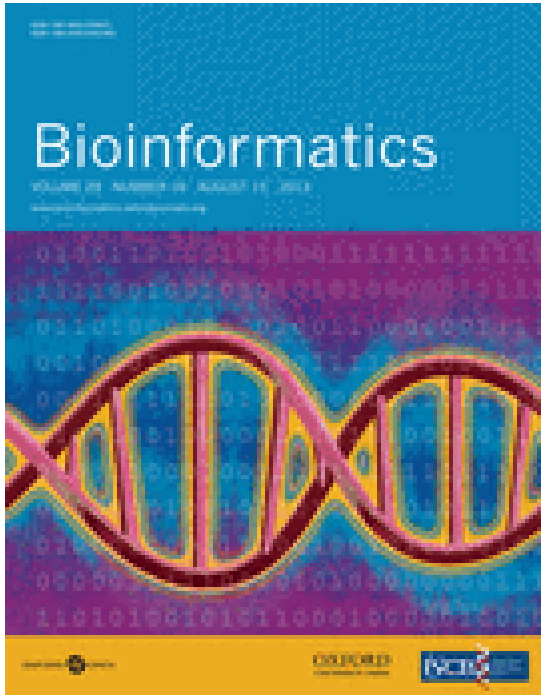
- Bioinformatics employs a wide range of computational techniques including
  - sequence and structural alignment,
  - database design and data mining,
  - macromolecular geometry,
  - phylogenetic tree construction,
  - prediction of protein structure and function,
  - gene finding, and
  - expression data clustering.
- The emphasis is on **approaches integrating a variety of computational methods and heterogeneous data sources.**
- Finally, bioinformatics is a **practical discipline**.

What is bioinformatics? A proposed definition and overview of the field.
http://www.ncbi.nlm.nih.gov/pubmed/11552348

BIOINFORMATICS RESEARCH CATEGORIES

BROAD CATEGORIES OF BIOINFORMATICS RESEARCH AND DEVELOPMENT

- Genome Analysis
- Sequence Analysis
- Phylogenetics
- Structural Bioinformatics
- Gene Expression
- **Genetic and Population Analysis**
- Systems Biology
- Data and Text Mining
- Databases and Ontologies
- Bioimage Informatics



http://www.oxfordjournals.org/our_journals/bioinformatics/for_authors/scope_guidelines.html

Making Discoveries from the Massive and Complex Genomics Datasets and Bioinformatics Results from H3Africa Projects

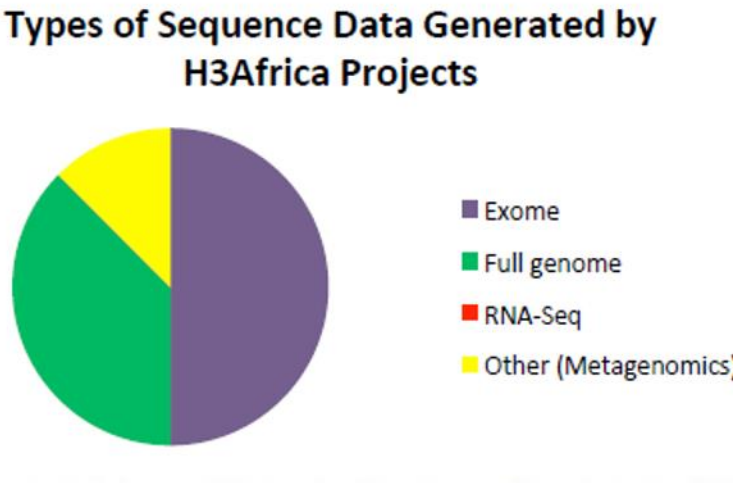
"The major bottleneck in genome sequencing is no longer data generation—the computational challenges around data analysis, display and integration are now rate limiting. New approaches and methods are required to meet these challenges".

National Human Genome Research Institute Strategic Plan:
Charting a course for genomic medicine from base pairs to bedside
http://www.genome.gov/Pages/About/Planning/2011NHGRIStrategicPlan.pdf

Examples of Projected Massive and Complex Datasets from H3Africa Projects (2013….

Type 2 Diabetes Project

- **12,000 Cases and 12,000 Controls**
- **Sequencing of known T2DM regions**
- **Genome-wide genotyping arrays**
- **Whole exome/genome sequencing**



Body Composition Project

- **African genome structure**
- **Phenotyping and sampling for Cohorts**
- **Genetic and environmental contribution to body composition (~12,000 individuals)**

These research investigations rely significantly on bioinformatics analysis and inferences from large and heterogeneous datasets obtained from populations inside and outside Africa.

Challenges in Data Science

Source: National Consortium for Data Science, USA

VISUAL ANALYTICS

ANALYTICAL REASONING

VISUAL REPRESENTATION & INTERACTIONS

DATA REPRESENTATIONS AND TRANSFORMATIONS

PRODUCTION, PRESENTATION & DISSEMINATION

Ability to Use Tools for Making Sense of Data for Biology

90%

Data Experts
Statistics
Bioinformatics
Informatics
Databases

10%

Subject Matter Experts
Genomics
Proteomics
Metabolomics

Inability of Domain Experts to make sense of Massive data is a major challenge for advancing translational research

WHAT IS VISUAL ANALYTICS?

"Visual analytics is the representation and presentation of data that exploits our visual perception abilities in order to amplify cognition."

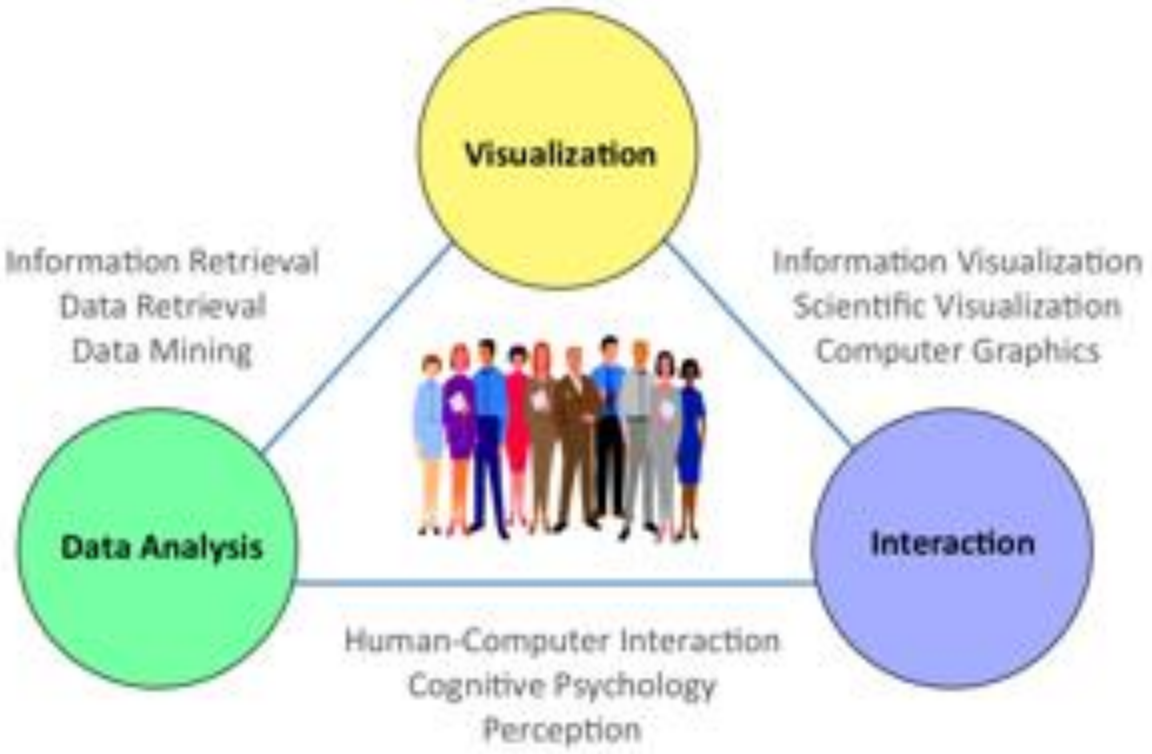
- Andy Kirk, author of "Data Visualization: a successful design process"

http://www.slideshare.net/TableauSoftware/visual-analytics-best-practices

WHAT IS VISUAL ANALYTICS?

- Multidisciplinary field defined as the science of analytical reasoning facilitated by interactive visual interfaces.

- An integrated approach combining fields such as visualization, human factors and data analysis which in turn integrates different methodologies as shown below:



http://www.visual-analytics.eu/faq//

Process of Visual Analytics

- Visual analytics is an iterative process that involves information gathering, data preprocessing, knowledge representation, interaction and decision making.

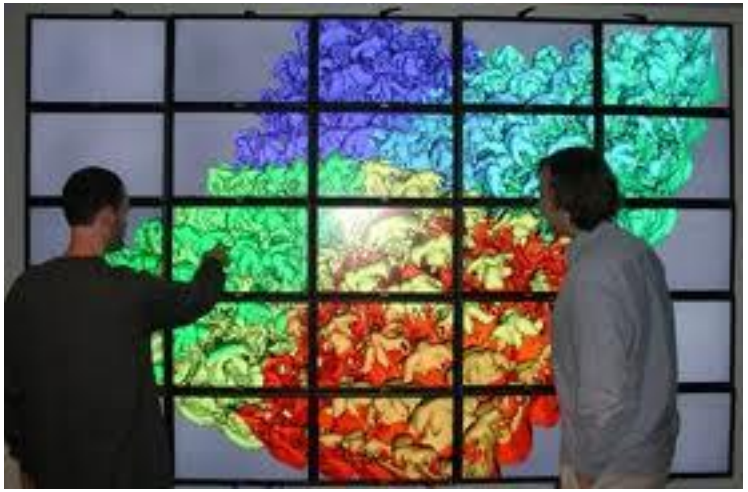
- The ultimate goal is to gain insight in the problem at hand which is described by vast amounts of scientific, forensic or business data from heterogeneous sources.

- To reach this goal, visual analytics combines the strengths of machines with those of humans.

- On the one hand, methods from knowledge discovery in databases (KDD), statistics and mathematics are the driving force on the automatic analysis side,
- while on the other hand human capabilities to perceive, relate and conclude turn visual analytics into a very promising field of research.

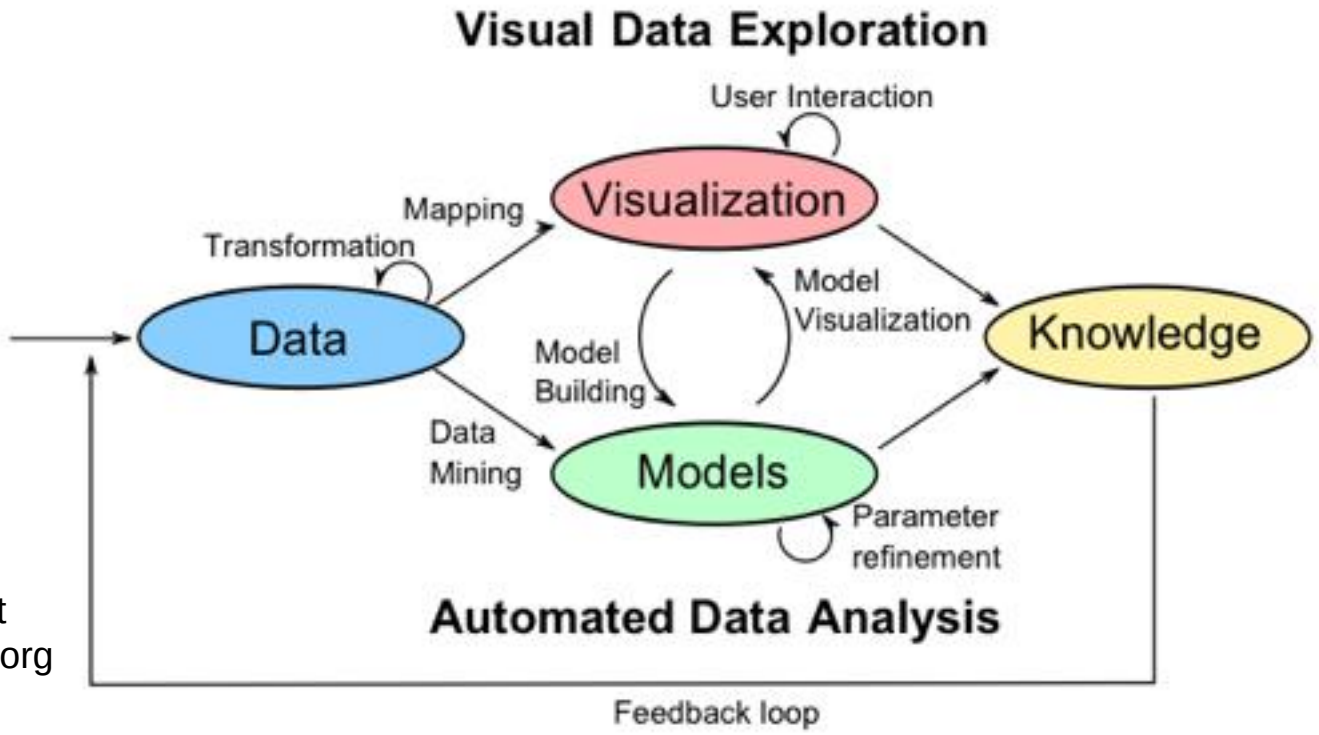
Source: Keim et al, 2008. Visual Analytics: Scope and Challenges Visual Data Mining, LNCS 4404, pp. 76–90, 2008

VISUAL INTERFACES



VISUAL ANALYTICS FOR MAKING SENSE OF "BIG DATA"

Visualizations are absolutely critical to our ability to process complex data and to build better intuitions as to what is happening around us. – Fox P and Hendler J. (2011) Changing the Equation on Scientific Data Visualization. Science 331: 705-708.

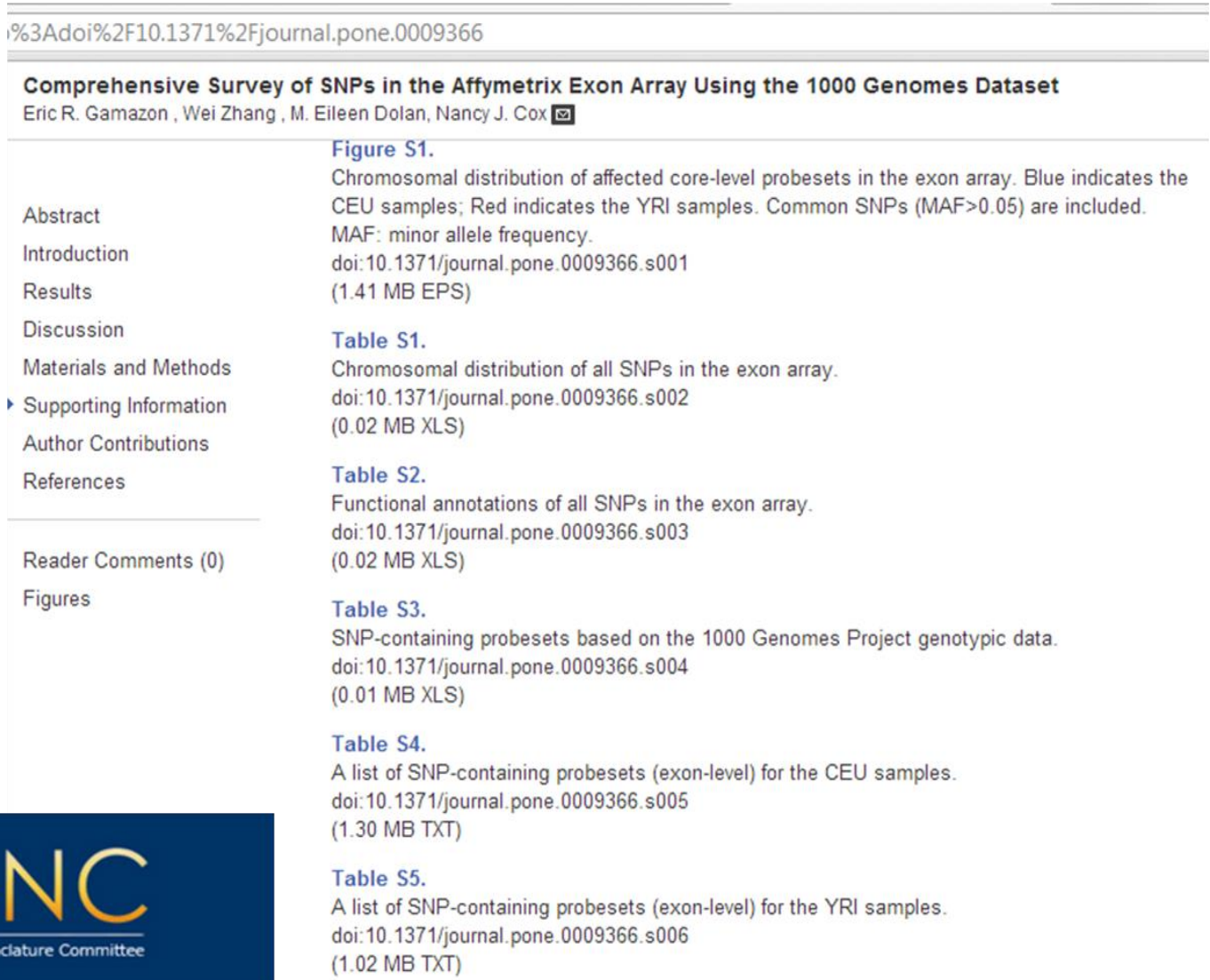


More Information at
www.vacommunity.org

http://www.vismaster.eu/faq/the-visual-analytics-process/

MOTIVATION FOR VISUAL ANALYTICS IN BIOMEDICAL RESEARCH

Availability in the public domain of datasets on human genomic variation that can be downloaded (as spreadsheet or text files) from web-based bioinformatics resources and supporting information of journal articles.



%3Adoi%2F10.1371%2Fjournal.pone.0009366

Comprehensive Survey of SNPs in the Affymetrix Exon Array Using the 1000 Genomes Dataset

Eric R. Gamazon , Wei Zhang , M. Eileen Dolan, Nancy J. Cox

Abstract
Introduction
Results
Discussion
Materials and Methods
Supporting Information
Author Contributions
References

Reader Comments (0)
Figures

Figure S1.
Chromosomal distribution of affected core-level probesets in the exon array. Blue indicates the CEU samples; Red indicates the YRI samples. Common SNPs (MAF>0.05) are included. MAF: minor allele frequency.
doi:10.1371/journal.pone.0009366.s001
(1.41 MB EPS)

Table S1.
Chromosomal distribution of all SNPs in the exon array.
doi:10.1371/journal.pone.0009366.s002
(0.02 MB XLS)

Table S2.
Functional annotations of all SNPs in the exon array.
doi:10.1371/journal.pone.0009366.s003
(0.02 MB XLS)

Table S3.
SNP-containing probesets based on the 1000 Genomes Project genotypic data.
doi:10.1371/journal.pone.0009366.s004
(0.01 MB XLS)

Table S4.
A list of SNP-containing probesets (exon-level) for the CEU samples.
doi:10.1371/journal.pone.0009366.s005
(1.30 MB TXT)

Table S5.
A list of SNP-containing probesets (exon-level) for the YRI samples.
doi:10.1371/journal.pone.0009366.s006
(1.02 MB TXT)

MOTIVATION FOR VISUAL ANALYTICS IN BIOMEDICAL RESEARCH

"Future, rapid, and transdisciplinary research advances will depend on our ability to harness bioinformatics and the resulting data that come from computational biology.

The challenge will be finding creative and more efficient ways to analyze, store, disseminate, and share data—both new and from older studies—widely and effectively.

This will require transdisciplinary and other researchers to create standardized ontologies and nomenclatures, harmonize data systems, and increase access to shared databases that also **provide innovative analytic tools."**

CONDUCT OF SCIENCE

The Scientific Vision of the National Institute of Child Health and Human Development (NICHD) for the next 10 years includes a section on the Conduct of Science.
NICHD Scientific Vision Next Decade.
https://www.nichd.nih.gov/publications/pubs/Documents/NICHD_scientific_vision120412.pdf

CONCLUSIONS – CURRICULUM FOR MASTER OF SCIENCE IN BIOINFORMATICS

H3ABioNet NABDA/FMST (Nigeria) Node

Year 1 Highlights – Education and Training

- 5 Training Workshops:
  - Video Conferencing for Bioinformatics Research & Collaboration (Nov. 2012)
  - Sequence Analysis and Bioinformatics Curriculum Development in Nigeria (July 2013)
  - Visual Analytics for Human Variation Datasets (July 2013)
  - Molecular Techniques (August 2013)
  - Pre-Conference Workshop for Biotechnology Society of Nigeria (August 2013)

H3ABioNet NABDA Node Training Course on Visual Analytics of Human Genome Variation Datasets 29th of July to 2nd of August 2013 National Biotechnology Development Agency (NABDA), Abuja, Nigeria

26th Annual International Conference of
Biotechnology Society of Nigeria
(26-30 August 2013)
Pre-Conference Workshop @ NABDA, Abuja, Nigeria

NextGen Molecular Biotechnology Workshop
National Biotechnology Development Agency
Abuja, Nigeria; 10-17 August 2013

H3ABioNet NABDA/FMST (Nigeria) Node

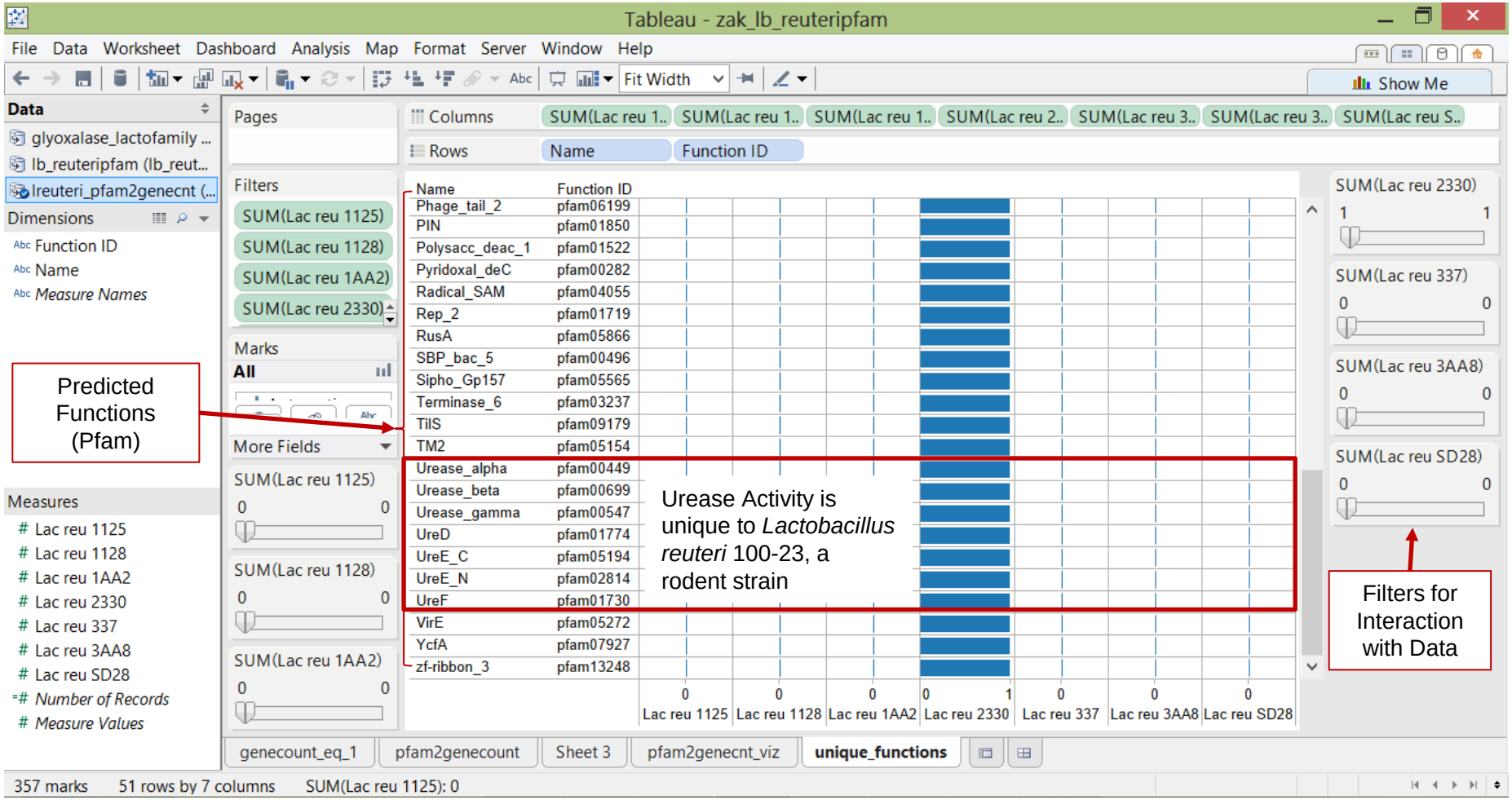
Year 1 Highlights – Outreach

Established Nigerian Bioinformatics Research and Education Network

H3ABioNet NABDA/FMST (Nigeria) Node

Year 1 Highlights – Bioinformatics Services

VISUAL DISCOVERY COMPARISON OF FUNCTIONS ENCODED IN LACTOBACILLUS REUTERI GENOMES



H3ABioNet NABDA/FMST (Nigeria) Node

Year 1 Highlights – Research Project (Probiotics Features)

- Protein secretion is important in biotechnological applications. Recombinant Proteins and Biopharmaceutical Proteins are of major importance in the biotechnology industry.
- A visual analytical integration of data sources of genes annotated for **enzymes, transmembrane and signal peptide functions** was designed.
- Additionally, the chromosomal alignment on the Integrated Microbial Genomes (IMG) system were determined for the genes of interest.
- The genome of *Lactobacillus casei* Lc-10 was the used for the research.

Lactobacillus casei Lc-10

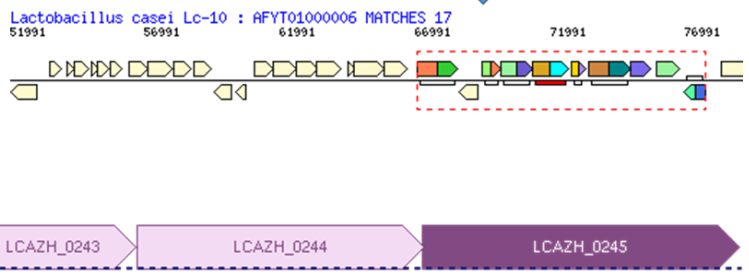
Genes total number	2888	100.00%
Protein coding genes	2813	97.40%
Protein coding genes with enzymes	687	23.79%
Protein coding genes coding signal peptides	95	3.29%
Protein coding genes coding transmembrane proteins	804	27.84%

Integrate Gene Lists with Visual Analytics Software

Gene Sets

Enzymes with:

1. Signal Peptide (SP)
2. Transmembrane Domain (TD)
3. SP + TD



- **Pipeline predicted gene cluster with gene for carboxypeptidases**
- **Carboxypeptidases are involved in the breakdown and reorganization of peptidoglycan**
- **Prediction: Biomolecular Network for Adapting to bile and acid stress by probiotic organisms.**

Master of Science in Bioinformatics
30 Credit-Hour Courses

Open to students from Life Sciences and Medicine as well as Physical Sciences and Engineering

- **Core Courses (18 Credit Hours):**

1. Advances in Bioinformatics (3 Credits)
2. Computing and Informatics Foundations for Bioinformatics (3 Credits)
    1. **Computational Thinking, Visual Literacy, Human-Computer Interactions and Visual Analytics**
3. Biological Databases and Bioinformatics Tools (3 Credits)
4. Research Methods in Bioinformatics (3 Credit)
5. Supervised Research (6 Credits)

Priorities for 21st Century Bioinformatics
Need to Gain Deeper Understanding of Biological Systems

Master of Science in Bioinformatics

- **Electives (12 Credit Hours)**

- 4 courses related to project from 10 bioinformatics categories:
- Bioimage Informatics; Data and Text Mining; Databases and Ontologies; Gene Expression; Genetic and Population Analysis; Genome Analysis; Phylogenetics; Sequence Analysis; Structural Bioinformatics; Systems Biology

Acknowledgments

- H3Africa Bioinformatics Network (H3ABioNet) -- National Institutes of Health (U41HG006941)
- National Biotechnology Development Agency. NABDA/FMST, Abuja, Nigeria
- Visual Analytics in Biology Curriculum Network