

THE WESTERN NEW YORK GENETICS IN RESEARCH PARTNERSHIP

Expanding Exposure, Career Exploration
and Interactive Projects in Basic Genome
Analysis and Bioinformatics

Stephen Koury, P.I.



Jacobs School of Medicine
and Biomedical Sciences

University at Buffalo

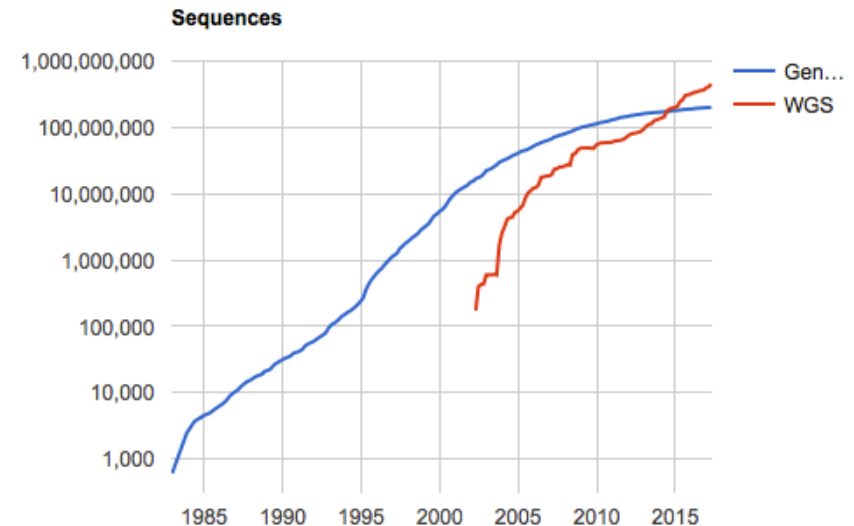
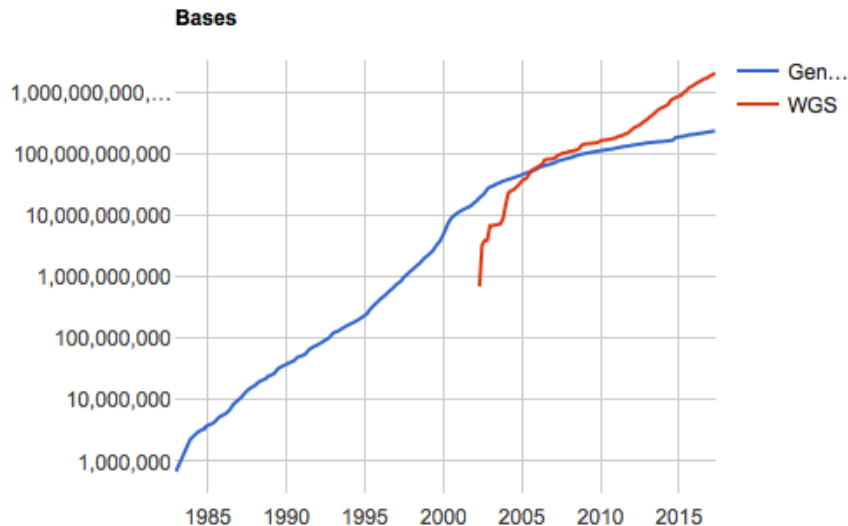


How is our project related to big data?

The size of the Genbank database of genomic information has increased dramatically over the last 35 years.

<https://www.ncbi.nlm.nih.gov/genbank/statistics/>

GenBank and WGS Statistics



Making sense of all of the genomic data

“....because genomics poses unique challenges in terms of data acquisition, distribution, storage, and especially analysis, waiting for innovations from outside our field is unlikely to be sufficient. We must face these challenges ourselves, starting with integrating data science into graduate, undergraduate, and high-school curricula to train the next generations of quantitative biologists, bioinformaticians, and computer scientists and engineers”

Taken from: Big Data: Astronomical or Genomical? Zachary D. Stephens, Skylar Y. Lee, Faraz Faghri, Roy H. Campbell, Chengxiang Zhai, Miles J. Efron, Ravishankar Iyer, Michael C. Schatz, Saurabh Sinha, Gene E. Robinson Published: July 7, 2015 <https://doi.org/10.1371/journal.pbio.1002195>

The STEM-related goals of the project are to:

- allow high school students and teachers to participate in scientific research
- stimulate the interest of students in pursuing careers in science and technology through the use of a unique, interactive learning environment combined with intensive support intervention (GENI-ACT)
- encourage teachers to include bioinformatics and genomics in their curriculum.
- Project website:
<http://ubwp.buffalo.edu/wnygirp>



ITEST Project Overview:

- A week long summer training workshop for teachers (30 per year) to learn the fundamentals of gene annotation using the GENI-ACT system (along with 3 additional subsequent refresher training days).
- 3 fall semester activities to build interest among student participants and to recruit students to participate in the spring semester genome annotation exercises.
- Spring semester teacher guided genome annotation projects for 5 or more students per teacher (150 total per year). A control group of students is included for comparison.
- A capstone symposium for student and teacher participants to present the results of their genome annotations
- 4th Year No-Cost Extension: Teachers with one year of training were supported for a second year to set up their own GENI-ACT courses to use independently with students going forward.

What is GENI-ACT?

[] [geni-act](#)

GENI-ACT allows collaborative genome annotation. Researchers or students can collectively suggest changes to an existing genome with supporting evidence. Changes can be ported back to genbank by exporting to a sequin file format.

GENI-ACT also has ported the education components from IMG-ACT. Teachers can assign students work to be completed in a lab notebook that is integrated with the classroom.

2775 genomes available

Sign Up

Login

Browse Genomes

<http://geni-act.org/genomes/browse/>

Available Modules

Basic Information

- **DNA Coordinates:** Use a [GENE Page](#) to enter coordinates
- **Nucleotide Sequence / Length:** Use a [GENE Page](#) to find the Nucleotide Sequence and Length
- **Protein Sequence / Length:** Use a [GENE Page](#) to find the Protein Sequence and Length

Sequence-based Similarity Data

- **BLAST:** Find the top hits in BLAST using NCBI BLAST
- **CCD:** Find COG Results from NCBI BLAST
- **T-Coffee:** Run a multiple sequence alignment using T-Coffee
- **WebLogo:** Find the sequence logo and analyze it

Cellular Localization Data

- **Gram Stain:** Research Pubmed to find the gram stain of the organism
- **TMHMM:** Plot the transmembrane topology
- **SignalP:** Plot the signal peptide graph
- **LipoP:** Predict lipoproteins and signal peptides
- **PSORT-B:** Predict protein localization
- **Phobius:** Plot the phobius posterior probabilities

Alternative Open Reading Frame

- **DNA Coordinates:** Change the DNA Coordinates if the ORF has been incorrectly determined



Structure-based Evidence

- **TIGRFAM:** Search TIGRFAM Hidden Markov Model database for hits
- **Pfam:** Find Pfam families within a sequence
- **PDB:** Align sequences using Protein Data Bank

Enzymatic Function

- **KEGG:** Find the KEGG Pathway and analyze the pathway map
- **MetaCyc:** Find the Metabolic Pathway and analyze
- **EC Number:** Use ExpASY ENZYME to locate the EC number and name

Duplication and Degradation

- **Paralogs:** Use NCBI BLAST to find paralogs
- **Pseudogene:** Research to find if the gene is a pseudogene

Horizontal Gene Transfer

- **Phylogenetic Tree:** Use Phylogeny.fr to plot the phylogenetic tree from sequences
- **Gene Context:** Find the Ortholog Neighborhood Region using IMG
- **Chromosome Viewer GC Heat Map:** Analyze the GC heat map to find the characteristic and average GC%

RNA

- **Rfam:** Find the Rfam number, score and pairwise alignment

Proposed Annotation

- **Note:** Propose a new annotation to be placed in the note section of GENI-ACT and Genbank GBK



2017 Capstone Symposium – 48 ITEST Poster Presentations from 26 Different Schools

