

MAKING MATH COUNT IN BIOMEDICINE

Bruce Conklin, Katherine Pollard, Alisha Holloway, and Alex Pico

Between 1856 and 1863, Gregor Mendel cultivated and tested some 29,000 pea plants. His careful analysis of the pea plants laid the foundation for modern genetics. Since Mendel's time, the biology and the mathematics needed to interpret genetic data have become much more complex. Gladstone investigators are not only using modern statistical and mathematical analyses, they are also deeply involved in developing methods to handle large amounts of genomic data.

"With the completion of the Human Genome Project and the delineation of several other genomes, we are awash in data," said Bruce Conklin, Gladstone senior investigator. "The results of a single genomics experiment can fill many telephone books. Now we need to find the tools that will enable us to extract meaningful information from all of these data."

Genomic Expression Data

Genome sequencing projects paved the way for large-scale experiments to study gene expression with DNA microarrays. However, a single microarray experiment can yield hundreds of pages of data on thousands of genes.

To help scientists cope with this flood of information, a team led by Dr. Conklin developed an innovative computer program called GenMAPP (<http://www.GenMAPP.org>). This free program is widely used by scientists around the world.

GenMAPP displays gene expression data on known biological pathways. In this familiar context and with color coding to show expression levels that rise or drop, the full extent of changes in levels of gene expression can be more easily understood. Since genes are viewed dynamically, the

user can easily switch criteria for the colors that illustrate the results.

GenMAPP works with metabolic pathways, signal transduction cascades, gene families, or any other organized collection of genes, proteins, and metabolites. A pathway in GenMAPP consists of boxes that represent individual components, such as genes, and arrows that represent the direction of the cascade of interactions that constitute a biological process. Each pathway component is linked to a database of compiled resources (e.g., Entrez Gene, Ensembl, GO, SwissProt). By clicking on a given gene or protein, users can access associated gene expression data, annotations, and hyperlinks to primary resources.

The success of GenMAPP has led the Conklin laboratory to tackle the problem of annotating and updating the biological pathways. As GenMAPP became

widely used, people began donating biological pathways to the project, and the small team in the Conklin laboratory was quickly overwhelmed. Instead of giving up, Alex Pico, a postdoctoral fellow with Dr. Conklin, worked with collaborators in the Netherlands to develop WikiPathways—the first Wiki that allows users to draw and annotate biological pathways. WikiPathways allows users all over the world to help in growing a large collection of biological pathways. Although this project was publicly announced in 2008, hundreds of contributions have already come in from all over the world.

“We have continued to refine the GenMAPP program to make it more accessible to researchers,” said Dr. Conklin. “We strongly believe that the content curated by multiple experts at WikiPathways provides users with the most efficient way to communicate information about these complicated genetic pathways.”

Another benefit of the GenMAPP project is that it resulted in a series of statistical problems that led Dr. Conklin to consult with the UC Berkeley Biostatistics Division, where he met a very bright graduate student named Katherine Pollard. “I have been following Katie’s spectacularly successful career ever since I first met her. It was a real coup for Gladstone to be able to attract her to join us.”

Analyzing Massive Genomic Data Sets

Biomedical scientists commonly examine relationships between specific outcomes (e.g., tumor grade, time to metastasis, or survival) and thousands of other variables. In her research, Dr. Pollard, now an associate investigator at Gladstone, explores ways of looking at very large sets of genomic data. For example, gene expression profiling is an established method for classifying patients into different disease subpopulations based on the examination of their mRNAs or microRNAs. A better understanding of disease subtypes would likely lead to improved diagnosis and customized treatments.

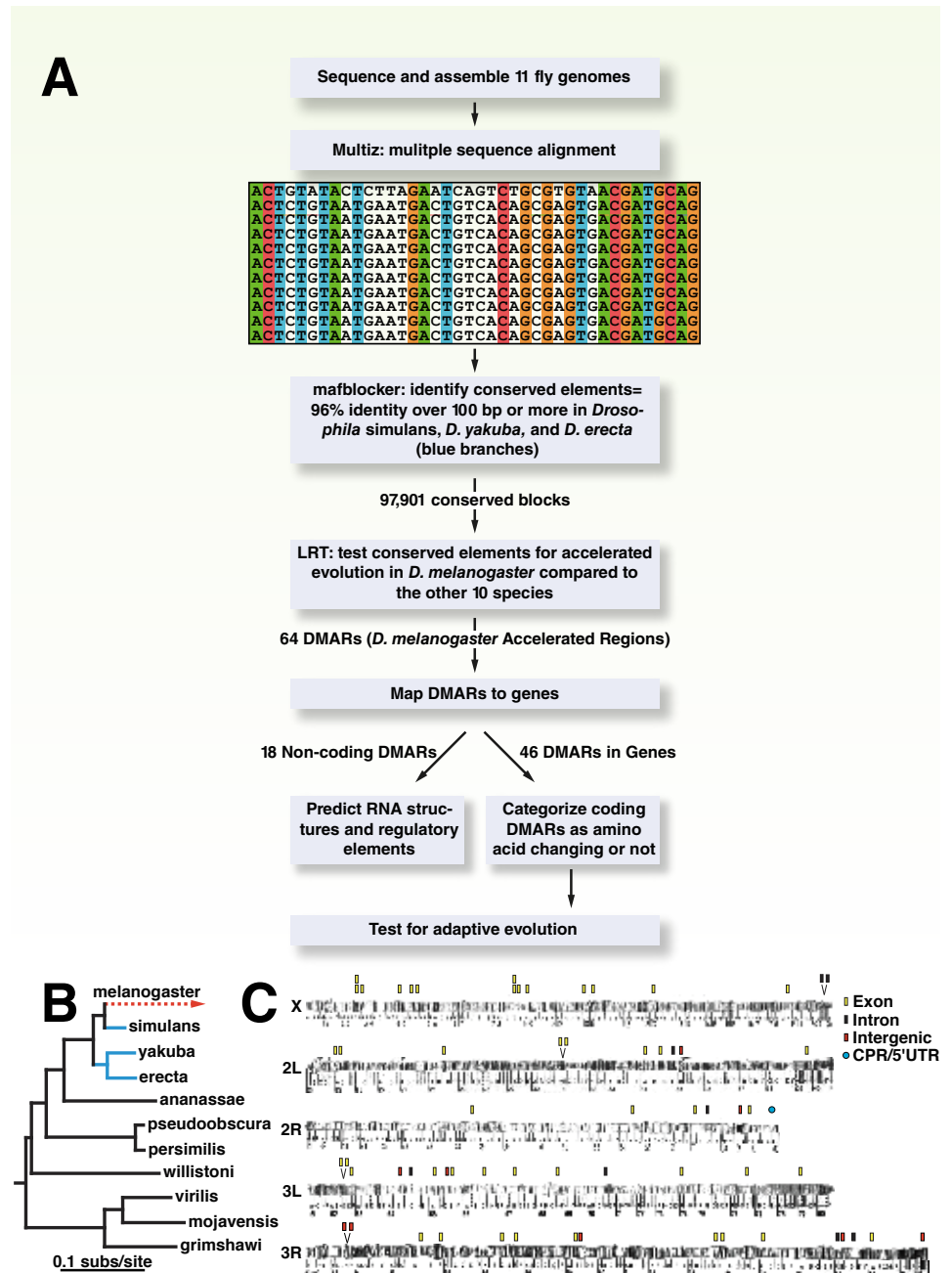
“Our goal is to develop statistical methods to estimate and understand causal and non-causal parameters of biological processes,” said Dr. Pollard.

Handling large data sets is difficult. Often many genes have roughly equal value for predicting outcomes. Dr. Pollard recently pioneered a new statistical tool—called a supervised distance matrix—that enables scientists to identify groups of genes whose expression profiles are all associated with a specific variable or outcome. Discovering these groups of genes is an important first

step in designing diagnostic tools and understanding disease mechanisms.

An Evolutionary Perspective

Another focus of Dr. Pollard’s research is to develop and apply probabilistic models of molecular evolution to detect DNA sequences that evolve uniquely in a specific species or population. Statistical modeling,



Rapidly evolving sequences in the fruitfly *Drosophila*. (A) Flowchart of methods for identifying genomic regions that have been conserved over long evolutionary time periods, but have recently experienced rapid rates of evolution in a single species. (B) Evolutionary relationships of *Drosophila* species. Branches in blue (*D. simulans*, *D. yakuba*, and *D. erecta*) were used to identify conserved elements. All other lineages (and the *D. melanogaster*–*D. simulans* ancestor) were used to infer whether *D. melanogaster* had a faster rate of evolution than expected. (C) Chromosomal locations of *D. melanogaster* accelerated regions (DMARs). Stacked bars indicate multiple DMARs within a single locus. Two bars above a “V” indicate two DMARs that were within the same chromosomal band. DMARs are found predominantly in exons (46/64) and are significantly over-represented on the X chromosome (16/64).



WIKIPATHWAYS Pathways for the People

bioinformatics, and experimental validation can then be used to determine how these changes are associated with changes in biological function.

“Our research focuses on the evolution of genomes and, in particular, on identifying genome sequences that differ significantly between or within species and how they relate to human disease,” said Dr. Pollard. “One of our top priorities is to implement the methods we develop in open source, freely available software packages.”

The statistical methods are highly technical. For example, PHAST—developed with Adam Siepel at Cornell University—is a collection of programs and utilities for use in comparative and evolutionary genomics. It can be used to identify novel functional elements, such as protein-coding sequences and sequences that have been conserved (or remained unchanged) throughout evolution. It is also useful for detecting selection and reconstructing ancestral sequences. The MULTTEST software package involves several testing procedures for controlling error rates when thousands of genes or genomic regions are analyzed in a single study.

By applying these methods, Dr. Pollard has made fascinating discoveries in her two major areas of interest—fast-evolving regions of the human genome and adaptive evolution in microbial communities. For example, she discovered two hundred noncoding sequences that evolved rapidly

in humans and might explain the great difference between humans and chimps, despite the fact that our proteins are nearly identical.

“Understanding the genetic basis for human biology and health is of fundamental interest,” said Dr. Pollard. “By also looking at the DNA of microbes that call the human body home, we gain insights into how they have helped to shape our genome and still interact with it, both in health and in disease.”

Dr. Pollard’s essay “What Makes Us Human” was featured on the cover of the May 2009 issue of *Scientific American*.

Using Biostatistics to Better Understand Data

Gladstone researchers have embraced the powerful technologies developed in the genomics era. The challenge of these technologies lies in integrating computer science and statistics with biomedical research. To facilitate this work, Gladstone hired a computational biologist, Alisha Holloway, and established a Bioinformatics Core in January 2009. The Bioinformatics Core provides experimental design and data analysis support for genomics and systems biology research at Gladstone and for the UCSF community.

The Core works with investigators to analyze gene expression, SNP chips, and behavioral data. In the case of gene expression and SNP-chip experiments, data are produced on many thousands of genes. The

experiments require computational power to process the massive amount of data and statistical expertise to aid in the analysis. Proper experimental design and statistical analysis provide an objective, unbiased method to identify significant results.

Dr. Holloway works with investigators throughout Gladstone to do just that. In the short time that the Bioinformatics Core has been operational, Alisha has been working with investigators from all three institutes to analyze large data sets ranging from whole genome gene expression data to complex behavioral data.

Conclusions

Genetics has come a long way since Mendel, but the challenges remain similar. For his day, Mendel collected huge data sets by recording the growth of pea plants by hand over many years. Mendel struggled to make sense of his data. Today scientists can produce many times that much data in a single afternoon.

As biological experiments continue to become more complex, the benefits of these methods of handling data will become even more important. The work of these Gladstone scientists is helping to make sense of it all.

“This area is absolutely critical to our work on many levels,” said Deepak Srivastava, GICD director. “It affects how we do and interpret our ongoing studies, and it gives us another tool to see more deeply into the human genome.”

