

September 24, 2003

Davis, CA & St. Louis, MO

Re: Genomes of *Drosophila simulans* and *yakuba*

Dear PI:

This letter has two purposes. The first is to tell you about the status of the *D. simulans* and *D. yakuba* genome projects and to invite your comments. Second, we want to tell you about a brief meeting we are interested in hosting sometime in the next few months, the goal of which is to facilitate communication among those interested in the analysis of the imminent data from these projects.

NHGRI/NIH recently granted the Genome Sequencing Center (GSC) at Washington University School of Medicine, St. Louis (<http://genome.wustl.edu/>) approval to proceed with the sequencing of the *Drosophila simulans* and *D. yakuba* genomes.

Rick Wilson (Director of the GSC) contacted Dave and Chuck about planning the project. Sandy Clifton is leading the effort at the GSC. Based on earlier discussion among a number of *Drosophila* population genetics colleagues, much of which occurred on list drosopogenomics@ucdavis.edu more than a year ago, Dave and Chuck had proposed that *simulans* might be shotgun sequenced to shallow (1X) coverage in 7 inbred strains and *yakuba* deeply (up to 8X) in one inbred strain (see white paper at www.dpgp.org). Based on that proposal and subsequent discussions a plan has been developed to achieve a solid *simulans* consensus (reference) sequence and extensive high quality polymorphism data as well as a high quality sequence of the *yakuba* genome. Fosmid libraries will be prepared from one *simulans* strain (*w⁵⁰¹*) and from *yakuba*. These fosmids would be arrayed, end-sequenced (0.3X) and re-arrayed to yield a tiling path of both genomes, which would be available to the community as clones.

Ana Llopart and Jerry Coyne (U. of Chicago) have created a *yakuba* stock that is inbred (10 generations of sib mating) and verified isokaryotypic. Inbred lines of *simulans* are from North America and from Tahiti, Madagascar, Kenya and New Caledonia (provided from recent collections made by Bill Ballard [U. of Iowa]). The genomic DNAs are being prepared at Davis with aid from Dan Barbash. The GSC team plans to start sequencing in December and be finished by the end of January. It is this rapidly approaching new data resource that prompts this email.

An important justification for moving these two sequencing projects forward was the active and creative research of the *Drosophila* population genetic and molecular evolution community. The large numbers of facts and concepts central to the study of genomic variation and the talent associated with our small community (including our esteemed theoretical colleagues) made the investment attractive to NHGRI.

How will all this sequence be assembled, organized, presented to various user communities and analyzed? Obviously the data will be available for all. However, initial analysis and interpretation of genome sequences have often been the result of a monolithic, genome-centered effort culminating in a large, single release and publication. An alternative model that is appealing to us is a broader, community-based approach that could be attractive to genome sequencers, assemblers, annotators and to the community of population geneticists and genome-

evolutionists. However, such a community-based effort requires considerable communication, coordination, and resources.

To that end we are writing to stimulate your interest, provoke your comments and invite your participation. A number of interacting levels of analysis are necessary to make the most of the *simulans* and *yakuba* genomes. First, sequence reads must be assembled into a representation of the genomic sequence that is of high quality (few errors and few gaps). Then, automated annotation algorithms must provide biological context for the sequence. This resource provides the foundation for functional, population genetic and molecular evolutionary analyses. Understanding and communication among individuals and groups with expertise from these areas will greatly improve the science that emerges.

Sequence assembly is a progressing "art form." This project offers (at least) two opportunities for assemblers. If the *melanogaster* reference sequence is to be used as a scaffold, say for *yakuba*, at what stage should this be done and what circularities might this introduce into subsequent comparisons between species? The design of the *simulans* project will produce a new type of population genetics data set and will require a formal algorithm for assembling a "consensus." Is there an objective process to compare and validate different sequence assembly approaches?

The *melanogaster* annotation provides the high quality framework for annotating these two genomes. But of course, deeper annotation of the *melanogaster* genome is also a motivation for the project. So, what are the different ways that the emerging *Drosophila* genomic sequences (and polymorphism data) can be used to produce high quality and detailed annotations of all the genomes? Can objective comparisons of different annotation approaches be fostered and supported?

Population geneticists will have these seven inbred stocks of *simulans* with partial genomic sequence for each. One can imagine many inventive approaches to analysis of these data and use of the corresponding stocks. But understanding how the sequence data were assembled and annotated is an essential first step. Communicating and interacting with the groups producing the assemblies and annotation will offer a unique opportunity to develop bridges between the population genetics and genomics communities.

Similarly those interested in comparative genomics and the integration of these *simulans* and *yakuba* genomes with those of *melanogaster*, *pseudoobscura*, mosquitoes and beyond will also want to interact with and understand the products of those doing assembly and annotation.

Thus, several rapidly impending issues merit our attention and discussion. We propose that labs (thus the contacting of PIs) having a strong interest in the early analyses of these sequences contact us. If there were sufficient interest we would try to get a number of PIs (a postdoc or senior graduate student could be sent in lieu of a PI) together at the GSC at Wash. U for a one- or two-day meeting on a Saturday/Sunday around the first of the year. This meeting would have two main goals. First, we want to facilitate sufficient coordination to ensure that those committed to working on the early data understand the processes and products. Second, besides the many scientific issues to be discussed, there are practical issues regarding common interests and potential collaborations. For example, we believe that the greatest excitement, fun and impact of these genome sequences will result if we choose a target date for simultaneous publication of several papers.

Please contact us if your lab will want to participate. Be thinking about specific scientific contributions you could make to these projects. Indicate whether a meeting would be advisable and possible. Please respond by October 6, 2003, so we can begin planning.

Regards,
Chuck, Dave, Sandy & Rick

Charles H. Langley
chlangley@ucdavis.edu
530.752.4085

David Begun
djbegun@ucdavis.edu
530.754.6362

Center for Population Biology & Section of Evolution and Ecology
Division of Biological Sciences, University of California
One Shields Avenue, Davis, CA 95616

Sandra W. Clifton
sclifton@watson.wustl.edu
314.286.1467

Richard K. Wilson
rwilson@watson.wustl.edu
314.286.1807

Genome Sequencing Center
Washington University School of Medicine
4444 Forest Park Boulevard
St. Louis, MO 63108