Comments and Suggestions for Improvement of the Archon Genomics X PRIZE Validation Protocol

Alexander Wait Zaranek*[1,2], Tom Clegg[2], Ward Vandewege[2], Joseph V. Thakuria[1,2,3]

[1] Personal Genome Project, Harvard Medical School, Boston, MA.
[2] Clinical Future, Somerville, MA.
[3] Center for Human Genetics Research, Massachusetts General Hospital, Boston, MA.

* E-mail: awaitz@post.harvard.edu

This document is a comment on the X PRIZE validation protocol written by Kedes et al. (2011). We propose several modifications which we think will improve the fairness and transparency of the contest while keeping the cost of the validation process under control.

Comment 1.
A key goal is to ensure the judging process is as transparent and open as possible. To this end, a sample public dataset ("Public Dataset") as well as free and open-source software for scoring genome assemblies against this sample Public Dataset, should be provided by the X PRIZE Foundation. This will allow contestants to score their own assemblies in advance of the actual competition. The Public Dataset should be generated from a set of 100 Public Genomes. The National Center for Supercomputing Applications (NCSA) should also provide access to the Public Dataset and open source software for scoring genome assemblies in advance of the actual contest.

Comment 2.
Use random sequencing of long DNA fragments for haploid phasing, structural variants, and sequence fidelity. Fosmid libraries are an established technique for this but other techniques are possible; see Fan et al., (2011), Kitzman et al. (2011), Zhang et al. (2006a,b). Use of long DNA fragments would also:

- Eliminate the need for trios to assess haplotype accuracy.
- Reduce or eliminate the need for targeted resequencing of highly polymorphic regions.
- Reduce or eliminate the need for deep distant pair-end/mate-pair resequencing for assessing structural variants.

Random fragments could be isolated from all of the 100 Public Datasets (see above) and Test DNA samples. Sequences from these fragments would be assembled de-novo to maximize the discovery of indels and structural variants. The exact configuration of reads would determine the ability to accurately de-novo assemble repetitive regions. If Fosmids are used, then assembled sequences would be anchored on both ends to a specified human reference genome to avoid the possibility of contamination with a

microbial sequence. These data would serve as a template to compare the accuracy, completeness, and diploid assembly (haploid phasing) of contestant genome assemblies for each cell line.

Comment 3.
The Archon Genomics X PRIZE Validation Protocol proposed that any assembled fragment that cannot be reasonably aligned to the extant reference genomes (possibly including primates) would not be used for validation. Such assembled fragments could be included if their sequences can be verified present in other human genomes (e.g., by PCR analysis). In any case, any such tests should be included in the open source software provided to contestants as part of the validation protocol and made available in conjunction with the Public Dataset.

In total, a minimum of 200 gigabases of alignable data would be generated for each of the Public and the Validation Test datasets (a minimum 30x coverage of 30 megabases from each sample and variable coverage of additional regions in each sample). In this way, one would expect that every position in the human genome would be interrogated by these data sets.

In particular, one would expect that every genome will have some assembled fragments from the HLA complex (and other repetitive regions) and that the whole region will be reasonably represented in both the Public and Validation Datasets. As an optional step to improve de-novo assembly of highly repetitive regions (such as those derived from the HLA complex), a few long DNA fragment libraries could be sequenced using a more expensive technology that delivered either longer reads and/or larger mate-pair gaps.

Comment 4.
Sequencing accuracy can be inexpensively assessed at defined genomic loci via genotyping “in duplicate”. Using the same platform and an additional reference DNA sample, one should detect copy number variants (CNVs). The sequence accuracy of at least one million common SNPs and all detectable CNVs should be determined for each test genome. Each sample should be assessed twice to minimize call errors and discordant results between duplicate runs should be discarded.

Comment 5.
Genotyping results can be combined with random sequencing of long DNA fragments to ensure internal validity. This can help ensure that artifacts are minimized and can drive automated identification of loci that require further PCR analysis and sequencing. Demonstrating this process on the Public Data set will help ensure that the final Validation Data set is of the highest quality.

Comment 6. Suggestions for Needed Software Development
It should be noted that the primary data produced by likely sequencing and genotyping facilities will not be in a format that allows a straightforward comparison with the contestant data. Along with the Public Dataset, open source software will need to be provided to score assemblies. This software will have to perform the following tasks:

- Comparison and cross-validation of the genotyping data obtained from different technical platforms
- De-novo assembly of long DNA fragments from raw reads (finding substitutions, indels and structural variations across the fragment—all of which should be in phase), cross-validation with genotyping and targeted sequencing (and if Fosmids are used, anchoring of contigs on the reference assembly).
- From all of the above, preparation of a sequence-based validation dataset that can be used for scoring the sequences submitted by contestants.

Comment 7. Suggestions for the Data Deposition and Format

The data submission requirements of the Validation Protocol should be modified. Contestants should submit data in a simple plain-text format that will be specified by Archon Genomics X PRIZE and specifically designed to facilitate scoring, subsequent publication, and clinical utilization. Software should be provided to manipulate this format. Example genomes in this format generated on the Illumina, SOLiD, and Complete Genomics and possibly other platforms should be provided to accompany the Public Dataset.

We agree that the data should be submitted to the Archon Genomics X PRIZE Judging Panel on two identical hard disks, each containing a full set of files but they should include a manifest of corresponding cryptographic hashes, to avoid potential data loss due to file corruption.

References

Fan HC, Wang J, Potanina A, Quake SR (2011) Whole-genome molecular haplotyping of single cells. Nature Biotech. 29:51-57. PMID: 21170043.

Kedes L, Sutton Gr, Liu E, and Jongeneel V. (2011) Archon Genomics X PRIZE Validation Protocol. Available from Nature Preceding <http://precedings.nature.com/documents/5731/version/1>

Kitzman JO, MacKenzie AP, Adey A, Hiatt JB, Patwardhan RP, Sudmant PH, Ng SB, Alkan C, Qiu R, Eichler EE, Shendure J. (2011) Haplotype-resolved genome sequencing of a Gujarati Indian individual. Nature Biotech. 29: 59-63.PMID: 21170042

Zhang K, Martiny AC, Reppas NB, Barry KW, Malek J, Chisholm SW, Church GM (2006a) Sequencing genomes from single cells via polymerase clones. Nature Biotech. Jun;24(6):680-6. PMID: 16732271

Zhang K, Zhu J, Shendure J, Porreca GJ, Aach JD, Mitra RD, Church GM (2006b) .Long-range polony haplotyping of individual human chromosome molecules Nature Genetics Mar; 38(3):382-7. PMID: 16493423