Archon Genomics X PRIZE Validation Protocol

Version 2-24-2011

Granger Sutton[1], Edison Liu[2], Victor Jongeneel[3], and Larry Kedes[4]

Preamble

The following document is a collective assembly of techniques designed to test the quality and accuracy of 100 whole human genome sequences resulting from the $10 Million Archon Genomics X PRIZE (AGXP) competition. The purpose of this article in *Nature Precedings* is to enlist constructive criticism from the genomic and genetic community on the outlined approaches. The intent for the final version of this Validation Protocol (VP) is to become a useful standard by which to gauge the capabilities of whole genome sequencing technologies that emerge even after 2012.

The authors of this posting will moderate the discussion, incorporate suitable suggestions and produce updated versions from time to time. We intend to close the discussion on or about April 15, 2011. Our intent is to publish the final version with as much community consensus as possible with all contributors to the final version being identified as such.

In making suggestions please keep in mind the overriding constraints implicit in such an endeavor: first, the final VP must be able to declare a winner or winners in the AGXP without controversy; second, any suggested changes should likely *reduce* the actual cost of carrying out the physical and bioinformatic procedures of the AGXP competition.

I. Commentary and Definitions

A. Goal:

This document provides a robust and routine approach to evaluating the quality, accuracy and completeness of producing 100 human genome assemblies as part of

[1] J. Craig Venter Institute, Rockville Md. GSutton@jcvi.org
[2] Genome Institute of Singapore, Agency for Science, Technology and Research, Singapore. liue@gis.a-star.edu.sg
[3] National Center for Supercomputer Applications, Urbana-Champaign Illinois. vjongene@illinois.edu
[4] CORRESPONDING AUTHOR: Department of Biological Chemistry, Geffen School of Medicine at University of California at Los Angeles and X PRIZE Foundation, Playa Vista, CA. kedes@usc.edu

the AGXP competition. It is the intent of the X PRIZE Foundation that the process by which the contest will be judged be as transparent and open as possible.

B. History of This Document

This document is the collaborative work of many individuals led by Granger Sutton, Edison Liu, Victor Jongeneel, and Larry Kedes. The general scheme of the validation protocol has grown out of a larger scale effort to enlist the ideas and opinions of a number of bioinformatics and genomic sequencing experts. That process began with a summit workshop at the J. Craig Venter Institute in November 2008 and a second, larger meeting at the National Center for Supercomputing Applications (NCSA) in March 2010. Lists of the attendees at those meetings appear in APPENDICES D and E.

C. Uniqueness of the Archon Genomics X PRIZE (AGXP) for Sequencing:

The Sequencing requirements of the AGXP are unusual in respect to the complexity of the judging criteria. Many X PRIZE competitions have a singular and definable threshold to be "crossed": e.g. flying to the moon and returning, or staying aloft for X period of time. The threshold/boundary to be surpassed has further value: i.e., the vessel can fly to Mars and back, or the vessel can stay aloft for X +1 period of time. However, the sequencing for the AGXP has an "asymptotic" goal of achieving a definition of perfection. With the human genome being a finite size and without a completely definitive standard of measuring this perfection, the judging is dependent on a definition of accuracy and completeness. For this reason, not only are clear and unequivocal Test Criteria needed, but also novel judging approaches may be required (see below).

Though the primary goal of the AGXP is to reward the most advanced sequencing technologies, we are also aware that the judging process and criteria, and the materials used for judging (i.e., the DNA samples and cell lines, mapping and comparison algorithms) will be important standards for clinical sequencing. Therefore, our goal will also be to develop protocols that can be used as industry benchmarks.

D. Test Criteria:

There are three primary evaluation criteria and minimum standards for winning in any category: 98% completeness, 99.999% accuracy (1 error per 100 kbp), and production of a full diploid genome (2 complete copies of each chromosome except in the case of a male sample where single copies of the X and Y chromosomes will be provided). Given that there are regions in the genome that remain impenetrable to contemporary sequencing, in defining the "denominator" for the accuracy and completeness criteria, we refer only to what is defined by the validation dataset. Lastly, the test criteria will be based on verifiable information extracted from the validation dataset, and not from the currently annotated human genome sequence. In this manner data that we receive will be experimentally verified.

The AGXP Rules require that ***each one*** of the 100 test genomes be sequenced by a competing Team to the same degree of accuracy and completeness.

E. Definitions:

The genomes that constitute the test substrate will be called the **Test Genomes**. The data from which completeness and accuracy will be judged will be called the **Validation Dataset**.

We will call the process by which we derive the Validation Dataset as the **AGXP Validation Protocol**.

The criteria by which we will judge the contestants will be called the **Test Criteria**.

F. Judging:

It is anticipated that there will need to be on-site judging and post sequencing judging through computational analysis. We define the protocols to govern the on-site and post sequencing judging as the **Judging Protocols**.

II. Validation Methods - Physical Analysis

A. Approach

We outline here approaches to a validation protocol that involves carefully choosing DNA samples and validation methods that will test the contestants' capability to sequence and assemble genomes in a consistent, accurate, and unbiased manner. The validation methods are constrained by the assumption that validation costs should be a small fraction of the AGXP award. For this reason, a sampling approach will be the foundation of the AGXP Validation Protocol that determines the Validation Dataset.

B. AGXP Validation Protocol.

Experimental validation methodologies that are inexpensive, robust and accurate and do not involve complete resequencing of some or all competition DNA samples must be established. A sampling approach will be used as a cost containment measure and to facilitate the analysis of a second sample set should one be required following one or more failed attempts.

C. Deriving the Validation Dataset

The Validation Dataset will be created by a sampling strategy using complementary technologies and judicious selection of DNA samples. The technologies to generate the validation dataset will be:

1. **DNA sample selection** for Haploid Phasing, structural variants, and sequencing fidelity: selection of trios/quartets where parents will be included

in the Test Genomes and progeny tested as part of the Validation Dataset. A single duplicate pair will be introduced in the Test Genomes to assess fidelity.

2. **Random Fosmid sequencing** for haploid phasing, structural variants, and sequence fidelity

3. **Targeted resequencing** of highly polymorphic regions for sequence fidelity

4. **Genotyping arrays** appropriate for SNP and CNV calling for structural variants and sequence fidelity

5. **Deep distant pair-end/mate-pair resequencing** for structural variants

These techniques will be applied to some, but not all of the samples. The range of test genomes examined by each of the techniques as described below reflects three alternative sampling approaches with decreasing intensities of inclusion of test genomes examined. A summary of the three alternative validation approaches is in APPENDIX B. The choice of which strategy is pursued is dependent on costs, resources and participation by technology centers and commercial vendors. A detailed summary of the three validation approaches can be found in APPENDIX C.

D. DNA sample selection.

A predefined set of 100 human DNA sources—the *X PRIZE DNA sample Test Genomes*—will be used by contestants. As of this writing, these will be identified from the sample set available at the Coriell Institute and will be chosen in the following manner:

1. **Related samples** – Parent-child trios and quartets will be employed since haplotypes from a child should be found in each parent. The intent is to use these trios to ascertain the accuracy of the haplotypes produced for each diploid genome. The production of full diploid genomes can be tested by the use of trios and quartets. The parents will be included in the 100 Test Genomes; the offspring(s) will be used to test the phasing. The progeny DNA will be withheld and not be included as Test Genomes, but will be used to derive the Validation Dataset. At least four families will be engaged in this manner.

2. **Duplicated samples** – This provides a simple comparison test of the accuracy of the sequencing and assembly processes. This has as a complication the possible use of this information for internal correction by the contestants. This however, can be resolved by the judging protocol. *Manipulation of data by a contestant relying on the duplication would be discoverable.*

3. **Ancestral variability** – Human genomes that exhibit varying degrees of sequence and structural variation and of recombination will be examined by selecting individuals with a range of ancestries including those expected to have the shortest haplotype blocks with most recombination.

E. Random fosmid sequencing validation[5]

The rationale for determining the fosmid sequencing strategy is described in APPENDIX A. Random clones (fosmids) will be isolated[6] from 10 to 20 of the 100 DNA samples. The number of fosmids sequenced from each DNA sample will be at least 6000[7]. The clones from each individual will be pooled without sample bar-coding of each fosmid clone and shotgun sequenced to deep (at least 50x) coverage. Only clones from the same DNA sample will be pooled. For determining sequence accuracy the reads will be aligned to the competitor's sequence.

The fact that fosmid alignments occur in localized ~40 KB regions of the genome will guide re-assembly of the sequence reads. All runs will be performed using paired end minimum 75 bp reads– or longer as new methods emerge. This data set, after assembly of each fosmid and alignment of assembled fosmids to the reference genome, will enable characterization of many MB of diploid sequence in each Test Genome. These data will serve as a template to compare the accuracy, completeness, and diploid assembly (haploid phasing) of contestant genome assemblies for each cell line[8].

Any clone that cannot be reasonably aligned to the extant reference genomes (possibly including primates) would not be used for validation. Such clones could be included if their sequences can be verified present in other human genomes (e.g., by PCR analysis).

F. Targeted sequencing via genome enrichment

Targeted sequencing will be performed on the same loci in the genome across between 50-99 Test Genomes. Targeted loci (e.g. the HLA complex) will be selected on the basis of polymorphic complexity in the genome and in population samples. This

---

[5] The tests will be distributed with limited overlap so that no single sample will have excessive weight in judging. This is for two reasons; first, the judging may be viewed as more fair if the samples for validation are distributed across the full set of 100, and second, in the event there is an error in a few samples, not all the validation criteria will be lost.

[6] Fosmid library creation, clone picking, propagation and pooling can be done at any institution with a well developed capacity for working in a high throughput fashion with fosmids and clone picking.

[7] The exact number of fosmid clones will be determined by budgetary constraints and required scoring sensitivity.

[8] One concern is artifacts in the clones such as rearrangements. We will address this concern by using the other validation methods as confirmation: where the clones do not appear to align well to any of the extant reference sequences the genotyping validation can be used to confirm some SNP variants in the clones and the long-range paired-end sequences to verify structural variants in the fosmid sequences.

will enable the ascertainment of sequencing accuracy of contestant's genomes in these regions. These loci will be enriched from each DNA sample and sequenced to generate ~100x average coverage per sample[9]. This is sufficient coverage to identify SNP and indels in these loci and will thus provide a measure of sequence accuracy in genomic loci bearing complex polymorphism.

G. Genotyping validation

We will assess haplotype-sequencing accuracy at defined genomic loci via genotyping in duplicate. Using the same platform and an additional reference DNA sample we will detect copy number variants (CNVs).

We will determine for each test genome the sequence accuracy of at least 2.5 million common SNPs and all detectable CNVs. Each sample will be assessed twice to minimize call errors and discordant results between duplicate runs will be discarded.[10]

H. Digital Karyotyping or Pair-End maps.

Distant pair-end sequencing approaches can allow for deep coverage specifically targeting the detection of structural variants. A single sequencing run on a Next-Generation instrument for a 10 KB distant pair end library will commonly provide 150X clonal coverage for a genome. In this manner, the near nucleotide mapping rearrangement breakpoints will be assessed on a number of the DNA samples. Two libraries will be constructed: a 10kb gPET and a 1kb gPET library from 4, 6 or 10 test genomes. The fosmid and pair end libraries will not be created from the same genomes.

---

[9] In order to optimize sequencing throughput and reduce cost the enriched loci from each sample will be bar-coded to retain sample identity. This will enable the pooling of the genomic enrichment products, minimize the number of Illumina machine runs required and retain sample identity. Thus, three machine runs will likely be sufficient to sequence all 100 samples generating paired end 75 bp reads.

[10] This will provide validation that heterozygous variants between the two haplotypes are being accurately detected and that homozygous variants are called correctly. This would not be a good measure of the required 99.999% accuracy as the genotyping chips are not themselves that accurate. However, with the double genotyping chip requirement the accuracy will be 99.999%.

III. Computational Validation and Scoring[11]

A. Processing of the Validation Dataset

The data will be deposited at a single analysis site for automated validation. This site will have the deep storage and analysis capabilities for speedy analysis of highly complex datasets. Given the capabilities of the NCSA at the University of Illinois, Champaign-Urbana, they are designated as the formal analysis site and data repository for the Archon Genomics X PRIZE.

The NCSA will hold, in a secure fashion, the entire Validation Dataset that will be provided before the start of the judging. These data will be the "basis set" for comparisons with contestant data.

It should be noted that the primary data produced by the sequencing and genotyping facilities will not be in a format that allows a straightforward comparison with the contestant data. NCSA and its potential collaborators will have to perform the following tasks:

1. Comparison and cross-validation of the genotyping data obtained from different technical platforms
2. Assembly of the capture sequences and identification of allelic variants
3. Assembly of fosmid sequences from raw reads
4. Assessment of structural variants from distant paired-end reads
5. Integration of genotyping and fosmid sequences to reconstitute haplotypes in parents from trios or quartets
6. From all of the above, preparation of a sequence-based validation dataset that can be used for scoring the sequences submitted by contestants

In addition, there will be primary data analysis by the NCSA performed on the contestant output from several specific cell lines. These analyses will address internal consistencies and reproducibility of the sequence information from the contestants. These will include: the comparison of at least one duplicate sample (to assess the degree of discrepancies), and at least one trio (mother-father-offspring).

---

[11] The validation criteria represent the spirit and intent of the AGXP Rules. The validation criteria are the final arbiters for scoring regardless of whether they are necessary or sufficient to meet the desired quality guidelines. The current quality criteria are quite strict. However, the validation proposed here should with high probability be easily satisfied if the assemblies meet the criteria; and assemblies of significantly lesser quality should fail the validation. Again, the validation criteria and not the desired quality guidelines will determine the final judgment.

B. Data Deposition and Format

The data on which the contestants will be judged will be for each analyzed genome a set of 46 sequence files in FASTA format, each containing the sequence of a single human chromosome. "N" characters will be used to represent undetermined nucleotide positions or gaps in the assembly. The contestants are also required to submit raw sequence data files in FASTQ format (or equivalent). While these will not be used for scoring the submissions, they may be required for verification purposes and in the case of a competitor challenge (see below).

The data will be submitted to the AXGP Jury on two identical hard disks, each containing a full set of files, to avoid potential data loss due to file corruption. The Jury will then hand them over to NCSA for comparison to the Validation Data and scoring.

C. Scoring[12]

Each one of the 100 genomic assemblies provided by the contestants must meet the Validation Criteria. If any one of the genomic assemblies provided by a contestant fails to meet any component of the Validation Criteria, then that contestant will have failed to meet the requirements of the Archon Genomics X PRIZE. The Validation Criteria define thresholds for **accuracy** and for **completeness**.

1. **Accuracy**

   A **mismatch** compared with the validation dataset is considered an error for the purposes of calculating accuracy. The worst 2% of the AGXP clones in terms of how well they align to the competitor's sequence will not be used in that validation in order to avoid penalizing contestants for possible errors in the Validation Dataset.

   a. A rearrangement or haplotype error counts as one error but insertion and deletion errors count the sum of each base in the indel. Missed insertions also count as bases missed for the purposes of calculating completeness (see below).

      The score for accuracy will be derived from the following comparisons:

      - Alignments between the competitor sequences and the fosmid sequences in the Validation Dataset
      - Alignments between the competitor sequences and the regions that have been subjected to targeted resequencing
      - Matching of the competitor sequences to the SNP genotypes
      - Comparison of the competitor sequences to the SNP haplotypes determined from genotyping trios and quartets

[12] The computational judging protocols will primarily be automated not only for speed but also for objectivity.

2. Completeness

The AGXP recognizes that there is no absolute standard against which to judge the completeness of a competitor submission. With this proviso in mind, the following criteria will be used for determining completeness:

- The extent of the sequences represented in the sequenced fosmids[13]
- The extent of the sequences represented in the resequenced polymorphic segments
- The length of the intervals determined in the paired-end sequences, especially for large indels and tandem duplications.
- The presence in the competitor sequence of all of the areas surrounding the SNPs for which an unambiguous call was obtained

D. Competitor challenges of validation results and retries

The AGXP intends to minimize the jeopardy of revealing the validation dataset in the case of a challenge or a future competition. A competitor will not have a right to challenge the decision of the judges but may request a review. A competitor will have no access to the Validation Dataset either before or after a contest. The AGXP may, at its sole discretion, make available to any contestant just sufficient data to demonstrate the reasons that their data failed to meet the validation criteria. Since any failure to meet completeness or accuracy goals must include a worst-case sample that falls below the passing criteria, the XPRIZE Foundation need only share the data from this one cell line when defending a "fail" decision. Thus all other X PRIZE data can be kept as a secret. New samples will ***not*** have to be prepared for a second trial except as replacement of the sample whose data was shared. The current rules call for competitors being charged for all costs related to a retry including samples, any new sequencing by X PRIZE, all judging costs, computational costs etc

Clearly, if a competitor maintained the DNA samples or passed them along to another competitor this would convey an unacceptable advantage. Contractual mechanisms will be placed in the Master Team Agreement to prevent this from occurring.

[13] Completeness might be the most contentious as a competitor could argue that many of the fosmid clones sequenced by AGXP fall in the 2% they are not required to cover. The probability is VANISHINGLY SMALL that a competitor who indeed did sequence 98% of a given genome will have failed to sequence 98% of the AGXP sequences obtained from that genome.

Appendix A: Derivation of Validation Datasets

Prepared by Edison Liu, Pauline Ng, Anbupalam Thalamuthu, JianJun Liu, Hidetoshi Inoko

Singapore Genetics Institute

Fosmid Sequencing:

The two uses for fosmid clone sequencing are to establish phasing and to have a measure of sequence coverage. The plan is to have fosmid libraries constructed from a number of the validation samples and to shotgun sequence a specific number of clones per individual. This means that the competitors should be able to define the variations that are linked within a haplotype block, but more importantly, define the relationship between haplotype blocks. If one considers that the human genome is in haplotype blocks, and competitors could use current HapMap data to reconstruct phase within a block, then the true test for phasing accuracy is the correct determination of the relationships between neighboring blocks.

Because YRI (Africans) have smaller haplotype blocks, then for a given region size (fosmid = 40 kb), such genomes provide more of these relationships to test. The mean size of a haplotype block is 16.3 kb (Caucasian = CEU), 13.2 kb (Asian = CHB+JPT), and (7.3 kb African = YRI). Thus for a 40 kb region that represents a fosmid clone, the expected number of breaks between haplotype blocks is 2, 3, 5 for CEU, CHB+JPT, and YRI respectively. Thus, by using African genomes, the validation test set would be at least 2x larger for the testing of phasing. This also means that for each 1500 clones, we have a possibility of testing between 3000 and 7500 breaks in haplotype blocks for Caucasian and African individuals respectively. (This does not however, ensure we will have this number of "phase test" possibilities/opportunities).

For phasing to be assessed, there are several ways to do this using a sampled fosmid library. One is to sequence paired-ends so that the sequence at each end are linked (i.e., phased). The other is to sequence to entirety, the entire fosmid clones. Actually there is a third, which is to sequence each individual fosmid clone using Sanger sequencing (in fact, this would be the most accurate, but of course the most costly and therefore eliminated from the discussion). For pair-end reads, length of the read fragment plays a huge role in the number of useful reads to figure out phasing. For 75 bp paired ends, only 0.5% of the mate pairs will have a SNP in both of the paired ends (for one to phase), if Africans are used. If Europeans are used, only 0.3% of the mate pairs will be useful for phasing. Hence, most of the reads will not be useful for phasing, but at least 68% more of the mate pairs will be useful for phasing if Africans are used. If technology will permit 150 bp paired ends, then 2% and 1% of the paired ends will be useful for African and European respectively (see above). Therefore relative to cost considerations, shotgun sequencing of pooled clones is the more efficient way to obtain both coverage and phasing. For phasing, unless each clone is bar coded, one would want to minimize the chance that the fosmids from

the same individual do not overlap because that would confound the reconstruction of the phasing. These calculations assume no bias in the fosmid library -which is unlikely -and might be as much as 20% misestimated. This is because in the random shotgun reads, in the regions of overlap there would be no way to discern whether individual sequence variants belonged to one allele or another. Therefore only the fosmid sequences that are from non-overlapping fragments could be used for phasing determination.

	Heterozygosity (taken from Nature 456:53-59)	Average variant per bp	**Length of read**	Poisson lambda: expected variants in a 75 bp read	Probability of having 1 or more variants in 75 bp read	Both paired ends having variants
European ancestry	0.00076	1315.789474	75	0.057	0.055405931	0.003069817
African ancestry	0.000994	1006.036217	75	0.07455	0.071838935	0.005160833
European ancestry	0.00076	1315.789474	150	0.114	0.107742044	0.011608348
African ancestry	0.000994	1006.036217	150	0.1491	0.138517038	0.01918697

By contrast, coverage would be best tested if as many fosmids as possible are sequenced as a validation reference/standard. This is then an issue of balancing the overlap possibilities that might limit the phasing with the desire to obtain the greatest coverage. Our goal therefore is to find the number of fosmid clones that should be sequenced to maximize the assessment of coverage yet minimizes the overlap that limits the ascertainment of phasing.

As an assessment, we simulated the sequencing of fosmid clones to ask the frequency of overlap related to the number of clones sequenced.

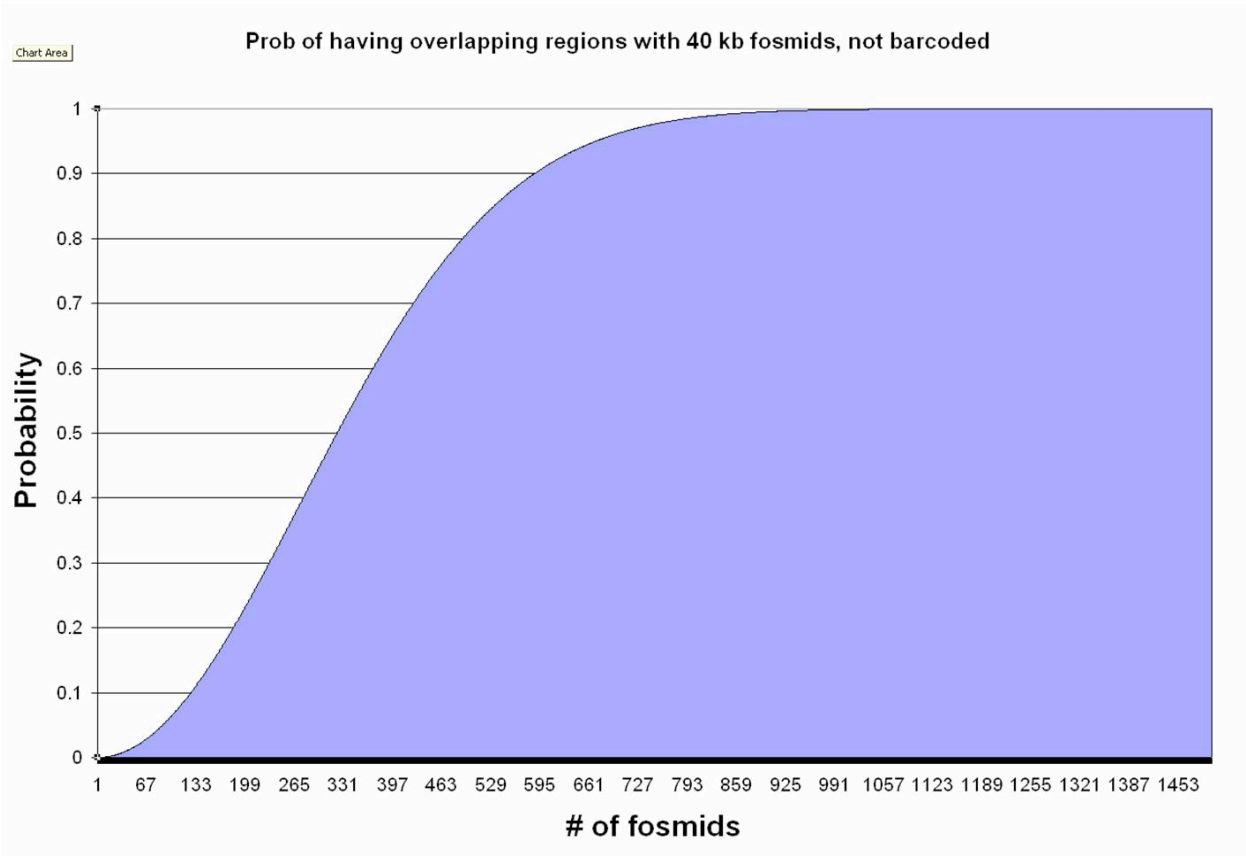
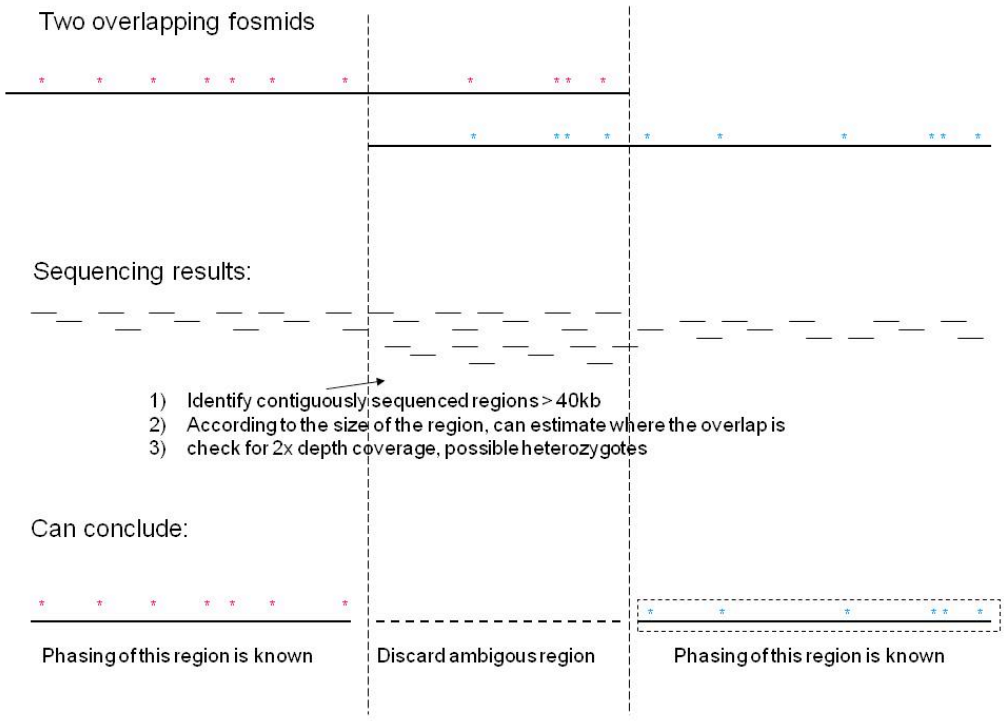


Chart Area
Prob of having overlapping regions with 40 kb fosmids, not barcoded
Probability
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
1 67 133 199 265 331 397 463 529 595 661 727 793 859 925 991 1057 1123 1189 1255 1321 1387 1453
# of fosmids

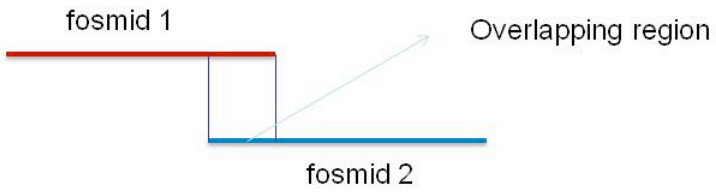

Two overlapping fosmids
Sequencing results:
1) Identify contiguously sequenced regions > 40kb
2) According to the size of the region, can estimate where the overlap is
3) check for 2x depth coverage, possible heterozygotes
Can conclude:
Phasing of this region is known
Discard ambigous region
Phasing of this region is known

Simulation

Sample N fosmids of size 40kb with replacement from 3Gb region.

Identify pairs of overlapping fosmids and compute the size of the overlap (see below).

Repeat the sampling 5 times to compute mean and standard deviation.



Size of overlap = length of overlapping region/40kb

Summary of overlap

N fosmids	>10 % overlap	>20 % overlap	>50 % overlap
500	2.2 (1.3)	2.2 (1.3)	1.2 (0.83)
1000	11.6 (3.05)	10.4 (2.79)	6.9 (3.11)
2000	47 (4.3)	41.6 (5.59)	26.8 (6.26)
3000	109.8 (8.22)	98.2 (7.8)	63.8 (5.4)
5000	297.6 (19.73)	263.6 (18.87)	159.8 (9.86)
*6000	412.2 (25.33)	370.2 (20.4)	241 (14.79)

*Approximate, extrapolated from simulation based on 1Gb

Average number of PAIRS of overlap (SD) is given in the above table. e.g. For 500 fosmids, 2.2 pairs overlap, and ~4 fosmids affected.

N fosmids	>10% overlap	Percent	>20% overlap	Percent	>50% overlap	Percent
500	0.0495 (0.0316)	0.25 (0.16)	0.0 495 (0.0316)	0.25 (0.16)	0.0373 (0.0269)	0.19 (0.13)
1000	0.2608 (0.0921)	0.65 (0.23)	0.2541 (0.0905)	0.64 (0.23)	0.2011 (0.0988)	0.50 (0.25)
2000	1.0451 (0.1601)	1.31 (0.20)	1.0126 (0.1746)	1.27 (0.22)	0.8026 (0.1733)	1.00 (0.22)
3000	2.4768 (0.1966)	2.06 (0.16)	2.4084 (0.202)	2.01 (0.17)	1.9323 (0.1580)	1.61 (0.13)
5000	6.4682 (0.3475)	3.23 (0.17)	6.2639 (0.3399)	3.13 (0.17)	4.7920 (0.3493)	2.40 (0.17)

Summary of overlap sizes

Average overlap size in Mb (SD) together with the percentage overlap (SD) to the total size of all the fosmids are given in the above table.

Roughly speaking, the level of overlap going from 1000 clones to 5000 clones ranges from ~0.5% to ~2.4% respectively, and about 3% for 6000 clones (estimated). To cover the entire genome with 1X coverage, we anticipate sequencing 75,000 clones. Therefore 5000-6000 clones appear to be a level with acceptable overlap and provide over 200 Mb of sequence to assess phasing. With the ends of the fosmid clones tagged, the overlap can be readily computed.

Recommendation: Fosmid libraries for 10-20 individuals at 6000 clones each. Recommend that the majority will be from African descent.

APPENDIX B: Alternative Validation Sampling Approaches

Numbers represent test genomes	**SNP Array**	**Capture Sequence**	**Fosmid Sequence** (6000 clones)	**Distant Pair-end Library** 10kb gPET	1kb gPET
Validation I	149	99	20	10	10
Validation II	131	75	12	6	6
Validation III	119	50	10	4	4

APPENDIX C: Summary of Three Validation Alternatives

Validation Protocol I:

One pair within the 100 test cases will be twins or the same cell line from the same individual. These will be used as "internal" controls and should yield near identical sequenced by the contestant.

Two samples will have had the full genome sequenced by two methods. The overlap of the two technologies will be used as the validation dataset

96 test cases each as part of a trio (minus fully sequenced genome and minus the twins, and adjusted for parentage). Therefore the adjunct cases will be N=48 and the total validation cases = 148 (100 + 48)

All relevant validation cases (N=148) will be assessed for SNPs as part of the test for sequence accuracy, for completeness, and for phasing. A 2.5 million SNP assay device will be used X2 and the concurrent SNPs will be used in the assessment. Phasing will be calculated on the samples

All test cases (N=98) will be assessed by capture sequencing of the HLA locus to test for sequence accuracy. Approximately 3.8Mb will be sequenced at 50-70X coverage.

20 test cases will be assessed by fosmid sequencing for phasing, for accuracy, and for completeness. 6000 clones will be sequenced per individual. Together this will mean 4.8 Gb will be sequenced

10 genomes will be assessed by distant pair end libraries to assess private structural variants. Each will be assessed by a 10Kb gPET library sequenced to 100-150X coverage, and a 1kb gPET library at 100X coverage.

Validation Protocol II:

One pair within the 100 test cases will be twins or the same cell line from the same individual. These will be used as "internal" controls and should yield near identical sequenced by the contestant.

One sample will have had the full genome sequenced by two methods. The overlap of the two technologies will be used as the validation dataset.

60 test cases will be part of a trio. Therefore the adjunct cases will be N=30 and the total validation cases = 130

All validation cases (N=130) will be assessed for SNPs as part of the test for sequence accuracy, for completeness, and for phasing. A 2.5 million SNP device will be used X2 and the concurrent SNPs will be used in the assessment. Phasing will be calculated on the samples

75% of the test cases (N=75) will be assessed by capture sequencing of the HLA locus to test for sequence accuracy. Approximately 3.8Mb will be sequenced at 50-70X coverage.

12 test cases will be assessed by fosmid sequencing for phasing, for accuracy, and for completeness. 6000 clones will be sequenced per individual. Together this will mean 2.9 Gb will be sequenced.

6 genomes will be assessed by distant pair end libraries to assess private structural variants. Each will be assessed by a 10Kb gPET library sequenced to 100-150X coverage, and a 1kb gPET library at 100X coverage.

Validation Protocol III:

One pair within the 100 test cases will be twins or the same cell line from the same individual. These will be used as "internal" controls and should yield near identical sequenced by the contestant.

One sample will have had the full genome sequenced by two methods. The overlap of the two technologies will be used as the validation dataset.

40 test cases will be part of a trio. Therefore the adjunct cases will be N=20 and the total validation cases = 120

All validation cases (N=120) will be assessed for SNPs as part of the test for sequence accuracy, for completeness, and for phasing. A 2.5 million SNP device will be used X2 and the concurrent SNPs will be used in the assessment. Phasing will be calculated on the samples

50% of the test cases (N=50) will be assessed by capture sequencing of the HLA locus to test for sequence accuracy. Approximately 3.8Mbfor each individual will be sequenced at 50-70X coverage.

10 test cases will be assessed by fosmid sequencing for phasing, for accuracy, and for completeness. 6000 clones will be sequenced per individual. Together this will mean 2.4 Gb will be sequenced.

4 genomes will be assessed by distant pair end libraries to assess private structural variants. Each will be assessed by a 10Kb gPET library sequenced to 100-150X coverage, and a 1kb gPET library at 100X coverage.

APPENDIX D: Attendees of Bioinformatics Summit

J. Craig Venter Institute, Rockville, MD (November 3-4, 2008)

NAME	AFFILIATION	TITLE
Mark Adams	Case Western Reserve University	Associate Professor PhD Training Faculty
Richa Agarwala	NIH – NLM (National Library of Medicine)	Computational Biologist
Serafim Batzoglou	Stanford	Associate Professor
Michael Brudno	University of Toronto	Assistant Professor & Canada Research Chair Computational Biology
Deanna M.Church	DHHS/NIH/NLM/NCBI	Staff Scientist Research Fellow
Rod Corriveau	Coriell Institute	Assoc Prof & Scientific Program Manager for the Coriell Cell Repositories
Robert Holt	BC Cancer Research Centre – Genome Sciences Centre	Head, Sequencing, Genome Sciences Centre
David Jaffe,	Broad – Genome Biology Program	Director, Computational R&D
Steven Scherer	The Centre for Applied Genomics The Hospital for Sick Children	Director
Larry Kedes		Sr. Advisor and Scientific Director
Andrew Wooten	X PRIZE Foundation	Senior Director, Archon Genomics X Prize
Barry Thompson	X PRIZE Foundation	Tervela Founder and CTO
Sam Levy	J. Craig Venter Institute	Director, Human Genomics
Yu-Hui Rogers	J. Craig Venter Institute	Vice President of Core Technology
Granger Sutton	J. Craig Venter Institute	Sr. Director, Informatics

APPENDIX E: Attendees Bioinformatics Workshop NCSA (March 2010)

Name	Organization
Bernie A'cs	NCSA
Loretta Auvil	NCSA
Serafim Batzoglou	Stanford
Chris Beitel	NCSA
Guillaume Bourque	ASTAR
Deanna M. Church	DHHS/NIH/NLM/NCBI
Andrew Davis	Monsanto
Thom Dunning	NSCA
Jennifer Eardley	UIUC
Adam Felsenfeld	NIH/NHGRI
Aaron Halpern	Complete Genomics
Jill Herschleb	Halcyon Molecular
Tim Hunkapillar	Discovery Biosciences
Victor Jongeneel	NCSA/UIUC
Scott Kahn	Illumina
Larry Kedes	X PRIZE Foundation
Jim Knight	454 Life Sciences
Denis Larkin	UIUC
Sam Levy	Scripps Health, San Diego
Harris Lewin	UIUC
Cristin Lindsay	X PRIZE Foundation
Ed Liu	HUGO
Havier Llora	NCSA
Jian Ma	UIUC
Elizabeth Mansfield	FDA
Francisco "Paco" Martinez-Murillo	FDA
Luke Nosek	Halcyon Molecular
Danny Powell	NCSA
Don Preuss	NIH/NCBI
Steve Skienna	Stony Brook
David Smith	Mayo Clinic
Jonathan Stark	Halcyon Molecular
Granger Sutton	J. Craig Venter Institute
Mike Welge	NCSA
David Tcheng	NCSA