Consensus Genotyper for Exome Sequencing (CGES): Improving the Quality of Exome Variant Genotypes

Vassily Trubetskoy 1,∗, Alex Rodriguez 2, Uptal Dave 2, Nicholas Campbell 3, Emily L. Crawford 3, Edwin H. Cook 4, James S. Sutcliffe 3, Ian Foster 2, Ravi Madduri 2, Nancy J. Cox 1 and Lea K. Davis 1 ∗

1Section of Genetic Medicine, University of Chicago Department of Medicine, Chicago, IL 60637, US
2Computation Institute, University of Chicago, Chicago, IL 60637, US
3Department of Molecular Physiology & Biophysics and Vanderbilt Brain Institute, Vanderbilt University School of Medicine, Nashville, TN 37232, US
4Department of Psychiatry, University of Illinois at Chicago, Chicago, IL 60608, US

Received on XXXXX; revised on XXXXX; accepted on XXXXX

ABSTRACT

Motivation: The development of cost effective next-generation sequencing methods has spurred the development of high throughput bioinformatics tools for detection of sequence variation. With many disparate variant-calling algorithms available, investigators must ask, “Which method is best for my data?” Machine learning research has shown that so-called ensemble methods that combine the output of multiple models can dramatically improve classifier performance. Here we describe a novel variant-calling approach based on an ensemble of variant-calling algorithms, which we term the Consensus Genotyper for Exome Sequencing (CGES). CGES employs a two-stage voting scheme among three algorithm implementations. While our ensemble method can accept variants generated by any variant-calling algorithm, we used GATKv2.0, FreeBayes, and Atlas-SNP2 in building CGES due to their performance, widespread adoption, and diverse but complementary algorithms.

Results: We apply CGES to 132 samples sequenced at the Hudson Alpha Institute for Biotechnology (Huntsville, AL) using the Nimblegen Exome Capture and Illumina sequencing technology. Our sample set consisted of 40 complete trios, two families of four, one parent-child duo, and two unrelated individuals. CGES yielded the fewest total variant calls (NCGES = 129,706), the highest Ts/Tv ratio (2.91), the lowest Mendelian error rate across all genotypes (0.057%), the highest rediscovery rate from the Exome Variant Server (82.9%) and 1000 Genomes (78.0%), and the highest positive predictive value (93.5%) for a random sample of previously validated de novo variants. We describe these and other QCs metrics from consensus data and explain how the CGES pipeline can be used to generate call sets of varying stringency including consensus calls present across all three algorithms, calls that are consistent across any two out of three algorithms, or a more liberal set of all calls made by any algorithm.

Availability: To enable accessible, efficient, and reproducible analysis, we implement CGES both as a stand-alone command line tool available for download in GitHub and as a set of Galaxy tools and workflows configured to execute on parallel computers.

Contact: trubetskoy@uchicago.edu

1 INTRODUCTION

Whole exome sequencing (WES) has quickly become an affordable approach to identifying rare variants contributing to disease. In 2013, 1,184 published papers were indexed in PubMed with the key words “exome sequencing”—more than twice the number published in 2012 representing a clear trend in human genetics. The utility of WES for revealing biological mechanisms depends upon the genetic architecture of the phenotype in question, the quality of the sequencing technology, and, to a significant extent, the analytic methods used to identify and genotype variations in sequence. In recent years, several methods have been developed to analyze raw WES data, including Atlas-SNP2 (Challis et al., 2012), GATK (DePristo et al., 2011; McKenna et al., 2010), SeqEM (Martin et al., 2010), FreeBayes (Garrison and Marth, 2012), SAMtools (Li et al., 2009), Dindel Albers et al. (2011), SOAPsnp (Li et al., 2009), and Varscan2 (Koboldt et al., 2012), among others.

These methods represent substantial effort and expertise in the analysis of next generation sequencing (NGS) data including both whole exome and whole genome sequencing. Here we present a natural extension of these individual algorithms that integrates their relative strengths into a consensus-calling approach, which we call Consensus Genotyper for Exome Sequencing (CGES). This algorithm, developed as a collaborative effort between the Department of Medicine and the Computation Institute at the University of Chicago, takes advantage of the diverse variant-calling strategies of three existing algorithms (GATKv2.0, Atlas-SNP2, and FreeBayes) in an open-source, freely-available, user-friendly analysis platform.

While all variant-calling programs seek to optimize performance relative to some core properties of sequencing data (such as read depth and allele count), they often differ along other dimensions.
We chose to base our consensus-based pipeline on GATKv2.0, Atlas-SNP2 and FreeBayes, because these three algorithms use complementary approaches and attempt to integrate as much information as possible, including prior variant observations, linkage disequilibrium (LD) structure, and structural variation, in order to reduce both type I and type II errors.

In brief, CGES first runs each algorithm separately and then combines the resulting three collections of genotype calls to create three possible output sets, typically of increasing size but lower average quality: consensus calls (i.e., calls made by all three algorithms), partial consensus calls (those made by two or more algorithms), and the union of all calls (calls made by one or more algorithms). CGES also harmonizes quality scores from each algorithm to provide quality control (QC) reports and publication quality plots (CGES-QC tool).

CGES and CGES-QC together form a multi-step pipeline and must perform multiple program invocations, as shown in Figure 1. We use the Galaxy platform (Goeccks et al., 2010) to combine implementations for each branch of the pipeline and for CGES itself. Our use of Galaxy has the benefit of democratizing NGS data analysis as Galaxy reduces the computational expertise needed to run an NGS pipeline and can run on public clouds such as those operated by Amazon Web Services, Google, and Microsoft. The links to the Galaxy workflows can be found in the supplementary table 3. We have applied the CGES consensus-calling approach to real world exome data collected on subjects with autism and their family members. We use the results of this study to demonstrate the power of the CGES approach and provide project-level, variant-level, sample-level, and family-based quality metrics across all algorithms. Additionally we provide known rare variant rediscovery rates, and an estimate of the positive predictive value (PPV) of each algorithm based on previously identified and lab-validated de novo variation.

2 METHODS

2.1 Samples

In order to test the robustness of the CGES consensus-calling algorithm in the context of real world data, we used raw aligned sequence data (BAM files) from 132 individuals representing 34 complete trios, two families of 4, and 22 additional unrelated individuals recruited from the Autism Center of Excellence study at the University of Illinois at Chicago, Vanderbilt University, or Tufts-New England Medical Center. Probands were assessed with the Autism Diagnostic Interview (ADI-R), the Autism Diagnostic Observation Schedule (ADOS), and clinical evaluation. We included families in this study if the probands met diagnostic criteria for autism or autism spectrum disorder (ASD) on both the ADI-R and ADOS-WPS (Le Couteur et al., 1989; Berument et al., 1999).

2.2 Next Generation Sequencing

Sequencing for the majority of samples was performed at the Hudson Alpha Biotechnology Institute (HABI, Huntsville, AL) as a part of the NIH ARRA Autism Sequencing Consortium (Liu et al., 2013; Neale et al., 2012) and for the remainder of samples at HABI as a part of the University of Illinois and Vanderbilt Autism Center for Excellence study. Methods used for whole exome sequencing are described in depth in supplementary materials and in previous publications (Neale et al., 2012). In brief, samples were sequenced at Hudson Alpha Biotechnology Institute using a paired-end approach with NimbleGen exome capture followed by Illumina HiSeq 2000 sequencing.

2.3 Determination of Parameter Values for Consensus Calling Algorithms

Figure 1 illustrates the variant calling schema utilized by CGES. The three variant calling algorithms used in this analysis are implemented in previously published programs GATK v2.0 (DePristo et al., 2011; McKenna et al., 2010), Atlas-SNP2 (Challis et al., 2012), and FreeBayes (Garrison and Marth, 2012). As these methods are described in depth in their primary publications, here we only describe their parameterization and implementation within the Galaxy framework.

As shown in Figure 1, we implement variant calling for each of the three programs within an independent sub-pipeline or branch. Within each branch, we select parameter values for exome sequence data as follows. For GATK, we followed best practices published by the Broad Institute (Institute, 2013). For FreeBayes and Atlas-SNP2, we determine optimal parameters empirically (supplementary methods) using quality control data from our project to iteratively develop a set of project-specific best practices. As the variant calls from each branch serves as the substrate for CGES, it is important that parameters for each branch are optimized for best performance. Thus, we strongly recommend that investigators review the parameters for each caller, and ensure that they are appropriately determined for a given dataset.
### 2.4 Consensus Genotyper for Exome Sequencing

#### Calling Pipeline

The first step in running the CGES pipeline is to run the GATKv2.0 Unified Genotyper across all chromosomes, with separate multi-threaded processes for each chromosome. The resulting VCF files are then combined and variant recalibration is applied. We train the GATK gaussian mixture model on 1000 Genomes, HapMap, and dbSNP135 data hosted by the Broad Institute. After performing variant quality score recalibrations (VQSRs), variants are then filtered down to a recommended VQSR-based quality tranche (99.0%). A simple filter is then applied to remove genotype calls that do not satisfy an on-target status criterion or that have a minimum QUAL score of less than 10.

The FreeBayes workflow is similar to the GATK workflow in that we first parallelize variant detection by chromosome. During this step we implement the chosen detection and genotyping parameters described in supplementary materials. This results in one multi-sample VCF file per chromosome. We then combine VCFs, and finally filter based on a minimum QUAL of 10 and on-target status.

Unlike the other callers, Atlas-SNP2 calls variation per sample instead of across multiple samples simultaneously. This required a sample-based parallelization scheme in which we spawn a separate Atlas-SNP2 process for each sample. The Galaxy Atlas-SNP2 tool was therefore developed with a Swift (Wilde et al., 2011) backend enabling parallelization across samples. We next follow the Atlas-SNP2 protocol for creating an initial multi-sample VCF file from individual sample VCF files Challis (2013). Finally we applied an on-target filter and compiled resulting variants into a multi-sample Atlas-SNP2 VCF file to produce the final multi-sample VCF file.

#### 2.5 Consensus Genotyper for Exome Sequence (CGES)

We use a two-stage voting scheme to generate consensus genotypes. First, we identify the variant positions (irrespective of genotype) that agree among a specified number of callers. At this step, a user can specify that either two or three of all three callers must agree. CGES considers variants to be uniquely identified by any difference from the reference sequence at a given chromosomal position. The algorithm then proceeds to consider genotypes within these consensus sites. By default, consensus requires that genotypes agree among all branches; however, users can also choose to require that only two out of three algorithms agree. Genotypes that do not agree are set as missing and flagged as discordant for downstream quality analysis.

Stringency thresholds for Samples concordance can be set independently for each stage, conditional on the fact that genotype concordance cannot be stricter than site concordance.

#### 2.6 Consensus Genotype for Exome Sequencing Quality Control (CGES-QC)

We have also developed CGES-QC, a tool for the calculation, comparison, and visualization of sample-based, variant-based, and project-based QC metrics across all branches of the consensus genotyper. CGES-QC incorporates QC calculations from PLINK, VCF-tools, and in-house developed scripts to perform analyses and output publication quality plots.

#### 2.7 Project Based Quality Control

Project-based QC results include the total number of variants called, Ts/Tv ratio, Exome Variant Server (EVS) variant rediscovery rate, 1000 Genomes (1KG) variant rediscovery rate, and the genotype Mendelian error rate (gMER). The total number of variants called is limited to the user specified settings and refers to the total number of variants present in the output VCF. The Ts/Tv ratio is a routinely reported QC measure for sequence data and refers to the ratio of transversions (G<C or A>T) to transversions (G>C or A>T). Based on previously reported analyses, the Ts/Tv ratio is expected to be 2.1 for whole genome sequencing and 2.6-3.3 for exome sequence data (DePristo et al., 2011). Low Ts/Tv ratios represent technical artifacts and a randomly generated set of variants yields a Ts/Tv ratio of 0.5 (Zook et al., 2013). The EVS and 1KG rediscovery rates represent the total number of variants in a VCF that has been previously identified by either the EVS or 1KG sequencing projects. Finally, for each variant site called there may be anywhere from one individual with a sequence variation to N individuals with variant genotypes (where N=sample size) present in the VCF file. Therefore, we calculate a “genotype Mendelian error rate” (gMER), which is the total number of MEs in a VCF file divided by the total number of genotypes with the potential for Mendelian inconsistency (i.e., offspring genotypes with parental genotypes known) in a VCF file. This measure describes the proportion of all offspring genotypes that are inconsistent with parent genotypes present in the VCF.

#### 2.8 Sample-Based Quality Control

Sample-based QC results include the F-statistic per sample, trio Mendelian error rate (tMER), and genotype concordance/discordance per sample. The F-statistic is calculated using the classic Wright formula one minus the ratio of observed heterozygote genotypes to expected heterozygote genotypes according to Hardy-Weinberg Equilibrium (Wright, 1950). This statistic provides a red flag for both sample contamination (extreme heterozygosity) and consanguinity (extreme homozygosity). We calculate Mendelian errors per trio, which we defined as the number of MEs in an offspring (given by the trio) divided by the total number of genotypes in the offspring. This metric is useful for determining if there are any trios that require further attention that may be due to sample mis-match or large copy number variants. In the case of a contaminated sample, the F-statistic and the tMER can be used jointly to determine if the contamination came from a relative or an unrelated sample. Finally, CGES genotype discordance rates are calculated per sample defined as the proportion of all genotypes in a given sample that are flagged as discordant across any two of the three calling algorithms.

#### 2.9 Variant-Based Quality Control

Variant-based quality control includes calculation of the variant site Mendelian error rate (vMER), variant site missingness distributions, and minor allele frequency distributions. The vMER is defined as the number of variant sites with at least one ME divided by the total number of variant sites in a project. This metric provides a bird’s eye view of the general sensitivity and specificity of each calling algorithm. A more inclusive approach to genotype calling will allow a higher number of sites to contribute at least one ME, while a stricter approach will result in fewer sites with at least one ME. At each variant site there are a small proportion of genotype calls that cannot be made (i.e., flagged as “missing”) because a given algorithm does not have enough information to accurately determine the genotype. In the CGES algorithm, the resulting conflicts between branches are also flagged and described as “missing”. Conflicts between branches consist of any conflicting genotype calls including scenarios in which one caller contributes a missing call. We defined the missingness rate per variant site as the number of missing genotypes at a site divided by the total number of genotypes at that site. We then calculated this missingness rate across all variant sites for all branches and provide the distributions. Finally, we provide the minor allele frequency distributions across all variant sites for each algorithm.

#### 2.10 Positive Predictive Value (PPV) estimated from De Novo Variant Calls

A set of predicted de novo variants was predicted based on the Broad Institute de novo filtering practices (supplementary methods) provided by the Hudson Alpha Genomic Services laboratory. A total of 55 variants from 31 samples were predicted de novo and validated as true positives upon resequencing with Sanger sequencing methods. A total of 15 variants (27.2%) were predicted de novo but did not validate with Sanger sequencing and were therefore classified as false positives. This set of laboratory validated true positive and false positive de novo variants then served as a benchmark to determine the positive predictive value (PPV) of CGES and
each branch algorithm. PPV was calculated as the ratio of true positives detected to the sum of true positives and false positives detected.

2.11 Availability of the CGES Pipeline to Investigators

There are three primary ingredients needed for the successful use of the CGES pipeline: 1) the user-supplied files necessary for analysis (i.e., raw BAMs, target BED file, and reference files) 2) the branch and CGES software, and 3) a computational infrastructure capable of handling the demands of the software. The input and reference files are user supplied and are routinely made available by NGS centers. There are multiple accessibility points for the CGES and branch software. First, we have made the enhanced version of these tools (e.g., CGES, FreeBayes, GATKv2.0, Atlas-SNP2) available in the public Galaxy toolshed (Blankenberg et al., 2010; Giardine et al., 2005; Goecks et al., 2010) so that community members can download the tools into their own respective Galaxy instances. We have also made the workflows (i.e., descriptions of the pipelines including various arguments used in the execution of the tools) available online (supplementary table 3). Second, the University of Chicago Computation Institute maintains and updates the pipelines under the Globus Genomics service offering (Madduri et al., 2013). Finally, the code for the CGES tool has been provided as a stand-alone command line tool (supplementary table 3). The computational infrastructure for the analysis performed in this manuscript was developed by the Globus Genomics initiative (http://www.globus.org/genomics/), led out of the Computation Institute (a joint institute between the University of Chicago and Argonne National Laboratory). The analysis of the autism trios described here was conducted on the Amazon Web Services public cloud. Investigators wishing to run the CGES software will require a local server capable of parallelization and analysis, or access to cloud computing space such as that offered by Amazon, Digital Ocean, or Azure.

3 RESULTS

3.1 CGES-QC Results

Quality control analysis and descriptive statistics of the consensus data set showed that the highest quality call set was obtained by using overlap of all callers together. Using the parameters for single callers described in supplementary materials we found that, as expected, the strict consensus of all calls made by CGES yielded the fewest total variant calls ($N_{CGES} = 129,706$; Table 1, Figure 2). CGES calls resulted in the highest $T_s/T_v$ ratio (2.91) (Table 1, Figure 3). CGES calls resulted in the lowest gMER (0.057%) (Table 2, Figure 3) across all genotypes and the lowest vMER (2.00%) (Supplementary Table 1, Figure 4). CGES results contained a total count of 17,121,192 genotypes representing 129,706 variant sites across 132 individuals, with 666,413 discordant genotypes resulting in CGES a discordance rate of 3.89%.

We also evaluated the distribution of calls across the MAF spectrum. We show spectra as empirical cumulative distributions in order to facilitate comparisons between CGES and the constituent branches (Supplementary Figure 3). We found that 82.9% of CGES consensus results were previously identified in the EVS project, and 78.0% of CGES consensus results were previously identified in the 1KG project (Table 1, Figure 5). CGES calls exhibited the highest proportion of variant rediscovery out of the total number of calls compared to any single caller data set.

During the course of our analyses we identified a sample with extreme deviation on the F-statistic ($F_{CGES} < -1.0$) suggesting that this sample showed extreme heterozygosity. Upon review of the branch call sets we found that the same sample deviated significantly from the rest of the samples according to every algorithm and showed evidence of extreme heterozygosity (Supplementary Figure 4). A review of laboratory records showed that this sample had been previously noted as possibly contaminated and contamination

![Fig. 2. Venn diagram of variant sites and their overlap between constituent call sets. CGES variants can be produced from the intersection of any two or three constituent sets, or from the union of all calls.](image1)

![Fig. 3. Transition/Transversion mutation ratio for different call sets. This ratio has been observed to lie between 3.0 and 3.5 for coding regions in the human genome (ref 1000 genomes as in poster). The ts/tv ratio was 2.91 for the CGES calling algorithm compared to 1.74 for GATK, 2.62 for Freebayes, and 2.12 for Atlas2.](image2)
Table 1. Set-Based Quality Control Results

<table>
<thead>
<tr>
<th>Call Set</th>
<th>Number of Variants</th>
<th>gMer</th>
<th>Ts/Tv</th>
<th>Exome Variant Server Rediscovery</th>
<th>1000 Genomes Rediscovery</th>
<th>PPV</th>
</tr>
</thead>
<tbody>
<tr>
<td>CGES</td>
<td>129,706</td>
<td>0.0572%</td>
<td>2.91</td>
<td>82.9%</td>
<td>78.1%</td>
<td>93.5%</td>
</tr>
<tr>
<td>Atlas-SNP2</td>
<td>214,149</td>
<td>0.280%</td>
<td>2.12</td>
<td>72.7%</td>
<td>69.0%</td>
<td>92.3%</td>
</tr>
<tr>
<td>FreeBayes</td>
<td>140,803</td>
<td>0.473%</td>
<td>2.62</td>
<td>78.9%</td>
<td>74.3%</td>
<td>81.1%</td>
</tr>
<tr>
<td>GATK v2.0</td>
<td>265,625</td>
<td>1.09%</td>
<td>1.74</td>
<td>61.3%</td>
<td>58.3%</td>
<td>84.4%</td>
</tr>
</tbody>
</table>

Fig. 4. The genotype Mendelian Error Rate (gMER). The gMER is calculated as the total number of MEs in a VCF file divided by the total number of genotypes with the potential for Mendelian inconsistency (i.e., offspring genotypes with parental genotypes known) in a VCF file. This measure describes the proportion of all offspring genotypes that are inconsistent with parental genotypes present in the VCF.

was subsequently confirmed using microsatellite markers. The contaminated sample was removed for the remainder of the analyses provided here, but is retained in supplementary figure 4 to illustrate the usefulness of the F-statistic as a QC measure.

3.2 Comparison of CGES Predicted and Laboratory Validated De Novo Calls

A Sanger sequencing validated set of de novo true positives and false positives was used to test the positive predictive value (PPV) of the CGES algorithm and its constituent branches. CGES demonstrated the highest PPV (93.5%), which was an improvement over constituent call sets (Table 1).

Fig. 5. Percentage of variants rediscovered using a 1000 Genomes reference set (a), and using an Exome Variant Server reference set (b). The rediscovery percentage is defined as the percentage of variants in a given set that are found in the reference. CGES produces a variant set with the highest percentage of previously discovered variants.

4 DISCUSSION

We have presented a novel variant calling approach based on an ensemble of variant calling algorithms, which we call Consensus Genotyper for Exome Sequencing (CGES). CGES employs a two-stage voting scheme among three algorithm implementations to identify variant sites and determine genotypes. In addition to presenting the consensus approach, we have described its application to real world exome data collected on a sample of autism trios and singletons. We provide project-based, sample-based, and variant-based quality metrics across all algorithms as well as an estimate of the positive predictive value of each algorithm and CGES. Finally, we provide a Galaxy-based implementation of CGES and its constituent parts. Taken together, the results show that Galaxy-CGES provides a robust, flexible, and user-friendly approach to exome sequence variant calling. Additionally, these results provide a strong rationale for further development of ensemble methodology in the analysis of NGS data.
The full CGES consensus-calling algorithm produced the highest quality output but the smallest number of genotypes: 129,706 SNVs in total in our example data. Leveraging the strengths of all callers produced a data set with the highest Tv/Tv ratio (2.91), the lowest vMER (1.52%), the lowest gMER (0.045%), the highest EVS rediscovery percentage (82.9%), the highest 1KG rediscovery percentage (78.0%) and the highest de novo PPV (93.5%). However, there is increased discordance for rare variants.

Consensus approaches for NGS variant detection can be particularly useful when the downstream analysis (i.e., rare variant TDT, pathway analysis, or de novo filtering) is reliant on a low false positive rate. However, there may be scenarios in which the preferred strategy is to maximize the rate of true positives even at the expense of a higher false positive rate: for example, when performing segregation analysis in large extended families. In that case, it may be more fruitful to use the union of all calls from all branches. Additionally, when identifying de novo variants one may wish to use the consensus of all calls in probands and the union of all calls in parents as an added stringency filter to reduce false positives.

It is important to stress that there is no one size fits all approach to sequencing analysis. The best approach for variant calling depends entirely on the type of data and the downstream analytic plans. As new methods are continually being developed it is our hope that this report in conjunction with other consensus efforts (Chapman, 2014) will help set the tone for an open discussion on the importance of unifying different approaches.

It is important to note the limitations of the analyses presented here. One important limitation is that the quality metrics from each branch are not directly comparable, as their optimization strategies differed. We optimized calling for each branch of CGES in order to reflect reasonable real-world parameter decisions and not for the sake of comparison among methods, which has been recently published (O’Rawe et al., 2013). For example, FreeBayes allowed us to set many parameters (supplementary methods) based on the raw data descriptive statistics and our previous sequencing experience. Atlas-SNP2, on the other hand, offered relatively fewer parameter options (supplementary methods). As best practices have been published for GATKV2.0, we used these guidelines verbatim. Ultimately the performance of each branch can differ dramatically based on parameters set by the user. Of course, the better the branch calls, the higher quality the final consensus calls will be. Additionally, it may be possible to use concordance between callers as a guiding metric when iterating to the optimal parameters for each branch.

In addition to providing the description of the pipeline and the resultant data, we have provided multiple accessibility modalities. The code for the CGES and CGES-QC algorithms is open-source and available through GitHub (supplementary table 3). For investigators who do not wish to invoke the command-line, we have provided CGES and its constituent branches in a user-friendly Galaxy environment. Finally, for researchers without institutional computational infrastructure, or simply for those who wish to outsource the computing but retain control over the scientific aspects of analysis, the pipeline is available and will be sustained through Globus Genomics (http://www.globus.org/genomics/).

ACKNOWLEDGEMENT

We wish to acknowledge the individuals who participated in research at the Autism Center for Excellence (University of Illinois, Chicago). We also wish to thank Steve Gutier (UIC) for assistance with data management, Bo Lui for early work on Galaxy wrappers, Kelley Moore (UIC) and Kathleen Hennessy (UIC) for expert technical assistance, Braden Boone (HAI), Jack Wimbish (HAI), and Shawn Levey (HAI) for sequencing and expert advice. This work was supported in part by P50 HD055751 (EHC), U19 GM61393 (NPC), P60 DK20595 (NPC), P50 MH04267 (NIC), R01 MH089482 (JSS), R24HL085343 (IF), by a Lever Award from the Chicago Biomedical Consortium, and by the U.S. Department of Energy under contract DE-AC02-06CH11357. We are grateful to Amazon Web Services, Inc., for an award of Amazon Web Services time that facilitated early experiments.

REFERENCES


