Workshop on Establishing a Central Resource of Data from Genome Sequencing Projects

Computational Requirements

Steve Sherry, Lisa Brooks, Paul Flicek, Anton Nekrutenko, Kenna Shaw, Heidi Sofia

High-density phased individual genotypes and haplotypes will become the standard basis for analysis in every sequencing project. The many software choices for alignment, variant detection, genotyping, and phasing will consolidate into community preferences in the next few years. Our consideration of the needs for computational resources covers data coordination, exchange, and access. We focus on unified presentations of data that could dramatically improve analysis across multiple sequencing projects.

Sequence data

The shape of next generation sequence (NGS) data

The world's corpus of human variation data can be imagined as a single global 2D table aligning all individual genotypes (in columns) by genome position (rows). Cell data are accessed by specifying a column and row index (e.g., sample ID and genome position). These indices are currently not normalized over all available data, making it impossible to perform global queries. International agreements could establish standard indices for individuals and genome positions, but both dimensions are heterogeneous and many technical details require consideration.

Columns (individuals) are naturally partitioned by the host data source, e.g., research archive, cloud repository, health care system, or company. Hosts operate within a framework of national law, regulations, ethical oversight, and organizational goals. Collectively, these requirements produce access policies and data access decisions that determine the accessibility of groups of columns (research studies or collections). Uniform principles of access would permit users to query across all sources and compile global query results.

Rows are ordered by genome position using a reference sequence coordinate system. Ideally, all data are mapped and organized in a standard order. The reference sequence needs a framework to describe alternative arrangements, unplaced sequence, and corrections/extensions.

Larger-scale patterns in the matrix: haplotypes and pedigrees

Kinship, the history of human demography, and natural selection have imprinted larger patterns onto this global matrix of genotypes. We identify patterns of allelic covariation as chromosome haplotypes. Each haplotype typically spans thousands to several hundred thousand bases. Columns can be organized by pedigree and ethnicity to group individuals

by genetic similarity and shared genetic history.

Sequence data in each cell likely have a triangular structure with the fundamental information axes of sequence, genotype, and haplotype. The axis of sequence would include multiple sequence reads at each base, sequence quality scores, alignment properties, orientation, and other technical features of the experiment in cSRA/CRAM format. The axis of genotype would include genotype likelihood, error mode, and functional annotation. Where possible, genotypes would be phased into the third axis of chromosome haplotypes, including frequency in reference samples and measures of quality/support.

Storage costs

There are costs to storing extensive data, and data producers will make different decisions about the data types they will store online (disk) and nearline (e.g., tape). Nearline hardware is about 50% of the cost of online storage and can re-provision archived data to disk for computation in minutes. This storage format is useful for sequence data that are important to retain but are rarely used.

What is needed?

1. An assurance that we capture all the analysis metadata correctly in the Variant Call Format (VCF), including analysis results and confidence of variant calls. Which regions of the genome are ignored in the VCF because calls cannot be made confidently? We need a standard and more expressive way to describe where the "known unknowns" are than we currently have.
2. SNP calling is still an issue for small research groups. Keep the primary sequence data.
3. Structural variant calling is far from mature, so the primary sequence data still need to be kept to allow SVs to be called when the methods are reliable.

Phenotype data

Phenotype data cover an immense range of measures, traits, responses, treatments, and medical conditions. While values are carefully defined and collected within a study, little coordination or standard definitions exist across studies. Projects like PhenX address this problem by constructing exact guidelines and coding standards for important measures. Data sources differ in the format and detail of term definitions, with solutions ranging from narrative text and study protocols to coded ontology terms for standard classification.

What is needed?

Sensible standards for basic phenotype data, including file formats that can be better harmonized across projects, would lower the barrier to data integration across studies. Too many efforts have tried to solve the entire phenotype ontology problem, and the absence of a general solution indicates that this is impossible or will take a long time, so solving smaller chunks is the way forward.

It would be useful to get data on the differences in cost and quality of data analysis when 1) using older phenotyped samples that need data harmonization, 2) rephenotyping to a common standard, or 3) using new samples and new phenotyping. Such data would help us estimate the percentage of existing data that would require exceptional treatment. Rephenotyping is costly, and should be considered only after community standards for the metrics are developed. We could standardize metrics for data types that we are not currently collecting at large scale, such as environmental or pharmaceutical data.

Queries

Most queries are likely to be of the form "give me data from specified individuals for specified genome regions that have certain phenotypes". Execution times for queries will vary by data source, the level of detail, and storage strategies. VCF-based genotype data files are relatively small and can be served rapidly, even as inputs to real-time analyses or interactive tasks. Performance for sequence data can be optimized with sufficient hardware and attention to engineering details. Simpler systems may be slower, which might be acceptable for many tasks that could be queued for remote execution.

What is needed?

A standard message protocol would help engineers build systems that return results in standard forms. The protocol should be able to specify individuals, genome positions, genotypes, haplotypes, and phenotypes in standard formats.

Open vs. controlled data access management

What is needed?

Data that have personal identifiers (PHI) require controlled access. Groups hosting data should provide access through coordinated and uniform methods. NIH leverages the eRA Commons database to provide access control after Data Access Committees (DACs) approve researchers for data access. This works for NIH-run systems, but does not work for data that will be stored and distributed outside the US, such as at the EBI. One solution would be to provide the list of approved researchers to other databases, as a central approach to approving data access.

Computational resources

Generally, an ideal data source would feature:

- efficient, cost-effective data storage;
- descriptions of the types of data for each project;
- access control systems to distribute data according to use conditions;
- centralized access mechanisms for collaborating analysts to perform data QC and exchange pre-publication results in a secure fashion;

- easy-to-use systems to make queries and download data;
- systems to allow queries across the data from multiple projects;
- data components that are archived as soon as possible;
- sufficient resources for timely analyses at reasonable costs;
- best practices for tools, formats, and analysis protocols.

Some classes of data can be summarized from individual genotypes and phenotype measures. These summary data sets could be distributed publicly or with a set of use conditions aligned with their risk to participant privacy and confidentiality. Summary data sets may present fewer risks compared to individual-level data sets. Compute resources placed next to the data could permit analyses to be performed in a firewalled area. The performance parameters of such a service would need clear specification: what types of analyses are permitted, who maintains the software and verifies correct execution on the hardware, who has access, and who is responsible for security of the system.

Archives can offer data reduction services such as compression, filtering, and slicing so users can download and compute locally with modest resources. Many questions involve genome regions much smaller than the full set of genomic data per sample. Questions at the largest scales require investment in more local hardware (for intensive use) or access to cloud environments (for occasional use).

Cloud solutions

NCI is using the cloud for data distribution (this works very well), running well-established pipelines (this is challenging to set up, but works), and exploratory computational work (this is still very difficult at large scale). In the latter two cases, it is much more efficient for users to work in-house. The cloud infrastructure is good for some things but not others.

The cloud is being explored for alignment, SNP calling, and data analysis. The cloud is likely to be important for large genome computations and data integration, but this area still requires development. EBI has been migrating parts of the Ensembl gene annotation pipeline to the Amazon Web Services cloud, where the infrastructure is most developed, and to other providers, which often have little support for the workflows that EBI uses. Cloud systems are well behind a standard compute cluster for large-scale internode communication and provisioning of large-memory machines, and the work required to adapt existing algorithms is significant. EBI and NIH are developing trials of cloud services. A set of validated tools for alignment, SNP calling, and analysis may be assembled for smaller projects that do not require extensive engineering expertise.

Clouds can charge for data transfers into the system, monthly data storage, and computation. The storage and transfer charges are generally significant for large data sets and the strategy makes sense only if many users would compute on the data. Generally, we cannot afford to keep a copy of SRA in every commercial cloud at taxpayer expense, since the data set is simply too large. Most problems will require only a fraction of SRA content,

although that fraction will differ among studies. Data access could be provided in several ways: 1) agencies, advocate groups, etc. could pay for storage and community access to common subsets of data to promote research in their area; 2) individual investigators could search the archives, identify datasets relevant to their questions, and transfer the data to the cloud at their expense; or 3) individual investigators could pay to transfer a dataset into a common area on the cloud and then get proportional rebates as other investigators use the same data.

All of the models that use data in the cloud are not applicable to dbGaP-protected data with the current policy of personalized data encryption for each approved user. However, it could work well if the shared data were encrypted once, with decryption keys issued to approved users, as EGA does successfully.

CPU costs associated with typical bioinformatics compute jobs are currently small. There are significant costs, however, to develop and deploy stable and cloud-friendly applications that are well written, tested, and maintained. There are additional costs to develop the platform layer discussed in the central analysis server position paper.

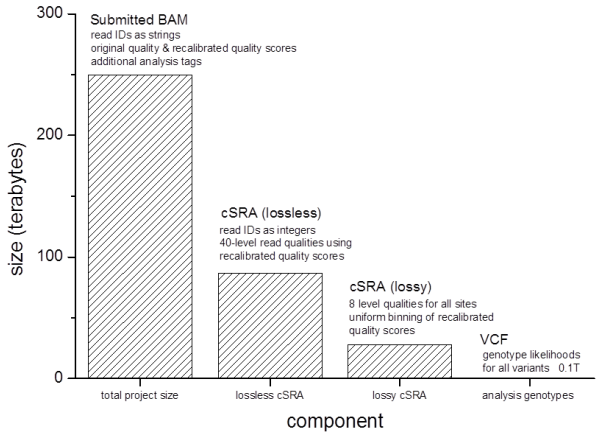
What is needed?

Development in this area is very important as many users simply can't download large data sets but have analyses to run and methods to test.

C apacity requirements for NGS projects

Project sizes vary for sequences with quality scores (full, reduced, none) and mate pair data, alignments and mappings (with overhead for secondary placements and unmapped sequence), variants, genotypes, phased haplotypes, and coverage graphs.

Here is an example of data for the 1000 Genomes Project as BAM, archive compressed cSRA, and VCF. The difference between the largest and smallest versions of the data is 3 orders of magnitude.



Questions about capacity planning

How should archive requirements (data amount, fidelity, persistence, accuracy) be predicted?

How much experience with large file management is necessary? This includes sufficient bandwidth to move data in and out, high performance infrastructure/network to manipulate data internally, backup/fault tolerance for hardware failures, familiarity with handshake protocols with sequencing centers, and sample/consent tracking on all data.

Resource requirements

The data access position papers describe possibilities for data access through certified users, analysis services near repositories or in the cloud, and models that permit private data to be integrated with repository data during analysis. It is impossible to estimate costs without having system requirements. While hardware, software, staff, and institutional requirements are outlined below, discussions should establish the breadth of needed functions and estimate potential interest as system load. Certified user access to a research commons would minimally include tasks to establish the partition, create a certified user authentication system, and deliver data slices. Adding computational services would require creating an interface website, implementing permitted use policies, supporting standard execution interfaces, and developing a platform for third-party application development. For analysis services in the cloud, a system would need to be established between the repositories and the servers to deliver analysis results or the individual-level data for computation. Finally, if users could upload private data for their analysis sessions, then additional tasks involving data management, security, and virtualized application hosts likely would be necessary. An ambitious program to provide all of these services in a robust, high-performance manner would likely need two to three dozen software developers.

Hardware: There is a significant cost to establish and maintain a large, local computing center. These environments cannot be replicated at every research institution. What are the possible alternatives?

- Lease the hardware (cloud). Does the cost advantage of clouds outweigh the challenges to computation for occasional users?
- Reduce the size of the data sets by data compression.
- Establish public resources that researchers can use.

Hardware costs include

- The CPUs.
- The memory capacity.
- The disk storage capacity.
- The power consumption.

Software: There are several choices for analysis tools that affect the cost, quality, and reproducibility of results. Software may be developed internally, be contributed to a community package, or be commercial. There are differences in reliability, fault tolerance, performance, and cost of maintenance between academic and production-grade tools.

Human: Costs include the number of dedicated staff and their expertise in dataflow, analysis, and QC by data type/product. There are additional costs for helpdesk service functions, collaboration/DCC, specialized analysis, and archive submission/operation.

Institutional: Requirements include managing access to data sets; security; restrictions/directives for industry collaboration; IP policies; operating regulations,

policies, and laws; and international coordination.

What is needed?

Depending on the design of an effort, the data (NGS, analysis products) may be compatible with NCBI/dbGaP and EBI/EGA. Are there common data formats to archive data? Are there common formats to exchange data with trusted partners or Data Coordination Centers (e.g., ICGC, CGHUB, deCODE)? What about data at consortium sites or at clinical/translational sites?

Data analysis tools

What tools will be needed to analyze the data? Some standard analyses (variant calling and imputation, loss-of-function variants, disease associations, gene x gene and gene x environment interactions) could be provided centrally. These results could be computed by the data host or submitted by a data coordination center. There are at least two analysis communities: the project analysts who analyze their data, and the users who ask other questions of the data. Deeper analyses are likely to be done locally by the first group. The second group will include sophisticated users facile with local computation and smaller groups who need guidance on best practices and tools to ask their questions using shared computing spaces like the cloud. The viability of particular pipelines or tools will depend on their ability to be distributed beyond their initial development environment.

The space of analysis tools will continue to outpace the growth in sequence data. We expect, however, that the space of computational tasks follows a power law and 80% of tasks could be accommodated with a manageably small list of tools. Can the community establish a subset of tools for pilot Research Commons efforts? Some questions to help improve the potential for long term success include:

- Is re-engineering needed to meet performance expectations? If so, are the resources available?
- Are resources committed to support a tool's lifecycle maintenance (upgrades, security, etc.)?
- Is there a user base for the tool? What is the environment to bring the tool online (an institutional investment, protocol requirement, open source community project, commercial)?
- Is there sufficient convergence on file formats, data quality, and exchange standards for the tool to work in an analysis environment without middleware conversion steps?

Large projects create best practices, but few groups follow them, so we need to provide tools in a standard harness. Steps such as recalibration, realignment, and variant filtering seem to have no widely adopted standards. Providing robust tools in standard analysis pipelines would provide more standard ways of doing analyses, which may be especially useful for small groups.

Appendix 1.

Tools: A partial list of tasks that require computational resources.

Sequence read alignment/realignment
Quality recalibration
De novo assembly
Variant detection (by variant class)
Variant integration (across classes)
Genotype calling
Cryptic relatedness detection
Phasing
Haplotype estimation within sample
Haplotype estimation from pooled data
Imputation of variants
Somatic mutation detection
De novo mutation detection
Functional classification
Functional prediction (analytical validity)
Network/pathway analysis
Genotype and phenotype associations
Clinical consequence (phenotypes, clinical significance, and clinical utility)
Data visualization (browsers, other modes)
Translational target prioritization