## Johns Hopkins Genomics Resources and Environment

### A. Laboratory:

**Johns Hopkins Genomics:** In preparation for and in response to the ever-expanding variety of genomic services required to support human genetics research, we established Johns Hopkins Genomics (JHG, jhgenomics.jhmi.edu), an initiative of the JHU School of Medicine to combine its research and clinical genomics expertise in a joint effort between the McKusick-Nathans Department of Genetic Medicine (DGM) and the Department of Pathology. The formation of JHG is a multimillion-dollar investment for the School of Medicine. A core group of more than 100 faculty and staff occupy the 2nd floor (25,000 square feet) of the 1812 Ashland building in the East Baltimore Science + Technology Park adjacent to the Johns Hopkins medical school campus. Wet laboratories occupy ~60% of the new floor and an open office suite ~40% of the total square footage.

This arrangement is a true melding of groups; all laboratory space is shared and CLIA compliant, and the open office space configuration has fostered collaboration among the clinical, research, informatics, and analytic faculty and staff (see Appendix 8 for space diagram). JHG consists of the staff and expertise of two major research labs, Center for Inherited Disease Research (CIDR) and the Genetic Resources Core Facility (GRCF), and three clinical genomics labs: the DNA Diagnostic Laboratory (DDL) for inherited germline diseases (in operation since 1979), the Pathology Molecular Diagnostics Laboratory (MDL) for cancer genomics (in operation since 1991), and the Cytogenomics Laboratory (Cyto, in operation since 1986). The CIDR/GRCF infrastructure allows for the processing of large sample numbers and the equipment and informatics expertise required for whole exome sequencing (WES) and whole genome sequencing (WGS) services. In total, the combined JHG laboratories have over 150 faculty and staff providing a variety of services, the most relevant of which are CLIA/CAP sequencing and genotyping for inherited and somatic mutations. Both existing diagnostic laboratories have a rich history of supporting clinical research studies and, by combining with CIDR and the GRCF, will be enabled to support much larger scale clinical translational research studies as required. The MDL has close ties to Kimmel Cancer Center researchers and has supported over 65 clinical translational research studies.

**1812 Ashland:** The 1812 Ashland laboratory space is approved for BioSafety Level 2 work and supports specimen intake and quality control, DNA and RNA isolations, tissue dissections, Illumina Infinium, Sequenom-based research and clinical genotyping, Sanger sequencing, and Ion S5, Illumina MiSeq, iSeq, NovaSeq6000, and NovaSeqXPlus research and clinical next-generation sequencing. In addition, long read sequencing is available on the Oxford Nanopore GridION or PromethION platform. All laboratory space is CLIA-compliant, physically separated and organized into color-coded levels, stratified down to the level needed to segregate the sequential common primer PCR steps during the sequencing library prep workflow. The blue labs are all pre-PCR/pre-amplification lab space for both genotyping and sequencing. The green labs are post-PCR/post- amplification/PCR (1) lab space for genotyping and sequencing. The orange lab is for the PCR (2) of library capture. The red labs are for post-PCR steps of sequencing. The MDL and DDL are all CAP- and CLIA-certified laboratories with a tremendous variety of clinical and research assays in use or being developed. Combined, they currently perform over 10,000 clinical tests per year and will be expanding their clinical exome, whole genome, and large panel based NGS services. Please see the JHG Equipment list for a complete list of available resources.

**Other Laboratory Spaces (Blalock, PCTB, Rangos, Nelson/Halsted):** The GRCF units that are not located at 1812 Ashland occupy ~6,000 square feet of laboratory and office space on the 10th floor of the Blalock building, ~2,000 square feet in the Preclinical Teaching Building (PCTB), and ~2,000 square feet of LN2 repository space in the Rangos Building. The JHU Cell Center tissue culture laboratory is a CAP-accredited facility on the Johns Hopkins Hospital campus. The facility includes 3 biological safety cabinets and 6 ThermoFisher humidified incubators in 2 negative air pressure rooms. Cells are quantified and quality control monitored by way of 3 Vi-Cell XRs. The Rees Environmental Monitoring System continually monitors the cell culture conditions. Single cell genomics is accomplished 10X Genomics Chromium platform. The Johns Hopkins BioBank is a CAP-accredited facility that houses 9 Taylor Warton LABS LN2 vapor phase freezers (40K and 80K), 3 MVE Vario LN2 vapor phase storage unit, 2 MVE High Efficiency LN2 vapor phase units, 2 –80C chest freezers, 1 –20C chest freezer, and 6 –80C upright freezers. For cryogenic transport of biospecimens, the Biorepository utilizes 2 portable LabRep Co. liquid nitrogen Cryocarts. Freezer security and integrity are maintained through the Rees Environmental Monitoring System, with an automated monitor and 24/7 surveillance. Repository inventories are maintained through the BSI Systems (a product of Information Management Systems, IMS) database. The JHU Nucleic Acid Technology lab (JHU-NAT) houses equipment for both STR and SNP genotyping, nucleic acid extraction, cDNA synthesis, RNAseq library prep, PCR support for Pyrosequencing and Sanger Sequencing, Bisulfite Conversion, Cell Line Authentication, and Mycoplasma detection services. Equipment includes a Tecan Freedom Evo150 liquid handler, Taqman7900HT, two AirClean 600 PCR workstation (BioExpress), four Veriti (Applied Biosystems) thermocyclers, one C1000 Touch thermocycler (Bio-Rad), two SureCycler 8800 (Agilent Technologies) thermocyclers, a Perkin Elmer chemagic MSM I Magnetic Particle Separator, a QIAcube Connect MDx that provides automated processing of Qiagen spin columns, a S2 Covaris Focused-ultrasonicator for nucleic acid extraction on FFFPE samples, Omni International Bead Ruptor 12, an Integra Biosciences VIAFLO 96, and a CheckScanner for reading mycoplasma MycoDtect™ microarrays. The lab also has three -80C freezers (24/7 monitored by the Rees Environmental Monitoring System) and two –20C freezers, a variety of electrophoresis equipment, centrifuges, a Nanodrop spectrophotometer, a Qubit 4 fluorometer, a Fragment

Analyzer (Agilent) for nucleic acid quality and sizing analysis, and a Synergy water purification system. The JHU DNA Services laboratory on Blalock 10 houses equipment for qPCR, digital PCR and medium throughput genotyping (2 QuantStudio 12K Flex instruments, a QIAgility liquid handler, two 8-plate QIAcuity dPCR instruments), Sanger sequencing (3730XL), and Pyrosequencing (Pyromark Q48).

The Cytogenetics Lab occupies ~10,000 square feet of space; located 2nd floor, Nelson/Halsted. Facilities are available for BioSafety Level 2 cell culture, microscopy, classical cytogenetics, FISH, DNA SNP microarrays and Optical Genome Mapping. Lecia CytoVison analysis software and other design and analysis applications are available on a secure, networked lab information system. The lab is CAP and CLIA certified, performing over 6200 tests per year.

Please see document JHG_equipment_Lab_IT_2024.xlsx for a full list of equipment.

## B. Informatics:

The informatics resources of Johns Hopkins Genomics are primarily comprised of the robust capacity of the existing JH CIDR group, supplemented by the clinical and research bioinformatics resources of the MDL and the DDL. All of these resources are currently located in the 1812 building.

**B.1. CIDR Informatics:** At JHU, CIDR has continuously expanded and improved informatics infrastructure, including data movement and storage, computational and network capacity, and server room space, cooling, and power. The dramatic increase in CIDR services and sample numbers over the last few years produced concomitant increases in data production rates and volume, requiring rapid expansion of computing and storage equipment. The server room in 1812 Ashland is electronically secured, with full environmental controls and fire suppression. Over 170kVA of UPS capacity provides continuously filtered power and, in the event of external power outages, several minutes of bridging power until the building generators activate.

The compute cluster consists of 36 fast compute nodes plus several many-core large-memory servers, totaling over 615 compute cores, 5TB RAM, and 52TB of local storage. It is administered via Bright Cluster Manager software, enabling superior management of computing jobs and resource distribution, along with bioinformatics software versioning for research use and version lockdown for the clinical informatics pipelines. A Microsoft Hyper-V virtualization platform consists of three Dell R740XD servers totaling 120 cores and 2.3 TB of RAM, running nearly 100 virtual machines. Storage includes 35.75 TB of mixed SSD and 7.2K NLSAS drives. Our production network-attached storage (NAS) hosts over 2.5 petabytes capacity. Included among these totals is a separate CAP-certified NGS bioinformatics analysis platform for the DDL, currently consisting of four Dell C6220 compute nodes (64 cores) and four 72NL storage nodes (260TB disk). A Quantum i500 tape platform running CommVault Sympana backup software, consisting of ten ultra-fast LTO-5 tape drives and 211 tape slots in a robotic library, serves all these systems.

The local network is built around six Cisco 10-gigabit optical core switches (140 ports total) and 13 1-gigabit copper edge switches, providing massive data movement capability protected by a Palo Alto Enterprise controlled Firewall. Our primary Internet connection is via multiple gigabit connections to the JHU network supplemented by a 10-gigabit connection (40Gb uplink) to the JHU research network for institutional data movement; JHU also has a 10-gigabit connection to Internet2. Currently, data distribution is via secure connections from an Aspera cluster at up to 300 Mbps per session.

The JHU research network also provides high speed access to the Advanced Research Computing at Hopkins (ARCH) on the Hopkins Bayview campus. This facility offers a massive compute cluster of 34,128 cores, including 24 GPU nodes, 13 petabytes storage with 8+PB in a fast IBM GPFS parallel filesystem. ARCH is readily accessible for our use should the need arise for processing at a scale beyond what our in-house infrastructure can accommodate, and/or for pilot projects so as not to impact production infrastructure.

The JH Research Information Technology also provides DISCOVERY Cluster, a NIST SP 800-171 compliant compute infrastructure. This consists of 1388 nodes with 9,120 cores, including 20 GPU nodes, 15BP storage with 5P WEKA SSD. Along with the the DISCOVERY Cluster are the Secure Analytic Framework Environment for Research (SAFER) virtualized desktops. These are also NIST SP 800-171 compliant systems with the following specifications: Windows 11 Small (8 CPU, 32GB RAM), Windows 11 Medium (16 CPU, 64GB RAM), Windows 11 Large 32 CPU, 128GB RAM, and Windows 11 Small GPU (4 CPU, 28GB RAM, 1x16 GPU).

**B.2. Software Development:** Custom software development, which is key to the successful function and improvement of our complex, high-throughput informatics environment, is conducted by a team of six CIDR software developers, guided by an IT Project Manager with sixteen years of experience. The rest of the team has fifty years of experience at CIDR among them. In addition to software development experience, two of the software development staff have previous experience working in the CIDR genotyping and sequencing labs. This experience grants these developers a better understanding of both lab protocols and software usage, and of the underlying biological processes. Large software projects are typically carried out by two or three developers, while small projects are developed individually with input from other developers at critical junctures. The team is highly collaborative, and cross-pollination of developers and ideas throughout the group is frequent.

Software developers at CIDR employ an agile methodology with continuous delivery, which facilitates close collaboration with and rapid feedback from colleagues in other groups at CIDR, particularly our bioinformaticians and statistical geneticists, who often create and refine early versions of software that are later re-engineered for production use. Because other staff have access to software early in the development process, they can make feature requests more often and more easily validate that software provides the desired functionality. The adoption of an agile methodology has reaped numerous other benefits: during the requirements gathering process, small face-to-face meetings between developers and their collaborators allow for effective communication, and removes roadblocks caused by a lengthy and formal specifications gathering process.

Software solutions are created using numerous technologies and languages. The most used programming language at CIDR is Java, with other production code written in Python, JavaScript, Bash, C#, C++, Scala, R and SAS. Many applications and pipelines are written as Docker containers, stored in a local Docker repository, and run on Singularity. Among the technologies used for various projects are JavaFX for GUI programming, Django and AngularJS for web development, Hibernate as an object-relational mapping (ORM) framework, Gradle for dependency management, Hikari for JDBC connection management, Jackson and XStream for XML parsing, JUnit and Mockito for unit testing, RabbitMQ for message passing, SL4J with Logback for application logging, and Apache Commons and Vavr for additional utilities. All CIDR-developed software uses MySQL as a database backend, which is maintained by dedicated database administration staff and all custom CIDR software is version-controlled using git via a cloud-based BitBucket instance, or in the case, of open-source tools, outward facing GitHub repositories. Software requests and emergent issues are tracked via Jira. Many JHG software projects have been presented at various meetings, including the annual American Society of Human Genetics (ASHG) conference.

In addition to our major in-house software applications, there are dozens of smaller applications written by our developers. All software is developed using third-party libraries, such as those from the Apache Foundation, as well as with libraries of in-house developer tools. These developer tools include many GUI-based utilities, encryption and decryption tools, e-mail and communication utilities, and a framework for interacting with traditional SQL databases fluidly. These developer tools, combined with the above shared practices, constitute a unified environment for the development and use of CIDR custom software, called CERUS (CIDR Ecosystem for Relationally Unified Software). This combined ecosystem simplifies discussions of current practices, allows for clear planning when existing tools or practices are replaced, facilitates collaboration between developers on different projects, and simplifies system analysis for regulatory and reporting purposes, such as for CAP inspection or Section 508 reporting. CERUS includes several APIs to allow for interoperability of major systems, detailed below, and including Phoenix, Cerberus, and LastCall. A detailed diagram of the most common software workflow for CIDR projects is presented as Figure 22.

## B.3. Software Systems (in alphabetical order)

### B.3.a CIDR- and JHG-Authored Software

**Bioinformatics Applications Manager (BAM)**: staff typically access our information systems through Web or Java client-based graphical user interfaces created by the software development team. Most custom software is executed via the Bioinformatics Applications Manager. Written using Java and JavaFX, the BAM provides a single consistent desktop GUI for many applications, regardless of programming language or other constraints. The BAM features search functionality and a "favorites" area for each user to facilitate ease of use and allows for deprecation of older applications. BAM-hosted applications include most of the GUI-based software listed here, as well as:

- **Barcode Printer Tools:** This suite of tools allows for interaction with a set of barcode printers throughout the lab, with the ability to use varying label sizes and printer locations from anywhere within the JH Genomics infrastructure. While primarily used to upload and print sets of barcodes, custom printing is also possible. Users are also provided with calibration tools to make sure that new rolls of labels and new label sizes are properly aligned for each printer. Barcodes to-be-printed are sent to a central server and then jobs are sent to individual printers.
- **Container Problem Mapper:** The Container Problem Mapper generates visual representations of sample problems and their resolutions, based on the original sample locations as sent by the PI. These visualizations are annotated with problem type, sample sex, and other data as desired. This tool allows PIs to identify issues across sets of samples or caused by specific sample origins, such as plating site or extraction method.
- **Data Manipulation Utilities:** These tools allow basic data manipulation: the Report Munger combines multiple similar reports, considering possible structural differences and allowing the user various options for navigating those differences; the Column Coordinator rearranges columnar data as specified by the user, via a provided order and metadata about the files, such as delimiter and header length.

**Cerberus LIMS:** The expansion of research and clinical genomic services from collaboration with other JH Genomics members has led to the in-house development and production use of a Laboratory Information Management System (LIMS) called Cerberus. The aims of this project were to add flexibility, give users more operational control and support better code organization and reuse. This LIMS

supports the breadth of current and future protocols and contains tools to reduce disruption to lab processing by limiting database schema complexity and create a catalogue of reusable graphical user interfaces (GUI) to flexibly build and extend workflows for any protocol driven technology. Cerberus uses Java and JavaFX language features to build the core framework GUI and database integration utilities. XStream APIs are used to marshal, and un-marshal user-submitted data from Java objects to XML that are inserted into a MySQL database. Using XML to represent java object submissions of user data allows for fields to be added and removed from the GUI without requiring changes to the backend MySQL database schema. The Cerberus framework is organized into a single source package to contain reusable GUI tools and database interaction utilities and multiple sub project packages to delineate between service specific code (i.e., Genotyping, Sequencing, Clinical, LIMS Administration and Reporting). Cerberus provides an API for use by other CERUS applications and uses the Phoenix, LastCall and Epic APIs to retrieve information during lab processing and report generation.

**Data Management Tools:** To facilitate management of the release of sequencing and genotyping data, developers created BAM-hosted applications to aid lab management in filtering, moving, copying, and deleting large quantities of files of diverse types. These applications feature a custom set of file manipulation modes that have been designed around the needs of sequencing lab managers. To match the dynamic requirements of lab data interpreters, the application's performance parameters are easily altered, with custom modes available to users. Via direct use of the Qumulo API, FileWatcher, an automated file system monitoring system, alerts users to unexpected changes across the files system via several custom alert types. Release data and raw data are archived to Azure for prescribed periods of time via BlobLobber, an archival blob management tool; these data are removed from the local file system via a corresponding tool, BlobLopper, following integrity checks on data archived by BlobLobber.

**Genotyping Data Release Tools:** The CIDR genotyping pipeline relies on tools for the generation of PLINK files and final genotyping reports, developed at CIDR and using the bpm and gtc parsers developed for LastCall and described below. These tools are used for generating project release data, having been developed to address both capacity and speed problems that arose from the use of Illumina's GenomeStudio for large projects. In head-to-head tests, these tools were 3-10 times faster than GenomeStudio for generating release files and allow for faster turnaround on large genotyping projects. In addition to these data generation tools, genetic analysts perform integrity checks on genotyping release data using the Final Genotyping Report Release Check and rename files using PI-preferred nomenclature using the Final Genotyping Reports Renamer. Control concordance for all genotyping project controls is performed via Calcordance, a tool that uses serialized compressed genotyping data from the 1000 genotypes project to quickly perform comparisons across all available genotypes. All genotyping data release tools described are hosted on the BAM.

**Known Variant Database:** The Known Variant Database (KVD) is a GUI-based tool developed by CIDR staff for use by the DDL for tracking recorded variants across all clinical sequencing performed by the DDL, as well as the ACMG classification of those variants. The KVD is built as a highly flexible GUI tool that allows users to specify detailed variant interpretation information, as well as to upload a variety of supporting information for the recorded classification. In line with clinical requirements, the KVD provides robust auditing of all activities and restricts all actions based on user roles. The KVD pulls data from the JHU Epic EHR systems and provides inputs to Emedgene and stores information provided by Emedgene.

**LastCall:** To allow for rapid generation and evaluation of genotypes without manual intervention, CIDR genotyping projects have relied on a succession of automated genotyping analysis pipelines. The most recent of these, LastCall, offers significant improvements in software architecture and design, and is built using Illumina Array Analysis Platform (IAAP), a Linux-based genotype calling tool, which allowed developers to transition from a Windows-based calling platform to the flexible, reliable, and expansible computation environment offered by a Linux-based implementation. LastCall is written using an actor-model, implemented using Akka, a Scala-based actor framework. Under this model, a variety of genotyping pipelines are defined, based on both project needs and on project stages. These include pipelines for initial calling of genotypes, initial methylation analysis, end-of-project genotype reanalysis, and the inclusion of optional steps, such as the generation of VCF files corresponding to called genotypes using highly tuned Picard commands and external QC analysis of those VCFs. Each pipeline step is independently executed via its own corresponding service and its associated actor model. Failed runs are automatically resubmitted, with a configurable number of resubmissions. In the case that repeated resubmission does not correct failure, lab staff are immediately notified via an automated message service integrated with Microsoft Teams, with email and Slack notifications available as alternative communication methods. By executing each pipeline step as a discrete step in a massively parallel actor-based infrastructure, available computational resources are fully saturated, leading to significantly faster genotyping analysis turnaround times, with no data generation bottlenecks preventing lab staff from reviewing generated data. Additionally, due to the implementation via a cluster-aware actor framework in a Linux environment, LastCall can burst into any available Linux servers in the case of short-term increases in genotyping data, or in the case of project-release reanalysis for extremely large projects. Interoperability with other projects in the CIDR Software Ecosystem (CERUS), is performed via the LastCall API, from which QC data and data location information are available. These are used for the generation of QC reports via the Cerberus LIMS and for release pipelines. Finally, LastCall makes use

of custom-implemented rapid binary file parsers for Illumina gtc (genotype call) files and bpm (beadpool manifest) files, making these parsers available for use in other CIDR software, such as the PLINK and genotyping final report files mentioned above.

**PhenoDB Ecosystem:** The CIDR Software development team maintains and augments the websites that make up the PhenoDB ecosystem, consisting of three tools: PhenoDB, GeneMatcher, and VariantMatcher. PhenoDB is a freely accessible website that allows researchers to store standardized phenotypic information, diagnosis, and pedigree data and then run analyses on VCF files from individuals, families, or cohorts with suspected Mendelian disease. GeneMatcher is a freely accessible web site developed to enable connections between patients, their families, clinicians and researchers from around the world who share an interest in the same gene or genes. The principal goal for making GeneMatcher available is to help solve 'unsolved' exomes. This may be done with cases from research or clinical sources. The tool allows individuals to post a gene (or genes) of interest and connects individuals who post the same gene. GeneMatcher is part of the MatchMaker exchange. Finally, the VariantMatcher tool enables public sharing of variant-level and phenotypic data from whole exome and genome sequencing. VariantMatcher is designed to enable connections between patients, their families, clinicians and researchers from around the world who share an interest in the same variant or variants. The principal goal for making VariantMatcher available is to help solve 'unsolved' exomes. This may be done with cases from research or clinical sources. To comply with patient privacy and security regulations, site users must register and be approved by site administrators.

**Phoenix LIMS:** Phoenix is a sample, problem and project tracking system developed in-house by the CIDR software development team. Designed to accommodate the volume and complexity of sample management for SNP genotyping and next-generation sequencing work, Phoenix has been used in production since 2014. Phoenix tools and workflows include sample accession, plate map creation, problem handling, project status tracking, and sample rearraying. An average of fifty distinct tools are used each month by thirty users across the CIDR and JHG organizations, including lab managers, lab technologists, project managers and genetic analysts. A web portal, PhoenixWeb, allowing investigators to communicate with Phoenix to upload files, check on the progress of their projects, and respond to problem reports is actively used by both CIDR project investigators and NIH staff. Phoenix provides flexibility and sophisticated problem handling that enables the delivery of the high-throughput, high-quality service that makes JHG a unique national resource to the life science research community. The Phoenix system also serves data to several other CERUS systems via an API, including many QC report generating applications, LIMS systems, and the genotyping data generation pipeline. Phoenix and PhoenixWeb workflows and functionality are further discussed throughout Section III.B.

**Sequencing Data Generation and Analysis Tools:** In addition to the sequencing analysis pipelines described in detail in Section III.B, the CIDR software team provides several tools ancillary to sequencing data generation, including AnnovarWrangler and AutoDemux. AnnovarWrangler allows users to perform annotations using ANNOVAR against a directory of VCF files and against any number of annotation databases (currently several dozen) via the use of ANNOVAR tables in conjunction with a set of pre- and post-analysis steps that regularize input databases, vcf files, and output data to address several common issues that arise from the use of ANNOVAR, such as duplicated database identifiers, mismatched columnar data, and ambiguous output identifiers. AutoDemux is a service that automatically demultiplexes sequencing runs as they come off the NovaSeq 6000 and notifies lab and bioinformatics staff that demultiplexing is complete, as well as of any errors that may occur.

**Sequencing Read Archive (SRA) Submission:** Many sequencing projects require submission to SRA at completion. This submission requires the creation of several comprehensive XML files prior to data transmission. Working directly with NCBI SRA staff, JHG developers have created an easy-to-use, wizard-based GUI tool for the creation of these files, which guides users through providing the required information via a simple drag-and-drop interface. These XML files are populated via data sources such as CRAM files, Checksum files and Sequencing QC Reports. The XML population utilities are directly based on the XML schema provided by the SRA. The SRA submission application has expanded to include additional data repositories, and now includes the creation of excel metadata files needed for other raw data repositories, such as CDS, the Cancer Data Service.

### B.3.b Integrated Third-Party Software Systems

**CIDR Troubleshooter:** The CIDR Troubleshooter is an electronic forum, based on phpBB and customized at CIDR, used for documenting equipment maintenance and service calls. Each new lab instrument is entered into the CIDR Troubleshooter upon arrival. Any subsequent service or problem with the piece of equipment is added as a new topic for that instrument. Service center and service technician contact information, serial number, and laboratory location of the piece of equipment are stored as well.

**Emedgene:** Emedgene is a commercial variant interpretation platform used by the DDL. Variants are loaded from Emedgene into the KVD and, subsequently, Emedgene variant data is accessed via an API and stored in the KVD, for easy central access by DDL staff.

Please see document JHG_equipment_Lab_IT_2024.xlsx for a full list of IT equipment.