The Center for Research Informatics **(CRI)** is a 30-person group dedicated to providing resources to University of Chicago faculty to enable biological research. The CRI is comprised of four service lines: (1) clinical research data warehousing, business intelligence, and analytics,
(2) clinical research programming, (3) bioinformatics, and (4) scientific computing, which includes high-performance computing, machine-learning, visualization, and research storage.

The Clinical Research Data Warehouse (CRDW) maintained by the CRI at University of Chicago has been approved by the IRB to operate as an enterprise resource to support research throughout the Biological Sciences Division (BSD). Data in the CRDW comes from Billing, Clinical Registries, Research Systems as well as our Electronic Medical Record (EMR) and comprises all encounters from 1/1/2006 to date. This includes over 12.8 million inpatient and outpatient encounters for at least 1.2 million distinct patients resulting in greater than 1 billion observations. The CRI is funded by the BSD at the University of Chicago. The CRI has all policies and procedures approved by the UChicago Clinical Research Data Stewardship Committee.

The **Clinical Research Programming** team develops and maintains a wide variety of custom applications for research both within and outside the University. These offerings are unique and tailored to each project's specific needs. Expertise includes: software development lifecycle for web-based applications, development in multiple platforms, with a focus on ASP.NET, design and implementation of transactional and operational data stores, patient registries that integrate data from multiple systems, multi-institution clinical research data networks, and unstructured text indexing and searching using natural language processing. Additionally, the team manages **REDCap**, which is a HIPAA- secure data collection tool that can be used to meet 21-CFR part 11 requirements. Databases can be quickly developed and customized for studies' needs. REDCap is useful for collecting and tracking information and data from research studies, scheduling study events (e.g., patient visits), and conducting surveys and collecting patient-reported outcomes.

The **Bioinformatics Core** work combines the use of powerful computing resources, advanced analytics tools, and a commitment to security to transform large amounts of raw data into meaningful results. They use high-performance computing cluster and large-scale storage resources, with which complex analytics can be executed in parallel or distributed environments to produce fast data processing rates, improving application performance and cost effectiveness. Through advanced high-throughput analytics solutions, the team digs down to the root of each computational challenge and designs the most direct path to a solution. the work is protected with the CRI's automated, resilient backup and security systems, ensuring data integrity and access controls that are aligned with standards required by HIPAA, FISMA, and other regulations. Services include bioinformatics analysis of high-throughput biological data including proteomics using well-defined analysis pipelines and consulting services for custom analyses beyond our standard pipelines, including genome-wide association studies.

The CRI supports a **High-Performance Computing** (HPC) Cluster comprised of 120 compute nodes organized in three tiers (based on RAM per node; 128 GB, 512 GB, and 1.28 TB) connected through an Infiniband FDR low latency network. Researchers have access to nVidia GPU nodes and Intel Xeon Phi Coprocessors. The system has a total 3,360 cores and perform at 97 TFLOPs (Rmax) with 400 TB GPFS scratch storage. The HPC environment contains over 100 advanced data analysis software modules for bioinformatics, including those for NGS and microarray analysis. In addition to the HPC environment the CRI has two **Linux Statistical Analytics Servers** with 32 cores and 768 GB RAM each and one **Windows Statistical Analytics Server** (48 cores and 1 TB RAM, nVidia A40 GPU).

The CRI, in partnership with BSD Information Services manages a ***Virtual Server Infrastructure*** with over 450 active virtual servers on VMWare VSphere (Dell VxRail hardware platform).
The CRI provides 6 petabytes of ***File Storage*** (IBM ESS storage system using GPFS) with available space to be provisioned for group shares assigned to BSD principal investigators. To ensure that access is only granted to authorized staff, all changes to group share access are verified with the principal investigator or designee. The storage systems are backed up to a SpectraLogic T950 tape library using IBM Storage Protect. The tape library is housed in a datacenter in a different location on campus to ensure separation from the original data. SpectraLogic Storcycle is used for ***Data Archiving*** functionality, where research data

is archived to Microsoft Azure Blob storage *and* on tape in the SpectraLogic T950 tape library. Research computing infrastructure and resources for the Center for Research informatics (CRI) are primarily located in the ***Kenwood Data Center*** which is in Hyde Park on the University of Chicago campus. The Data Center is designed and tested to withstand extended power outages without system or service interruption. CRI servers are physically secure in locked racks in a facility fitted with electronic entry and alarm systems. The data center is divided in two sections, each designed for different use profiles: POD-A (2,500 square feet) and POD-B (2,100 square feet). POD-A is designed to house mission-critical workloads and meets the Uptime Institute's Tier 2 rating. POD-B is designed for compute-heavy High-Performance Compute (HPC) with a power draw of up to 25kW per cabinet. The Kenwood data center is managed by University of Chicago IT Services and monitored by staff 24/7. The datacenter is equipped to house systems that may fall under certain federal guidelines, including Health Insurance Portability and Accountability Act (HIPAA) and the Federal Information Security Management Act (FISMA). The ***1155 Data Center*** is also located in Hyde Park on the University of Chicago campus, which is 4,144 square feet and managed by University of Chicago IT Services. The CRI utilizes this facility to house the tape library that is used for data backups.