

Compute Cluster

In the framework of the ([Scientific Computing Infrastructure \(SCI\)](#) project) we operate a compute cluster for data processing and downstream analysis. It is used by the ODCF and other groups at the DKFZ for clinical and scientific projects. The cluster runs the IBM Spectrum LSF workload management software and is based on CentOS. As of November 2020, it consists of over 150 servers and provides access to more than 4500 CPUs and almost 800'000 CUDA cores in 200 GPUs, a high-memory node with 1 TB of memory and associated development systems (RStudio and JupyterHub servers). All cluster nodes are directly connected to the central storage systems of the DKFZ IT Core Facility.

The cluster is expanded based on demand. Our goal is to provide a cluster based on the needs of our users, so please feel free to contact us if you have special requirements.

Data Management

We offer structured storage of various raw omics data, with a focus on various Next Generation Sequencing (NGS) data, based on our in-house software OTP (One Touch Pipeline, <https://otp.dkfz.de>). We are happy to help you find a suitable way to structure your data.

Sequencing data (FASTQ files): For a regular ODCF project your sequencing metadata is imported into our database, your FASTQ files are transferred to our filesystems and your data gets processed (if available). All information is visualized via our graphical user interface. Your project's raw data and the data processed by ODCF will be stored in a project specific share, provided by us. This share is secured

by a project specific UNIX Group and you will only get read permission for that folder.

In addition, an analysis share is provided on the filesystem where you can store related analysis data. Here you get read and write permissions and are free to manage the data yourself. Both folders can be accessed via the LSF-cluster and OpenStack-cloud however, there is no possibility to access them via Windows.

Other research data can also be stored at the ODCF. Only the user access management is currently available for these projects, and no automated data storage. Nevertheless, an analysis share on the filesystem will be created in order to store the data. It is your responsibility to manage and store the data in this folder (see: <https://odcf-guide.dkfz.de>; only available from the DKFZ intranet).

All data managed by ODCF is under strict surveillance so that only authorized people can access them. Data access can be managed by the PI of each project or a delegated user via the graphical user interface (GUI) of OTP.

Processing of NGS Data

Data registered in **OTP** is automatically processed using all available workflows.

Our OTP supported workflows include alignments, variant calling (SNVs, INDELS, CNVs, SVs), signature analysis, feature counts and other downstream analyses. All results are stored on the filesystem and can also be accessed via the GUI. All processing steps are book kept to enable reproducibility. Additionally, sequencing coverage and quality checks are provided.



Moreover, we offer the **HUSAR** (Heidelberg Unix Sequence Analysis Resources) program package which contains several workflows for NGS data and other tools to process sequencing data. It is a self-service system which can be used via a GUI or command line. Bioinformatic guidance may be provided in special cases. Start Husar's web interface (W2H) within the DKFZ network (or via VPN) using the following link: <http://husar2.dkfz-heidelberg.de/menu/cgi-bin/w2h/w2h.start>.

In order to register please contact us under odcf-service@dkfz.de

In case any **further analyses** are required, several different tools are provided on the cluster and can be used.

IGV

The IGV Linker (<https://otfiles.dkfz.de>) allows streaming of all ODCF-managed data into your local IGV-session and can be accessed inside the DKFZ-intranet. This is a visualization tool, which can replace the need for filesystem mounts or network drives.

It provides a patient-based sorted view of the visualized data in your project and allows loading into IGV with a single click.

After a one-time setup and fine-tuning, based on your project's needs, the overview page is updated daily depicting the latest data from the filesystem.



The **Omics IT and Data Management Core Facility (ODCF)** provides data management, high-throughput computing and basic bioinformatics services to groups at the DKFZ and their collaborating partners.

Services

Data Management

- Structured data storage for NGS data (filesystem)
- Metadata handling and storage (in database)
- Data access management
- Support for EGA-Uploads
- Support for data sharing with external collaboration partners

Processing & Visualization

- Automatic alignment, variant calling & signature analysis (if available for respective sequencing type) in OTP.
- Automatic quality control of raw data (FastQC) and analyses in OTP.
- Self-service NGS analysis pipelines: ChIP-Seq, smallRNA, RNA-Seq, methylation, single-cell, differential expression
- Self-service sequence comparisons, alignments and annotation, secondary structure analysis of RNA & proteins
- IGV-Crawler

Compute Cluster

- Access to over 4500 CPU cores and around 800'000 CUDA cores in 200 GPUs
- High-memory node (1 TB RAM)
- Worker node for interactive analysis and testing
- RStudio server for development and interactive analysis
- JupyterHub server for development and interactive analysis

Courses / Workshops

We are happy to schedule a meeting to help you set up a project and discuss the data storage and processing possibilities.

Moreover, we offer different courses for novice users to become familiar with our systems and to discover their possibilities. The courses are held on a regular basis at the DKFZ (please register via the HCM portal <http://logaportal/maportal>). For details and dates on courses and workshops check the HCM portal.

CONTACT

Cluster issues: odcf-itsupport@dkfz.de
All other issues: odcf-service@dkfz.de

OFFICE ADDRESS

German Cancer Research Center (DKFZ)
Omics IT and Data Management (ODCF)
Mathematikon - Berliner Straße 41
D-69120 Heidelberg, Germany

All information in this flyer is subject to change. Please contact odcf-service@dkfz.de for up-to-date information about our services.

Version: 2020-11

dkfz.

GERMAN
CANCER RESEARCH CENTER
IN THE HELMHOLTZ ASSOCIATION

.....
Research for a Life without Cancer



**Omics IT and Data Management
Core Facility**