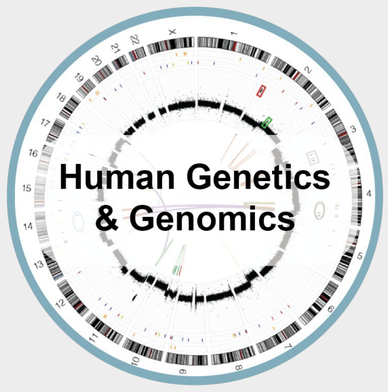
Heidelberg Center for Human Bioinformatics: Human Genetics and Genomics

M. Smith[1], P. R. Kensche[3], Zuguang Gu[2], T. Rausch[1],
I. Naveed[2,5], I. Buchhalter[3], W. Huber[1], J. Korbel[1], R. Eils[5], R. Russell[4]

[1] European Molecular Biology Laboratory, Genome Biology Unit, Heidelberg
[2] Division of Theoretical Bioinformatics / [3] Omics IT & Data Management Core Facility
German Cancer Research Center (DKFZ), Heidelberg
[4] BioQuant, Heidelberg University, Heidelberg
[5] Centre for Digital Health, Charité / Berlin Institute of Health (BIH), Berlin



Fkz 031A537A/B/C/D

Project Description

The Heidelberg Center for Human Bioinformatics (HD-HuB) combines expertise in the area of Human Genetics and Genomics from groups at three established institutions in Heidelberg - the German Cancer Research Center (DKFZ), European Molecular Biology Laboratory (EMBL), Heidelberg University - with colleagues at Charité / Berlin Institute of Health (BIH).

The goal of this project is to provide bioinformatics knowledge and resources for efficiently analyzing complex data from a wide range of 'omics experimental techniques, most specifically human genome sequencing.

The Genetics and Genomics resources supported by HD-HuB span a range from software infrastructure, through standalone tools focused on a specific analysis task, to frameworks for processing data via a number of steps. Some are highly focused on human genomics, while others are more widely applicable bioinformatics tools that can be used across the research spectrum.

Progress Report

Here we describe several of the key HD-HuB Genetics and Genomics services and highlight the developments over the course of the last year that have been enabled by **de.NBI**.

- HD-HuB has maintained support for 17 CRAN & Bioconductor packages. Total downloads-per-month have increased year-on-year.
  - We have introduced additional widely used, genomics focused, packages e.g. **ComplexHeatmap**, **circlize** to the HD-HuB portfolio in the past year.
  - **rhdf5** has been updated to work with additional compression filters, which are being tested for their efficacy with single-cell sequencing data.
  - **biomaRt** usage continues to grow; we have added new documentation focused on the most widely seen use cases.
- **OTP** now supports Cell-Ranger execution for single-cell RNA-seq data and is being modularized for better server security and for supporting more workflow management systems via a Workflow Execution Services (WES). To this aim we started developing a Snakemake WES server.
- **Roddy** is in maintenance mode and its workflows are being migrated to alternative systems. New developments include:
  - new production features, such as Post-Bisulfite Adapter Tagging (PBAT), sample-swap detection, and rare genomic variant detection.
  - ongoing optimization for cancer variant detection with the **GRCh38** human assembly.
- **GEAR** is a web service providing a number of tools for working with genomic sequencing data. These include primer design, trace alignment, and BAM file statistics amongst others.
- Members of HD-HuB and CIBI have joined forced to establish a python API for the **seqAn3** library for fast an efficient sequence analysis. This exciting venture aims to increased accessibility to the **seqAn3** framework.

Training

Details on training and outreach are provided on the HD-HuB overview poster.

Publications

Amezquita R.A. et al. (2019) *Orchestrating single-cell analysis with Bioconductor.* ***Nature Methods***

Park J. et al (2019). *Segmentation-free inference of cell types from in situ transcriptomics data.* **bioRχiv**

Selected Tools and Workflows

Butler

Framework for scalable cloud-based scientific data analysis

Carried out variant discovery and joint genotyping for ≈ 3,000 tumour/normal sample pairs on multiple clouds

One Touch Pipeline (OTP)

Automation platform for managing next generation sequencing data

Data and workflow management for almost 30,000 samples from > 18,000 individuals

biomaRt

HD-HuB personnel maintain the code and provide support for biomaRt, the most used programmatic access route to the Ensembl database.

≈ 14,000 unique downloads per month
≈ 100 support requests answered each year

ComplexHeatmap & Circlize

These two packages provide highly flexible solutions for creating publication quality heatmaps and circular visualisations

Combined ≈ 10,000 unique downloads per month from www.bioconductor.org and CRAN

DELLY

Workflow for the discovery of germline and somatic structural variants in whole genome sequencing data.

≈ 10,000 downloads per year from https://anaconda.org/bioconda/delly and https://github.com/dellytools/delly

rhdf5 & Rhdf5lib

Providing an R interface to HDF5, allowing R software to work on datasets larger than available memory. Underpins many single-cell analysis tools.

Combined ≈ 10,000 unique downloads per month from www.bioconductor.org

DESeq2 & DEXSeq

A suite of R/Bioconductor packages for differential expression analysis of high-throughput sequencing assays.

Combined ≈ 8,000 unique downloads per month from www.bioconductor.org

Roddy-based genomics workflows

Includes alignment for Whole Exome, Genome, and Bisulfite Datasets and variant callers like SNV & Indel Calling, ACEseq, and CNVkit.

Roddy is available on www.github.com/TheRoddyWMS/Roddy and genomics workflows at www.github.com/DKFZ-ODCF

Mechismo

Predicting how genetic variants impact biomolecular function using collections of protein interactions.

Accessed via mechismo.russelllab.org

SMART

SMART (Simple Modular Architecture Research Tool) is a web-based resource for protein domain annotation, analysis and visualization.

> 500,000 web page visits / month on smart.embl.de and 485 citations in 2017

GEAR

Genome analysis server providing an array of tools for working with genomic seqencing data

Access via https://www.gear-genomics.com/

GEFÖRDERT VOM Bundesministerium für Bildung und Forschung
Projektträger Jülich Forschungszentrum Jülich

www.denbi.de
www.hd-hub.de