# Biologist-friendly analysis software for NGS data

**Aleksi Kallio[1], Taavi Hupponen[1], Massimiliano Gentile[2], Jarno Tuimala[3], Kimmo Mattila[1], Ari-Matti Saren[1], Petri Klemelä[1], Ilari Scheinin[4], Eija Korpelainen[1]** ✉

[1]CSC – IT Center for Science, Helsinki, Finland
[2]Blueprint Genetics Oy, Helsinki
[3]SPR Veripalvelu, Helsinki
[4]VU University Medical Center, Netherlands

## Motivation and Objectives

NGS technology offers unprecedented possibilities for life science, motivating efforts to develop new data analysis tools and techniques. However, available tools are scattered and often require some programming skills, leaving them out of reach of non-computational researchers. Chipster provides a clear and biologist-friendly interface to analysis tools for NGS data. It has a graphical user interface that connects to server environment for heavy data processing. Chipster is a free and open source software, and it is available as a virtual machine for easy server installation.

## Methods

Chipster (http://chipster.csc.fi) provides data analysis tools for many NGS applications, including DNA-, RNA-, miRNA-, ChIP-, methyl- and CNA-seq. Users can easily save and share analysis workflows, and built-in genome browser allows seamless viewing of reads and results.

Users can perform their whole data analysis in Chipster from quality control to downstream applications such as pathway enrichment and motif discovery. Popular tools such as FastQC, FASTX, PRINSEQ, SAMtools, BEDTools, Bowtie, BWA, TopHat, HTSeq and Cufflinks are included, and care has been taken to serve them in a biologist-friendly manner. Also several R/Bioconductor packages have been integrated, including edgeR, DESeq and MEDIPS.

Chipster's built-in genome browser allows visualization of reads and results in their genomic context using Ensembl annotations. Users can zoom in to nucleotide level, highlight SNPs and view the automatically calculated coverage. Cross-talk between the genome browser and BED, VCF and GTF files allows users to quickly inspect genomic regions by simply clicking on the data row of interest.

Technically Chipster is a Java-based client-server system (Kallio *et al.*, 2011). Recently system's data handling capabilities have been enhanced to cope with NGS scale data. The system is capable of tracking data copies across the distributed system (both server and client side), minimizing data transfers and always using the closest copy for best performance. We are also working with the Hadoop MapReduce framework so that large jobs can be run in a massively parallel way (Niemenmaa *et al.*, 2012).

To maintain good user experience with large scale NGS datasets, Chipster server allows users to save their analysis sessions on the server side. For the administrators it also provides tools to monitor disk space usage on the server environment.

New analysis tools can easily be added using a simple mark-up language. Chipster places no restrictions on what type tools can be integrated. Currently we are introducing intuitive graphical interfaces also for tool development and server administration.

## Results and Discussion

Taken together, Chipster provides an easy way to serve NGS data analysis tools in a biologist-friendly manner. In our national role of providing bioinformatics services for whole country we have noticed that a user-friendly software together with training on data analysis methods is a powerful combination for enabling life scientists to analyze their own data.

The complete Chipster server system is freely available under the open source GPL license at http://chipster.sourceforge.net, and it has been adopted by many institutes worldwide. The recommended option is to download virtual machine images that contain all tools and databases, bundled together with a ready-to-run Chipster installation. Virtual machine images support all major virtualization platforms.

## Acknowledgements

This work was supported by the Intelligent Monitoring of Health and Well-being program of SalWe – the Strategic Centre for Science, Technology and Innovation in Health and Well-being, and by the Cloud Software Program of the Finnish Strategic Centre for Science, Technology and Innovation TIVIT.

## References

Kallio, Tuimala, Hupponen *et al*. (2011) Chipster: User-friendly analysis software for microarray and other high-throughput data. *BMC Genomics* **12**, 507.

Niemenmaa, Kallio, Schumacher *et al.* (2012) Hadoop-BAM: Directly manipulating next generation sequencing data in the cloud. *Bioinformatics* **28**(6), 876.