

The Stem Cell Commons: An HSCI Resource for Stem Cell Genomics Data and Reproducible Analyses

Ho Sui, Shannan J.^{1}, Das, Sudeshna^{2*}, Gehlenborg, Nils^{3*}, Merrill, M. Emily², Sytchev, Ilya⁴, Park, Richard W.³, Haseley, Psalm³, Hutchinson, John N.⁴, Corlosquet, Stephane², Hansen, Terah A.⁴, Hofmann, Oliver¹, Park, Peter J.³, Hide, Winston¹*

¹Harvard Stem Cell Institute, Harvard School of Public Health, Boston, MA, USA,

²MIND Informatics, Massachusetts General Hospital, Cambridge, MA, USA,

³Center for Biomedical Informatics, Harvard Medical School, Boston, MA, USA,

⁴Biostatistics, Harvard School of Public Health, Boston, MA, USA

** These authors contributed equally to this work*

The Stem Cell Commons (stemcellcommons.org) has been developed by the Harvard Stem Cell Institute to create a community for stem cell bioinformatics. This open source environment for sharing, processing and analyzing stem cell data brings together genomics data sets with tools for curation, searching, dissemination and analysis. Standardizing the experimental metadata and analysis approaches provides us with opportunities to perform data integration and comparison of data sets at the HSCI.

The Stem Cell Commons database, built on the eXframe system, contains data from 20 HSCI laboratories and 89 public datasets made up of over 3000 assays. These include microarray and next generation sequence data spanning multiple cell types and tissues from human, mouse, rat and zebrafish. We have both raw data and results of standardized pipelines, to maximize reproducibility and data provenance. Custom analysis results are made available to researchers as well.

We have linked this data with the Refinery analysis and visualization platform (refinery-platform.org), a flexible environment for managing data and running diverse analysis workflows. We showcase an early version of a new ChIP-seq analysis pipeline and invite new users to provide feedback on the user interface and new workflows to be added to the system. This system will ultimately serve as a platform for the development of novel visual exploration tools that can directly access large and complex datasets and analysis results, and trigger new analyses on these data.