Published in final edited form as:
Nat Methods. 2013 October ; 10(10): 926. doi:10.1038/nmeth.2643.

Building an ENCODE style data compendium on a shoestring

David Ruau[1], **Felicia S. L. Ng**[1], **Nicola K. Wilson**[1], **Rebecca Hannah**[1], **Evangelia Diamanti**[1], **Patrick Lombard**[1], **Steven Woodhouse**[1], and **Berthold Göttgens**[1]

[1]Department of Hematology, Cambridge Institute for Medical Research and Wellcome Trust and MRC Cambridge Stem Cell Institute, Cambridge University, Hills Road, Cambridge, UK

One perhaps unintended consequence of the unquestionable success of the human genome project has been a shift in the biomedical research funding landscape towards large scale programs, commonly involving several hundred scientists and budgets of hundreds of millions of dollars. This emphasis on large-scale projects however is sometimes questioned, as illustrated by recent debates following last year's publications from the ENCODE project [1,2]. Here we have explored an alternative approach. Rather than making advanced decisions about the datasets that should be generated for a given research community, as large scale projects have to do, we have instead compiled all datasets produced by that community, as soon as they are deposited in public databases. We demonstrate that such real-time curation can exceed large consortia efforts, which constitutes a highly topical contribution to the ongoing 'small vs big science' debate.

We created HAEMCODE, a repository for transcription factor (TF) binding maps in mouse blood cells, generated by chromatin immunoprecipitation sequencing (ChIP-seq) . Using a standardized analysis pipeline we manually curated more than 300 TF ChIP-Seq studies from a wide range of primary mouse haematopoietic cells and major cell line models. As of September 2013, the HAEMCODE compendium covered 84 TFs across 24 major blood cell types. Haemopoiesis is also a major focus of ENCODE, yet the currently available mouse ENCODE data covers less than half of HAEMCODE (36 TFs, May 2013), with only 9 ENCODE TFs not available elsewhere.

We next developed a web interface (http://haemcode.stemcells.cam.ac.uk) to provide data access as well as a range of online analysis tools, designed to be useful to both experimentalist and computational biologists. The classical use case consists of selecting experiments within HAEMCODE before being directed to a workspace, which offers pre-computed options to inspect and/or download selected ChIP-Seq datasets. Additional online tools can compute global similarity between selected experiments, investigate overrepresentation of a user-submitted gene list in any subset of ChIP-Seq experiments[3], inspect pre-computed results from de-novo motif discovery, and output all ChIP-Seq experiments with binding peaks for a user-supplied gene locus.

Integration of publicly available data represents a powerful approach to make novel discoveries across diseases, species and platforms that would be impossible to achieve from

Corresponding authors: David Ruau, djr62@cam.ac.uk; Berthold Göttgens, bg200@cam.ac.uk.

single projects[4]. Successful completion of the HAEMCODE project on a small budget highlights this approach as a potentially widely applicable complement to multi-million dollar research initiatives.

Acknowledgments

This work was funded by the Biotechnology and Biological Sciences Research Council, Leukaemia and Lymphoma Research, the Medical Research Council, Cancer Research UK, and core support grants by the Wellcome Trust & MRC Cambridge Stem Cell Institute. FSN is the recipient of a Yousef Jameel scholarship.

References

1. Alberts B. Science. 2012; 337:1583. [PubMed: 23019614]
2. The Encode Project Consortium. Nat Cell Biol. 2012; 489:57–74.
3. Joshi A, Hannah R, Diamanti E, Göttgens B. Exp Hematol. 2013; 41:354–366.e14. [PubMed: 23220237]
4. Butte AJ, Kohane IS. Nat Biotechnol. 2006; 24:55–62. [PubMed: 16404398]
5. Zhang Y, et al. Genome Biol. 2008; 9:R137. [PubMed: 18798982]