A functional genomics explorer for visualizing and comparing large amounts of datasets retrieved in the public domain

Marco-Antonio Mendoza-Parra[1], Pierre-etienne Cholley[1], Matthias Blum[1], ,Anissa Djedid[1],Jordan Friess[1] and Hinrich Gronemeyer[1]

[1]Dept. of Functional Genomics and Cancer; IGBMC, Strasbourg, France

ABSTRACT

With the democratization of high throughput approaches, interrogating genome functions in a comprehensive manner will constitute in the following years the driving force for the development of personalized / predictive medicine. In fact, while data assessment is becoming largely accessible thanks to the democratization of massive parallel sequencing, the major challenge in the present big genomics data era resides in the necessity of developing computational solutions for interrogating huge numbers of datasets in a comparative manner. We have previously developed a quality control system dedicated to ChIP-sequencing and related datasets, which has been used for qualifying more than 30,000 publicly available data (www.ngs-qc.org). More rescently, we have expanded the concept of quality assessment to long-range chromatin interaction assays, such that more than 250 publicly available HiC (including several variants of the original protocol, like in situ or capture HiC), but also several ChIA-PET (>50) and 4C-seq (>900) datasets were qualified and made available via a dedicated database: www.ngs-qc.org/logiqa.

Here we present a functional genomics datasets explorer providing a large flexibility for the retrieval of publicly available datasets; access to their quality grades in a global and local manner, as well as to visualize their associated enrichment patterns. Furthermore, it provides computational solutions for multiple dataset comparisons in a large scale manner (several hundreds) and in either a global or local genomic context, such that datasets retrieved in a variety of studies could suddenly be integrated. For the first time, users may have the possibility of using the large amounts of public data as a resource for inferring predictions without requiring to download and preprocess the available raw data.
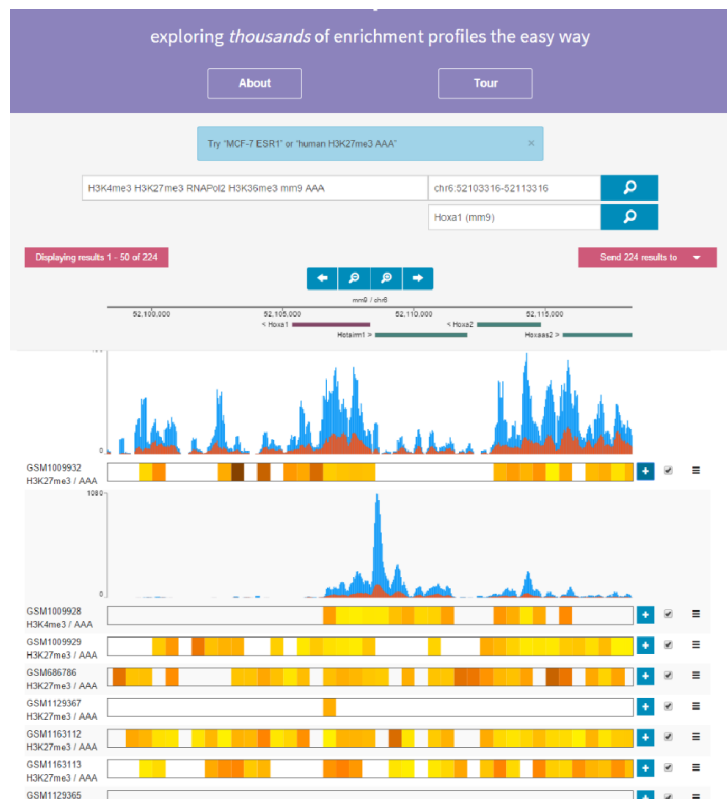
Figure 1. A functional genomics datasets explorer providing a large flexibility for the retrieval and visualization of publicly available datasets in the context of their associated quality scores. A simple query panel is used for providing users the largest flexibility for the retrieval of datasets associated to terms like target molecule (e.g. H3K4me3), organism, celltype/tissue or even the minimal required quality (e.g. AAA). If wished, users can provide in addition a genomic coordinate or a gene name, thus the display will be centered around the requested region, in which a heatmap bar will reflect the presence of enriched regions with a minimal confidence. The displayed heatmap can also be expanded to display the raw count intensities (blue: total reads; red; unique reads). This explorer environment can then send the retrieved datasets to several analytical tools in which a comparative analysis among retrieved datasets is performed either in a global or local genomic context.