High-throughput gene to knowledge mapping through massive integration of public sequencing data

Brian Tsui[1,2] , Hannah Carter[1]

[1]Department of Medicine, Division of Medical Genetics, University of California San Diego, La Jolla, CA 92093, USA [2]Bioinformatics and Systems Biology Graduate Program, University of California San Diego, La Jolla, CA 92093, USA

Abstract

Sequencing Read Archive (SRA) contains more than one million runs of publicly available sequencing data (Fig. 1, top). However, the lack of consistently preprocessed summary and molecular quantification data (for example, gene expression quantification for RNAseq) for each sequencing run hinders efficient Big Data interpolation. Here, we introduce Skymap, a standalone database that offers a single data matrix per species in each data layer incorporating all public sequencing studies. The omic layers include expression quantification, allelic read counts, microbes read counts, chip-seq. We reprocessed petabytes of sequencing data to generate the data matrix for each data type (Fig. 1, bottom). We also offer a reprocessed biological metadata file that describes the relationships between the sequencing runs and the associated keywords, extracted from over 3 million freetext annotations using natural language processing. The processed data can fit into a single hard drive (<500GB). In https://github.com/brianyiktaktsui/Skymap, we showcase how one can (1) retrieve and analyze the SNPs and expression of a genetic variant across >200k runs in less than a minute and (2) increase the temporal resolution for tracking gene expression in mouse developmental hierarchy.

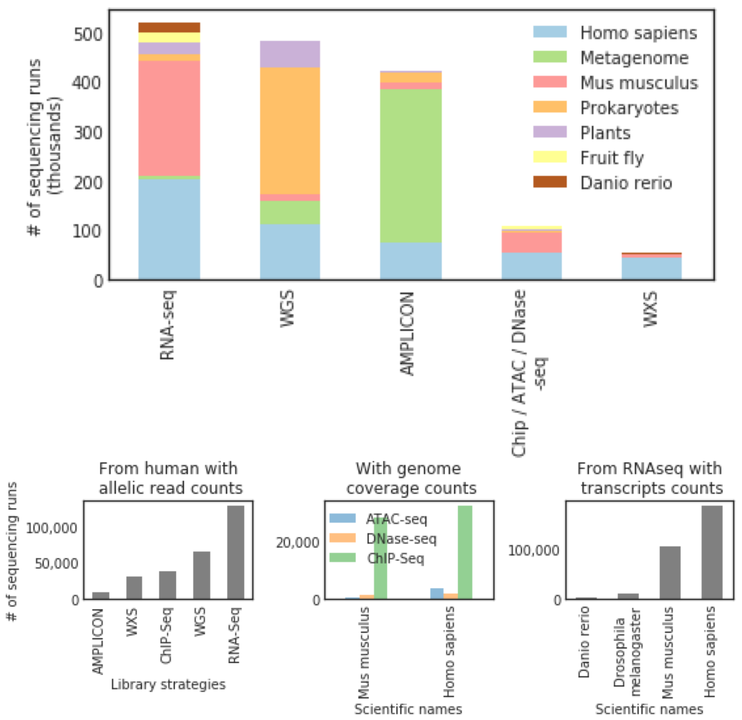

Fig. 1 SRA data landscape. Top: # of data available in SRA; Bottom: # of data processed in SRA

Example 1: High resolution mouse developmental hierarchy expression map

Aggregating studies (node) can form a smooth mouse developmental hierarchy map (Fig. 2), which sometime we can see a more transient expression dynamics in different tissues over development, like Trp53 (Fig. 3).

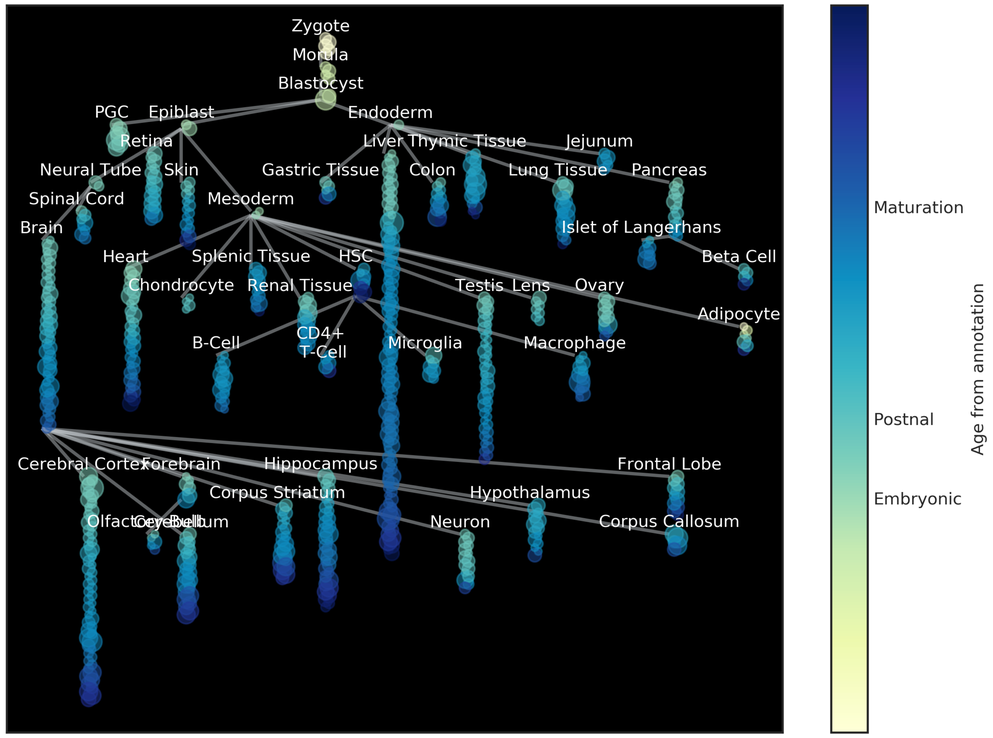

Fig. 2 High resolution mouse developmental hierarchy

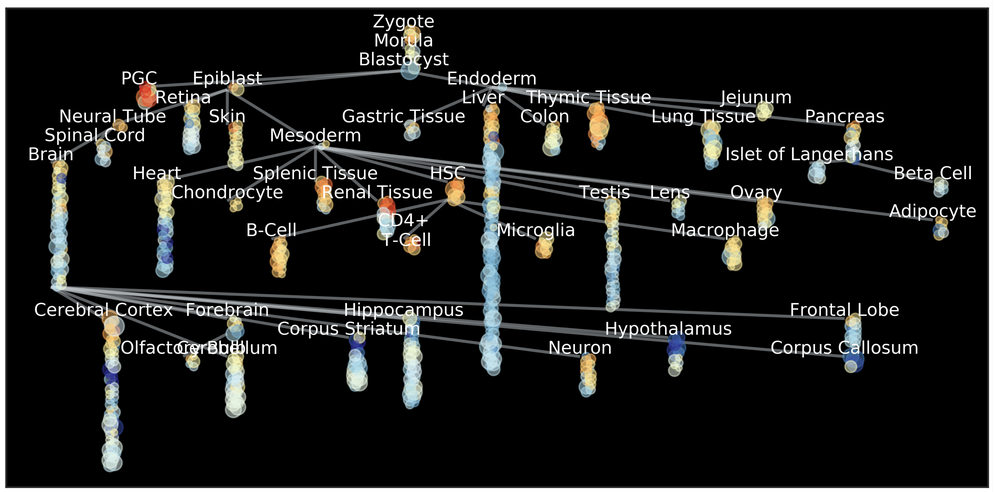

Fig. 3 Dynamic of Trp53 expression over time and spatial locations

Example 2: Slice >200k SNP and expression profiles in under a minute for EQTL

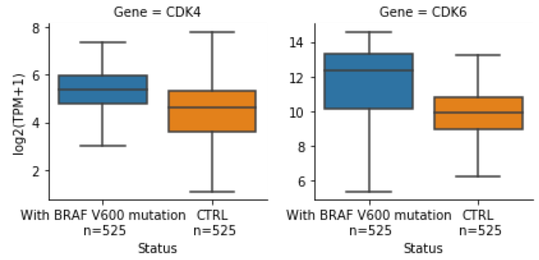

Fig. 4: BRAF V600E mutation is associated with CDK4 and CDK6 expression levels increase

Example 3: Slice any public data for viral load evaluation

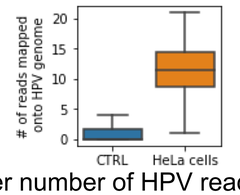

Fig. 5: Higher number of HPV reads detected in Hela cells as compared to random control

Future development: Completing the data driven gene to knowledge pipeline by having a data driven NLP engine

Use deep learning and natural language processing (NLP) techniques, we automated the biomedical named entity recognition process in biospecimen annotation data without the need of curation (Fig. 6, https://github.com/brianyiktaktsui/DEEP_NLP). With this, we might be able to capture the human reasoning process in omic related research and thus enable automated various omics to experimental conditions associations in the future.

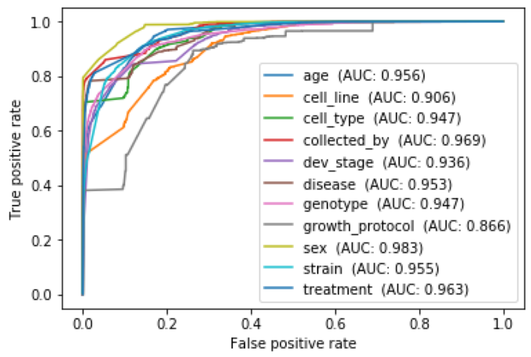

Fig. 6 Performance of deep learning based NLP model in named entity recognition