Informatics for Encyclopedia of DNA Elements (ENCODE3)

DENIS N. SALINS, NATHANIEL K. WATSON, ESTHER CHAN, SOMALEE DATTA, TRUPTI KWALI
Stanford University

Encyclopedia of DNA Elements (ENCODE) project is a massively large scale collaborative research effort, to create a comprehensive public catalog of functional elements in the human genome. All data generated from production labs are submitted and presented by the Data Coordination Center (DCC). The DCC handles the data validation, tracking, storage, distribution, and visualization to the research community. Stanford is amongst the largest contributors to the overall project with submissions of tested Antibodies, Transcription Factors, and Biosamples to produce over 10,000 of data and metadata points.

The data generated to achieve these datasets is also immensely important to understand the process that was used to test and arrive at the published results. Typically, testing 100 antibodies will yield a 5-15% (Testing over 200,000) success factor of the transcription factor passing all the desired tests and experiments.

Stanford’s – Sequencing Workflow

The Stanford’s – Sequencing Workflow is designed to facilitate the alignment of reads with the reference genome for further analysis. The workflow involves several steps including the input of raw sequencing data, alignment, and identification of regulatory elements. The workflow is depicted in a diagram showing the sequence of events from input to output, ensuring a streamlined process for data analysis.

Stanford’s – ENCODE 3 Data Scale (GB)

The diagram represents the data scale for Stanford’s ENCODE 3 project. It illustrates the amount of data generated over time, showcasing the growth in data production from Q1 2013 to Q3 2015. The data scale is depicted using a bar chart with a y-axis indicating the volume of data in GB.