Information Sciences 000 (2016) 1–20

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

Indexing Next-Generation Sequencing data

Vahid Jalili*, Matteo Matteucci, Marco Masseroli, Stefano Ceri

Dipartimento di Elettronica, Informazione e Bioingegneria (DEIB) Politecnico di Milano, Piazza Leonardo da Vinci, 32, Milan, Italy

ARTICLE INFO

Article history:
Received 16 February 2016
Revised 11 July 2016
Accepted 26 August 2016
Available online xxx

Keywords:
Genomic computing
Domain-specific data indexing
Region-based operations and calculus
Data integration

ABSTRACT

Next-Generation Sequencing (NGS), also known as high-throughput sequencing, has opened the possibility of a comprehensive characterization of the genomic and epigenomic landscapes, giving answers to fundamental questions for biological and clinical research, e.g., how DNA-protein interactions and chromatin structure affect gene activity, how cancer develops, how much complex diseases such as diabetes or cancer depend on personal (epi)genomic traits, opening the road to personalized and precision medicine.

In this context, our research has focused on *sense-making*, e.g., discovering how heterogeneous DNA regions concur to determine particular biological processes or phenotypes. Towards such discovery, characteristic operations to be performed on region data regard identifying co-occurrences of regions, from different biological tests and/or of distinct semantic types, possibly within a certain distance from each others and/or from DNA regions with known structural or functional properties.

In this paper, we present Di3, a 1D Interval Inverted Index, acting as a multi-resolution single-dimension data structure for interval-based data queries. Di3 is defined at data access layer, independent from data layer, business logic layer, and presentation layer; this design makes Di3 adaptable to any underlying persistence technology based on key-value pairs, spanning from classical B+ tree to LevelDB and Apache HBase, and makes Di3 suitable for different business logic and presentation layer scenarios.

We demonstrate the effectiveness of Di3 as a general purpose genomic region manipulation tool, with a console-level interface, and as a software component used within MuSERA, a tool for comparative analysis of region data replicates from NGS ChIP-seq and DNase-seq tests.

1. Introduction

Next-Generation Sequencing (NGS) is a family of technologies for precisely, quickly and cheaply reading the DNA or RNA of biological samples [49,50], producing huge amounts of data. Large-scale sequencing projects are spreading and very numerous genomic features, produced by processing NGS raw data, are collected by research centers, often organized through world-wide consortia, e.g., ENCODE [17], TCGA [58], 1000 Genomes Project [1], Roadmap Epigenomics [47], and others.

The availability of NGS data has opened the possibility of a comprehensive characterization of genomic and epigenomic landscapes. Answers to fundamental questions for biological and clinical research are hidden in these data, e.g., how DNA-protein interactions and chromatin structure affect gene activity, how cancer develops, how much complex diseases such as

* Corresponding author.
E-mail address: vahid.jalili@polimi.it (V. Jalili).

http://dx.doi.org/10.1016/j.ins.2016.08.085
0020-0255/

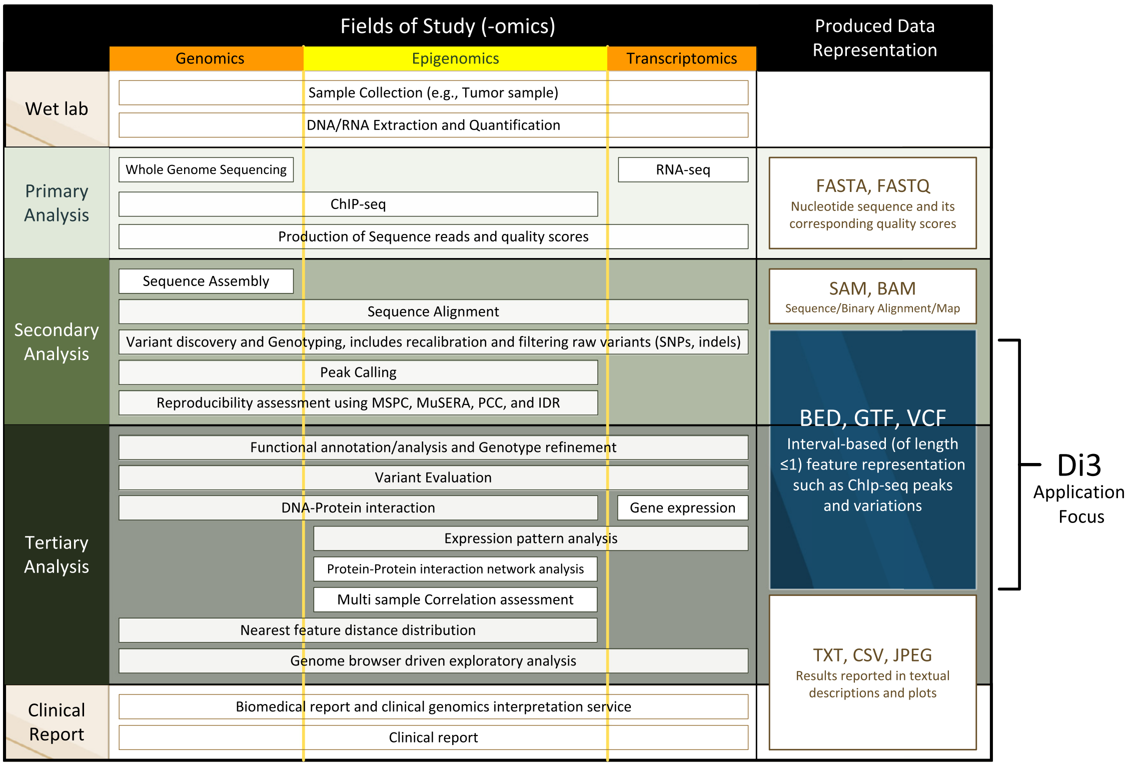

Fig. 1. Di3 application focus.

diabetes or cancer depend on personal (epi)genomic traits. Personalized and precision medicine based on genomic information is becoming a reality; the potential for data querying, analysis and sharing may be considered as the biggest and most compelling big data problem of mankind.

NGS technologies [9,48] allow collecting genome-wide genomic and epigenomic features, including DNA mutations or variations (DNA-seq), transcriptome profiles (RNA-seq), DNA methylations (BS-seq), DNA-protein interactions and chromatin characterizations (ChIP-seq and DNase-seq) [12,42]. The processing of raw data (i.e., NGS reads) produced by these technologies returns lists of regions of cellular DNA, characterized by some common properties; such regions, often referred as *peaks* (of NGS reads), are defined through their linear genomic coordinates and they are usually associated with several attribute values, including a statistical significance score, i.e., a *p*-value [44,62].

Fig. 1 summarizes the various kinds of data which are produced by NGS technologies. Data analysis is partitioned into three phases; *primary analysis* focuses on the production of sequence reads; *secondary analysis* focuses on alignment of raw data (short reads) to reference genomes and on feature calling. *Tertiary analysis* is responsible of sense-making, e.g., discovering how heterogeneous features/regions synergically concur to determine particular biological processes or phenotypes. Fig. 1 shows that the various analysis phases apply to genomics, epigenomics and transgenomics, and also shows the most relevant types of data representations used at each phase.

Our main research focus is on tertiary analysis; we recently proposed a new holistic approach to genomic data modeling and querying[1] that takes advantage of cloud-based computing to manage heterogeneous data produced by NGS technologies. In [37], we introduced the novel *GenoMetric Query Language* (GMQL), built on an abstract model for genomic data; we sketched out its main operations and demonstrated its usefulness, expressive power and flexibility through multiple different examples of biological interest (including finding ChIP-seq peaks in promoter regions, finding distal bindings in transcription regulatory regions, associating transcriptomics and epigenomics, and finding somatic mutations in exons).

We also developed methods for secondary data analysis, with a focus on data integration. Indeed, NGS experimental protocols recommend the production of at least two replicates for each sequenced sample, in order to reduce the number of false-positive calls and “rescue” (i.e., call) regions with low significance score which would probably be discarded in a single sample evaluation, but that are supported by a sufficiently strong evidence when combined across multiple replicate samples. To perform such task, and assess the reproducibility in high-throughput experiments, we recently proposed a novel

[1] http://www.bioinformatics.deib.polimi.it/genomic_computing/

Please cite this article as: V. Jalili et al., Indexing Next-Generation Sequencing data, Information Sciences (2016), http://dx.doi.org/10.1016/j.ins.2016.08.085