Chapter 6
Genomics Data

Ephrem Habyarimana and Sofia Michailidou

Abstract *In silico* prediction of plant performance is gaining increasing breeders' attention. Several statistical, mathematical and machine learning methodologies for analysis of phenotypic, omics and environmental data typically use individual or a few data layers. Genomic selection is one of the applications, where heterogeneous data, such as those from omics technologies, are handled, accommodating several genetic models of inheritance. There are many new high throughput Next Generation Sequencing (NGS) platforms on the market producing whole-genome data at a low cost. Hence, large-scale genomic data can be produced and analyzed enabling intercrosses and fast-paced recurrent selection. The offspring properties can be predicted instead of manually evaluated in the field . Breeders have a short time window to make decisions by the time they receive data, which is one of the major challenges in commercial breeding. To implement genomic selection routinely as part of breeding programs, data management systems and analytics capacity have therefore to be in order. The traditional relational database management systems (RDBMS), which are designed to store, manage and analyze large-scale data, offer appealing characteristics, particularly when they are upgraded with capabilities for working with binary large objects. In addition, NoSQL systems were considered effective tools for managing high-dimensional genomic data. MongoDB system, a document-based NoSQL database, was effectively used to develop web-based tools for visualizing and exploring genotypic information. The Hierarchical Data Format (HDF5), a member of the high-performance distributed file systems family, demonstrated superior performance with high-dimensional and highly structured data such as genomic sequencing data.

E. Habyarimana (✉)
CREA Research Center for Cereal and Industrial Crops, Via di Corticella 133, 40128 Bologna, Italy
e-mail: ephrem.habyarimana@crea.gov.it

S. Michailidou
Center for Research and Technology Hellas - Institute of Applied Biosciences, 6th Km Charilaou Thermis Road, 57001 Thessaloniki, Greece

C. Södergård et al. (eds.), *Big Data in Bioeconomy*,
https://doi.org/10.1007/978-3-030-71069-9_6

6.1 Introduction

The array of techniques for probing complex biological systems such as (crop) plants is continuously expanding, providing unprecedented data on multiple phenotypic layers as well as multiple omics layers (genome, proteome, metabolome, epigenome or methylome, and more). Furthermore, new and cheap local sensor techniques as well as advances in remote sensing and geo-information systems provide extensive descriptions of the environmental conditions under which plants grow. This allows *in silico* prediction of plant performance (e.g. traits like yield, abiotic and biotic resistance) depending on genotype, environment and crop management. Several statistical, mathematical and machine learning methodologies for analysis of phenotypic, omics and environmental data typically use individual or a few of these data layers. Genomic selection is one of the applications, where heterogeneous data, such as those from genomics, metabolomics and phenomics technologies, are handled also accounting for several genetic models of inheritance [1].

Genomic selection is a new paradigm in plant breeding allowing to bypass the costly and time-consuming phenotyping step by selecting superior lines based on DNA information according to the workflow in Fig. 6.1 [2, 3].

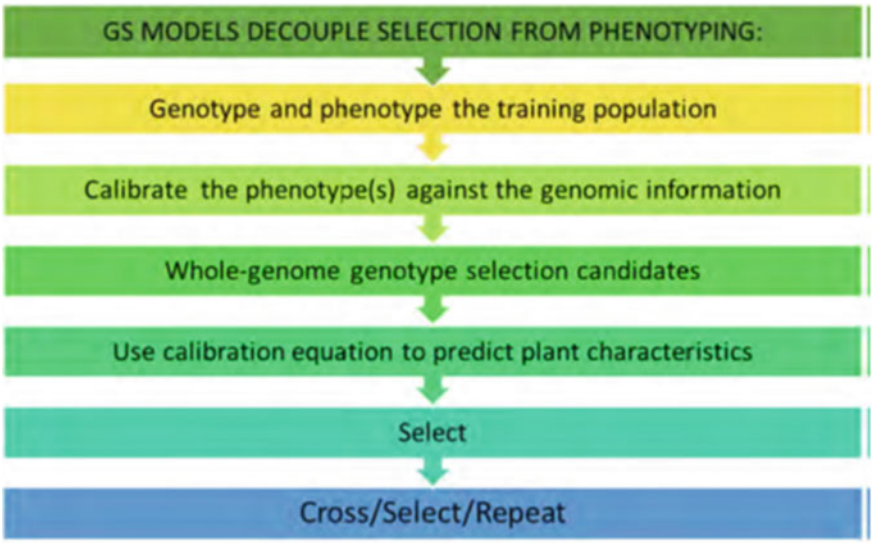

Fig. 6.1 Implementation of the routines of the genomic models

6.2 Genomic and Other Omics Data in DataBio

Genomics and other omics data were produced in sorghum (*Sorghum bicolor (*L.*)* Moench) and tomato (*Solanum lycopersicum* L.) crops (Fig. 6.2) evaluated in DataBio Genomics pilots; four categories of data were produced including (Tables 6.1 and 6.2): (1) in situ sensors and farm data, (2) genomic data from plant breeding efforts in greenhouses and in open field produced using Next Generation Sequencers (NGS), (3) biochemical data produced by chromatographs (LC/MS/MS, GS/MS, HPLC), wet chemistry and NIRS (near infrared spectroscopy) (Tables 6.1 and 6.2), and (4) genomics modelling output represented by integrative analytics information. In situ sensors/environmental outdoor generated wind speed and direction, evaporation, rain, light intensity, UVA and UVB data. In situ sensors/environmental indoor generated air temperature, air relative humidity, crop leaf temperature (remotely and in contact), soil/substrate water content, crop type, and several other data. Farm Data generated in situ measurements comprising soil nutritional status, farm logs (work calendar, technical practices at farm level, irrigation information), and farm profile (Static farm information, such as size).

Fig. 6.2 Tomato accessions in glasshouses (top) and sorghum pilot fields (bottom) used genomic models platform

Table 6.1 Genomic, biochemical and metabolomic data tools, description and acquisition

Data	Mission, Instrument	Data description and acquisition
Genomic data	• To characterize the genetic diversity of sorghum and tomato varieties and lines used for breeding (Fig. 2) • To identify novel variants in the sorghum and tomato genomes, associated plant characteristics of interest • To use the genomic information to guide breeding strategies (as a selection tool for higher performance) and develop a model to predict the final breeding result in order to rapidly achieve with the minimum financial burden varieties of higher performance • Data were produced using the MiSeq and NextSeq 500 sequencing platforms (Illumina Inc., San. Diego, CA, USA)	• Data were produced from plant biological samples (leaf and fruit) • Collection was conducted in two different plant stages (plantlets and mature plants) • Genomic data were produced using standard and customized protocols at CREA and CERTH facilities • Data produced from Illumina platforms were stored in compressed text files (fastq) • Genomic data, although in plain text format, are big volume data and pose challenges in their storage, handling and processing • Analysis was performed using CREA and CERTH's HPC computational facilities
Biochemistry, agronomy, metabolomics	To characterize the biochemical profile of fruits from tomato varieties used for breeding. Data were produced from different chromatographs, mass spectrometers, wet lab, NIRS	Data was mainly proprietary binary sets converted to XML or other open formats. Data were acquired from biological samples of tomato fruits
IoT, sensor, and environmental data	To characterize growing environments and crop management	Environmental indoor/outdoor, farm data/log/profile

Table 6.2 Phenomics, metabolomics, genomics and environmental datasets

Field	Value
Name of the dataset/API provider	Phenomics, metabolomics, genomics and environmental datasets
Short description	This dataset includes phenomics (sensor data), metabolomics, genomics, environmental (IoT) data, as well as genomic predictions and selection data
Data type	Raw text, CSV data
Dataset/API owner/responsible contacts	ephrem.habyarimana@crea.gov.it, argiriou@certh.gr
Data Volume	30 TB (5 TB/year/institution)
Geographical coverage	Regions of Emilia Romagna (Italy) and Thessalia (Greece)

Genomics data used in the DataBio project resulted from genomic DNA (Deoxyribonucleic acid) of the plant species of interest resequenced using Illumina sequencing platform consisting of high-throughput Next Generation sequencers. The genomic data included SNPs (Single Nucleotide Polymorphisms), InDels (Insertions / Deletions), SVs (Structure Variations), and CNVs (Copy Number Variation). A Single Nucleotide Polymorphisms is a variation caused by changing of a single nucleotide (A, T, C or G) in the genome. The SNPs, including switch and reverse of single nucleotide bases, are responsible for genome diversity between species and between individuals of the sample species. InDel refers to insertion mutation, deletion mutation or both, including what happened in the early stage of evolution. CNVs, a form of structural variations, are alterations of the DNA of a genome that results in the cell having an abnormal number of copies of one or more sections of the DNA. CNVs correspond to relatively large regions of the genome that have been deleted (fewer than the normal number) or duplicated (more than the normal number) on certain chromosome. Structural Variation includes deletion, insertion, duplication, inversion and transposition of long fragment (at least 50 bp) in genome.

In the process of whole-genome resequencing, genomic DNA (gDNA) libraries are prepared (Fig. 6.3) and sequenced; Images generated by sequencers are converted by base calling into nucleotide sequences, which are called raw data or raw reads and are stored in FASTQ format.

FASTQ files are text files that store both read sequences and their corresponding quality scores. Each read is described in four lines as follows [4, 5]:

@FCB068CABXX:6:1101:1403:2159#TAGGTTAT/1

GTAGAAGACTTATAGATTAAAATTCTCCAACATATAGATGTCCTTACA

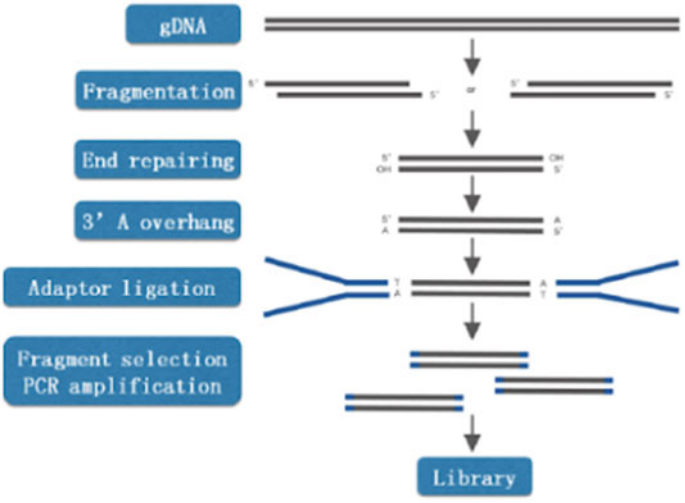

Fig. 6.3 Genomic DNA library construction workflow

CCGTTTTCCTTTGCTCAGCAGGCTCCGTGTTTGCTTGTCCTT
+
c‘bcc_cˆccde_df\c_aeff‘ffcfffdfedadcaˆ‘b_eed‘fe\fed\babdbaˆ
Yeebeccfdeae_eecˆdbXbda‘]bcbebc

where line 1 is the DNA sequence identifier and description, lines 1 and 3 are sequence names generated by the sequencer; line 2 is the DNA sequence letters; line 4 is sequencing quality scores, in which every letter corresponds to a base in line 2; the base's sequencing quality is the ASCII value that the letter in line 4 refers to minus 64 (Specification). For example, the ASCII value of c is 99, so the corresponding sequencing quality value is 35. In this work, the quality value of sequencing bases ranged from 2 to 35; the higher the sequencing quality, the lower the sequencing error rate. For instance, the sequencing qualities of 13 and 30 correspond to error rates of 5% and 0.1%, respectively.

The generated raw reads were processed through bioinformatics analysis to filter the raw data and generate clean (reads) data. The filtered reads are subsequently aligned to the reference sequence, the alignment processed and the variation (SNPs, InDels, SVs, and CNVs) detected according to the standard Workflow (Fig. 6.4), which constitute the genomics data used in genomic prediction and selection models.

6.3 Genomic Data Management Systems

Generation of DNA data requires laboratories equipped with molecular biology infrastructure for basic techniques (e.g. DNA extraction, library construction), along with advanced technologies such as Next Generation Sequencing (NGS) and computational facilities. To date, there are many new high throughput NGS platforms available on the market producing sequence data at a very low cost per sequenced base, affordable even for small-scale laboratories [6]. Hence, large-scale genomic data can be produced and analyzed by many scientists, providing the breeder accurate information at the genomic level, for selection of candidates before crosses, in a short time. Among the advantages these technologies offer is accelerating breeding by genomic selection, thus, bypassing time-consuming cultivation and field testing. Additional advantages are the implementation of genomic selection to inform intercrosses and recurrent selection, and predicting instead of field evaluating the offspring.

In the real world, breeders often have a short window of time to decide and take actions on their breeding schemes by the time they receive phenotypic and genotypic data, and this is among the major challenges for many commercial agriculture applications. In order to implement genomic selection routinely as part of breeding programs, data management systems and analytics capacity have to be in order. In short, infrastructures and software that will enable scientists to design and analyse multi-phenotype and multi-omics experiments for maximal data-to-information conversion, are required. This is the major challenge in order to efficiently exploit the huge volume and complexity of the information produced.

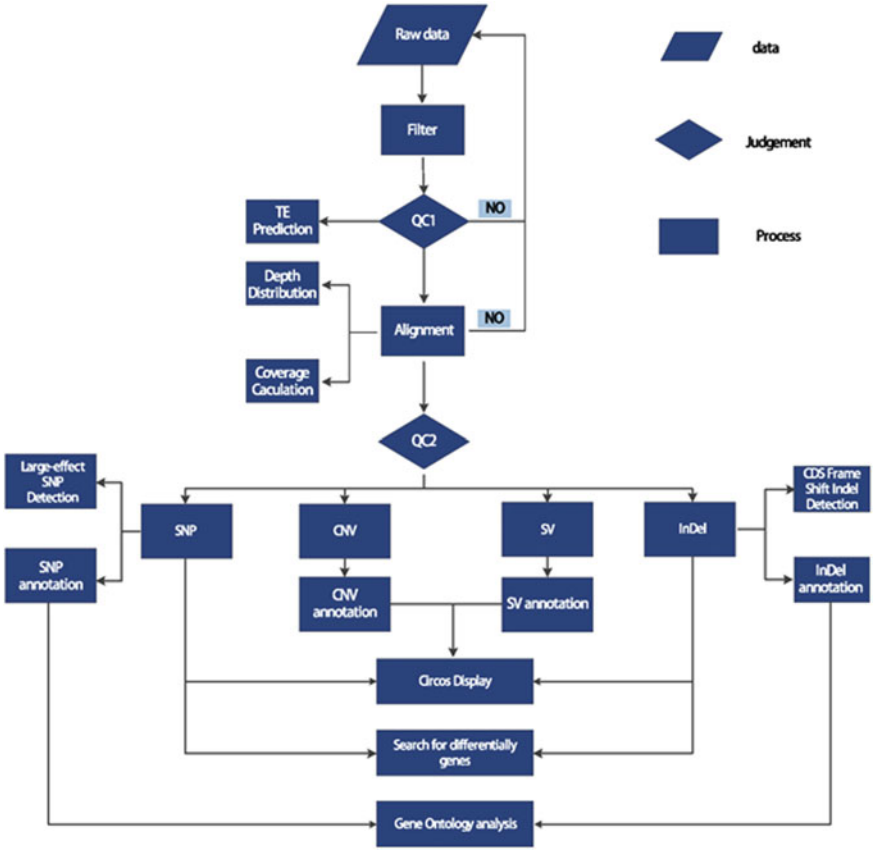

Fig. 6.4 Workflow of standard bioinformatics analysis

The genomic data management system must be able to efficiently store and retrieve huge volumes of genomic information with high complexity and provide rapid data extraction for computation. The system must be scalable and flexible for large breeding programs while being able to run effectively in situations with limited access to large computational clusters. For this purpose, traditional relational database management systems (RDBMS) offer many appealing characteristics. The RDBMS systems are designed and built to store, manage and analyze large-scale data. However, performance can be problematic, when dealing with large matrix data like those commonly encountered in genomic research. To address this performance issue, many RDBMS were upgraded with the capabilities for working with binary large objects (BLOBs). In addition, NoSQL systems have been considered more recently as effective tools for managing high dimensional genomic data [7]. NoSQL systems for distributed file storage and searching represent scalable solutions comparable to RDBMS, when dealing with semi-structured data types. MongoDB system, for instance, is a document-based NoSQL database, which has been used to

develop web-based tools for visualizing and exploring genotypic information. The Hierarchical Data Format (HDF5) is a member of the high-performance distributed file systems family. It is designed for flexible, efficient I/O and for high-volume and complex data. It has demonstrated superior performance with high-dimensional and highly structured data such as genomic sequencing data making it an appealing option for a hybrid system approach.

References

1. Habyarimana, E., Lopez-Cruz, M. (2019). Genomic selection for antioxidant production in a panel of sorghum bicolor and S. bicolor × S. halepense Lines. Genes 10:841. https://doi.org/10.3390/genes10110841.
2. Habyarimana, E. (2016). Genomic prediction for yield improvement and safeguarding genetic diversity in CIMMYT spring wheat (Triticum aestivum L.). *Australian Journal of Crop Science, 10,* 127–136.
3. Habyarimana, E., Parisi, B., & Mandolino, G. (2017). Genomic prediction for yields, processing and nutritional quality traits in cultivated potato (Solanum tuberosum L.). *Plant Breeding, 136,* 245–252. https://doi.org/10.1111/pbr.12461.
4. Mount, W. D. (2004). *Bioinformatics: Sequence and genome analysis* (2nd ed.). Cold Spring Harbour Laboratory Press.
5. Gibas, C., Jambeck, P. (2001). *Developing bioinformatics computer skills* (1st ed.). O'Reilly Media, Beijing.
6. Habyarimana, E., Lopez-Cruz, M., & Baloch, F. S. (2020). Genomic selection for optimum index with dry biomass yield, dry mass fraction of fresh material, and plant height in biomass sorghum. *Genes, 11,* 61. https://doi.org/10.3390/genes11010061.
7. Nti-Addae, Y., Matthews, D., Ulat, V. J. et al. (2019). Benchmarking database systems for genomic selection implementation. Database (Oxford) 2019. https://doi.org/10.1093/database/baz096.