What is Multi-Omic Data Data Integration?

Author: Joaquim Ollé López[1], Tutor: Antonio Barbadilla Prados[1]
[1]Bioinformatics of genomics diversity, Institute of Biotechnology and Biomedicine, Universitat Autònoma de Barcelona, Barcelona, Spain

Introduction

The project objectives are: [1] The **definition of concepts** to understand Data Integration. [2] Study the **classification** of different methods. [3] A bibliographic study of some **applications**.

The **Hypothesis** of this report is: Data Integration is an efficient analysis to understand globally the biological phenomes, more than the separately analysis of omic data.

The **Big Data** is the large amounts of complex data that are generated, manipulated, processed and stored in applied in several fields as Education and Healthcare.

Omic Sciences are the global study of the different molecular levels where Big Data is applied (e.g. the study of Genome, Genomics).

The **Data integration** (Fig. 1) is a computer holistic approach that study multi-omic data to answer biological questions building complex network models and identifying key factors.

There are basically two assumptions when a complex character is tested:

- **Assumption A**, based on central dogma of molecular biology, that genetic variables entail differential gene expression, which affect the protein expression and finally the phenotype.
- **Assumption B,** the existence of interactions in and between different molecular levels, which may affect to the phenotype.

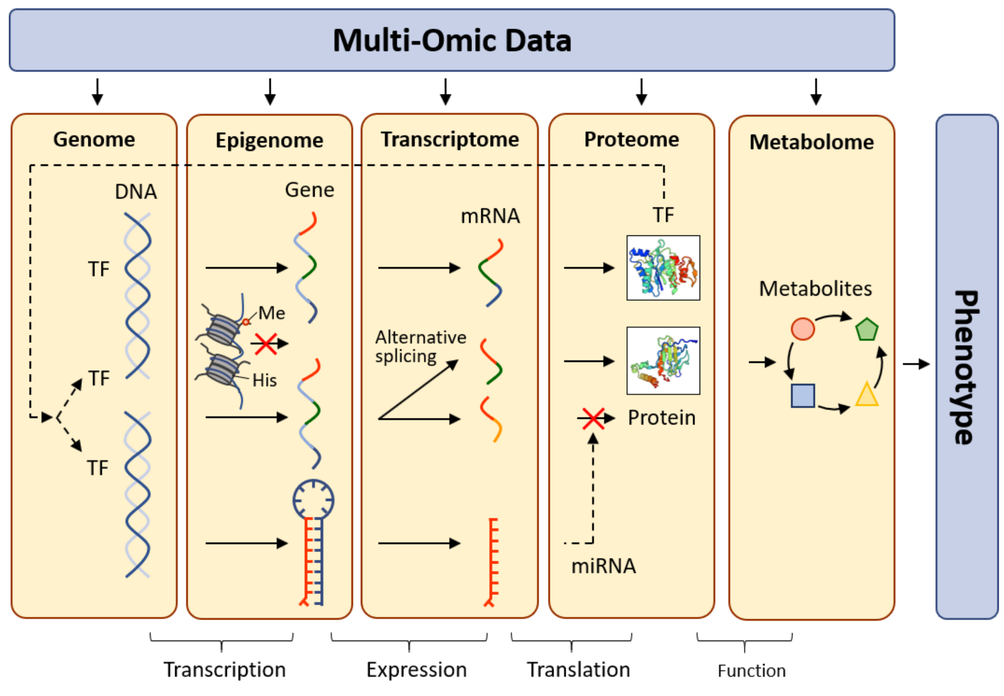

Figure 1: Biological systems multi-omics from the genome, epigenome, transcriptome, proteome and metabolome to the biological phenome or phenotype. Adapted from Ritchie MD et. al. (2015).

Methods (Ritchie et al. 2015)

Multi-staged analysis

Based on Assumption A, types of **sequential analysis** (Fig. 2):

Genomic variation analysis approaches

Also called the **Triangle Method**:

[1] Filtering of SNPs that pass certain genome-wide significance threshold. [2] Test these variants against expression, methylation or protein levels obtaining Quantitative Trait Loci: eQTL, mQTL or pQTL. [2] Infer correlation of both data groups with the phenotype.

Allele-specific expression approaches

[1] Identification of each parental allele molecular product. [2] Association with level expression (eQTLs) and transcript variants. [3] Test the functional differences and infer correlation with phenotype.

Domain knowledge-guided approaches

The approach is a derivation of the genomic variation analysis. [1] Triangle Method. [2] Test the results against previous knowledge from specific databases. [3] Infer the correlation with phenotype.

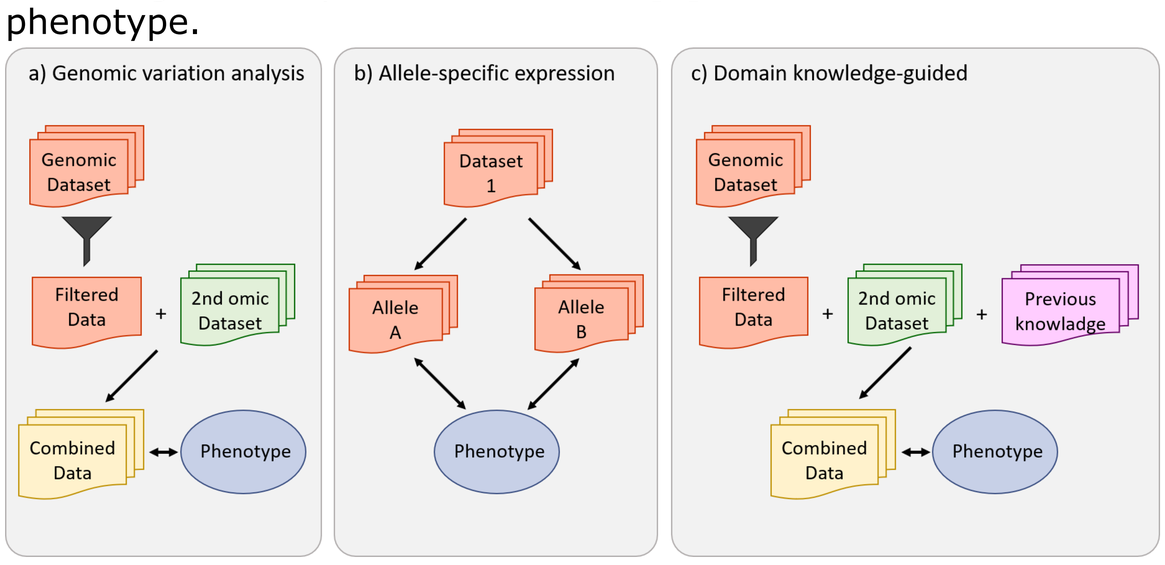

Figure 2: Categorization of Multi-staged analysis. a) Genomic variation analysis approaches. b) Allele-specific expression approaches. c) Domain knowledge-guided approaches.

Meta-dimensional analysis

Based on Assumption B aim to **analyse** different types of data **at the same time** to avoid losing information (Fig. 3):

Concatenation-based Integration

[1] Merge the different omic data in a **single matrix**. [2] Apply biostatistics and probabilistic models more efficiently. [3] Build a Final Model to test against phenotype.

Transformation-based Integration

[1] Group and transform specific data to **graphs or Kernel matrices**. [2] Join all the information in a final analysis. [3] Build a Final Model to test against phenotype.

Model-based Integration

[1] **Build models** separately from different omic experiments. [2] Combine the best models of each prior analysis. [3] Build a Final Model to test against phenotype.

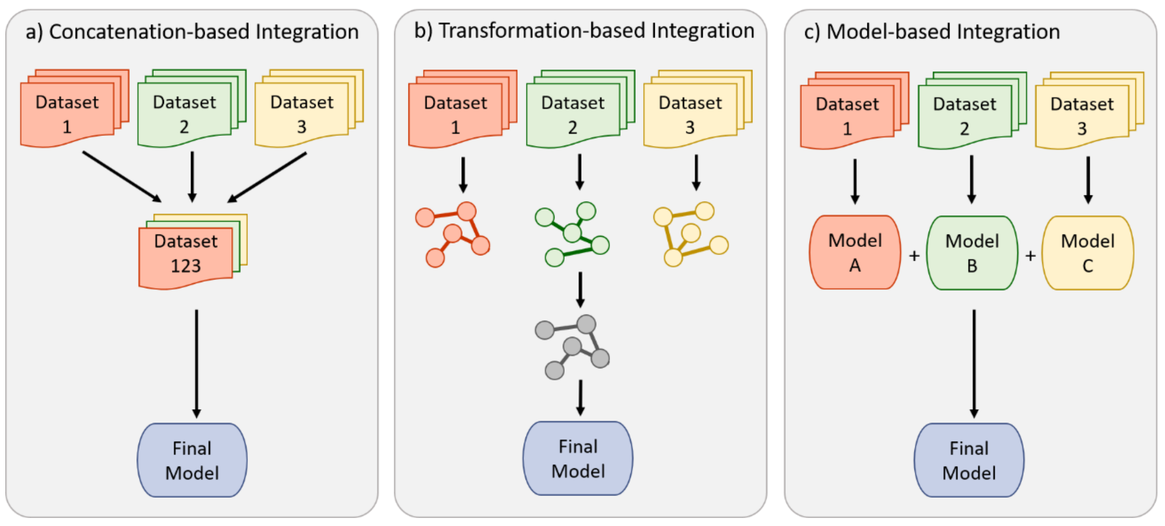

Figure 3: Categorization of Meta-Dimensional analysis. a) Concatenation-based Integration. b) Transformation-based Integration. c) Model-based Integration. Adapted from Ritchie et al. (2015).

Applications

Animal improvement

The objective is **increase the efficiency and the sustainability** of animal production. Improving non-complex character, like muscle mass, by genetic selection is possible studying QTLs. However, performing genetic improvement for complex characters remains a challenge.

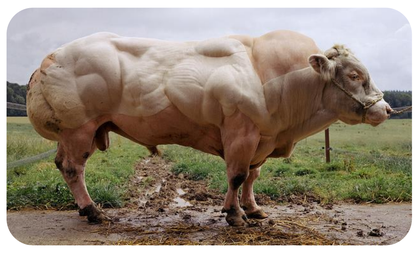

Figure 4: Belgian Blue Cattle, 8-21% of the muscle mass heritability is explained for QTLs (Druet et al. 2014).

Personal Medicine

The aim is enable **personal treatments** in humans by three approaches (Hasin, Seldin, and Lusis 2017): [1] **Genome First,** are Multi-staged analysis type C. [2] **Phenotype first**, are Multi-staged analysis type B (Fig. 2). [3] **Environment first**, are studies based on the relation of disease mechanism and environmental factors. They are especially complex because of the large variability between subjects.

Conclusions

1. The Data Integration is a **complex bioinformatic process** that is described for a vast quantity of authors and there isn't a consensus classification yet.
2. There must be a compromise of a quantity of **tested variables** to have an efficient computation but detecting **complex networks** to understand the biologic phenome.
3. There is not a general method applicable for all cases. However, Data Integration is **more efficient** analysis **for complex characters** than single omic analysis.
4. Price and technical **limitations will be reduced** and multi-omic data will be applied even more, so constant analysis improvement is required

Bibliography

1. Ritchie, M. D., Holzinger, E. R., Li, R., Pendergrass, S. A. & Kim, D. Methods of integrating data to uncover genotype–phenotype interactions. *Nat. Rev. Genet.* **16,** 85–97 (2015).
2. Druet, T. *et al.* Selection in action: dissecting the molecular underpinnings of the increasing muscle mass of Belgian Blue Cattle. *BMC Genomics* **15,** 796 (2014).
3. Hasin, Y., Seldin, M. & Lusis, A. Multi-omics approaches to disease. *Genome Biol.* **18,** 1–15 (2017).