



UNIVERSITAT
POLITÈCNICA
DE VALÈNCIA

Statistical methods development for the multiomic systems biology

Author: Manuel Ugidos Guerrero

Supervisors: Dr. Sonia Tarazona Campos
Dr. Ana Conesa Cegarra
Dr. Alberto J. Ferrer Riquelme

Contents

1 Introduction	1
1.1 The concept of Systems Biology	3
1.2 Omic technologies	5
1.3 Nature and structure of omic data	11
1.4 Statistics for systems biology	15
1.4.1 Single-omic data analysis	15
1.4.1.1 Omic data pre-processing and normalization	16
1.4.1.2 Batch effect correction	18
1.4.1.3 Identification of differentially expressed/quantified features	20

1.4.1.4	Identifying differentially activated pathways by functional enrichment analyses	22
1.4.2	Multiomeric integration	25
1.4.2.1	Level 1: Element-based integration	26
1.4.2.2	Level 2: Pathway-based integration	33
1.4.2.3	Level 3: Model driven approaches	34
2	Motivation, Objectives and Contributions	41
2.1	Motivation	43
2.2	Objectives of this thesis	46
2.3	Contributions	48
2.3.1	Articles in peer-reviewed journals	49
2.3.2	Conference contributions	51
2.3.3	Software	51
2.3.4	Teaching	52
3	Understanding and removing batch effects on a multiomic scenario	53
3.1	Introduction	55
3.2	Data	59
3.2.1	A yeast multiomic dataset obtained at different laboratories	59

3.2.2	Proof of concept data.....	60
3.2.3	Simulated data.....	62
3.3	BECAs usage: estimating the batch effect.....	66
3.3.1	Methodology behind BECAs.....	67
3.3.1.1	Limma.....	67
3.3.1.2	ComBat.....	68
3.3.1.3	ARSyN method.....	68
3.3.2	Comparison of BECAs' performance.....	70
3.4	BECAs for multiomic data.....	72
3.4.1	MultiBaC: A multiomic BECA.....	72
3.4.2	Other multiomic batch effect correction approaches.....	76
3.4.2.1	Missing data imputation strategy:.....	77
3.4.2.2	Product transfer model:.....	77
3.5	MultiBaC validation.....	79
3.5.1	Validation of MultiBaC PLS models.....	79
3.5.2	Multiomic BECAs comparison on simulated data... ..	84
3.5.2.1	Latent space concordance.....	84
3.5.2.2	Differential expression analysis.....	85
3.5.3	MultiBaC validation using "Proof of Concept" data.....	87
3.5.4	MultiBaC application to a real problem.....	88

3.6 MultiBaC implementation as an R package	95
3.6.1 ARSyN batch effect correction	99
3.6.2 MultiBaC correction	101
3.6.3 Visualization of results	102
3.7 Discussion	106
4 Generating a multiomic dataset	111
4.1 Introduction	113
4.2 Experimental design	116
4.3 Statistical methods	118
4.3.1 Sequencing data pre-processing	118
4.3.2 Differential expression/quantification analysis	122
4.3.3 Time-series data modeling	124
4.3.4 Multiple testing correction	126
4.3.5 Multiomic integration	127
4.4 Data acquisition and preprocessing	129
4.4.1 RNA-seq	129
4.4.2 Metabolomics	130
4.4.3 Histone modifications	132
4.5 Technical validation	135
4.6 Omic-wise differential expression/quantification analysis	135

4.7 Multiomic data integration138

 4.7.1 PEA identifies significant differences between strains supported by all omics.....138

 4.7.2 Multi-Block PLS finds consistent changes across omics142

4.8 Discussion.....144

5 Development of a model-driven multiomic integration approach 147

5.1 Introduction149

5.2 Data and computational details153

5.3 Description of the approach154

 5.3.1 Flux Balance Analysis.....154

 5.3.2 Integration of gene/protein associated information into GEMs157

 5.3.3 Formalizing multiomic-based constraints in MAMBA 157

 5.3.3.1 Gene/protein associated data 159

 5.3.3.2 Metabolomics data 162

 5.3.4 Evaluation of MAMBA method166

 5.3.4.1 Sensitivity and Robustness analysis . . 166

5.3.4.2	Evaluation of metabolite prediction accuracy	168
5.4	Results	169
5.4.1	MAMBA model	169
5.4.2	Application of MAMBA to a yeast heat-shock dataset improves metabolic prediction accuracy	171
5.4.3	Deciphering differential behavior between strains . . .	177
5.4.4	Evaluating the effect of mip6 affinity on metabolic changes	182
5.4.5	Metabolic control by ChIP-seq signal	186
5.5	Discussion	187
6	Conclusions and Future Work	193
6.1	Conclusions	195
6.1.1	Objective 1: To develop the different specific pre-processing pipelines for each omic data type	195
6.1.2	Objective 2: To develop a batch effect correction algorithm for multiomic integration strategies	196
6.1.3	Objective 3: To develop novel multiomic integration approaches	197
6.2	Research relevance	198
6.3	Future research lines	199

Appendix 1: Material	203
Appendix 2: Scripts for public data preprocessing	207
Appendix 3: MultiBaC R package vignette	219
Appendix 4: List of differential reactions and their activation state	243
Appendix 5: Pathway Activation Score (PAS) for the list of relevant pathways.	291
References	347

