Framework for the integration of genomics, epigenomics, and transcriptomics in complex diseases

S Pineda (1,2), P Gomez-Rubio (1), A Picornell (1), K Bessonov (2), M Márquez (1), M Kogevinas (3), FX Real (4,5), K Van Steen (2,6), N Malats (1)

(1) Genetic and Molecular Epidemiology Group, Spanish National Cancer Research Centre (CNIO), Madrid, Spain.

(2) Systems and Modeling Unit, Montefiore Institute, University of Liége, Liége, Belgium

(3) Centre for Research in Environmental Epidemiology (CREAL), Barcelona, Spain; Institut Municipal d'Investigació Mèdica - Hospital del Mar, Barcelona, Spain.

(4) Epithelial Carcinogenesis group, Spanish National Cancer Research Centre (CNIO), Madrid, Spain.

(5) Departament de Ciències Experimentals i de la Salut, Universitat Pompeu Fabra, Barcelona, Spain.

(6) Bioinformatics and Modeling, GIGA-R, University of Liege, Avenue de l'Hôpital 1, Liége, Belgium

Corresponding authors:

N Malats, MD, MPH, PhD
Genetic and Molecular Epidemiology Group, Spanish National Cancer Research Centre (CNIO), C/Melchor Fernández Almagro, 3, 28029 Madrid, Spain, Phone: +34-912-246-900 (ext. 3330), Fax: +34-912-246-911, E-mail: nmalats@cnio.es

K Van Steen, Prof., PhD, PhD
Systems and Modeling Unit, Montefiore Institute, University of Liege, Bât. B28 Bioinformatique, Grande Traverse 10, Liége 4000, Belgium. Tel: +32 4 366 2692; E-mail: kristel.vansteen@ulg.ac.be

Abstract

Objectives: Different types of 'omics' data are becoming available in the post genome era; still a single 'omics' assessment provides limited insights to understand the biological mechanism of complex diseases. Genomics, epigenomics and transcriptomics data provide insight into the molecular dysregulation of neoplastic diseases, among them urothelial bladder cancer (UBC). Here we propose a detailed analytical framework necessary to achieve an adequate integration of the three sets of 'omics' data to ultimate identify previously hidden genetic mechanisms in UBC. *Methods*: We build a multi-staged framework to study possible pairwise combinations and integrate data in three-way relationships. SNP genotypes, CpG methylation levels, and gene expression levels were determined for a total of 70 individuals with UBC and with available fresh tumor tissue. *Results*: We suggest two main hypothesis-based scenarios for gene regulation based on the "omics" integration analysis where DNA methylation affects gene expression and genetic variants co-regulate gene expression and DNA methylation. We identified several three-way trans-association "hotspots" that are found at the molecular level and that deserve further studies. *Conclusions*: The proposed integrative framework allowed us to identify relationships at the whole genome level providing some new biological insights and highlighting the importance of integrating 'omics' data.

Introduction

Big data at the molecular field (‘omics’ data) is being generated at an unprecedented pace, this including genome, methylome, transcriptome, and microbiome, among others. There is a growing interest in combining the different types of ‘omics’ datasets that are becoming available since a single ‘omics’ assessment provides limited insights into the understanding of the underlying biological mechanisms of a physiological/pathological condition. For example, even when many genome-wide association studies (GWAS) have identified several Single Nucleotide Polymorphisms (SNP) involved in complex diseases, the functional implications of the susceptibility loci are still poorly understood and they only partially account for the phenotype variability. Combining different ‘omics’ data types seems to be a more suitable approach, as it will likely reveal previously hidden information.

The simplest form of data integration involves the combination of two different data types, common examples being genetic variants and gene expression or, more recently, genetic variants and DNA methylation [1]. DNA methylation involves the addition of a methyl group to the 5’ position of the cytosine at a Cytosine-phosphate-Guanine (CpG) site. Genomic regions with high density of CpG dinucleotides are denominated CpG islands; they are often located in gene promoters and have important roles in gene regulation. CpG sites located up to 2kb from the island’s boundaries are called CpG shores and it has been demonstrated that they are also very important for gene regulation and that they are implicated in cancer [2]. Both CpG islands and shores, when hypermethylated and located in the promoter region of a gene, negatively regulate gene repression [3]. Therefore, it is important to take into account the relationship between DNA methylation and gene regulation in order to better understand complex diseases [4]. For example, it has been shown that hypermethylation of CpGs located in the promoter region of some tumor suppressor genes (*INK4A, Rb, VHL, hMLH1, BRCA1,* etc) contribute to cancer development [5]. Therefore, analyzing gene expression data without considering epigenetics provides an incomplete genomic explanation of the transcriptome. Moreover, as DNA methylation regulates gene expression, genetic variants affecting CpG sites might, in turn, affect gene expression too. It is well known that genetic variants can alter gene expression levels and hence the importance of connecting the DNA sequence to the RNA level. The identification of these expression quantitative trait loci (eQTL) relationships may help to identify regulators of gene expression [6]. These eQTLs have been extensively studied to find associations between common genetic variants and gene expression levels [7–11]. By contrast, the study of potential associations between common variants, DNA methylation levels (methylation QTLs, methQTLs), and gene expression has generated less interest, so far [1,12–15].

Genome, transcriptome, and methylome data offer unique opportunities when combined in the same analyses. This strategy has been applied to HapMap cell lines [14], whole blood from healthy human subjects [16], and human monocytes [17]. Furthermore, some studies have combined these types of data to better understand complex diseases, such as breast cancer [18] or type 2 diabetes [19]. As DNA methylation is tissue-specific, these analyses have also been applied to different types of tissues, such as human brain [12] or adipose tissue [15]. It is worth noting that the majority of these studies have only assessed *cis*- relationships, but *trans*- effects deserve further study within the 'omics' context, especially as the complex organization of chromatin in the nucleus is better understood.

In the present study we built and propose a multi-staged analytical framework to integrate 'omics' data. We tested it in an urothelial bladder cancer (UBC) model using common genetic variants, DNA methylation, and gene expression transcripts data from 70 cancer patients. We proved the ability of the framework to identify some "multi-omics" relationships that provided further knowledge to better understand the biological mechanisms underlying the disease.

Material and Methods

Study Subjects: SNP genotypes, CpG methylation levels, and gene expression levels were measured for a total of 70 individuals with available fresh tumor tissue that were recruited as part of the pilot phase of the EPICURO study. All of them were histologically confirmed UBC cases recruited in 2 hospitals in Spain during 1997-1998. Tumor DNA and RNA were extracted and used for 'omics' assessment. SNP data was available for 46 patients, CpG methylation for 46 patients and gene expression for 43. The overlapping of patients between the three 'omics' was 31 for the expression-methylation relationship, 27 for the eQTL, and 46 for the methQTL studies.

SNP genotype data: Genotyping was performed using Illumina HumanHap 1M array in tumor samples. A total of 1,047,101 SNPs were genotyped in 46 individuals. For genotype calling, we used the cluster file obtained when the same array was applied to germline DNA from 2,424 subjects included in the main EPICURO study. We considered SNPs with <5% of missing values and with a minor allele frequency (MAF) ≥ 0.01. Standard Quality Control (QC) was performed using BeadStudio and R. From BeadStudio, the genotypes (AA, Aa, aa) were obtained in forward strand for those samples having a call rate higher than 90%.

DNA methylation data: After bisulphite modification of 46 tumor DNA samples using EZ-96 DNA METHYLATIONGOLD KIT (Zymo Research, Irvin, CA, USA), CpG methylation data was generated using the Infinum Human Methylation 27 BeadChip Kit that detected the CpG sites with two probes, one designed against the unmethylated site (signal U) and the other against the methylated site (signal M). The level of methylation was determined at each locus by the intensity of the two possible fluorescent signals [20]. At each CpG site, the methylation levels were measured with the β-value, defined as:

$$\beta = \frac{\max(M,0)}{\max(U,0) + \max(M,0) + 100}$$

The maximum between signal intensity and 0 is used for β calculation to avoid the negative numbers caused by background subtractions, consequently, β-values rank between 0 (unmethylated) and 1 (methylated). The constant 100 was used to regularize the β-values when they were very small. Although β-values are useful under some circumstances, it has been demonstrated that M-values are more statistically valid than β-values due to a better approximation of the homocedasticity [21]. This property is important when applying regression models that require this assumption. The M-value is calculated as follows:

$$M = log_2\left(\frac{\max(M,0) + 1}{\max(U,0) + 1}\right)$$

It ranges between -∞ (unmethylated) and +∞ (methylated). In our study, M-values were used when applying linear regression models, while β-values were used in the rest of the analyses.

The initial number of CpGs in the studied array was 27,578. We then applied BeadStudio software and R to preprocess the data. Background normalization was performed minimizing the amount of variation in background signals between arrays and, as recommended by Illumina, CpGs were rejected when detection p-value was > 0.05. The β-values < 0 or > 1 were also excluded. CpGs with SNPs (N=908) or cross reactive probes (N=2,985) were deleted based on earlier reports for the 27K array [22]. After QC, a total number of 23,034 CpGs were kept for analysis. These were classified in 3 categories for subsequent analyses: CpG islands (located in the promoter region of a gene), CpG island shores (in a sequence up to 2Kb from an island) and CpGs outside of an island or a shore.

Gene expression data: Gene expression data were obtained from 43 tumor samples using the Affymetrix DNA Microarray Human Gene 1.0 ST Array with 32,321 probes. This array was based on 2006 (UCSC hg19, NCBI build 37) human genome sequence with coverage of RefSeq, Ensembl and putative complete CDS GenBank transcripts (www.affymetrix.com). QC was performed using Bioconductor libraries in R (www.bioconductor.org/). The arrayQualityMetrics package [23] was used to implement a background correction and to carry out normalization of expression levels across arrays. Application of QC steps resulted in 20,899 probes and 37 individuals. The affy library in R [24] was used to annotate the probes.

Statistical Analysis

First, tumoral DNA methylation levels in CpG sites and gene expression levels were compared using Spearman's rank correlation for non-normally distributed variables. Second, we assessed eQTLs and methQTLs, via linear regression modeling for those expression-methylation pair probes that were strongly associated in the previous step. To perform these analyses, we obtained a linear regression model for each SNP as:

$$Gene\ Expression_i = \alpha + \beta * SNP_i$$

$$Methylation\ CpG\ _i = \alpha + \gamma * SNP_{\hat{\imath}}$$

Prior to analysis, we excluded those SNPs that had less than two individuals per genotype due to the imbalance that may produce a highly differential gene expression values, i.e: an individual with rare homozygous genotype and with an extreme gene expression value that could produce an artificial high significant p-value.

Expression-methylation probe pairs and eQTLs and methQTLs were classified in three categories according to possible genomic distance effects: *cis*-acting, if probes were located within 1Mb; *trans*-acting, if probes were on the same chromosome but located more than 1Mb apart; and *trans*-acting-outside, if they were on different chromosomes. To control the analyses for multiple testing we applied the Benjamini & Yekutieli [25] FDR method that allows for panel dependencies between tests. We applied this correction taking into account the number of tests performed in the eQTL and the methQTL study independently. Finally, we checked the regions of the trait-associated SNPs already published for UBC.

Third, in line with the study, we integrated the results obtained from pairwise analyses on genome, epigenome and trascriptome data. We checked the SNPs that were common in the eQTL and methQTL analysis based on those probes-CpGs that were previously correlated in order to have a complete view of the genome in individuals with UBC. We obtained the distribution of the triplets (SNP-CpG-Gene expression) that were significantly associated in the same relationship.

Statistical analyses were performed with R and results were visualized with Circos software [26].

Results:

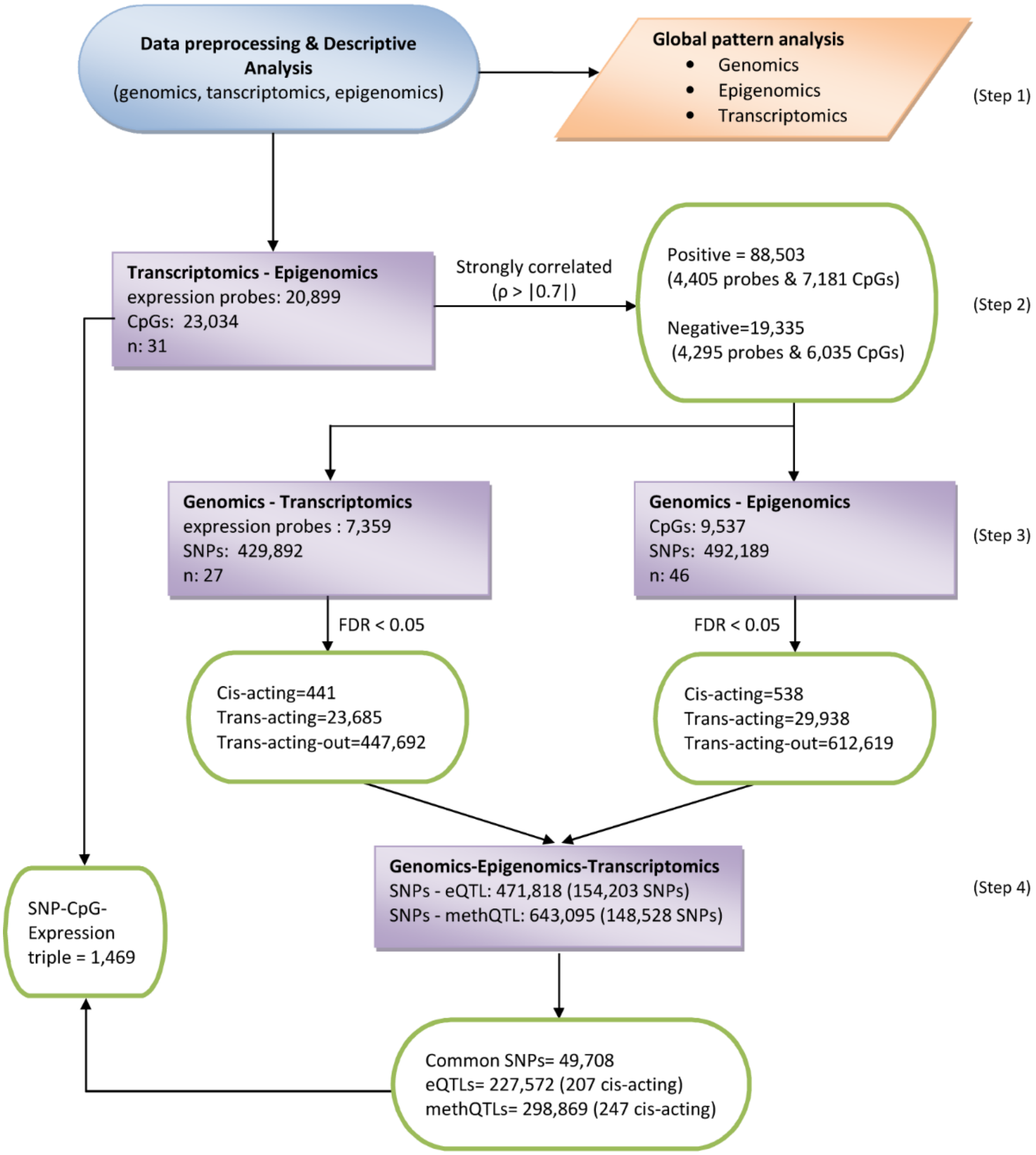
The majority of the individuals included in our study were male (93%) and current (50%) or former (36%) smokers. According to established criteria based on tumor stage (T) and grade (G) for UBC, individuals were classified as having low-risk non-muscle invasive tumors (45%), high-risk non-muscle invasive tumors (22%) or muscle-invasive tumors (29%) (Table 1).

Table 1. Characteristics of the studied patients

Characteristics	N (%)
Total	72
Gender	
Male	67 (93)
Female	5 (7)
Age	
Mean (SD)	65.6 (9.5)
Min-max	41-80
Region	
Barcelona	31 (43)
Elche	41 (57)
Smoking status	
Non-smoker	8 (11)
Current	36 (50)
Former	26 (36)
Unknown	2 (3)
Tumor-stage	
Low-grade-NMIBC	32 (45%)
High-grade-NMIBC	16 (22%)
MIBC	21 (29%)
Unknown	3 (4%)

The description of the study results is organized in four sections following the framework steps proposed (Figure 1): (1) Description of the patterns of individual 'omics' data, globally and according to epidemiological data, (2) Correlation analysis between methylation and expression probes, (3) Identification of *cis*- and *trans*- eQTLs and methQTLs, and (4) Integration of results derived from the previous pairwise analysis.

Figure 1. Framework for data integration showing the steps to integrate genetic variants, DNA methylation levels, and gene expression levels. Step 1 corresponds to the preprocessed data, quality control and global patterns individually per data set. Steps 2, 3 and 4 are represented for purple boxes corresponding to the analysis performed and the input data, and green oval boxes correspond to the results and the input of the next step.



1. Patterns of individual 'omics' data. Table 2 shows the distribution of the genotypes according to their MAF; 14% had a MAF of 0 and were excluded from the analysis, 11% ranged between (0.01-0.05], 30% between (0.05-0.2] and 31% between (0.2-0.4]. Missingness <5% was observed in 84% of the SNPs.

Table 2. Summary of SNPs genotyped

SNPs	N (%)
Total number	1,047,101
MAF	
[0.0]	150,548 (14)
(0.0 – 0.01]	0 (0)
(0.01 – 0.05]	108,496 (11)
(0.05 – 0.2]	312,220 (30)
(0.2 – 0.4]	327,762 (31)
(0.4 – 1.0]	148,075 (14)
Missingness	
No missing	488,288 (47)
5% missing	400,918 (38)
20% missing	147,732 (14)
> 20% missing	10,163 (1)

MAF = 0.0 means that all individuals are common homozygous for the measured SNP.

The patterns for DNA methylation according to the β- and M-values were different for autosomal chromosomes and X-chromosomes in females due to the X-chromosome inactivation in females. The majority (71%) of CpGs in autosomal chromosomes were unmethylated ($\beta < 0.3$) while, as expected, the majority of the CpGs (66%) in the X-chromosomes showed β-values in the range ($0.3 \leq \beta < 0.7$). While the M-values for autosomal chromosomes displayed a bimodal distribution, X-chromosomes approximated a normal distribution (Supplementary Figure 1). No significant different methylation patterns were found according to the clinical/epidemiological data considered, i.e. smoking status, tumor stage, age, and sex (Pearson's χ^2-test, data not shown).

The expression of the gene probes after background correction and normalization followed a normal distribution (Supplementary Figure 2). We did not find any significant difference according to the clinical/epidemiological data by applying student's *t*-test (data not shown).

2. Correlation between gene expression and DNA methylation. While it is well established that DNA methylation may affect the expression of a gene, mainly when the relationship is in *cis*-, little is known when it is in *trans*-. We investigated a total of 481,387,566 possible correlations between gene expression and methylation both in *cis*- and in *trans*-. The number of comparisons performed was based on data derived from 31 individuals (Table 3). We obtained 19,335 strong-negative (ρ < -0.7) and 88,503 strong-positive (ρ > 0.7) associations between gene expression and methylation corresponding to 7,359 expression traits and 9,537 CpG sites. The distribution of the stronger relationships according to the CpG location and direction is shown in Table 4: 5,414 (28%) were located in CpG islands, 1,690 (59%) in CpG shores and 2,433 (57%) outside of CpG islands/shores. There were 263 (0.03%) *cis*-acting correlations, 6,177 (0.02%) *trans*-acting correlations within the same chromosome, and 101,398 (0.02%) *trans*-acting outside the chromosome (*trans-out* correlations). A whole list of CpGs with significant *cis*- association with a gene can be found in Supplementary Table 1.

Table 3. Strength of correlations between gene expression and DNA methylation

Spearman's rho	**Strength of correlation**	**Nº of combinations**
(-0.9 : -1.0]	Very Strong-negative	0
(-0.7 : -0.9]	Strong-negative	19,335
(-0.4 : -0.7]	Moderate-negative	9,266,544
(-0.0 : -0.4]	Weak-negative	238,601,864
[0.0]	No correlation	380,834
(0.0 : 0.4]	Weak-positive	223,165,638
(0.4 : 0.7]	Moderate-positive	9,864,848
(0.7 : 0.9]	Strong-positive	88,503
(0.9 : 1.0]	Very Strong-positive	0

Table 4. Strong correlation for *cis*-acting and *trans*-relationships between CpG methylation and gene expression

		Negative correlation	Positive correlation
		N (%)	N (%)
Cis-acting (same gene)	CpG island/shore	37 (80)	9 (20)
	CpG outside	3 (37)	5 (63)
Cis-acting (dif. gene)	CpG island/shore	41 (26)	116 (74)
	CpG outside	11 (21)	41 (79)
Trans-acting	CpG island/shoe	757 (17)	3,736 (83)
	CpG outside	412(24)	1,272 (76)
Trans-acting-outside chromosome	CpG island/shore	11,860 (16)	63,054 (84)
	CpG outside	6,214 (23)	20,270 (76)

3. Identification of *cis*- and *trans*- eQTLs and methQTLs. In order to detect genetic variants affecting gene expression or DNA methylation, we investigated a total of 7,359 expression traits and 9,537 CpG sites that were strongly correlated in the previous step. The number of SNPs considered here after QC was 429,892 for the eQTL and 492,189 for the methQTL analyses, resulting in a total of 3,163,575,228 eQTLs in 27 individuals and 4,694,006,493 methQTLs explored in 46 individuals. After correction for multiple testing (FDR<0.05), we obtained 471,818 significant eQTLs involving 154,203 SNPs, and 643,095 methQTLs involving 148,528 SNPs. These results pointed to the fact that multiple expression probes and CpGs were significantly associated with more than one SNP. We refer to this phenomenon as "hotspots" (Supplementary Figure 3). We show the distribution of QTLs classified by genomic distance and MAF of the relationship for eQTLs in Table 5 and methQTLs in Table 6. When classifying the QTLs by genomic distance we observed 441 *cis*-eQTLs (0.02%), 23,685 *trans*-eQTLs (0.01%) and 447,692 *trans-out*-eQTLs (0.01%); and 538 *cis*-methQTLs (0.01%), 29,938 *trans*-methQTLs (0.01%), and 612,619 *trans-out*-methQTLs (0.01%). When classifying the QTLs in terms of MAF the majority had a MAF ≤ 0.2 (0.006%), while 0.003% and 0.002% had MAFs of (0.2-0.4] and ≥ 0.4, respectively. Detailed information regarding the *cis*- relationship is provided in supplementary tables 2 and 3. When we checked how the significant findings are distributed in terms of the direction of the relationship, there were more QTLs positively than negatively (60% vs. 40% eQTL, 63% vs. 37% methQTLs) associated implying that having more copies of the rare allele increases the levels of the gene expression or the levels of methylation. Lastly, we investigated, for QTL associations in our study, how many of the SNPs involved have been previously reported as a trait associated SNPs for UBC. We found that the SNP rs401681-*TERT/CLPTM1L* on chromosome 5 was associated with the expression of *FRMD6* located on chromosome 14 (p-value = 3.7*10^{-5}), and with the cg18368125-*TMED6* on chromosome 16 (p-value = 4.8*10^{-5}). Also, the SNP rs1495741-*NAT2* on chromosome 8 was associated with the expression of *C19orf73* located in chromosome 19 (Figure 2).

Table 5: Significant (FDR<0.05) *cis*-eQTLs and *trans*-eQTLs by MAF and sign of the association

MAF	Sign	cis-eQTL N (%)	trans-eQTL N (%)	Trans-out-eQTL N (%)
(0.01-0.2]	Positive	106 (0.005)	7,026 (0.005)	127,177 (0.004)
	Negative	56 (0.002)	2,857 (0.002)	61,134 (0.002)
(0.2-0.4]	Positive	95 (0.003)	4,759 (0.003)	88,213 (0.003)
	Negative	66 (0.002)	3,220 (0.002)	65,457 (0.002)
> 0.4	Positive	57 (0.003)	2,930 (0.002)	54,087 (0.002)
	Negative	61 (0.003)	2,893 (0.002)	51,624 (0.002)

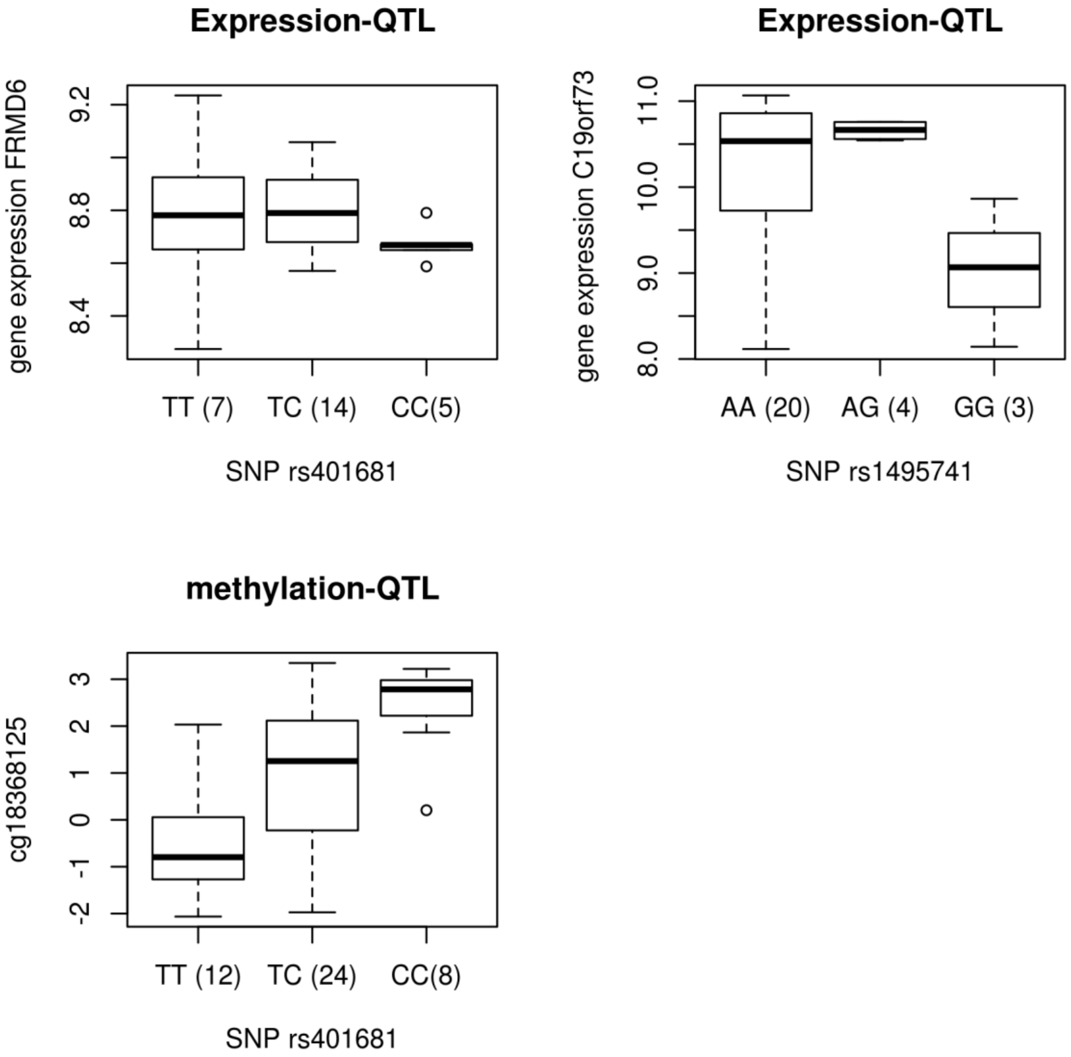
%: Percentage of significant eQTLs after multiple testing correction over the total number of *cis*- (2,331,808), *trans*- (151,738,928) and *trans*-out (3,009,504,492) eQTL

Table 6: Significant (FDR<0.05) *cis*-methQTLs and *trans*-methQTLs by MAF and sign

MAF	Sign	cis-methQTL N (%)	trans-methQTL N (%)	trans-methQTL-out N (%)
(0.01-0.2]	Positive	137 (0.004)	8,576 (0.004)	190,221 (0.004)
	Negative	61 (0.002)	3,554 (0.002)	72,611 (0.002)
(0.2-0.4]	Positive	118 (0.003)	6,864 (0.003)	139,830 (0.003)
	Negative	139 (0.004)	5,230 (0.002)	98,068 (0.002)
> 0.4	Positive	39 (0.001)	3,090 (0.001)	57,476 (0.001)
	Negative	44 (0.001)	2,624 (0.001)	54,413 (0.001)

%: Percentage of significant methQTLs after multiple testing correction over the total number of *cis*- (3,499,636), *trans*- (224,328,090) and *trans*-out (4,466,178,767) methQTL.

Figure 2. GWAS-reported SNPs significantly associated with gene expression levels and/or DNA methylation levels in UBC.



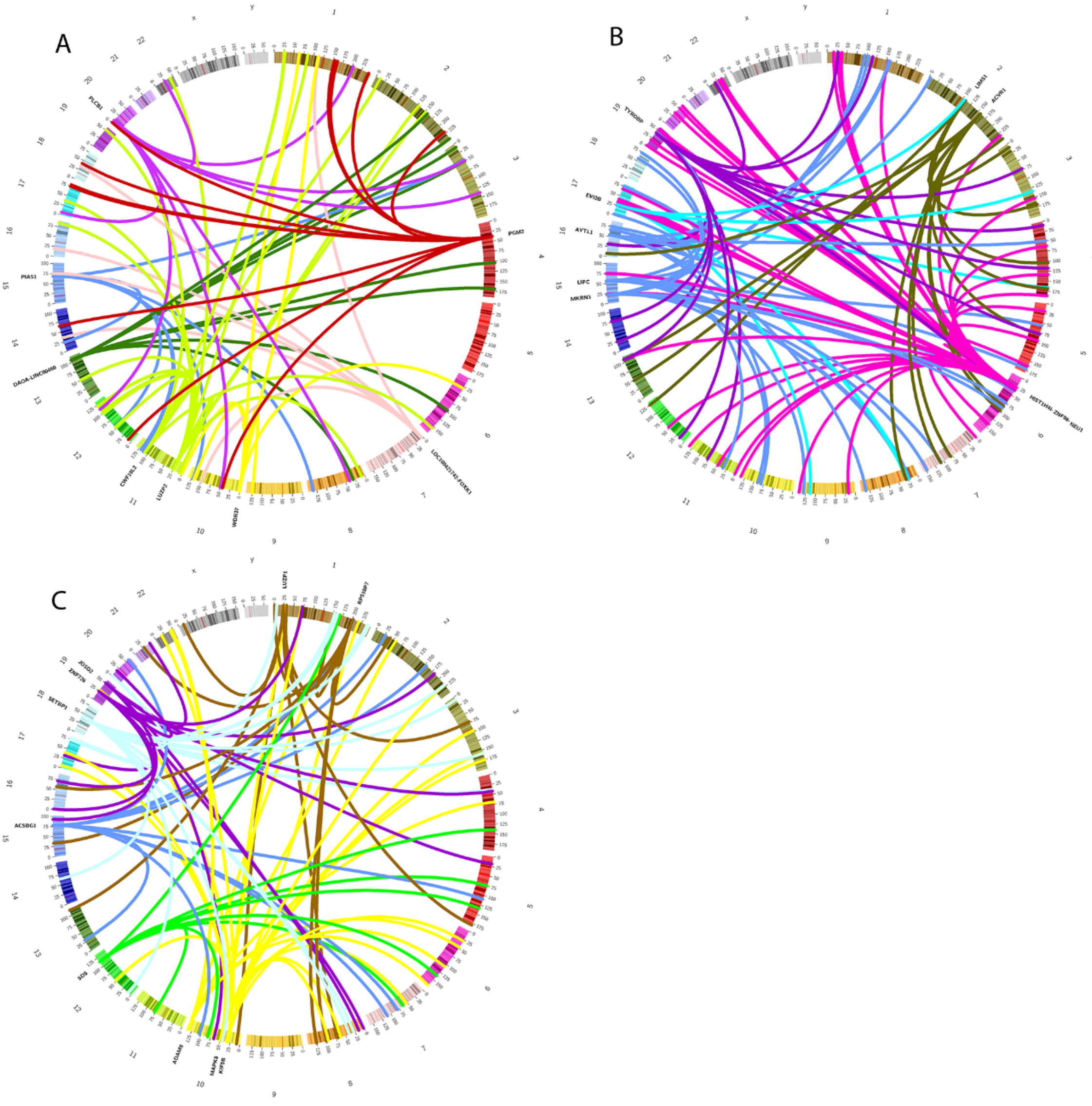
4. Integration of results derived from the pairwise analysis. From the final subset of eQTLs and methQTLs, we obtained 49,708 common SNPs (50% from the total SNPs for eQTLs and methQTLs), affecting a total of 227,572 eQTLs (207 *cis*-acting) and 298,869 methQTLs (247 *cis*-acting). Multiple expression probes and CpGs were significantly associated with more than one SNP and vice versa. We found that 1,469 QTLs belonged to a triple relationship (SNP-CpG-Gene expression) (Supplementary Table 4). Regarding the association patterns, majority (29%) of these 1,469 triplets show a positive association pattern, that is, the higher the methylation the higher the expression, where the rare allele is classified with higher expression and methylation levels. A second pattern (19%) regarded to "the higher the methylation the lower the expression", where the rare allele is associated with high expression levels and low methylation levels. When restricted to *cis*-relationship, no triplets were found but there were 19 pairs (1 eQTL, 1 methQTL and 17 CpG-Gene expression pairs) that were in *cis*. The distribution of these triplets was completely different than that of the rest of the triplets. The most frequent pattern (32%) show a positive association between the SNP and methylation and negative for the association of both (SNPs and CpGs) with the expression. All the possible patterns with their percentages are shown in Table 7. Lastly, we checked for the "hotspots" in these triplets and we found some of them for SNPs, CpGs and Gene Expression probes (Figure 3).

Table 7: Distribution of the 1,946 triple relationships directions per pairwise analysis

eQTL	methQTL	Expr-methy	N[1] (%)	N[2] (%)
+	+	+	419 (29)	1 (5)
-	-	-	58 (4)	3 (16)
+	-	-	276 (19)	4 (21)
-	+	+	78 (5)	1 (5)
-	+	-	262 (18)	6 (32)
+	-	+	62 (4)	3 (16)
-	-	+	250 (17)	1 (5)
+	+	-	64 (4)	0 (0)

[1] The total distribution for the 1,469 triplets
[2] The distribution only for the ones that had one pair in *cis*-effect

Figure 3. Circular representation of the "hotspots" found for SNPs (A), CpGs (B) and gene expression probes (C) extracted from the relationships on the triplets. Each chromosome is represented with a different color and the color of the lines corresponds to the SNPs, CpGs or gene expression probes that are located in the chromosome that share the color with. The name of the genes is located in the gene with the "hotspot".



Discussion

The post genome era delivers a wealth of 'omics' data allowing to explore the relationships between genetics, epigenetics and gene expression being of great importance to better understand the biological mechanism underlying a disease. In the cancer field, this integrative approach becomes particularly crucial on the basis of the knowledge indicating that SNPs, CpGs, and gene expression play an important role in the development of these complex diseases [27,28].

In this work, we propose an 'omics' integrative analytical framework based on a multi-staged strategy and we apply it to explore the relationships between three sets of data measured at a genome-wide level in UBC tumor samples. We provide further evidences on how common genetic variation and DNA methylation are statistically associated with the regulation of gene expression. Based on the knowledge that DNA is looped, allowing the interaction between two DNA regions located far away from each other, we not only studied *cis*- but also *trans*- relationships [29]. Here, we show that some SNPs are associated with DNA methylation, that the latter is associated with gene expression, and that some SNPs associate with both DNA methylation and gene expression.

Individual and pairwise analysis

The global pattern for methylation observed in our study (Supplementary Figure 1) parallels that reported previously for germline (blood) [14]. Consistently with previous studies performed in blood [14,16] and human brain samples [13], we found that - when located in an island/shore - the correlations between DNA methylation and gene expression from the same gene are predominantly negative, supporting the known biological mechanisms of gene regulation (80%). DNA methylation occurs near the Transcription Start Site (TSS) of a gene, blocking the initiation of gene expression

(Review in [3]). To highlight relevant results, four different CpGs (cg01354473, cg07778029, cg25047280, cg26521404) located in a CpG island of *HOXA9* gene on chromosome 8 were negatively correlated with the expression of the gene. It was reported that *HOXA9* acts as a tumor suppressor gene in oral cancer [30] while methylation of this gene has been associated with the regulation of its expression in UBC [31] and with risk of different cancers such as breast [32], oral cavity [33], and ovarian [34], as well as with risk of recurrence in UBC [35]. The observed negative association between four CpGs and *HOXA9* expression in our study suggests that the inhibition of *HOXA9* expression may affect the development of UBC and supports the approach applied in this study.

On the other hand, the ENCODE Project provided some clues in the understanding of the biological behavior of *trans*- relationships and of the CpGs belonging to *cis*-relationships when located in a different gene [36]. In our study, we mainly observed positive correlations (79%) in all of these scenarios, meaning that increasing levels of methylation correlates with increasing levels of gene expression or the other way around, suggesting either a direct mechanism or an indirect mechanism where methylation affects expression of a gene repressor, thus leading to apparent association with increased gene levels. These results warrant further mechanistic studies explaining the complex association between DNA methylation and gene expression.

Little is known about the relationship between genetic variants and DNA methylation. Heyn *et al*. [1] recently published a methQTL analysis using the cancer genome atlas data but only with SNPs detected in GWAS studies and *cis*-acting methQTLs. They detected one methQTL in UBC where the SNP rs401681 in *TERT_CLPTM1L* was associated with cg06550200 located in *CLPTM1L*; unfortunately we have not been able to replicate this association as this CpG is not present in the 27K methylation array. Nonetheless, for the first time we have performed *cis*- and *trans*- acting methQTL analysis in UBC tumor tissue samples using CpGs that were previously correlated with gene expression. From this assessment, we found 538 *cis*- relationships listed in the supplementary Table 3 with all necessary information for further studies and validation. More frequently, *cis*-relationships between genetic variants and gene expression levels have been assessed. We also performed eQTL association studies in *cis*- and *trans*- in the same conditions that for methQTLs and found 441 *cis*-eQTLs (Supplementary Table 2). We performed these analyses on significant expression-methylation correlated probes identified in the first step upon the assumption that epigenetics interferes with the gene expression levels.

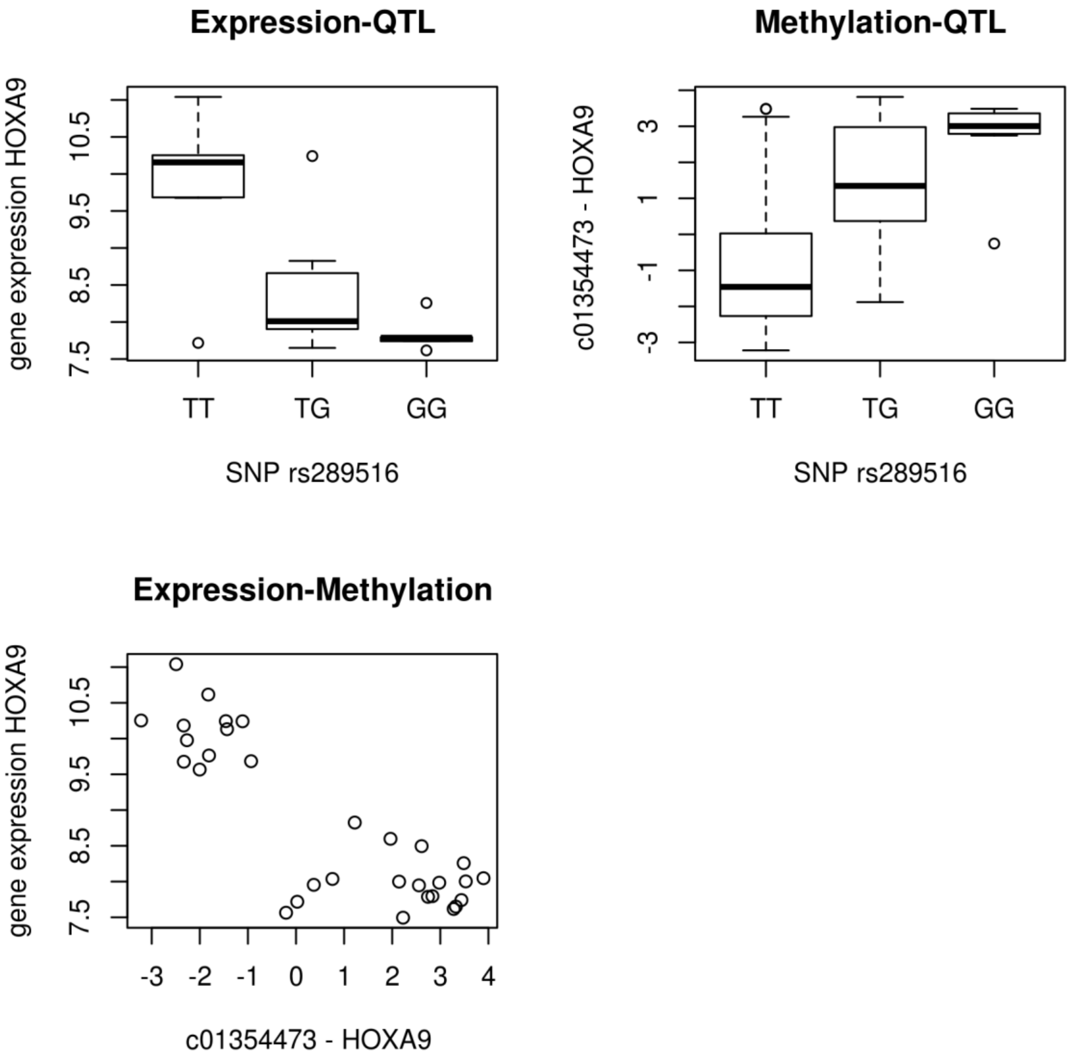
The proportion of eQTLs (0.01%, 471,818) and methQTLs (0.01%, 643,477) was similar, although more SNPs were involved in eQTLs (32.6%, 154,203) than in methQTLs (22.7%, 148,528), possibly because of the smaller sample size of the former. Similarly, we found no major differences in the percentages of QTL associations classified as *cis*-, *trans*- and *trans-out* according to the genomic distance defined before. Nevertheless, when considering the MAF distribution, a higher number of QTLs were observed for SNPs with MAF ≤ 0.2. While these results should be interpreted cautiously, due to the possibility of false positives, it is worth highlighting that we found a greater number of positive than negative QTLs relationships, meaning that having the rare allele is associated with increased gene expression or methylation levels.

Some studies have related SNPs associated with complex diseases at genome-wide significance level to gene expression or methylation levels [1,10,37]. Out of the 14 GWAS UBC SNPs [38], two showed to be associated with gene expression and methylation in *trans*-relationships (Figure 2). Interestingly, rs401681-*TERT/CPTL1M,* a variant strongly associated with low grade and low risk UBC [38], was found associated with a lower expression of *FRMD6* in our study, a gene that was reported to be involved in the inhibition of proliferation in human cells [39].

Integrative analysis

We observed an enrichment of significant associations of genetic variants with methylation and gene expression with 49,708 SNP related to 227,572 eQTLs and 298,869 methQTLs (207 *eQTLs* and 247 methQTL in *cis*-) suggesting a co-regulated expression and methylation. The percentage of enrichment associated with eQTLs (11.5%) and methQTLs (10.0%) was similar to that found by Wagner et al. [40] who detected an enrichment of 9.5% in fibroblasts. Bell et al. [14] also found an enrichment in lymphoblastoid cell lines. By contrast, Gibbs et al. [12] found only a modest overlap between both data in brain tissues, while Drong et al. [15] found no enrichment in adipose tissue. This highlights the fact that a specific genetic variants may show tissue-specific effects and that little is known about them at a genome wide level. We also found a total of 1,469 QTLs where the same SNP was significantly associated with both eQTL and methQTL in previously identified gene expression-CpG significant pairs. This three-way type relationship between SNP-CpG-Gene expression supports the notion that the three data sets implemented in this study are closely related in regulating part of the genome, an observation that may provide new insight into the genetics of this complex disease. Furthermore, we observed that the most frequent pattern (29%) in these three way relationships is a positive association pattern, suggesting that hypermethylation may act through a direct mechanisms or affect a repressor gene associated with an over-expression of gene levels. In addition, having the rare allele is associated with hypermethylation and over-expression pattern. This finding together with the fact that, in our study, we have demonstrated that 82% of the CpGs that are related with gene expression in *trans*-effect are positively correlated suggest that if one SNP is co-regulating both, this relation should be positive. Thus, we could hypothesize that the rare allele of the SNP associates with hypermethylation that, at the same time, associates with over-expression, as a possible regulation scenario in *trans*-effect. When inspecting the *cis*-relationships, no triplets were found, but there were 19 pairs (1 eQT, 1 methQTL and 17 CpG-gene expression pairs) that were in *cis*. In this scenario, the most frequent pattern (32%) suggests that having the rare allele is associated with hypermethylation and under-expression where the expression and methylation are associated inversely. This fact suggests another possible regulation scenario based on previous findings. We demonstrated that the 79% of the CpGs located in the promoter region of the gene are negatively correlated in *cis* with the gene expression levels; meaning that higher methylation levels may affect to a decrease in the gene expression levels. An example of this scenario is shown in Figure 4 where the SNP rs289516 located in gene *DLC1* is negatively associated in *trans* with the expression of *HOXA9* (β = -1.1; p-value = 3.7*10^{-5}) and positively with the cg01354473 located in the island of the *HOXA9* gene (β = 1.8; p-value = 9.9*10^{-5}). The relationship between the expression and the methylation levels in *HOXA9* gene was already reported as negatively correlated (r^2 = -0.7; p-value = 1.4*10^{-5}). It has been already published that the methylation of *HOXA9* is negatively correlated with the gene expression in UBC [31] as we observed in our study. We added a new step on this complex scenario, since the SNP rs289516 is also involved in this triple relationship. This SNP belongs to the *DLC1* gene considered as a tumor suppressor gene and the particular SNP has been picked up in two GWAS, one for asthma [41] and one for breast cancer [42], but any of them passed the GWAS significant threshold. Other examples with biological support are the triplet composed by the SNP rs29658399 located in gene *DNAH11*, the gene expression of *HSPA1A*, and the cg00929855 located in gene *HSPA1A*. It has been published that the *HSPA1A* promoter methylation underlies the defect in gene expression reduction observed in UBC cell lines [43]. In addition we found some “hotspots” in these triplets regarding SNPs, CpGs and gene expressions probes. In the circos plot (Figure 3A) we observed a predominant relation for one SNP (rs10569 located in the gene *PGM2*) in chromosome 4. *PGM2* is a protein-coding gene and is associated with diseases such as pneumonia and hypoxia. While alterations in this gene have not yet been directly associated with cancer, hypoxia is a known relevant process for tumor survival. This SNP was positively associated with the expression of *SETBP1,* coding for an important cancer gene located in chromosome 18 that is observed also as a predominant "hotspot" in Figure 3C. Somatic mutations in *SETBP1* [44], as well as its expression patterns [45], are related with myeloid leukemia disease. Moreover in Figure 3B we observed a very predominant "hotspot" regarding three CpGs belonging to three different genes but close located in chromosome 6; Two of them (cg02622316 located in the gene *ZNF96* and cg02599464 located in the gene *HIST1H41)* were already published as hypermethylated in individuals with muscle invasive bladder cancer [46]. The first one is associated positively with many SNPs and gene expression probes and the second is associated positive and negative with some SNPs and positively with some gene expression probes. A more detailed discussion of the potential biological findings than involved the triple relationships is beyond this particularly study and detailed results about all the combinations are provided in Supplementary Table 4.

Figure 4. Example of one triple relationship where integrated common genetic variants with DNA methylation and gene expression in one of the main possible scenarios for regulation.



The integrative framework

We built and propose a multi-staged ‘omics’ integration framework that its application does not require a strong methodological knowledge, being easy and effective to use. The multi-staged framework we applied has the advantage of analyzing data of all subjects that overlap among pairs of data and has not to restrict only to the few individuals with a complete overlap among all the data types. Thus, we take advantage of more samples using this framework than integrating the data in a multi-dimensional model. Therefore, we show here the application for the first time of multi-staged framework that allowed us to (1) integrate more than two ‘omics’ data for the same set of individuals, (2) dissect the biological relationships that may point to new mechanisms involved in the development/progression of UBC through a hypothesis-based models built step by step, and (3) to envision the complexities of the general scenario of genomic regulation.

Conclusions

While these results are exciting, we acknowledge the following limitations. First, in this study we use the 27K methylation array that only covers a selection of CpG sites making infeasible to replicate previous reported findings using the 450k array. Second, statistical power is a commonplace in any QTL analysis given the extensive amount of data analyzed and the small sample size. While this limitation needs to be considered in the interpretation of the results, it is worth mentioning that a large enough size will unlikely be available to meet the standard criteria of statistical power; therefore, our study represents a proof of concept in the integrative ’omics’ field. In addition, while we might not be able to address for unmeasured confounding factors, no differences were found between demographic factors and methylation and gene expression in our series. Validation of these results to discard false positive findings is not trivial due to the multiple genomic factors, the models considered, and the characteristics of the series. Despite these limitations, this study has several strengths. We have performed the study in tumor samples what gave us the opportunity to study in detail the regulation of three types of ‘omics’ data in UBC providing some evidences on the genomics regulation of the tumor. We have applied an easy, reproducible, and detailed framework to perform an integrative study of the relationships between genetic variations, DNA methylation and gene expression, showing a whole spectrum of the associations between them. We have shown that ‘omics’ data integration helps unraveling biological mechanisms involved in UBC. All these relations may help in the identification of new molecular targets to be further explored in detail, mainly regarding *trans*- relationships.

In conclusion, this study provides the scientific community with a pipeline to integrate more than two sets of ‘omics’ data that can be applied in future analyses seeking to better understand the biology behind the complex diseases. In addition, we highlight the importance of integrating ‘omics’ data to identify new genetic mechanisms in UBC. While several pieces of evidences support these findings, they still require of experimental validation to be considered conclusive.

REFERENCES

1. Heyn H, Sayols S, Moutinho C, Vidal E, Sanchez-Mut J V, et al. (2014) Linkage of DNA methylation quantitative trait Loci to human cancer risk. Cell Rep 7: 331–338. doi:S2211-1247(14)00194-6 [pii] 10.1016/j.celrep.2014.03.016.

2. Irizarry RA, Ladd-Acosta C, Wen B, Wu Z, Montano C, et al. (2009) The human colon cancer methylome shows similar hypo- and hypermethylation at conserved tissue-specific CpG island shores. Nat Genet 41: 178–186. doi:10.1038/ng.298.

3. Jones PA (2012) Functions of DNA methylation: islands, start sites, gene bodies and beyond. Nat Rev Genet 13: 484–492. doi:10.1038/nrg3230.

4. Portela A, Esteller M (2010) Epigenetic modifications and human disease. Nat Biotechnol 28: 1057–1068.

5. Esteller M (2008) Epigenetics in cancer. N Engl J Med 358: 1148–1159. doi:358/11/1148 [pii] 10.1056/NEJMra072067.

6. Cheung VG, Spielman RS (2009) Genetics of human gene expression: mapping DNA variants that influence gene expression. Nat Rev Genet 10: 595–604. doi:10.1038/nrg2630.

7. Nica AC, Montgomery SB, Dimas AS, Stranger BE, Beazley C, et al. (2010) Candidate causal regulatory effects by integration of expression QTLs with complex trait genetic associations. PLoS Genet 6. doi:10.1371/journal.pgen.1000895.

8. Nicolae DL, Gamazon E, Zhang W, Duan S, Dolan ME, et al. (2010) Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS. PLoS Genet 6: e1000888. doi:10.1371/journal.pgen.1000888.

9. Pickrell JK, Marioni JC, Pai AA, Degner JF, Engelhardt BE, et al. (2010) Understanding mechanisms underlying human gene expression variation with RNA sequencing. Nature 464: 768–772. doi:10.1038/nature08872.

10. Westra HJ, Peters MJ, Esko T, Yaghootkar H, Schurmann C, et al. (2013) Systematic identification of trans eQTLs as putative drivers of known disease associations. Nat Genet 45: 1238–1243. doi:ng.2756 [pii] 10.1038/ng.2756.

11. Zhernakova D V, de Klerk E, Westra H-J, Mastrokolias A, Amini S, et al. (2013) DeepSAGE reveals genetic variants associated with alternative polyadenylation and expression of coding and non-coding transcripts. PLoS Genet 9: e1003594. doi:10.1371/journal.pgen.1003594.

12. Gibbs JR, van der Brug MP, Hernandez DG, Traynor BJ, Nalls MA, et al. (2010) Abundant quantitative trait loci exist for DNA methylation and gene expression in human brain. PLoS Genet 6: e1000952. Available: http://dx.plos.org/10.1371/journal.pgen.1000952.

13. Zhang D, Cheng L, Badner JA, Chen C, Chen Q, et al. (2010) Genetic control of individual differences in gene-specific methylation in human brain. Am J Hum Genet 86: 411–419. doi:S0002-9297(10)00087-X [pii] 10.1016/j.ajhg.2010.02.005.

14. Bell JT, Pai AA, Pickrell JK, Gaffney DJ, Pique-Regi R, et al. (2011) DNA methylation patterns associate with genetic and gene expression variation in HapMap cell lines. Genome Biol 12: R10. doi:10.1186/gb-2011-12-1-r10.

15. Drong AW, Nicholson G, Hedman AK, Meduri E, Grundberg E, et al. (2013) The presence of methylation quantitative trait loci indicates a direct genetic influence on the level of DNA methylation in adipose tissue. PLoS One 8: e55923. doi:10.1371/journal.pone.0055923 PONE-D-12-31174 [pii].

16. Van Eijk KR, de Jong S, Boks MPM, Langeveld T, Colas F, et al. (2012) Genetic analysis of DNA methylation and gene expression levels in whole blood of healthy human subjects. BMC Genomics 13: 636.

17. Liu Y, Ding J, Reynolds LM, Lohman K, Register TC, et al. (2013) Methylomics of gene expression in human monocytes. Hum Mol Genet 22: 5065–5074. doi:ddt356 [pii] 10.1093/hmg/ddt356.

18. Li Q, Seo JH, Stranger B, McKenna A, Pe’er I, et al. (2013) Integrative eQTL-based analyses reveal the biology of breast cancer risk loci. Cell 152: 633–641. doi:S0092-8674(12)01556-5 [pii] 10.1016/j.cell.2012.12.034.

19. Greenawalt DM, Sieberts SK, Cornelis MC, Girman CJ, Zhong H, et al. (2012) Integrating genetic association, genetics of gene expression, and single nucleotide polymorphism set analysis to identify susceptibility Loci for type 2 diabetes mellitus. Am J Epidemiol 176: 423–430

20. Bibikova M, Le J, Barnes B, Saedinia-Melnyk S, Zhou L, et al. (2009) Genome-wide DNA methylation profiling using Infinium(R) assay. Epigenomics 1: 177–200. doi:10.2217/epi.09.14.

21. Du P, Zhang X, Huang C-C, Jafari N, Kibbe WA, et al. (2010) Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis. BMC Bioinformatics 11: 587. doi:10.1186/1471-2105-11-587.

22. Chen Y, Choufani S, Ferreira JC, Grafodatskaya D, Butcher DT, et al. (2011) Sequence overlap between autosomal and sex-linked probes on the Illumina HumanMethylation27 microarray. Genomics 97: 214–222. doi:10.1016/j.ygeno.2010.12.004.

23. Kauffmann A, Gentleman R, Huber W (2009) arrayQualityMetrics--a bioconductor package for quality assessment of microarray data. Bioinformatics 25: 415–416.

24. Gautier L, Cope L, Bolstad BM, Irizarry RA (2004) affy--analysis of Affymetrix GeneChip data at the probe level. Bioinformatics 20: 307–315. doi:10.1093/bioinformatics/btg405 20/3/307 [pii].

25. Benjamini Y, Yekutieli D (2001) The control of the false discovery rate in multiple testing under dependency. Ann Stat 29: 1165–1188.

26. Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, et al. (2009) Circos: an information aesthetic for comparative genomics. Genome Res 19: 1639–1645. doi:10.1101/gr.092759.109.

27. You JS, Jones PA (2012) Cancer genetics and epigenetics: two sides of the same coin? Cancer Cell 22: 9–20.

28. Kanwal R, Gupta S (2012) Epigenetic modifications in cancer. Clin Genet 81: 303–311. doi:10.1111/j.1399-0004.2011.01809.x.

29. Bickmore WA, van Steensel B (2013) Genome architecture: domain organization of interphase chromosomes. Cell 152: 1270–1284. doi:10.1016/j.cell.2013.02.001.

30. Uchida K, Veeramachaneni R, Huey B, Bhattacharya A, Schmidt BL, et al. (2014) Investigation of HOXA9 promoter methylation as a biomarker to distinguish oral cancer patients at low risk of neck metastasis. BMC Cancer 14: 353. doi:10.1186/1471-2407-14-353.

31. Reinert T, Modin C, Castano FM, Lamy P, Wojdacz TK, et al. (2011) Comprehensive genome methylation analysis in bladder cancer: identification and validation of novel methylated genes and application of these as urinary tumor markers. Clin Cancer Res 17: 5582–5592. doi:10.1158/1078-0432.CCR-10-2659.

32. Gilbert PM, Mouw JK, Unger MA, Lakins JN, Gbegnon MK, et al. (2010) HOXA9 regulates BRCA1 expression to modulate human breast tumor phenotype. J Clin Invest 120: 1535–1550. doi:10.1172/JCI39534.

33. Guerrero-Preston R, Soudry E, Acero J, Orera M, Moreno-López L, et al. (2011) NID2 and HOXA9 promoter hypermethylation as biomarkers for prevention and early detection in oral cavity squamous cell carcinoma tissues and saliva. Cancer Prev Res (Phila) 4: 1061–1072. doi:10.1158/1940-6207.CAPR-11-0006.

34. Wu Q, Lothe RA, Ahlquist T, Silins I, Tropé CG, et al. (2007) DNA methylation profiling of ovarian carcinomas and their in vitro models identifies HOXA9, HOXB5, SCGB3A1, and CRABP1 as novel targets. Mol Cancer 6: 45. doi:10.1186/1476-4598-6-45.

35. Reinert T, Borre M, Christiansen A, Hermann GG, Ørntoft TF, et al. (2012) Diagnosis of bladder cancer recurrence based on urinary levels of EOMES, HOXA9, POU4F2, TWIST1, VIM, and ZNF154 hypermethylation. PLoS One 7: e46297. doi:10.1371/journal.pone.0046297.

36. The ENCODE (ENCyclopedia Of DNA Elements) Project. (2004). Science 306: 636–640. doi:10.1126/science.1105136.

37. Fu Y-P, Kohaar I, Rothman N, Earl J, Figueroa JD, et al. (2012) Common genetic variants in the PSCA gene influence gene expression and bladder cancer risk. Proc Natl Acad Sci U S A 109: 4974–4979. doi:10.1073/pnas.1202189109.

38. Rothman N, Garcia-Closas M, Chatterjee N, Malats N, Wu X, et al. (2010) A multi-stage genome-wide association study of bladder cancer identifies multiple susceptibility loci. Nat Genet 42: 978–984. doi:ng.687 [pii] 10.1038/ng.687.

39. Visser-Grieve S, Hao Y, Yang X (2012) Human homolog of Drosophila expanded, hEx, functions as a putative tumor suppressor in human cancer cell lines independently of the Hippo pathway. Oncogene 31: 1189–1195. doi:10.1038/onc.2011.318.

40. Wagner JR, Busche S, Ge B, Kwan T, Pastinen T, et al. (2014) The relationship between DNA methylation, genetic and expression inter-individual variation in untransformed human fibroblasts. Genome Biol 15: R37. doi:10.1186/gb-2014-15-2-r37.

41. Moffatt MF, Gut IG, Demenais F, Strachan DP, Bouzigon E, et al. (2010) A large-scale, consortium-based genomewide association study of asthma. N Engl J Med 363: 1211–1221. doi:10.1056/NEJMoa0906312.

42. Hunter DJ, Kraft P, Jacobs KB, Cox DG, Yeager M, et al. (2007) A genome-wide association study identifies alleles in FGFR2 associated with risk of sporadic postmenopausal breast cancer. Nat Genet 39: 870–874. doi:10.1038/ng2075.

43. Qi W, White MC, Choi W, Guo C, Dinney C, et al. (2013) Inhibition of inducible heat shock protein-70 (hsp72) enhances bortezomib-induced cell death in human bladder cancer cells. PLoS One 8: e69509. doi:10.1371/journal.pone.0069509.

44. Piazza R, Valletta S, Winkelmann N, Redaelli S, Spinelli R, et al. (2013) Recurrent SETBP1 mutations in atypical chronic myeloid leukemia. Nat Genet 45: 18–24. doi:10.1038/ng.2495.

45. Makishima H, Yoshida K, Nguyen N, Przychodzen B, Sanada M, et al. (2013) Somatic SETBP1 mutations in myeloid malignancies. Nat Genet 45: 942–946. doi:10.1038/ng.2696.

46. Ibragimova I, Dulaimi E, Slifker MJ, Chen DY, Uzzo RG, et al. (2014) A global profile of gene promoter methylation in treatment-naïve urothelial cancer. Epigenetics 9: 760–773. doi:10.4161/epi.28078.