Genomics 111 (2019) 1387–1394

Contents lists available at ScienceDirect

Genomics

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

Original Article

A powerful method to integrate genotype and gene expression data for dissecting the genetic architecture of a disease

Sarmistha Das[a], Partha Pratim Majumder[b], Raghunath Chatterjee[a], Aditya Chatterjee[c], Indranil Mukhopadhyay[a,*]

[a] Human Genetics Unit, Indian Statistical Institute, Kolkata, India
[b] National Institute of Biomedical Genomics, Kalyani, India
[c] Department of Statistics, University of Calcutta, Kolkata, India

ARTICLE INFO

Keywords:
GWAS
Multi-locus association test
Latent variable
Data integration

ABSTRACT

To decipher the genetic architecture of human disease, various types of omics data are generated. Two common omics data are genotypes and gene expression. Often genotype data for a large number of individuals and gene expression data for a few individuals are generated due to biological and technical reasons, leading to unequal sample sizes for different omics data. Unavailability of standard statistical procedure for integrating such datasets motivates us to propose a two-step multi-locus association method using latent variables. Our method is powerful than single/separate omics data analysis and it unravels comprehensively deep-seated signals through a single statistical model. Extensive simulation confirms that it is robust to various genetic models as its power increases with sample size and number of associated loci. It provides *p*-values very fast. Application to real dataset on psoriasis identifies 17 novel SNPs, functionally related to psoriasis-associated genes, at much smaller sample size than standard GWAS.

1. Introduction

To decipher the genetic architecture of a human disease, various types of omics data are generated. Two common omics data types are (a) genotypes at a large number of marker loci, and (b) expression levels for a large number of genes. Often these data are generated at the genome scale using microarrays. Although traditionally data generated by various omics platforms were analysed separately, in recent times data integration and joint statistical analysis have been emphasised to obtain robust inferences. One problem in joint statistical analysis of multi-type omics data is that sample sizes of different data types vary. There are many reasons for such variation. A major reason is that RNA and protein are less stable than DNA. Therefore, commonly in a genetic association study genotype data are available from a larger number of individuals than data on gene expression. Further, assays for generating gene expression data are expensive, that leads the researcher to generate and analyse genotype data first, arrive at a set of tentative inferences pertaining to the research question based on the genotype data, and then select a non random subset of individuals for gene expression assay to spawn additional information on disease association. The non random selection may be due to degradation of the RNA samples and/or other biological and technical reasons. Clearly, the individuals whose gene expression values are missing will depend only on expression profiles of itself and not on the genetic profile of other subjects. Therefore, we can assume that the "missing" gene expression values in the entire dataset are missing not at random (MNAR). However, the variation in the sample sizes of different types of omics data poses a major challenge in multiomics data analysis. This challenge motivated us to develop a statistical method to integrate the available subset of gene expression information with the complete genotype data, generated in a case-control study. Our method does not rely on reference transcriptome data for imputation of missing gene expression data that induces population stratification bias [7,11]. We show that our proposed method is statistically powerful and can identify disease associated variants, that remain undetected by analysis of genotype data alone.

Note that MNAR data that is often governed by sample selection bias [8,15] could be analysed using a framework developed by Little and Rubin [18]. For MNAR data, failure to account for the true missingness mechanism may result in biased parameter estimates [5] that can be adjusted using the latent variables [12]. Although MNAR data may be analysed using a pattern mixture model, the assumption of this model [17] is violated by the data under consideration.

Here, we propose an integrated Genotype and gene Expression

[*] Corresponding author.
E-mail address: indranil@isical.ac.in (I. Mukhopadhyay).

https://doi.org/10.1016/j.ygeno.2018.09.011
Received 20 March 2018; Received in revised form 14 August 2018; Accepted 17 September 2018
Available online 01 October 2018
0888-7543/ © 2018 Elsevier Inc. All rights reserved.

Method (iGEM), which is a novel two-step multi-locus association method using a latent variable conjointly with logistic regression to integrate genome-wide marker information with gene expression data on a subset of individuals. We apply our method to 902 psoriasis cases and 676 healthy controls. Genotype data for about half a million of single nucleotide polymorphisms (SNPs) for all individuals and gene expression data for both cases and controls on small subset of individuals were generated (dbGaP; phs000019.v1.p1). Using these real data, we identified 17 novel SNPs, in addition to those identified in the genome wide association study (GWAS) (that is using a single marker test, viz. Chi square test or logistic regression followed with multiple testing correction) by applying our method. We have also verified that these novel SNPs confirm strong functional association with enhancing psoriasis risk. To avoid computation intensive permutation or other techniques, we have derived the asymptotic distribution of our proposed test statistic that will aid in fast calculation of *p*-value in the real dataset to detect associated loci. We have also identified other statistical properties of our method using extensive simulations.

2. Results

2.1. Simulations

We carried out extensive simulations to assess the gain in statistical power by our proposed data integration method. We simulate datasets for various combinations of sample sizes for gene expression and genotype data, considering different genetic models with different values of relative risk (RR), allele frequency, penetrance function and other parameters. We generate genotypes for cases and controls at $K(=10)$ independent marker loci and gene expression data for small subsets of genotyped individuals. HWE at each locus is assumed for controls. Cases are ascertained based on 1 to 4 causative SNP genotype(s) among the 10 markers conjointly with an appropriate penetrance function and RR as 1.25 or 1.5 at each causative locus under specific genetic model assumption [20]. We consider three genetic models for risk such as additive, multiplicative, and recessive. Minor allele frequency (MAF) at each causative locus is assumed to be 0.05, while for a non-causative locus it is assumed to range from 0.1 to 0.5. We have also simulated causative and non-causative loci with comparable MAFs such as 0.1 (0.2) for all 10 markers and some other combinations of comparable MAFs. We also simulated scenarios when there is no SNP effect at all but the gene expression values have significant effect on the disease status of the individuals. We simulated gene expression data using an additive model for the effects of causative marker(s) for each individual (Xiong et al. [30]). In our simulations, sample size for each case and control group, was taken as 500, 700, and 1000 for genotype data and 100, 150, and 200 for gene expression data. We also compare the performance of iGEM with the test based on genotype data alone, for different sample sizes under different genetic models.

Under the null hypothesis, the empirical distribution of our test statistic under each combination of sample sizes of gene expression and genotype data, matches with the theoretically derived asymptotic distribution, which is a χ^2 distribution with $K + 1$ degrees of freedom. QQ plots based on 10000 datasets generated under null hypothesis elucidates this fact (Fig. 1a, b). In each case, the plot shows strong resemblance with theoretical asymptotic distribution. Using the asymptotic distribution of the iGEM statistic, we calculated type I error rate based on 10000 datasets. In all cases this rate was below 5% (Table 1, S1-S4), suggesting that our test statistic is conservative in controlling false positives. QQ plot also confirmed that one can use the asymptotic distribution of the iGEM statistic directly for real data analysis to calculate the *p*-value associated with the observed value of the test statistic.

We calculate the power of the iGEM statistic based on 1000 datasets with various sample sizes for gene expression data and genotype data under different genetic models. We found that in each scenario the power of our test statistic is increased substantially compared to the test based on genotype data only; (1) as the number of causative SNPs increased (Fig. 1c, d, S1a, S1b), (2) as the sample size of genotype data increased, for a fixed sample size of gene expression data, and vice versa (Tables 2, 3, S5–S8), and (3) when the combined sample size of gene expression and genotype data increased (Fig. 1e, f). Increment of power, upon inclusion of a larger set of gene expression data along with GWAS data is clearly discernible (Figs. 1e, 1f); similar trend is observed for other genetic models as well (Figs. S1c, S1d). Moreover, in the absence on any information from gene expression data, the power of our test statistic remained the same as the test based on genotype data only, for all simulation scenarios (Fig. 1d, S1b). These results clearly indicate that iGEM is able to capture information through gene expression, only when it is available. Results based on more extensive simulations, strengthen our claim that iGEM performs better than a genotype based association test for case-control data.

Interestingly we find that our method also identifies associated genes based on gene expression only, even when there is no effect of SNP. We performed extensive simulations where we generated data with null SNP effect but significant effect of gene expression. iGEM seems to be powerful in such cases also, keeping type I error rate below 5% level (Table S9). Moreover we observed that power of the test increases as sample size increases (Table 4).

2.2. Application of iGEM on Psoriasis data

We apply our proposed method on genome wide association data obtained from dbGaP (phs000019.v1.p1) pertaining to 902 psoriasis cases and 676 healthy individuals, with expression profiles of 148 differentially expressed genes for a subset of nearly 30 psoriasis patients and nearly 30 normal individuals for each gene. Integrative analysis using iGEM identified association of psoriasis with some SNPs and/or differentially regulated genes. In this dataset, we have applied our proposed method to each single SNP to investigate whether it is able to detect any SNP that remained unidentified by application of standard GWAS methods. As described in Section 4.3, using iGEM, we find the set (S_I) of all associated SNPs including the ones that remain unidentified through single marker test. We see that $\mathbb{C}(S) = 135$, $\mathbb{C}(S^* \cap S^{**}) = 25$ and $\mathbb{C}(S \cap (S^* \cap S^{**})) = 8$, where $\mathbb{C}(A)$ denotes the number of elements in a set A and S, S^*, and S^{**} are defined in Section 4.3. Hence the total number of associated SNPs is 152 out of which 17 novel SNPs are identified by our method.

Table 5 represents the functional annotation of 17 novel iGEM SNPs. Interestingly, we find some of reported variants involved in psoriasis pathogenesis, are in strong LD with some of the novel iGEM SNPs (Table 6). Moreover, we find that all novel SNPs are strongly linked with variants carrying important functional information (Table S10). Although the sample size of the psoriasis dataset analysed here is much less compared to other GWA studies (Stuart et al. [25]; Tsoi et al. [28]; Villarreal-Martínez et al. [29]; Strange et al. [24]), with the little extra information on gene expression, our method identified disease associated SNPs successfully. We further investigated the regulatory functions of these 17 novel SNPs and their linked SNPs using HaploReg v4 (http://www.broadinstitute.org/mammals/haplo-reg/haploreg_v4.php) (Table S10). Interestingly, we find that some SNPs overlap with binding sites of transcription factors (TFs) viz. nuclear factor kappa-light-chain-enhancer of activated B cells (NF-κB), that play many vital roles in psoriasis pathogenesis. Moreover, NF-κB regulate both innate and adaptive immune systems.

Boxplots of gene expression corresponding to genotypes of the discovered loci strengthen the evidence of functional relationship of the disease with these variants. Expression of Solute Carrier Family 16 Member 10 (SLC16A10) clearly depicts a functional relationship with *rs*3132496 (T/G) (Fig. 2a) under a dominant model. Calcium Regulated Heat Stable Protein 1 (CARHSP1) and MAX Dimerization Protein 1 (MXD1) genes of psoriasis patients illustrate distinct functional relationships with *rs*3817151 (C/T) (Fig. 2b) and *rs*13026755 (C/T) (Fig.

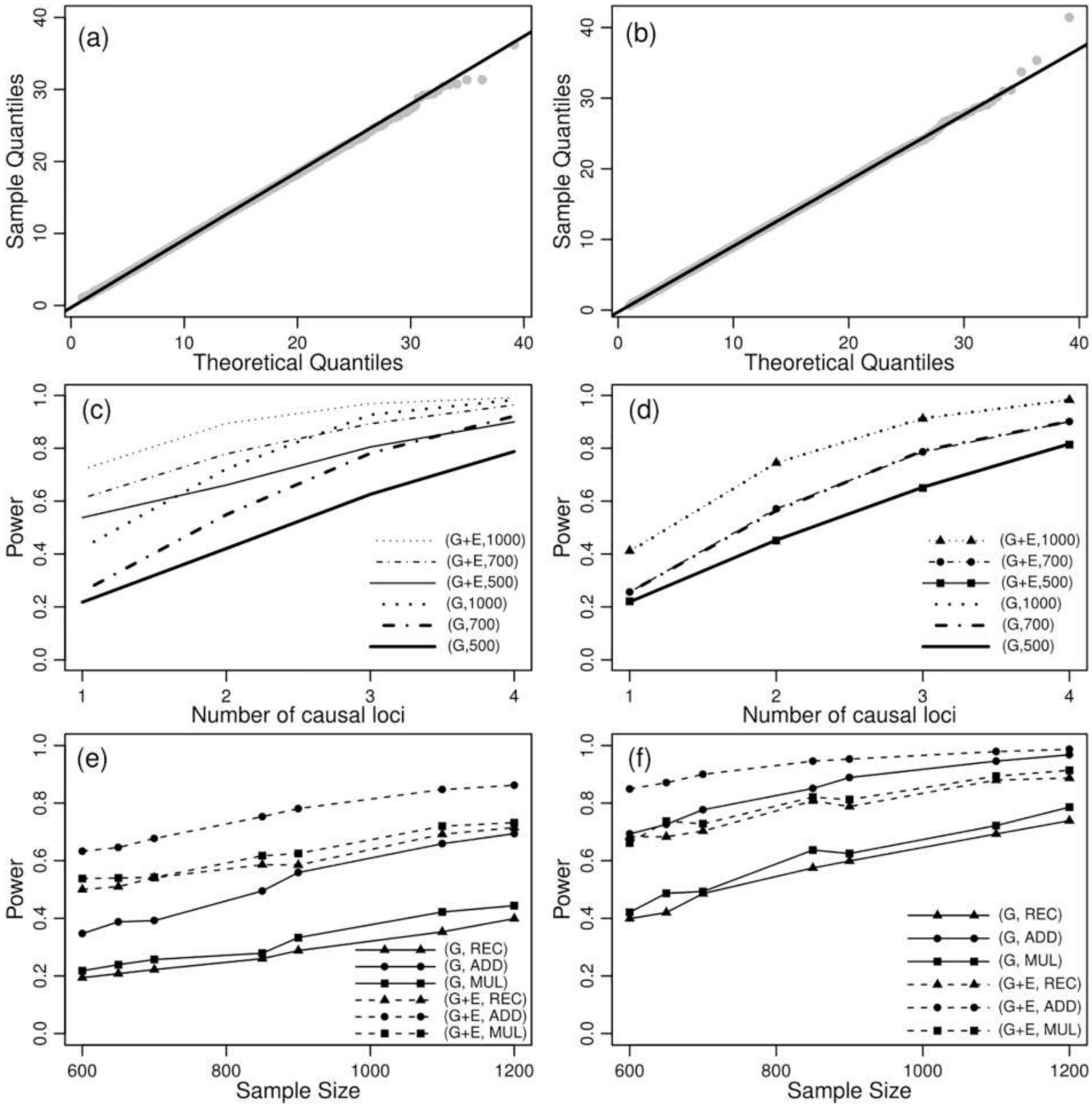

Fig. 1. Performance of iGEM based on simulated data: (a) QQ-plot with sample sizes for gene expression and genotype as 100 and 500 respectively; (b) QQ-plot with sample sizes for gene expression and genotype as 200 and 1000 respectively; (c) Comparison of power for iGEM and test based on genotype only, for multiplicative model with RR 1.5, in presence of expression QTL (eQTL), with respect to number of causal genotype loci; (d) Comparison of power for iGEM and genotype based test for multiplicative model with RR 1.5, without any effect of eQTL, with respect to number of causal genotype loci; (e) Power comparison of iGEM and genotype based test for different genetic models with RR 1.5 and 1 causal genotype locus with eQTL effect; (f) Power comparison of iGEM and genotype based test for different genetic models with RR 1.5 and 2 causal genotype loci with eQTL effect.

(G, n): Genotype based test using n samples; $(G + E, n)$: iGEM using n samples; $(method, disease\ model)$ where 'disease models' are ADD (Additive), MUL (Multiplicative), and REC (Recessive) and 'method' is either genotype based test (G) or iGEM (G + E).

Table 1

Type I error rate under different combination of sample sizes of gene expression and genotype data based on 10,000 simulations (five out of ten have MAF = 0.05 and rest are 0.1, 0.2, 0.3, 0.4, 0.5).

SS_G	500	700	1000
SS_{GE}			
100	0.0283	0.0310	0.0295
150	0.0293	0.0306	0.0290
200	0.0267	0.0265	0.0301

SS_G: sample size for genotype data.
SS_{GE}: sample size for gene expression data.

S2a) respectively under dominant model. We find a trend in gene expression with respect to genotypes for each of the above three SNPs. Expression of Kynureninase (KYNU), Coiled-Coil Alpha-Helical Rod Protein 1 (CCHCR1), and Peptidase Inhibitor 3 (PI3) also show functional relationships with *rs*2083482, *rs*3873386, and *rs*609932 respectively (Figs. S2b, S2c, S2d), under a recessive model. Similar relationship exists for *rs*607331 as it is in strong LD with *rs*609932. We found an interesting functional difference in gene expression of CARHSP1 with respect to genotype of *rs*2054213 (A/G) (Fig. 2c) among cases under both dominant and recessive models. Simultaneously, there is a significant difference between the homozygotes under a codominant model. Therefore, increase in sample size may provide a more clear

Table 2
Power of iGEM and only genotype based test (given in brackets) for different genetic models with varying sample sizes, relative risk as 1.5 and one causal locus, based on 1000 simulated datasets (five out of ten loci have MAF = 0.05, rest are 0.1, 0.2, 0.3, 0.4, 0.5).

	Additive			Multiplicative			Recessive		
	SS_G								
SS_E	500	700	1000	500	700	1000	500	700	1000
100	0.633 (0.347)	0.726 (0.465)	0.847 (0.659)	0.538 (0.218)	0.611 (0.258)	0.720 (0.422)	0.499 (0.194)	0.591 (0.255)	0.692 (0.353)
150	0.646 (0.388)	0.753 (0.495)	0.860 (0.648)	0.539 (0.239)	0.617 (0.279)	0.718 (0.416)	0.500 (0.208)	0.586 (0.260)	0.709 (0.396)
200	0.676 (0.392)	0.781 (0.559)	0.862 (0.694)	0.542 (0.257)	0.625 (0.333)	0.732 (0.444)	0.541 (0.222)	0.585 (0.288)	0.715 (0.399)

SS_G: sample size for genotype data; SS_{GE}: sample size for gene expression data.

Table 3
Power of iGEM and test based on genotype only (given in brackets) for different genetic models with varying sample sizes, relative risk as 1.5 and one causal locus based on 1000 simulated datasets with comparable frequencies of causal and non causal loci (all MAF = 0.1).

	Additive			Multiplicative			Recessive		
	SS_G								
	500	700	1000	500	700	1000	500	700	1000
SS_E									
100	0.818 (0.694)	0.913 (0.844)	0.979 (0.95)	0.602 (0.387)	0.728 (0.525)	0.841 (0.733)	0.577 (0.304)	0.666 (0.455)	0.769 (0.598)
150	0.840 (0.721)	0.910 (0.854)	0.969 (0.949)	0.646 (0.424)	0.732 (0.582)	0.871 (0.761)	0.575 (0.343)	0.684 (0.479)	0.823 (0.661)
200	0.852 (0.757)	0.939 (0.894)	0.987 (0.978)	0.650 (0.457)	0.785 (0.612)	0.892 (0.767)	0.593 (0.378)	0.688 (0.522)	0.818 (0.689)

SS_G: sample size for genotype data; SS_{GE}: sample size for gene expression data.

Table 4
Power of iGEM and test based on genotype only (given in brackets) under different combination of sample sizes of gene expression and genotype data based on 1000 simulations in absence of any SNP effect (five out of ten have MAF = 0.05 and the rest are 0.1, 0.2, 0.3, 0.4, 0.5).

	SS_G		
	500	700	1000
SS_{GE}			
100	0.334 (0.034)	0.346 (0.046)	0.3708 (0.048)
150	0.343 (0.051)	0.359 (0.054)	0.376 (0.045)
200	0.352 (0.049)1	0.361 (0.056)	0.395 (0.047)

SS_G: sample size for genotype data.
SS_{GE}: sample size for gene expression data.

picture about the genetic model. Moreover, the boxplot reveals that the gene expression clearly follows a trend with respect to genotypes for rs2054213.

3. Discussion

iGEM identified *rs*3132496 and *rs*3873386 to be associated with SLC16A10. *rs*3132496 is located very close to the MHC class I−C (HLA−C) gene, known as psoriasis susceptibility gene 1 (PSORS1). *rs*3873386 is located in 6p21.33, which is reported to be associated with CCHCR1 [4], that encodes a protein that is over-expressed in lesional skin of psoriasis patients and contains non synonymous substitutions across many populations [1]. SLC16A10 is involved in transportation of inorganic ions and amino acids pathway associated with psoriasis risk. Genetic variation in this pathway potentially increases the exposure to psoriasis development by functional modulation of T cells [2]. SLC16A10 is located very close to another psoriasis susceptible gene TRAF3 Interacting Protein 2 (TRAF3IP2) (6q21) that contains risk SNPs [6,13]. TRAF3IP2 encodes a protein involved in IL-17 signaling pathway and activates NF-κB and mitogen-activated protein kinase (MAPK) pathways ([6,13]; Strange et al. [24]). These pathways provide insight into biological mechanism associated with psoriasis susceptibility.

We identify *rs*13026755, located close to long noncoding RNA AC010733.4 and REL (2p16.1) and is in strong LD with two already reported SNPs *rs*62149416(Tsoi et al. [28]) and *rs*702873(Strange et al. [24]) having $r^2 = 0.85$ and $r^2 = 0.59$ respectively. REL encodes a member of the NF-κB family of TFs and aids in NF-κB signaling pathway, that is associated with psoriasis pathogenesis (Strange et al. [24]).

Further, we identified *rs*2054213 in intronic SET Domain Containing 1A (SETD1A) (16p11.2) and is in strong LD ($r^2 > 0.80$) with *rs*10782001 located in F-Box And Leucine-Rich Repeat Protein 19 (FBXL19) (16p11.2)(Stuart et al. [25]), *rs*12445568 (Tsoi et al. [28]) residing in intronic Syntaxin 1B (STX1B) (16p11.2), and *rs*12924903 replicated in Mestizo population(Villarreal-Martínez et al. [29]). FBXL19 activates NF-κB as a putative inhibitor and hence aids in NF-κB signaling pathway(Stuart et al. [25]).

We note that nearly all SNPs detected by our method and their linked SNPs are enriched in transcriptional activities, viz., TF binding sites and/or DNase hypersensitive site (DHS) and/or histone modification marks (H3K4me1, H3K27ac chromatin marks in enhancer region, and H3K4me3, H3K9ac chromatin marks in promoter region) in blood, skin cells, etc. (Tables 5 and S10).

Among other SNPs identified by our method *rs*7195745 is located in psoriasis associated region 16q23 (Nair et al. [21]). Three SNPs *rs*607331, *rs*609932, and *rs*13045901 are associated with PI3, which is reported to be misregulated in psoriasis(Ruano et al. [23]). The last one is located near small noncoding Y_RNA, that is increasingly gaining importance for specific cellular functions and has been recently detected as an abundant part in the blood and tissues of humans [14]. In our analysis the findings of strong link with already reported SNPs, in a way highlights the strength of our method to detect disease associated SNPs. Two other SNPs, *rs*2296633 and *rs*857369 are associated with C10orf99 (10q23.1), which is an upstream component of growth signal transduction pathway associated to psoriasis [10]. Among other SNPs identified by iGEM, *rs*16859665, *rs*7122993 and *rs*12946388 are associated with Wnt Family Member 5A (WNT5A), ETS Homologous Factor (EHF) and Keratin 16 (KRT16) respectively. While, the first two genes are reportedly associated with psoriasis pathogenesis ([9]; Swindell et al. [27]), the SNP *rs*12946388 resides in Mitogen-Activated Protein Kinase Kinase 6 (MAP2K6), that functionally contributes to the disease

Table 5
Functional annotation of 17 SNPs identified by iGEM.

SNP found by iGEM	Chr	Assoc. Gene	PEHMD[a]	Remark of iGEM SNP
rs2083482 (12 kb 3′ of FIGN)	2	STAT1		
rs13026755	2	RRM2	Bl	H3K27ac_Enh in Bl,FK; H3K4me1_Enh in Bl
rs16859665 (intron of C3orf70)	3	WNT5A		
rs3873386 (4.3 kb 5′ of XXbac-BPG248L24.13)	6	SLC16A10	Bl (Bl)	H3K9ac_Pro in Bl; H3K4me1_Enh in FF
rs6947649 (133 kb 3′ of AC006322.1)	7	AKR1B10	SKIN	H3K27ac_Enh in Bl
rs10815803	9	GDA		
rs2296633 (DOCK1, intronic)	10	C10orf99		H3K4me1_Enh in FF, M; H3K27ac_Enh in FF
rs7122993 (BARX2, intronic)	11	EHF	SKIN (SKIN)	H3K4me1_Enh in FK H3K27ac_Enh in FK
rs1864335 (RYR3, intronic)	15	RAB27A	SKIN (SKIN)	H3K4me1_Enh in FM; H3K27ac_Enh in FM
rs7195745 (SLC38A8, intronic)	16	GDPD3		H3K27ac_Enh in FF
rs12946388 (MAP2K6, intronic)	17	KRT16	Bl	H3K4me1_Enh in Bl
rs13045901 (19 kb 3′ of Y·RNA)	20	PI3		
rs607331, rs609932 (RP5-839B4.7, intronic)	20	PI3	SKIN	H3K4me1_Enh in FM; H3K27ac_Enh in FM

rs2054213 (SETD1A, intronic)	16	CARHSP1	Bl,
			ADF,
			SKIN,
			EK cells,
				H3K4me1_Enh in Bl, FF, FM;
				H3K9ac_Pro in ADF;
				H3K4me3_Pro in Bl, FF, FK;
				H3K27ac_Enh in Bl, FF, FM;

rs3132496 (28 kb 3′ of HLA-C)	6	SLC16A10	Bl, SKIN	
rs857369 (PCDH15, intronic)	10	C10orf99		

[a] PEHMD denotes the presence of promoter and enhancer histone mark in tissues; DNase hypersensitive sites are shown in brackets. Bl: Blood, FF: Foreskin Fibroblast, FM: Foreskin Melanocyte, FK: Foreskin Keratinocyte, ADF: Adult, Dermal Fibroblast, EK: Epidermal Keratinocyte, IPSC: Induced pluripotent stem cells; Enh: Enhancer, Pro: Promoter.

Table 6
Novel iGEM SNPs in LD with reported psoriasis susceptible SNPs.

SNP	Chr	Reported SNP(s)	LD values (r^2)
rs13026755	2	*rs*62149416(Tsoi et al. [28])	0.85
		*rs*702873(Strange et al. [24])	0.59
rs2054213	16	*rs*10782001(Stuart et al. [25]),	0.96
		*rs*12445568(Tsoi et al. [28])	0.85
		*rs*12924903(Villarreal-Mart´ ınez et al. [29])	0.92

phenotype [19].

We find another SNP *rs*2083482, close to Fidgetin (FIGN) (2q24.3), that is associated with psoriasis susceptible Signal Transducer And Activator Of Transcription 1 (STAT1) (2q32.2). Interferon Induced With Helicase C Domain 1 (IFIH1) that lies very near to FIGN, encrypts an inherited receptor that is engaged in triggering type I Interferon (IFN) in response to infection caused by microbes and is associated to psoriasis susceptible loci *rs*17716942(Strange et al. [24]). Other studies ([16,31]) report missense IFIH1 SNPs associated with decreased risk of psoriasis. On the other hand, expression of STAT1 is increased in psoriatic skin. It also regulates the expression of interferon-responsive genes. Besides, studies (Sun and Zhang [26]) also reveal the pivotal role of STAT1 in transcriptional regulatory network for psoriasis.

Again, the available gene expression data can also be utilised for another purpose. To detect any possible functional impact of a SNP identified by GWAS, we regress gene expression on genotype and find a few SNPs are associated with gene expression. To examine how our newly proposed method works in this situation, we apply the part of our model that handles the specification bias due to non-random missing gene expression data instead of the standard regression approach. Our method identifies one additional SNP *rs*2844627. Table S11 presents the detailed results of this analysis.

Compared to some other methods, our proposed multi-locus association method is free from any error that usually accumulates due to the use of summary statistics [11] or imputing gene expression data using genotype information [7].

We have developed a precise test statistic *T*, with a simple asymptotic distribution, for fast calculation of *p*-value in real datasets. Simulation study shows that the empirical distribution of the test statistic under null hypothesis is approximately same as the distribution of a χ^2 variable with appropriate degrees of freedom. Even for small samples, we can use bootstrap or a permutation technique to calculate the p-value.

Simulation studies reveal that iGEM is robust, consistent and more powerful than methods based on analysis of genotype data alone (Tables 2, 3, S5-S8). The robustness is reflected in Tables 2, 3, S5-S8, when iGEM extracts relevant information from genotype data (of fixed sample size) and integrates it with gene expression data, as the sample size of the latter varies from low to moderately low, under various genetic models. The consistency of our method is established in Theorem 2 while, increase in power with increase in (*a*) the sample size and, (*b*) the number of causal loci are evident from Tables 2, 3, S5-S8.

Our method identified 17 additional loci along with all the SNPs that could be identified using standard genome-wide association (GWA) method in the given sample. Interestingly, we found that a few of these 17 SNPs, were in strong LD with some reported psoriasis susceptible SNPs. The method proposed here obviously is more efficient than the traditional GWAS in detecting even low-effect SNPs.

The specification bias that arises due to non-random subset selection is tackled statistically using a firm theory underlying our method. Application of iGEM to case-control data could provide new directions for further biological exploration. Our method holds promise for extending the integration paradigm using more than two different types of omics data. Even when the sample size for gene expression data is not very large, iGEM is able to extract some information that increases the power of the test.

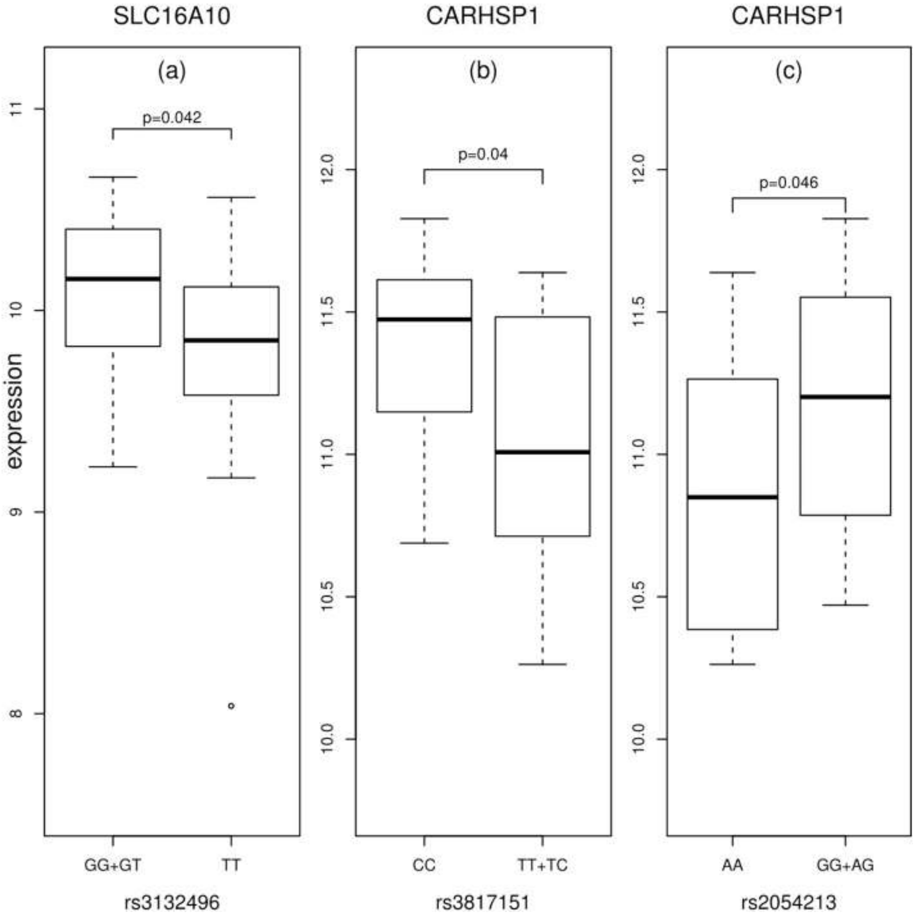

Fig. 2. Boxplot showing gene expression and genotype relationship under a dominant model for SNPs *rs*3132496, *rs*3817151, and *rs*2054213.

4. Materials and methods

4.1. Overview of the data structure

We apply our integrated approach to a dataset containing genome-wide genotype information from 902 psoriasis patients and 676 healthy controls of European ancestry, with expression of 148 functionally annotated genes for a subset of about 30 cases and 30 controls for each gene(Nair et al. [22]) (dbGaP; phs000019.v1.p1). About 440K SNPs, genotyped using high density oligonucleotide arrays by Perlegen Sciences (Mountain View, CA, USA) passed data quality control checks for all individuals. Markers not in Hardy-Weinberg Equilibrium (HWE) (p-value $< 10^{-6}$) and minor allele frequency (MAF) < 0.05 were excluded. Imputed genotypes with $R^2 > 0.3$ (between true and imputed SNP genotypes) were included for genotyping. Further, RNA samples from skin biopsies of the smaller subset of samples, from University of Michigan, Department of Dermatology, were analysed using Affymetrix U133 Plus 2.0 arrays to evaluate expression of probes. Raw data from microarrays were processed appropriately and adjusted for batch and sex effects before further analysis. Average gene expression was calculated from multiple probes and served as gene expression for each gene (Nair et al. [22]). Here selection of subjects for gene expression profiling was not random; hence these data conform to our assumption of MNAR for the incomplete gene expression dataset.

We introduce the general analytical framework under which such data may be integratively analysed. Suppose that a large group of I individuals comprising patients and controls are genotyped at K marker loci, while a small subset of I_1 individuals are assayed for gene expression profiles. These two types of data contain more information than a single type of data. So, we propose a test statistic that explores the disease-gene association by integrating available gene expression and entire genotype information. Our method can accommodate unrestricted number of markers belonging to a particular gene or a user defined genomic region, provided that K is smaller than the sample size, so that parameters are estimable.

4.2. Sub-sample selection criteria

Following Heckman [12], we define a latent variable Y_2 to model the non-random missing pattern of gene-expression (Y_1) for genotyped individuals. The sub-sample selection criterion assures availability of Y_1 for individual i. Y_{1i} is observed if $Y_{2i} \geq a$, and not observed if $Y_{2i} < a$, for the i^{th} individual for a particular gene and "a"is some constant. We introduce I latent variables $Y_{21}, Y_{22}, \ldots, Y_{2I}$ that follow $N(0, \sigma_{22})$ independently and identically. Thus $Y_{2i} \geq a\,(< a)$ implies expression is available (unavailable) for individual i. Using simple algebra, it can be shown that the choice "a"may be replaced by 0.

4.3. Integrated genotype expression method (iGEM)

We propose a two-step integrated Genotype Expression Method (iGEM) to identify some novel loci along with those found in standard GWAS. The first step generates a list of strongly associated loci after Benjamini-Hochberg (BH) correction [3] using logistic regression of case-control status on genotypes. The second step combines additional information from gene expression. For individual i, the model is,

$$Y_{1i} = \beta_0 + \beta_1 X_{cci} + \sum_{k=1}^{K} \beta_{2k} g_{ik} + U_{1i}, \quad i = 1, \ldots, I_1 \tag{1}$$

$$Y_{2i} = \beta_0^* + \beta_1^* X_{cci} + \sum_{k=1}^{K} \beta_{2k}^* g_{ik} + U_{2i}, \quad i = 1, \ldots, I \tag{2}$$

$$logit\,(P\,(X_{cci} = 1)) = \gamma_0 + \sum_{k=1}^{K} \gamma_{1k} g_{ik}, \quad i = 1, \ldots, I \tag{3}$$

where, Y_{1i} is the gene expression for gene G ($i = 1,\ldots,I_1$), X_{cci} is the disease status and g_{ij} is the genotype of individual i at the j^{th} SNP ($i = 1, \ldots, I_1, I_1 + 1, \ldots, I; j = 1, \ldots, K$). g_{ij} assumes values 0,1,2 for the 3 possible genotypes at a biallelic SNP locus. X_{cci} takes 0 for controls and 1 for cases. Eq. (3) is a standard logistic regression model for genotype association that is integrated with information from gene expression data (Eq. (1) and (2)). We assume that U_1 and U_2 are random error components.

To facilitate the description of our proposed method, we write Eqs. (1)–(3) using matrix notation. Thus, for individual i, we have,

$$Y_{1i} = X_{1i}\boldsymbol{\beta} + U_{1i}, \; i = 1, \ldots, I_1 \tag{4}$$

$$Y_{2i} = X_{2i}\boldsymbol{\beta} \; + U_{2i}, \; i = 1, \ldots, I \tag{5}$$

$$logit\,(P\,(X_{cci} = 1)) = X_{3i}\boldsymbol{\gamma}, \; i = 1, \ldots, I \tag{6}$$

where $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_{21}, \ldots, \beta_{2K})'$, $\boldsymbol{\beta}^* = (\beta_0{}^*, \beta_1{}^*, \beta_{21}{}^*, \ldots, \beta_{2K}{}^*)'$, and $\gamma = (\gamma_0, \gamma_{11}, \ldots, \gamma_{1K})'$. We assume that,

(i)$E\,(U_{ji}) = 0, \quad j = 1,2,3$

(ii)(U_{1i}, U_{2i})jointly follows a bivariate normal distribution

$$\text{(iii)}E\,(U_{ji}U_{j'i'}) = \begin{cases} \sigma_{jj'} & \text{if } i = i' \\ 0 & \text{otherwise}; \quad j' \neq j = 1, 2 \end{cases}$$

If U_{1i} and U_{2i} are independent, data on Y_{1i} would be missing randomly and in that case $E(U_{1i}|\,\text{sub - sample selection rule}) = 0$. But here, U_{1i} and U_{2i} are correlated. Therefore, given the sub-sample selection rule, the conditional distribution of U_{1i} will depend on X_{2i}*i. e.* $E(U_{1i}|\,\text{sub - sample selection rule}) = E(U_{1i}|\,Y_{2i} \geq 0) = E(U_{1i}|\,U_{2i} \geq \; - X_{2i}\boldsymbol{\beta}^*)$. The sub-sample regression function, depending on X_{1i} and X_{2i} will be,

$$E\,(Y_{1i} \mid \text{sub-sample selection rule}, X_{1i}) = E\,(Y_{1i} \mid Y_{2i} \geq 0, X_{1i})$$
$$= X_{1i}\boldsymbol{\beta} + E\,(U_{1i} \mid U_{2i} \geq -X_{2i}\boldsymbol{\beta}\;, X_{1i}) \tag{7}$$

Our hypotheses of interest are: $H_0 : \beta_1 = 0, \; \gamma_1 = \mathbf{0}$, against $H_1 : H_0$ is not true. Naturally rejection of H_0 would indicate the presence of association. To test this, we propose a new statistic as:

$$T = \frac{(\widehat{\beta}_1 - \beta_1)^2}{\widehat{V}\,(\widehat{\beta}_1)} + (\widehat{\boldsymbol{\gamma}}_{\mathbf{1}} - \boldsymbol{\gamma}_{\mathbf{1}})'\widehat{\Sigma}^{-1}(\widehat{\boldsymbol{\gamma}}_{\mathbf{1}} - \boldsymbol{\gamma}_{\mathbf{1}}) \tag{8}$$

where $\widehat{\Sigma}$ and $\widehat{V}\,(\widehat{\beta}_1)$ are consistent estimators of variance-covariance matrix of $\widehat{\boldsymbol{\gamma}}_{\mathbf{1}} = (\gamma_{11}, \ldots, \gamma_{1K})'$ and $\widehat{\beta}_1$, respectively. Explicit expression of the required estimators are given in Appendix.

Let $P_1(j)$, $P_2(j)$, and $P_I(j)$ be the BH-corrected p-values for the j-th locus based on GWAS (equation 3), test identifying significant gene expression adjusted for genotype (equations 1 and 2), and null distribution of T. First we screen loci with $P_1 < 0.1$ to include loci with moderate to weak effect. We select a set of loci S_I associated with phenotype where $S_I = S \cup (S^* \cap S^{**})$ with $S = \{j : P_1(j) < 0.05\}$, $S^* = \{j : P_I(j) < \min\,(P_1(j), P_2(j))\}$, and $S^{**} = \{j : P_I(j) < 0.05\}$. This *min*$(P_1, P_2)$-criterion (1) eliminates unusually large effect of gene expression that may be due to other epigenetic mechanism and (2) includes weakly associated loci undetected by GWAS, carrying additional information from expression data.

Now the problem reduces to calculating P_I using the distribution of T under H_0. We derive the asymptotic distribution of T under H_0 as it is extremely difficult, if not impossible, to get its exact distribution in a compact form. This helps in faster calculation of p-value for the observed value of the test statistic and reduces computational burden to a great extent, compared to computation intensive procedure like permutation technique. In these genetic studies, sample sizes are usually large enough so as to apply large sample theory to develop a suitable test for the null hypothesis under consideration. We summarize a few large sample properties of the above statistic in the following two theorems.

Theorem 1. *Under the assumptions of the model described in* Eqs. (1)–(3)

$$T \xrightarrow{L} \chi^2_{K+1} \, \text{as} I_1 \to \infty \tag{9}$$

Theorem 2. The test procedure using T is consistent i.e. power of the test increases as sample size increases.

Proofs of the above theorems are given in the Appendix. Theorem 1 justifies the accuracy of p-value calculated on the basis of a $\chi_{K+1}{}^2$ variable while Theorem 2 ensures the high power associated with this test.

iGEM captures information from genotype data and gene expression data together. Simulation confirms that power based on combined genotype and gene expression is much greater than that based on genotype data only. In Theorem 3, we present a result ensuring that in order to match the power based on iGEM, the sample size for test based on genotype only must be greater than that for iGEM. Proof of this theorem is given in the Appendix.

Theorem 3. Let n^* be the sample size for the test based on genotype data only. Also let n and n_1 be the sample sizes of genotype data and gene expression data respectively when we use a test for association based on iGEM. To achieve approximately same power by these two tests, it is necessary that $n^* > n$ under the assumptions of the model as given in Eqs. (1)–(3).

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ygeno.2018.09.011.

Acknowledgement

We are grateful to the Department of Biotechnology, Govt. of India, for their partial support to this study through SyMeC. SD gratefully acknowledges the support from the Department of Science and Technology, Government of India (SR/WOS-A/PM-1014/2015(G)). PPM acknowledges his Sir J.C. Bose National Fellowship.

Conflict of interest

The authors do not have any conflict of interest.

References

[1] K. Asumalahti, C. Veal, T. Laitinen, S. Suomela, M. Allen, O. Elomaa, M. Moser, R. De Cid, S. Ripatti, I. Vorechovsky, et al., Coding haplotype analysis supports hcr as the putative susceptibility gene for psoriasis at the mhc psors1 locus, Hum. Mol. Gen. 11 (5) (2002) 589–597.

[2] A. Aterido, A. Julià, C. Ferrándiz, L. Puig, E. Fonseca, E. Fernández-López, E. Dauden, J.L. Sánchez-Carazo, J.L. López-Estebaranz, D. Moreno-Ramírez, et al., Genome-wide pathway analysis identifies genetic pathways associated with psoriasis, J. Invest. Dermatol. 136 (3) (2016) 593–602.

[3] Y. Benjamini, Y. Hochberg, Controlling the false discovery rate: a practical and powerful approach to multiple testing, J R Stat Soc Series B Stat Methodol (1995) 289–300.

[4] F. Capon, A.D. Burden, R.C. Trembath, J.N. Barker, Psoriasis and other complex trait dermatoses: from loci to functional pathways, J. Invest. Dermatol. 132 (2012) 915–922.

[5] L.M. Collins, J.L. Schafer, C.-M. Kam, A comparison of inclusive and restrictive strategies in modern missing data procedures, Psychol. Methods 6 (4) (2001) 330.

[6] E. Ellinghaus, D. Ellinghaus, P.E. Stuart, R.P. Nair, S. Debrus, J.V. Raelson, M. Belouchi, H. Fournier, C. Reinhard, J. Ding, et al., Genome-wide association study identifies a psoriasis susceptibility locus at traf3ip2, Nature Genet. 42 (11) (2010) 991–995.

[7] E.R. Gamazon, H.E. Wheeler, K.P. Shah, S.V. Mozaffari, K. Aquino-Michaels, R.J. Carroll, A.E. Eyler, J.C. Denny, D.L. Nicolae, N.J. Cox, et al., A gene-based association method for mapping traits using reference transcriptome data, Nature Genet. 47 (9) (2015) 1091–1098.

[8] R. Gronau, The effect of children on the housewife's value of time, Economics of the Family: Marriage, Children, and Human Capital, University of Chicago Press, 1974,

pp. 457–490.
[9] J.E. Gudjonsson, A. Johnston, S.W. Stoll, M.B. Riblett, X. Xing, J.J. Kochkodan, J. Ding, R.P. Nair, A. Aphale, J.J. Voorhees, et al., Evidence for altered wnt signaling in psoriatic skin, J. Invest. Dermatol. 130 (7) (2010) 1849–1859.
[10] P. Guo, Y. Luo, G. Mai, M. Zhang, G. Wang, M. Zhao, L. Gao, F. Li, F. Zhou, Gene expression profile based classification models of psoriasis, Genomics 103 (1) (2014) 48–55.
[11] A. Gusev, A. Ko, H. Shi, G. Bhatia, W. Chung, B.W. Penninx, R. Jansen, E.J. De Geus, D.I. Boomsma, F.A. Wright, et al., Integrative approaches for large-scale transcriptome-wide association studies, Nature Genet. 48 (3) (2016) 245–252.
[12] J.J. Heckman, Sample selection bias as a specification error, Econometrica 47 (1979) 153–161.
[13] U. Hüffmeier, S. Uebe, A.B. Ekici, J. Bowes, E. Giardina, E. Korendowych, K. Juneblad, M. Apel, R. McManus, P. Ho, et al., Common variants at traf3ip2 are associated with susceptibility to psoriatic arthritis and psoriasis, Nature Genet. 42 (11) (2010) 996–999.
[14] M.P. Kowalski, T. Krude, Functional roles of non-coding y rnas, Int. J. Biochem. Cell Biol. 66 (2015) 20–29.
[15] H.G. Lewis, Comments on selectivity biases in wage comparisons, J. Polit. Econ. 82 (6) (1974) 1145–1155.
[16] Y. Li, W. Liao, M. Cargill, M. Chang, N. Matsunami, B.-J. Feng, A. Poon, K.P. Callis-Duffin, J.J. Catanese, A.M. Bowcock, et al., Carriers of rare missense variants in ifih1 are protected from psoriasis, J. Invest. Dermatol. 130 (12) (2010) 2768–2772.
[17] R.J. Little, Pattern-mixture models for multivariate incomplete data, J. Am. Stat. Assoc. 88 (421) (1993) 125–134.
[18] R.J. Little, D.B. Rubin, Statistical analysis with missing data, J. Wiley, 2002.
[19] A. Mavropoulos, E.I. Rigopoulou, C. Liaskos, D.P. Bogdanos, L.I. Sakkas, The role of p38 mapk in the aetiopathogenesis of psoriasis and psoriatic arthritis, Clin. Dev. Immunol. 2013 (2013).
[20] I. Mukhopadhyay, E. Feingold, D.E. Weeks, A. Thalamuthu, Association tests using kernel-based measures of multi-locus genotype similarity between individuals, Genet. Epidemiol. 34 (3) (2010) 213–221.
[21] R.P. Nair, T. Henseler, S. Jenisch, P. Stuart, C.K. Bichakjian, W. Lenk, E. Westphal, S.-W. Guo, E. Christophers, J.J. Voorhees, et al., Evidence for two psoriasis susceptibility loci (hla and 17q) and two novel candidate regions (16q and 20p) by genome-wide scan, Hum. Mol. Gen. 6 (8) (1997) 1349–1356.
[22] R.P. Nair, K.C. Duffin, C. Helms, J. Ding, P.E. Stuart, D. Goldgar, J.E. Gudjonsson, Y. Li, T. Tejasvi, B.-J. Feng, et al., Genome-wide scan reveals association of psoriasis with il-23 and nf-κb pathways, Nature Genet. 41 (2) (2009) 199–204.
[23] J. Ruano, M. Suárez-Fariñas, A. Shemer, M. Oliva, E. Guttman-Yassky, J.G. Krueger, Molecular and cellular profiling of scalp psoriasis reveals differences and similarities compared to skin psoriasis, PLoS ONE 11 (2) (2016) e0148450.
[24] A. Strange, F. Capon, C.C. Spencer, J. Knight, M.E. Weale, M.H. Allen, A. Barton, G. Band, C. Bellenguez, J.G. Bergboer, et al., Genome-wide association study identifies new psoriasis susceptibility loci and an interaction between hla-c and erap1, Nature Genet. 42 (11) (2010) 985–990.
[25] P.E. Stuart, R.P. Nair, E. Ellinghaus, J. Ding, T. Tejasvi, J.E. Gudjonsson, Y. Li, S. Weidinger, B. Eberlein, C. Gieger, et al., Genome-wide association analysis identifies three psoriasis susceptibility loci, Nature Genet. 42 (11) (2010) 1000–1004.
[26] L. Sun, X. Zhang, The immunological and genetic aspects in psoriasis, Applied Informatics 1 (1) (2014) 1.
[27] W.R. Swindell, M.K. Sarkar, P.E. Stuart, J.J. Voorhees, J.T. Elder, A. Johnston, J.E. Gudjonsson, Psoriasis drug development and gwas interpretation through in silico analysis of transcription factor binding sites, Clin. Transl. Med. 4 (1) (2015) 1.
[28] L.C. Tsoi, S.L. Spain, J. Knight, E. Ellinghaus, P.E. Stuart, F. Capon, J. Ding, Y. Li, T. Tejasvi, J.E. Gudjonsson, et al., Identification of 15 new psoriasis susceptibility loci highlights the role of innate immunity, Nature Genet. 44 (12) (2012) 1341–1348.
[29] A. Villarreal-Martnez, H. Gallardo-Blanco, R. Cerda-Flores, I. Torres-Muñoz, M. Gómez-Flores, J. Salas-Alans, J. Ocampo-Candiani, L. Martnez-Garza, Candidate gene polymorphisms and risk of psoriasis: A pilot study, Exp. Ther. Med. 11 (4) (2016) 1217–1222.
[30] Q. Xiong, N. Ancona, E.R. Hauser, S. Mukherjee, T.S. Furey, Integrating genetic and gene expression evidence into genome-wide association analysis of gene sets, Genome Res. 22 (2) (2012) 386–397.
[31] X. Yin, H.Q. Low, L. Wang, Y. Li, E. Ellinghaus, J. Han, X. Estivill, L. Sun, X. Zuo, C. Shen, et al., Genome-wide meta-analysis identifies multiple novel associations and ethnic heterogeneity of psoriasis susceptibility, Nat. Commun. 6 (2015).