Gene 593 (2016) 225–234

Contents lists available at ScienceDirect

Gene

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

Research paper

Weighted gene co-expression based biomarker discovery for psoriasis detection

Sudharsana Sundarrajan, Mohanapriya Arumugam *

Department of Biotechnology, School of Biosciences and Technology, Vellore Institute of Technology University, India

ARTICLE INFO

Article history:
Received 1 March 2016
Received in revised form 5 August 2016
Accepted 10 August 2016
Available online 11 August 2016

Keywords:
Psoriasis
Coexpression
WGCNA
System biology
Random forest
Classifier
Biomarker

ABSTRACT

Psoriasis is a chronic inflammatory disease of the skin with an unknown aetiology. The disease manifests itself as red and silvery scaly plaques distributed over the scalp, lower back and extensor aspects of the limbs. After receiving scant consideration for quite a few years, psoriasis has now become a prominent focus for new drug development. A group of closely connected and differentially co-expressed genes may act in a network and may serve as molecular signatures for an underlying phenotype. A weighted gene coexpression network analysis (WGCNA), a system biology approach has been utilized for identification of new molecular targets for psoriasis. Gene coexpression relationships were investigated in 58 psoriatic lesional samples resulting in five gene modules, clustered based on the gene coexpression patterns. The coexpression pattern was validated using three psoriatic datasets. 10 highly connected and informative genes from each module was selected and termed as psoriasis specific hub signatures. A random forest based binary classifier built using the expression profiles of signature genes robustly distinguished psoriatic samples from the normal samples in the validation set with an accuracy of 0.95 to 1. These signature genes may serve as potential candidates for biomarker discovery leading to new therapeutic targets. WGCNA, the network based approach has provided an alternative path to mine out key controllers and drivers of psoriasis. The study principle from the current work can be extended to other pathological conditions.

List of abbreviations

Gene	Description
SPRR2C	Small proline-rich protein 2C
TGM1	Transglutaminase 1
ZC3H12A	Zinc finger CCCH-type containing 12 A
S100 A12	S100 calcium binding protein A12
AKR1B10	Aldo-Keto reductase family 1, member B10
RHCG	Rh hamily, C glycoprotein
HPSE	Heparanase
IL36G	Interleukin 36, gamma
TPBG	Trophoblast glycoprotein
ADAP2	ArfGAP with dual PH domains 2
APLP2	Amyloid beta (A4) precursor-like protein 2
LAMA4	Laminin, alpha 4
OLFML3	Olfactomedin-like 3
PMP22	Peripheral myelin protein 22
FYN	Proto-oncogene, Src family tyrosine kinase
AEBP1	AE binding protein 1
PCOLCE	Procollagen C-endopeptidase enhancer
ZHX2	Zinc fingers and homeoboxes 2
CDK14	Cyclin-dependent kinase 14
SNX7	Sorting nexin 7
RCOR1	REST corepressor 1
XPOT	Exportin, TRNA
UBE2D2	Ubiquitin-conjugating enzyme E2D 2
MRPL13	Mitochondrial ribosomal protein L13
STAM	Signal transducing adaptor molecule (SH3 domain and ITAM Motif) 1
WSB2	WD repeat and SOCS box containing 2
G3BP1	GTPase activating protein (SH3 domain) binding protein 1
TXNDC9	Thioredoxin domain containing 9
CPSF6	Cleavage and polyadenylation specific factor 6
PLK2	Polo-like kinase 2
CNOT7	CCR4-NOT transcription complex, subunit 7
RAB1A	Member RAS oncogene family
MRPL42	Mitochondrial ribosomal protein L42
CCT2	Chaperonin containing TCP1, subunit 2 (Beta)
CAPRIN1	Cell cycle associated protein 1
TMX1	Thioredoxin-related transmembrane protein 1
NPEPPS	Aminopeptidase puromycin sensitive
BZW1	Basic leucine zipper And W2 domains 1
EIF5	Eukaryotic translation initiation factor 5
PWP1	Endonuclein
SNRPG	Small nuclear ribonucleoprotein polypeptide G
RPL26L1	Ribosomal protein L26-like 1
COX7B	Cytochrome C oxidase subunit VIIb
PSMB1	Proteasome (prosome, macropain) subunit, beta type, 1

(continued on next page)

* Corresponding author at: Vellore Institute of Technology University, Vellore, Tamil Nadu, India.
E-mail address: mohanapriyaa@vit.ac.in (M. Arumugam).

http://dx.doi.org/10.1016/j.gene.2016.08.021
0378-1119/

PSMB3	Proteasome (prosome, macropain) subunit, beta type, 3
PSMB6	Proteasome (prosome, macropain) subunit, beta type, 6
UQCRQ	Ubiquinol-cytochrome C reductase, complex III subunit VII
NDUFAB1	NADH dehydrogenase (ubiquinone) 1, alpha/beta subcomplex, 1
SLIRP	SRA stem-loop interacting RNA binding protein
TIMM8B	Translocase of inner mitochondrial membrane 8 homolog B

1. Introduction

Psoriasis is a complex disorder mediated by the interplay between immune system and skin (Cai et al., 2012). The disease is characterized by hyperproliferation of keratinocytes, infiltration of immune components such as T cells, dendritic cells, macrophages and neutrophils. The characteristic lesions of psoriasis are well defined, red, indurated plaques with micaceous scale. Psoriasis commonly presents before the age of 35 and has a major impact on the quality of life of patients affected by it (Finlay and Kelly, 1987). The therapeutic management of the psoriasis includes conventional treatments and new generation inhibitors targeting the immune response elements such as TNF-α, IL12-23, IL-17, JAK and PDE4 (Winterfield et al., 2005; Yeilding et al., 2011; Gooderham et al., 2015; Kwatra et al., 2012; Wittmann and Helliwell, 2013), however the disease presents us with the complex therapeutic challenge due to treatment failures. Many efforts are focussed in the development of novel therapeutics yet the complete cure still remains obscure. The identification of new biomarkers for the disease is a pioneering research area in psoriasis (Cordiali-Fei et al., 2014) and its translation into routine clinical practise can minimize the time of diagnosis and treatment.

Gene expression profiling has been widely used in varied research areas which include cancer research (Van't Veer et al., 2002), neurological disorders (Scherzer et al., 2007) and other infectious diseases (Gu et al., 2010). Majority of the microarray analysis focuses on the comparison between the normal and diseased states (Bragde et al., 2011). The identification of the differentially expressed genes is widely used type of analysis which screens the potential markers in the diseased condition (Charafe-Jauffret et al., 2006). Despite its great contribution to the field, a single microarray data are prone to false findings and the cross validation of datasets would significantly reduce those false findings and increase sensitivity. Gene co-expression network, the recent advancement in network biology has emerged as a holistic strategy for microarray analysis (Oldham et al., 2006). Coexpression networks investigate the relationships between genes and identify intrinsic modules of co-ordinately expressed genes (Horvath et al., 2006). The hub genes correlating with many other module genes within the modules have shown to play key roles in orchestrating the module behaviour and disease phenotype.

Several research groups around the globe utilized psoriasis transcriptome to find out genes related to disease pathogenesis. A meta-analysis was carried out by Tian et al., using microarray data from 5 studies involving 386 paired samples from 193 patients (Tian et al., 2012). The study identified candidate genes involved in psoriasis pathogenesis. Gene expressions from 6 different datasets involving 215 patients were utilized by Swindell et al., to identify genes that can be targeted for therapeutic interventions (Swindell et al., 2013).

In the present work, to find out the common threads in psoriasis by prevailing over the inevitable artifacts, we combined independent datasets from six different experiments using the established meta-analytic method. Integration of data from two platforms was done to improve gene signature selection (Bigler et al., 2013). Previous workers listed candidate genes by comparing the expression profiles of genes from psoriatic involved and uninvolved skin samples. In contrast to the previous works, the present work incorporates samples from multiple platforms and the gene signatures are identified by comparing the diseased sample with healthy normal samples. A gene co-expression network was built to uncover genes that are induced in the diseased state. The psoriasis specific gene modules are reduced into psoriasis specific hub signatures (PSHS) which includes highly interconnected genes from five enriched modules. A binary classifier is built using random forest method employing the normalized gene expressions of the PSHS to validate the effectiveness of these hub genes in distinguishing the diseased state from the non-diseased state.

2. Materials and methods

2.1. Data acquisition

Microarray datasets obtained from Gene expression Omnibus (GEO) (Barrett et al., 2013) was used for the analysis. The psoriasis expressions were retrieved from GSE6710, GSE13355, GSE14905, GSE34248, GSE30999 and GSE41662 (Table 1) and the normal skin gene expressions were retrieved from GSE13355, GSE14905, GSE16161 and GSE7553.

2.2. Data processing

All the expression profiles were normalized independently and identically for consistency. The gene expressions were processed using microarray specific packages available from Bioconductor (Gentleman et al., 2004) (http://www.bioconductor.org/) in R (http://cran.r-project.org/). The raw data (.cel files) were loaded using affy library. Pre-processing affymetrix expression arrays involves background adjustments, quantile normalization and summarization based on a multi-array model fit robustly using the median polish algorithm which was carried out using robust multichip average method (RMA) (Gautier et al., 2004). The probes not associated with any known genes were removed from further analysis. If multiple probes match a single gene, the probes with the highest interquartile range (IQR) were selected as recommended by previous researchers (Wang et al., 2012). The datasets were merged together to get a global data set. The batch effect was eliminated using COMBAT method (empirical bayes) and cross platform normalization was carried out using XPN method implemented using inSilicoMerging package in R. The information on merging through the removal of batch effects can be found in (Shabalin et al., 2008). Limma package was used to mine out genes with a threshold of logFC >2 and adjusted p-value (*t*-test) <0.05 analyzed using Benjamini and Hochberg method.

2.3. Network analysis and module detection

The WGCNA (Weighted gene correlation network analysis) library of R was used for coexpression network analysis (Langfelder and Horvath, 2008). GSE13355, the psoriatic transcriptome with highest number of samples (58 diseased and 64 normal) was considered as a primary source for the analysis. The network analysis was restricted to top 5000 most varying, common and highly connected genes in all the expression profiles. An adjacency matrix was built based on the Pearson correlation coefficients of the gene pairs establishing an unsupervised

Table 1
Psoriatic gene expression datasets used in the study.

Dataset	Size[a]	Age	Gender		Remarks
			Male	Female	
GSE6710	26	32–76	11	2	–
GSE13355	116	21–69	Not specified		–
GSE34248	28	19–71	10	4	Sub series belonging to same experiment
GSE41662	48	19–71	Not specified		
GSE30999	170	12–60	66	23	3 samples were excluded from the analysis due to low quality control measures
GSE14905	61	Not specified	Not specified		–

[a] Includes both lesional and non-lesional samples.