Journal of Dermatological Science xxx (2017) xxx–xxx

Contents lists available at ScienceDirect

Journal of Dermatological Science

journal homepage: www.jdsjournal.com

Deciphering psoriasis. A bioinformatic approach

Juan L. Melero, Sergi Andrades, Lluís Arola, Antoni Romeu*

Department of Biochemistry and Biotechnology, Rovira i Virgili University, Tarragona, Spain

ARTICLE INFO

Article history:
Received 21 July 2017
Received in revised form 25 September 2017
Accepted 18 November 2017

Keywords:
Psoriasis
Keratinocyte
Gene expression
Epigenetics
Bioinformatics

ABSTRACT

Background: Psoriasis is an immune-mediated, inflammatory and hyperproliferative disease of the skin and joints. The cause of psoriasis is still unknown. The fundamental feature of the disease is the hyperproliferation of keratinocytes and the recruitment of cells from the immune system in the region of the affected skin, which leads to deregulation of many well-known gene expressions.
Objective: Based on data mining and bioinformatic scripting, here we show a new dimension of the effect of psoriasis at the genomic level.
Methods: Using our own pipeline of scripts in Perl and MySql and based on the freely available NCBI Gene Expression Omnibus (GEO) database: DataSet Record GDS4602 (Series GSE13355), we explore the extent of the effect of psoriasis on gene expression in the affected tissue.
Results: We give greater insight into the effects of psoriasis on the up-regulation of some genes in the cell cycle (CCNB1, CCNA2, CCNE2, CDK1) or the dynamin system (GBPs, MXs, MFN1), as well as the down-regulation of typical antioxidant genes (catalase, CAT; superoxide dismutases, SOD1-3; and glutathione reductase, GSR). We also provide a complete list of the human genes and how they respond in a state of psoriasis.
Conclusion: Our results show that psoriasis affects all chromosomes and many biological functions. If we further consider the stable and mitotically inheritable character of the psoriasis phenotype, and the influence of environmental factors, then it seems that psoriasis has an epigenetic origin. This fit well with the strong hereditary character of the disease as well as its complex genetic background.

What's already known about this topic?

- Psoriasis influences a wide variety of genes with a wide range of biological functions.
- Psoriasis is also known to have a strong hereditary character and a complex genetic basis.

What does this study add?

- Cyclins B1, A2, E2, CDK1, and B-type lamin genes are highly expressed in psoriatic tissue.
- Dynamin system is up-regulated in psoriasis.
- Catalase, superoxide dismutase and glutathione reductase a down-regulated in psoriasis.
- Psoriasis is associated with kallikreins.
- The effect of psoriasis extends across all regions of the chromosomes of the human genome. This effect on gene expression affects not only a single gene but also a gene cluster.
- Combining the strong hereditary character of the psoriasis with the results we show here, it appears that psoriasis could be a disease of epigenetic origin.

1. Introduction

Psoriasis is a chronic autoimmune skin disease [1,2]. Although the cause of the disease is still unknown, psoriasis is a phenotype of skin biochemical and immune disorder. Some hereditary factors predispose to psoriasis, which has been linked with at least nine chromosomal loci (PSORS1 through PSORS9). The molecular mechanisms responsible for psoriasis still have to be fully elucidated. The disease is multifactorial and noteworthy involves the hyperproliferation of keratinocytes in the epidermis. Various reviews have been published of the advances in our understanding of psoriasis [3–5].

Gene expression, although representing a specific site of regulation, is only one step in the complex cascade from an initial stimulus to a final phenotype. In this context, we were keen to explore the extent of the effect of psoriasis on gene expression in

* Corresponding author.
E-mail address: antoni.romeu@urv.cat (A. Romeu).

https://doi.org/10.1016/j.jdermsci.2017.11.010
0923-1811/

Please cite this article in press as: J.L. Melero, et al., Deciphering psoriasis. A bioinformatic approach, J Dermatol Sci (2017), https://doi.org/10.1016/j.jdermsci.2017.11.010

the affected tissue. First, at a quantitative level, that is, how many genes there are (or can be) that in a healthy state are silenced and awakened by the disease. And, vice versa, how many genes are highly expressed in a healthy state but totally silenced in psoriasis. Second, at a more qualitative level, we were interested in studying the extent of psoriasis in functional genomics, as well as the regions of the genome that are affected. To do this, we used freely available data from the NCBI-Gene Expression Omnibus (GEO) database of whole human genome arrays with paired lesional and non-lesional skins from psoriasis patients. Using this data, we made an analysis of variance to identify differences in gene expression profiles among groups. From this standpoint, then, this study uses a bioinformatic approach to data mining.

2. Methodology

The sources of information were the current version of the Human Genome (NCBI, GRCh37.p13) and the freely available NCBI Gene Expression Omnibus (GEO) database: DataSet Record GDS4602 (Series GSE13355) based on an analysis of skin punch biopsies from lesional skin and non-lesional skin from a group of 58 psoriatic patients. The platform used in these whole Human Genome arrays was the Affymetrix GPL570 (HG-U133_Plus_2, Affymetrix Human Genome U133 Plus 2.0 Array). For this study, we created our own pipeline of scripts in Perl and MySql for data management. The rationale of this work is based on the following tasks:

2.1. Task 1. Samples download

In accordance with the experimental design of the GDS4602 record, gene expression profiles in uninvolved and involved skin from 58 psoriasis affected individuals were characterised. We used these two array groups. Each array (or sample) record consists of a text file of 54,675 rows, in which each "probe" has its "probe-RNA hybridization signal". These values are considered arbitrary units. We downloaded the 116 array (sample) records.

2.2. Task 2. Probe selection

The GPL570 platform has a 54,675 Affymetrix probe set. However, we only selected the probes whose associated Entrez_id number and gene symbol match those annotated in the current Human Genome release. A total of 46,002 probes were selected. Some of the probes in this platform are designed to recognize only transcripts (ending with '_at') and while others are designed to recognize multiple transcripts from the same gene family (ending with '_s_at' or '_a_at'). In order to further adjust the analysis of gene expression, here we have only considered the transcript-specific probes. On the other hand, with respect to the genes in the GPL570 platform, 19,873 had a match in the Human Genome release. Of these, 12,927 genes are associated with a single probe, 4387 are associated to two probes, 1611 genes are associated to three probes, 599 genes are associated to four probes, etc.

2.3. Task 3. From "probe signal value" to "gene signal value"

Our goal was to assign each gene from each array (sample) with a "gene signal value" based on the signal values of the probes associated with the genes. When a gene had one, and only one, associated probe, the signal value of that probe was the signal value of the gene.

When a gene had more than one associated probe, we performed a descriptive statistical analysis of the signal values of the associated probes. If the standard error was less than ten percent of the mean value, we considered that this mean value of the probe signal values was consistent enough to be accepted as the corresponding gene signal value. However, if the standard error was greater than ten percent of the mean, we removed an extreme value (the minimum or maximum depending on the relative values of the mean and the median), which were expected to be outliers, and repeated the analysis with the remaining values. We used the same criteria to accept the "gene signal value". If the error value was still large after this second analysis, no gene signal value was accepted and the gene in question was not included in the gene list of the array (sample) analysed. We took each sample record individually, so not every sample had the same number of genes because we discarded genes from some samples for being ambiguous.

Once we had calculated the "gene signal value" for each array and each gene, and to facilitate the data analysis, we also calculated the "gene rank order" of signal measurements. All the gene signal values of one array were sorted, and the percentage calculated. This gave us an indication of where the signal of one gene fell with respect to all the other genes in one array.

2.4. Task 4. Descriptive statistics of gene signal values from each experimental group

We had two experimental groups of whole human genome arrays (N = 58): skin biopsies from lesional skin and non-lesional skin. So, for each group, the data set consists of 58 variables (each array or sample), and the cases were the genes (about 20,000 rows). The units of the variables were the "gene signal values". For each group, we used a Perl script to make an automatic descriptive statistical analysis and calculate the signal value for each gene.

2.5. Task 5. Statistical significance of mean gene signal values between groups

We were interested in gathering information about these groups so that we could compare them. We focused on the genes common to both groups. For each gene, we automatically used the test of significance for two known means, with known standard deviations and "N" (number of variables). The null hypothesis was that the means of the gene expression value were equal between groups. We calculated the two-sample z statistic. The significance level we considered was $\alpha = 0.01$. Therefore, there were significant differences if $z < -2.58$ or $z > 2.58$. All the results were analysed in detail using a local MySQL database.

3. Results and discussion

The discussion is based on the statistical analysis of specific patterns of mRNA expression values from gene expression profiling by arrays of healthy and lesional skin biopsies of psoriasis. Because environmental factors are important in psoriasis [6], we avoided mixing data from different array experiments from the NCBI-GEO series because they involved different population groups. According to the experimental design, the cell population in both types of skin biopsies (control and lesional) is notoriously different. In healthy skin biopsies, the cell population is typical of the epidermis, dermis and hypodermis, in which keratinocytes are the majority [7]. However, in lesional biopsies there is also a population of immune cells [1]. Thus, in the present data analysis of the gene array signal of probe-RNA hybridization, it should be borne in mind that this is not the analysis of a single cell but of a cluster of cells.

Table S1 shows the complete list of human genes in both experimental groups (control and lesional skin biopsies), which have been through the filtering procedures described above. Table S1 shows the mean group values of the probe-RNA signals for

Please cite this article in press as: J.L. Melero, et al., Deciphering psoriasis. A bioinformatic approach, J Dermatol Sci (2017), https://doi.org/10.1016/j.jdermsci.2017.11.010