Manually curated and harmonised transcriptomics datasets of psoriasis and atopic dermatitis patients

Antonio Federico1,2,5, Veera Hautanen1,2,5, Nils Christian3, Andreas Kremer3, Angela Serra1,2 & Dario Greco1,2,4 ✉

We present manually curated transcriptomics data of psoriasis and atopic dermatitis patients retrieved from the NCBI Gene Expression Omnibus and EBI ArrayExpress repositories. We collected 39 transcriptomics datasets, deriving from DNA microarrays and RNA-Sequencing technologies, for a total of 1677 samples. We provide quality-checked, homogenised and preprocessed gene expression matrices and their corresponding metadata tables along with the estimated surrogate variables. These data represent a ready-made valuable source of knowledge for translational researchers in the dermatology field.

Background & Summary
Psoriasis (PSO) and Atopic dermatitis (AD) are among the most common inflammatory skin disorders associated with immunologic impairment. While the first signs of AD tend to appear in the early childhood, the manifestation of PSO is most common during the third decade of life. Both the diseases have a substantial negative impact on the quality of life of affected patients. Although a number of therapeutic approaches have been developed in the last two decades to mitigate PSO and AD symptoms, their pathophysiology is still not completely understood. AD is believed to be driven by epidermal barrier disruption, activation of specific T-cell subsets, and dysbiosis of the commensal skin microbiome while psoriatic inflammation is sustained by uncontrolled responses of the innate and adaptive cutaneous immune system, which lead to intense keratinocyte proliferation and dysfunctional differentiation.

Transcriptomics technologies, such as DNA microarray and RNA Sequencing (RNA-Seq), have been used to characterise the molecular alterations of human diseases, including PSO and AD. To date, only marginal efforts have been carried out in order to collect, quality-check and harmonize PSO- and AD-related transcriptomics data in order to make them easily reusable by the research community. Therefore, the motivation behind this study was to create a source of ready-to-use data of gene expression profiles of PSO and AD patients derived from both DNA microarray and RNA-Seq publicly available datasets.

The preprocessed and harmonized microarray data provided in this study were collected from the NCBI Gene Expression Omnibus (GEO) and EBI ArrayExpress public repositories, while the RNA-Seq datasets were retrieved from the European Nucleotide Archive (ENA). Overall, 26 microarrays datasets were collected, for a total of 991 samples, 632 of which from patients affected by psoriasis and 70 by atopic dermatitis. Some of the microarray datasets contain samples collected from patients affected by other skin diseases such as psoriatic arthritis, psoriasis sebaceous hyperplasia, palmoplantar pustulosis, lichen planus and discoid lupus. These datasets were generated with commercially available Affymetrix and Agilent platforms. All of the analytical steps performed in this work were carried out through the use of the eUtopia software. We also retrieved 13 RNA-Seq datasets, for a total of 686 samples, 392 of which from patients affected by psoriasis and 94 by atopic dermatitis. RNA-seq data were mostly produced through Illumina platforms, while a minority of datasets were produced...
through other platforms. All the datasets underwent meta-data curation and harmonisation, data quality check and preprocessing with standardised procedures. The curation and harmonisation of the meta-data consisted in the definition and usage of a common data model for all of the collected datasets. The data models, to which the raw meta-data were mapped to, are reported in the data dictionary files (enclosed with the preprocessed data). The data dictionary describes all the variables reported in the final metadata tables. For each variable, the description, type and allowed values are reported. At the same time, this work is aimed at homogenising the preprocessing procedures in order to improve the comparability of the gene expression data across different studies and platforms. Therefore, in this work we provide meta-data tables, along with the inferred surrogate batch variables, as well as the preprocessed gene expression estimates.

Our analysis significantly increases the FAIRness of publicly available PSO and AD transcriptomics data and represents a valuable “ready-to-use” resource available to the scientific community.

Methods

Microarray data. Data collection and homogenization. Transcriptomics data generated by DNA microarrays of psoriasis and atopic dermatitis patients were retrieved from NCBI GEO (GEO - https://www.ncbi.nlm.nih.gov/geo/) and EBI ArrayExpress (https://www.ebi.ac.uk/arrayexpress/) repositories by using the R packages GEOquery and ArrayExpress, respectively. For each dataset, a table specifying the disease (psoriasis/atopic dermatitis) and the origin of biopsy (lesional/non-lesional sample) in addition to other phenotypic information was also retrieved. Since the phenotypic information was heterogeneous across the datasets, rigorous harmonization procedure was performed. The GEO and Array Express identifiers of the retrieved datasets are reported in Tables 1–3.

Data quality check. The retrieved datasets were thoroughly quality checked. In particular, each sample was evaluated by visual inspection of the array pseudo-images, quality check reports and density plots of probe intensities by using the eUTOPIA software. Further, outlier detection step, based on the sample distributions, was performed within each dataset by using ad hoc R scripts (see Code Availability section). Moreover, for the Affymetrix datasets, outlier samples were detected by computing the Normalized Unscaled Standard Error (NUSE) and the Relative Log Expression (RLE) from the affyPLM v1.64.0 R package, and the RNA degradation curves (RNADeg) from the affy v1.64.0 R package (Fig. 1).

The distributions of the values of these three metrics were investigated by means of boxplots and the sample outlierness was evaluated for each measure based on the data distribution. Eventually, a concordance outlierness score was computed across the three metrics. In particular, a sample was removed from the analysis if considered an outlier in at least two out of three metrics, one of them being the RNA degradation curve.

Normalization. Data normalization was performed by using the eUTOPIA software. Affymetrix-based studies were normalized by using the justRMA from the R affy v1.66.0 package. For the Agilent datasets, the probes were mapped to the Ensembl gene IDs and the expression matrix was aggregated by computing the median of the expression of the Agilent probes mapping to the same Ensembl transcript ID. The entire DNA microarray data preprocessing is depicted in Fig. 1.

Surrogate variable analysis. In order to investigate the effect of unknown batches that might mask biological variability, Surrogate Variable Analysis (SVA) was performed with the eUtopia software, which implements the sva R package. The analysis was performed by using origin of biopsy or diagnosis as variable of interest. The other biological variables (if present and if not confounded with the variable of interest) were used as covariates. The estimated surrogate variables for each dataset are included in the meta-data tables.

Probe annotation. Custom annotation files (CDF files) were downloaded from Brainarray (http://brainarray.mbni.med.umich.edu/Brainarray/Database/CustomCDF/CDFi_download.asp) for Affymetrix-based microarrays. The latest version of Agilent probe annotation was retrieved from the Agilent website (https://earray.chem.agilent.com/earray). The probes were mapped to the Ensembl gene IDs and the expression matrix was aggregated by computing the median of the expression of the Agilent probes mapping to the same Ensembl transcript ID. The entire DNA microarray data preprocessing is depicted in Fig. 1.

RNA Sequencing data. Data collection and homogenization. Raw files in “.fastq” format were retrieved from the European Nucleotide Archive (ENA). Along with the raw data files, the metadata tables reporting the samplewise clinical features for each dataset were also collected. As for the DNA microarray data, the metadata tables of RNASeq data were carefully harmonized to improve the across-datasets comparability. Phenotypic
information for each dataset is reported along with the gene expression tables. GEO and ENA identifiers of the retrieved datasets are reported in Tables 1–3.

**Quality control.** All the RNA-Seq datasets underwent quality check through the use of FastQC v0.11.7 (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/). Reads were trimmed for low-quality ends in addition to adapters removal by TrimGalore v0.4.4_dev (http://www.bioinformatics.babraham.ac.uk/ projects/trim_galore/). In particular, the reads were trimmed if the Phred score was lower than 20 and discarded if the number of undetected nucleotides was greater than 50. The trimmed and adapter-clipped raw files were further quality checked with FastQC v0.11.7.

**Read alignment.** RNA Sequencing reads were then aligned against the human reference genome assembly GRCh38. The alignment was performed through the use of the HISAT2 algorithm17,18 using the genome indexes built for usage with HISAT2 (retrieved from https://ccb.jhu.edu/software/hisat2/manual.shtml).
Conversions between .sam and .bam file formats, sorting and extraction of uniquely mapped reads were performed through the use of samtools version 1.8-27-g089626219.

Read counts extraction. Transcript abundance was computed by using the featurecounts function from the Rsubread v2.2.3 R package20. To accomplish this task, the Gencode version 31 annotation was downloaded from https://www.gencodegenes.org, and then utilized for read counts extraction.

Low counts filtering. In order to filter out the transcripts with low expression levels in all the samples of each dataset, the proportion test strategy was used as implemented in the function filtered.data of the R package NOISeq v2.31.021.

Normalization. RNASeq expression data were normalized using the upper quantile method22 implemented in the R package NOISeq v2.31.0.

Surrogate Variable Analysis. As for the DNA microarray data, in order to identify unknown sources of technical variability, a Surrogate Variable Analysis (SVA) was performed through the use of the svaseq function implemented in the sva v3.36.0 R/Bioconductor package16. The analysis was performed by using disease state or diagnosis as variable of interest. The other biological variables (if present and if not confounded with the variable of interest) were used as covariates6. The estimated surrogate variables for each dataset are included in the meta-data tables, along with the gene expression tables. The entire RNA-Seq data preprocessing is depicted in Fig. 2.

Data Records
The complete list of DNA microarray and RNA Sequencing datasets discussed in this work is reported in Tables 1–3. All the preprocessed transcriptomics data, along with harmonised meta-data, were submitted to Zenodo23.

Technical Validation
DNA microarray and RNA-Seq data are linked to clinical meta-data, reporting multiple information such as gender, age or the treatment (including e.g. drug dose). Additionally, sample meta-data is recorded, such as the tissue type a sample was taken from, or whether the tissue derives from a lesional or nonlesional sample.

In order to ensure that the data is recorded in a consistent and well-formed way, we created data dictionaries describing each of these variables. The data dictionaries contain detailed information describing the content of a variable, the data type (numeric, categorical, text, date, etc), the allowed values of categorical data or ranges of numeric variables.

The data was validated by checking compliance with the rules encoded in the data dictionaries. Data that was found not to comply with the rules was manually curated by consulting the original data sources. In fact, a large proportion of the datasets were found not to meet the requirements encoded in the data dictionaries. For
instance, big heterogeneity was found in the description of the skin status. “Involved skin,” “psoriatic skin” were reported in order to describe the “lesional” status of the skin. “Normal,” “ctrl,” “Non-involved skin of healthy individual” were used to describe the “healthy control” samples. Yet, to define the gender, “m,” “f,” “male” and “female” were used across the datasets. All of these variables were mapped to the allowed values reported in the data dictionaries to improve the comparability across the datasets.

Usage Notes

The transcriptomics data presented in this article is an unprecedented source of preprocessed, harmonized, “ready-to-use” and FAIR datasets, made available to the scientific community. Data derived from both DNA microarray and RNASeq technologies can be exploited in order to uncover the molecular mechanisms underlying psoriasis and atopic dermatitis. Differential expression analysis can be carried for instance by the limma package[15] for the microarray data, and the edgeR[24], DESeq 2[25] or NOISeq[21] packages for RNA-Seq data, respectively. Functional analysis of differentially expressed genes can be performed by using FunMappOne[26], the R/Bioconductor package ReactomePA[27] or Ingenuity Pathway Analysis (Qiagen, http://www.ingenuity.com/products/ipa). The inference and analysis of co-expression networks can be performed, for instance, by using the INfORM tool[28]. Altogether, these analyses can aid the stratification of PSO and AD patients, the identification of relevant biomarkers and novel therapeutic targets.

Code availability

R scripts for the analysis of DNA microarray and RNA-Seq transcriptomics data are available for download at: https://github.com/Greco-Lab/psoriasis-dermatitis-analysis.

Received: 10 July 2020; Accepted: 3 September 2020;
Published: xx xx xxxx

References

1. Bowcock, A. M. & Cookson, W. O. The genetics of psoriasis, psoriatic arthritis and atopic dermatitis. Hum. Mol. Genet. 13, 43–55 (2004).
2. Tsoi, L. C. et al. Progression of acute-to-chronic atopic dermatitis is associated with quantitative rather than qualitative changes in cytokine responses. J. Allergy Clin. Immunol. 5, 1406–1415 (2019).
3. Tsoi, L. C. et al. Large scale meta-analysis characterizes genetic architecture for common psoriasis associated variants. Nat. Commun. 8, 15382 (2017).
4. Rendon, A. & Schäkel, K. Psoriasis Pathogenesis and Treatment. Int. J. Mol. Sci. 6, 1475 (2019).
5. Casamassimi, A., Federico, A., Rienzo, M., Esposito, S. & Ciccodicola, A. Transcriptome profiling in human diseases: new advances and perspectives. Int. J. Mol. Sci. 8, 1652 (2017).
6. Marwah, V. S. et al. eUTOPIA: solution for Omics data Preprocessing and Analysis. Source Code Biol. Med. 14 (2019).
7. Wilkinson, M. D. et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci. Data 1, 160018 (2016).
8. Barrett, T. et al. NCBI GEO: archive for functional genomics data sets update. Nucleic Acids Res. 1, 991–995 (2012).
9. Davis, S. & Meltzer, P. S. GEOquery: a bridge between the Gene Expression Omnibus (GEO) and BioConductor. Bioinformatics 14, 1846–1847 (2007).
10. Kauffmann, A. et al. Importing ArrayExpress datasets into R/Bioconductor. *Bioinformatics* **16**, 2092–2094 (2009).
11. Athar, A. et al. ArrayExpress - update from bulk to single-cell expression data. *Nucleic Acids Res.* **1**, 711–715 (2018).
12. Bretttschneider, J., Collin, F., Bolstad, B. M. & Speed, T. P. Quality assessment for short oligonucleotide microarray data. *Technometrics* **3**, 241264 (2008).
13. Fasold, M. & Binder, H. AffyRNAdegradation: control and correction of RNA quality effects in GeneChip expression data. *Bioinformatics* **1**, 129–131 (2013).
14. Gautier, L., Cope, L., Bolstad, B. M. & Irizarry, R. A. affy-analisis of Affymetrix GeneChip data at the probe level. *Bioinformatics* **3**, 307–315 (2004).
15. Ritchie, M. E. et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **7**, e47 (2015).
16. Leek, J. T. et al. sva: Surrogate Variable Analysis. R package version 3.32.1. [https://bioconductor.org/packages/release/bioc/html/sva.html](https://bioconductor.org/packages/release/bioc/html/sva.html) (2019).
17. Pertea, M., Kim, D., Pertea, G. M., Leek, J. T. & Salzberg, S. L. Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown. *Nat. Protoc.* **9**, 1650–1667 (2016).
18. Kim, D., Langmead, B. & Salzberg, S. L. HISAT: a fast spliced aligner with low memory requirements. *Nat. Methods* **4**, 357–360 (2015).
19. Li, H. et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **1**, 2078–2079 (2009).
20. Liao, Y., Smyth, G. K. & Shi, W. The R package Rsquared is easier, faster, cheaper and better for alignment and quantification of RNA sequencing reads. *Nucleic Acids Res.* **8**, e47 (2019).
21. Tarazona, S. et al. Data quality analysis of differential expression in RNA-seq with NOISeq R/Bioc package. *Nucleic Acids Res.* **21**, e140 (2015).
22. Bullard, J. H., Purdom, E., Hansen, K. D. & Dudoit, S. Evaluation of statistical methods for normalization and differential expression in mRNASeq experiments. *BMC Bioinf.* **11**, 94 (2010).
23. Federico, A. et al. Preprocessed and Harmonised Transcriptomics Datasets for Psoriasis and Atopic Dermatitis. *Zenodo* [https://doi.org/10.5281/zenodo.4009497](https://doi.org/10.5281/zenodo.4009497) (2020).
24. Robinson, M. D. & Oshlack, A. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol.* **3**, R25 (2010).
25. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **12**, 550 (2014).
26. Scala, G., Serra, A., Marwah, V. S., Saarimäki, L. A. & Greco, D. FunMappOne: a tool to hierarchically organize and visually navigate functional gene annotations in multiple experiments. *BMC Bioinf.* **1**, 79 (2019).
27. Yu, G. & He, Q. Y. ReactomePA: an R/Bioconductor package for reactome pathway analysis and visualization. *Mol. Biosyst.* **2**, 477–479 (2016).
28. Marwah, V. S. et al. INFORM: Inference of NetwOrk Response Modules. *Bioinformatics* **12**, 2136–2138 (2018).

**Acknowledgements**

This study was supported by the EU IMI2 Biomap Project (Grant agreement 821511).

**Author contributions**

A.F. and D.G. conceived and designed the study; A.F. and V.H. retrieved the data and performed the data preprocessing; A.F. and N.C. quality checked and harmonised the meta-data; A.K. supervised the quality check and the data harmonisation. A.F. and N.C. drafted the manuscript. A.S. and D.G. supervised the activities and revised the manuscript. All authors have read and agreed to the published version of the manuscript.

**Competing interests**

The authors declare no competing interests.

**Additional information**

**Correspondence** and requests for materials should be addressed to D.G.

**Reprints and permissions information** is available at [www.nature.com/reprints](https://www.nature.com/reprints).

**Publisher’s note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit [http://creativecommons.org/licenses/by/4.0/](http://creativecommons.org/licenses/by/4.0/).

The Creative Commons Public Domain Dedication waiver [http://creativecommons.org/publicdomain/zero/1.0/](http://creativecommons.org/publicdomain/zero/1.0/) applies to the metadata files associated with this article.

© The Author(s) 2020