*Perspective*

# Large Scale Advanced Data Analytics on Skin Conditions from Genotype to Phenotype

**Maryam Panahiazar [1,*], Darya Fadavi [2], Jihad Aljabban [3], Laraib Safeer [4], Imad Aljabban [5] and Dexter Hadley [1]**

[1]   Institute for Computational Health Sciences, School of Medicine, University of California San Francisco, San Francisco, CA 94143, USA; Dexter.Hadley@ucsf.edu
[2]   Department of Bioengineering, University of California Berkeley, Berkeley, CA 94720, USA; daryafadavi@berkeley.edu
[3]   College of Medicine, Ohio State University, Columbus, OH 43210, USA; jihadal1992@gmail.com
[4]   Baylor College of Medicine, Houston, TX 77030, USA; Laraib.Safeer@bcm.edu
[5]   College of Medicine, Pennsylvania State University, Hershey, PA 16801, USA; imadaljabban1992@gmail.com
*   Correspondence: MARYAM.PANAHIAZAR@UCSF.EDU

check for updates

**Abstract:** A crucial factor in Big Data is to take advantage of available data and use that for new discovery or hypothesis generation. In this study, we analyzed Large-scale data from the literature to OMICS, such as the genome, proteome or metabolome, respectively, for skin conditions. Skin acts as a natural barrier to the world around us and protects our body from different conditions, viruses, and bacteria, and plays a big part in appearance. We have included Hyperpigmentation, Postinflammatory Hyperpigmentation, Melasma, Rosacea, Actinic keratosis, and Pigmentation in this study. These conditions have been selected based on reasoning of big scale UCSF patient data of 527,273 females from 2011 to 2017, and related publications from 2000 to 2017 regarding skin conditions. The selected conditions have been confirmed with experts in the field from different research centers and hospitals. We proposed a novel framework for large-scale available public data to find the common genotypes and phenotypes of different skin conditions. The outcome of this study based on Advance Data Analytics provides information on skin conditions and their treatments to the research community and introduces new hypotheses for possible genotype and phenotype targets. The novelty of this work is a meta-analysis of different features on different skin conditions. Instead of looking at individual conditions with one or two features, which is how most of the previous works are conducted, we looked at several conditions with different features to find the common factors between them. Our hypothesis is that by finding the overlap in genotype and phenotype between different skin conditions, we can suggest using a drug that is recommended in one condition, for treatment in the other condition which has similar genes or other common phenotypes. We identified common genes between these skin conditions and were able to find common areas for targeting between conditions, such as common drugs. Our work has implications for discovery and new hypotheses to improve health quality, and is geared towards making Big Data useful.

**Keywords:** Data Analytics; Large Scale; Meta Analysis

## 1. Introduction

Individual studies have been performed on each skin condition, from gene discovery to finding new molecular signatures to improve health quality for different skin conditions [1–4]. This is due to complexity and challenges to using different sources on a large scale at the same time [5]. In this study, we take advantage of avaibale large-scale public data such as literature and OMICS Data to find the common

genotypes and phenotypes of skin conditions. We have included Hyperpigmentation, Postinflammatory hyperpigmentation, Melasma, Rosacea, Actinic keratosis, and Pigmentation in this study. This selection is based on reasoning over all female patients at UCSF (527,273 patients) from 2011–2017 and all related publication from 2000–2017 which is almost 20 years of publications. For the last step which is required for any medical study, we got confirmation from the experts in the field as well.

Pigmentation disorders involve alterations in the number of melanocytes or of melanin production and ultimately result in either hyperpigmentation or hypopigmentation. Postinflammatory hyperpigmentation occurs at sites of previous skin inflammation [6]. Melasma is hyperpigmentation of the skin, but it may also occur in association with pregnancy, or with ingestion of hormonal contraceptives and certain medications [6]. Rosacea is a chronic, inflammatory disorder of the skin that is associated with increased reactivity of capillaries leading to skin flushing [6]. Over time, chronic inflammation in Rosacea results in edema and fibrosis, leading to a rubbery thickening of the skin of the nose, cheeks, forehead, or chin [6]. Actinic keratosis results from prolonged and repeated sun exposure and is a precursor of cutaneous squamous cell carcinoma [6]. We excluded cancer conditions from this study due to the specific considerations required and their complexity, which warrant a separate study that we will consider for a future work. Our objective is to analyze several different skin conditions using Advanced Data Analytics (ADA) as a facilitator Tools and platform for our research. In this study we are not trying to explain IPA platform specifically and how to integrate different underlying data types as its knowledge base. They are several papers and resources that explain how IPA works by details [7]. Advanced Data Analytics (ADA) in ingenuity pathway analysis (IPA) looked at what genes and proteins are being expressed, overexpressed or underexpressed in each sample. Then using this information, we found common genotypes and phenotypes for the skin conditions.

In this study, we take a step forward and look to find patterns and commonalities between these different skin conditions with our results from ADA. The novelty of this work is that it is a meta-analysis of different features on different skin conditions. It means instead of looking at individual conditions with one or two features, which is how most of the previous works have been conducted, we looked at several conditions with different features to find the common factors between them. Our hypothesis is that by finding the overlap in genotype and phenotype between different skin conditions, we can suggest using a drug that is recommended in one condition, for treatment in the other condition which has similar genes or other common phenotypes. We took all the data for different features and conditions and looked at where there is an overlap between them, what genes are commonly expressed between different conditions, and what drugs are used and effective for these different conditions. By analyzing the data from a more zoomed-out view, we are able to get a complete picture of skin and see common areas for targeting between conditions. In the following sections, we will explain the related work, materials and method, results, and we will conclude with the discussion and suggestions for future works.

## 2. Related Work

The gene expression profiles of dermatological conditions such as actinic keratosis, Rosacea, and Hyperpigmentation remain largely unknown. Skin pigmentation, for instance, is a complex pathway with multiple regulatory elements controlling the production of melanin and its transport to neighboring keratinocytes [8]. These regulatory elements may be potential targets in the treatment of Hyperpigmentation disorders, and a genomics approach can help identify such key regulators. Microarray studies evaluating the genomics of skin disorders are typically limited by small sample size [9]. Furthermore, most genomics studies focus on one or two skin conditions [10]. One study employed a Big Data approach by performing meta-analysis on five data-sets from the NCBI Gene Expression Omnibus (GEO) database [11] to elucidate the pathogenesis of multiple subtypes of hyperpigmentation: pigmentation following UV exposure, long-lasting pigmentation, Postinflammatory pigmentation, age spots, and ethnic skin [10]. A limitation of this study is that the microarray data was limited to the limited studies from the GEO database, resulting in small sample sizes. Another limitation is that the

researchers only examined microarray data of skin hyperpigmentation, and did not consider other dermatologic conditions. A second study analyzed publicly available microarray data of 16 diverse skin conditions in order to gain insight into disease pathogenesis, but was again limited to data-sets from the GEO database [12]. It is clear, however, that even with such limitations, genomic studies have successfully identified pigmentation control targets and effective therapies, and are a stepping stone to personalized medicine [13].

In this study, we employ large-scale data across multiple microarray databases to examine six skin conditions in order to demonstrate that a genomics approach incorporating large-scale data produces generalizable results that may help identify biologically robust gene expression and key regulators of disease.

## 3. Material and Method

This work is based on Ingenuity Knowledge Base and Ingenuity Pathway Analysis (IPA) to Advanced Data Analytics (ADA) [7,14]. Data were analyzed through the use of IPA [15]. IPA reveals information on biological pathways, drugs, and diseases underlying OMICS data and scientific literature. Genomics, proteomics, metabolomics, and other life science data are puzzle pieces to complex biological networks and pathways. Ingenuity knowledge Base leverages OMICS data to access this complex network.

In the following section, we explain data collection, feature selection, Ingenuity Pathway Analysis, and Advanced Data Analytics. We used these to find genes, drugs, all synonymous gene names Entrez Gene ID for Human, Entrez Gene ID for Mouse and Rat, Functions and Functional Domains, Subcellular Locations, Canonical Pathways, and miRNA Functional Cluster for each identified gene. An example of the data gathered for a protein is shown in Table 1 using the protein Tyrosinase. Other findings from Ingenuity Knowledge Base from the literature include proteins and pathways the identified gene or protein regulates, proteins and pathways it is regulated by, targets it binds to, its role in the cell, and the related diseases. We also found GO Annotations including Molecular Function, Biological Process, and Cellular Component. And finally, we introduce Targeting Drugs for each condition (all detail info which is hundreds of pages could be sent by request to the reader and/or supplementary materials if applied).

### 3.1. Data Collection

The Ingenuity underlying Knowledge Base icontains over five million facts extracted from scientific publications and databases such as OMICS so that each relationship between molecule, diseases, and phenotype is characterized and searchable. Data derived from gene expression experiments include RNA-Seq, microRNA and SNP microarrays, metabolomics, and proteomics generate genes, chemical lists and molecular signatures of Ingenuity. IPA is based on the QIAGEN Knowledge Base which has been built as a horizontally and vertically structured database that pulls in relevant scientific and medical information and describes it consistently, making this data interoperable and computable so you can efficiently and effectively interpret your results [7].

This study is based on all accessible knowledge base used in IPA and ADA. In fact, one of the novelties of this study is the use of this big scale data for the analysis, which is very difficult for individual studies that are looking at a limited number of resources. IPA support different identifiers such as Genbank, SwissProt, Affy probe ID, RefSeq, etc. Therefore, different data types have been used in each individual resources, which need to be mapped or converted. The detail infrastructure for connecting different resources and mapping them has been discussed in IPA by details [7]. The outcomes include biological findings, millions of which have been curated by QIAGEN directly. All information—whether it's from content curation efforts or from publicly available or licensed databases is structured into ontology and exhaustively reviewed.

**Table 1.** Example Data for Tyrosinase Protein, TYR (Partial List of all Data).

| | |
|---|---|
| **Synonyms:** | albino, ATN, CMM8, Dopa oxidase, OCA1, OCA1A, OCAIA, SHEP3, … |
| **Entrez Gene ID for Human:** | TYR EG:7299 |
| **Entrez Gene ID for Mouse:** | TYR EG:22173 |
| **Functions/ Functional Domains:** | copper ion binding, cytoplasmic domain, cytosolic tail domain, enzyme, … |
| **Subcellular Location:** | cell periphery, cell membrane, cytoplasm, endosomes, melanosomes, … |
| **Canonical Pathway:** | Aryl Hydrocarbon Receptor Signaling; Eumelanin Biosynthesis, … |
| **Targeted By miRNA Cluster:** | miR-1208 (miRNAs w/seed CACUGUU), miR-1229-3p (miRNAs w/seed UCUCACC), … |
| **Regulates:** | melanin, L-tyrosine, TYR, L-dopa, KLRD1, KLRC1, … |
| **Regulated By:** | POMC, MITF, forskolin, ASIP, PD98059, cyclic AMP, ciglitazone, KITLG, … |
| **Binds To:** | CANX, CALR, HSPA5, TYRP1, SYVN1, … |
| **Role in the Cell:** | proliferation, pigmentation, melanogenesis in, shape change, … |
| **Related Diseases:** | oculocutaneous albinism, melasma, rosacea, acne vulgaris, cancer, melanosis, … |
| **Molecular Function:** | copper ion binding; monooxygenase activity; oxidoreductase activity; … |
| **Biological Process:** | eye pigment biosynthetic process; melanin biosynthetic process; pigmentation; … |
| **Cellular Component:** | cytoplasm; cytosol; Golgi-associated vesicle; melanosome; nucleus; … |
| **Targeting Drug:** | azelaic acid, hydroquinone |

As a first step, we extracted a group of genes for each skin conditions from Gene expression Omnibus (GEO). We found over 1000s of genes with the Search Tag Analyze Resource for GEO (STARGEO) platform [16]. STARGEO has been explained in our previous work by details [7]. The National Center for Biotechnology Information (NCBI) Gene Expression Omnibus (GEO) is an open database of more than 2 million samples of functional genomics experiments. The Search Tag Analyze Resource for GEO (STARGEO) platform allows for meta-analysis of genomic signatures of disease and tissue through tagging of biological samples from multiple experiments [16]. We extracted over 1000s of genes for skin conditions based on studies and samples from GEO with STARGEO platform. Then we analyzed the signature (genes) from our STARGEO analyses in Ingenuity Pathway Analysis (IPA), restricting genes that showed statistical significance ($p < 0.05$) and an absolute experimental log ratio greater than 0.1 between conditions and control samples. These selected genes have been used for the next step analysis in IPA to extract drugs, biological process and other factors, which we will explain in the result section.

*3.2. Feature Selection*

The six skin conditions have been selected based on prevalence in the literature review and UCSF patient data, and finally double checked with expert consulting from different medical schools around the USA such as Ohio, Pennsylvania, Texas, Berkley, and UCSF. Our cohort was composed of female patients. They are more female's history in skin conditions compare with men in electronic medical records, which constituted 54 percent of all patients. Figure 1 shows the diversity of skin conditions from 527,273 female patients in the UCSF patient data for these skin conditions. We also looked at the publications from 2000 to 2017. As it shows in Figure 2 during the last almost 20 years, there have been

several studies in the conditions such as UV, solar and cancer (excluded in our study), but the other conditions are neglected in comparison with these three conditions. Still, so many individual studies have been published. Therefore, we decided to focus on skin conditions including hyperpigmentation, Melasma, Rosacea, Postinflammatory hyperpigmentation, Actinic Aeratosis, and Pigmentation.

Therefore, the decision of which skin conditions to be included was made based on these two valuable pieces of information before confirming with experts. For the last step which is required for any medical study, we got confirmation from the experts in the field as well. We found this way very straightforward in the medical domain. Feature engineering, selection, and attribute selections in machine learning approaches take different approaches, which are different from the nature of this study, in which we looked up the statistics of the big scale data.
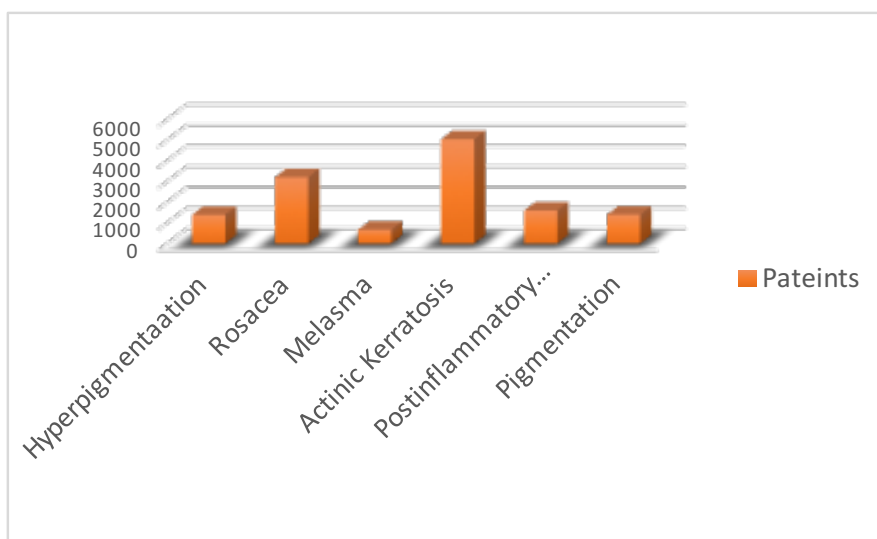


**Figure 1.** Distribution of patients at UCSF for each skin condition in our study.
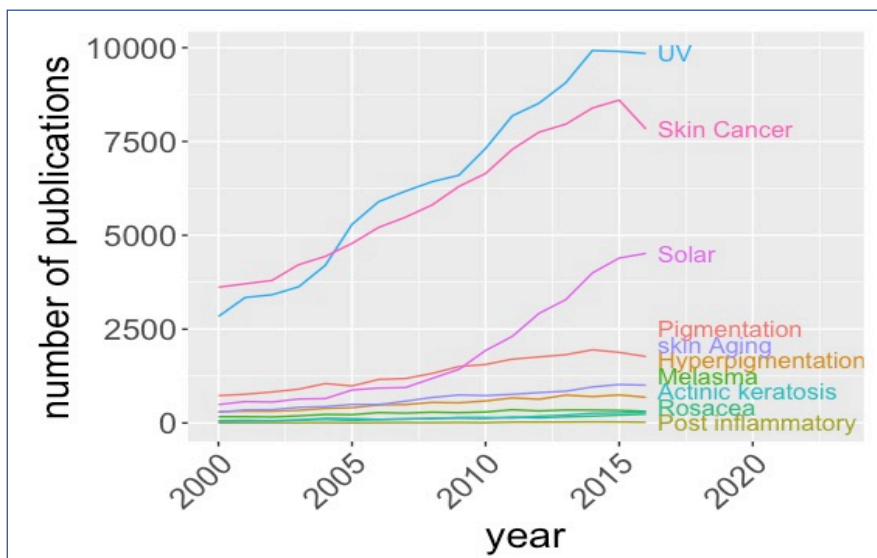


**Figure 2.** Distribution of the number of publications for more common skin conditions.

*3.3. Ingenuity Pathway Analysis (IPA) and Advanced Data Analytics (ADA)*

We used the Ingenuity Pathway Analysis (IPA) [7] to query and obtain the data from genotype and phenotype of skin conditions. The purpose of this manuscript is not to explain the detailed infrastructure of the IPA software, in which reader could find in related articles [7,14]. IPA is a

software application for the analysis, integration, and interpretation of data sourced from different resources [7]. Ingenuity Downstream Effects Analysis predicts cellular functions, disease processes, and other phenotypes impacted by patterns in an analyzed dataset. Upstream Regulator Analysis identifies regulators (transcription factors, cytokines, kinases, etc.). This data is directly linked to the targets in analyzed data and whose activation or inhibition may account for observed changes [7].

With using ADA, we could get most of the data for our analysis. There are several innovative features of IPA Advanced Analytics such as the ability to a. Generate novel hypotheses for mechanisms of action or drug targets b. Prioritize predicted regulatory networks by connection to disease or phenotype of interest and c. Uncover causal relationships relevant to the experimental data [7]. We used Advanced Analytics toward extracting more detail knowledge as shown above for each individual skin condition from Genotype to Phenotype. BioProfiler probes the repository of scientific information which we discussed in Study Cohort to generate molecular profiles of diseases, phenotypes, and biological processes (e.g., apoptosis), listing all the genes and compounds that have been associated with the profiled term.

The power of this tool lies in the intuitive and comprehensive layout of the results, which includes all data extracted from different resources, enabling the user to find, filter, and prioritize genes and compounds based on the research question at hand. The user can focus on molecules of interest, find causally relevant genes, filter for specific genetic evidence or for species, and explore associations with similar diseases or phenotypes [7]. The surfaced data can then be examined further in the context of pathways using all available IPA features, with supporting evidence from published data and powerful analytics. IPA Advanced Analytics goes beyond standard analyses so that we can focus on novel insights about the causes of disease, respond of a specific gene to a specific drug or other phenotypes of interest [7]. We will explain the output of ADA in the result section. In our framework, we used all individual results per condition to find the common factors between conditions.

## 4. Results and Discussion

In this section, we first show and discuss the individual discoveries for each condition based on Advanced Data Analytics from genotype to phenotype. We then discuss our framework and results for common factors and follow that with a discussion on the possible hypotheses based on common factors. Hundreds of genes were extracted for the skin conditions as we explain above, including 32 genes for Hyperpigmentation, 63 genes for Actinic Keratosis, 21 genes for Melasma, 42 genes for Rosacea, 5 genes for Postinflammatory, and 87 genes for Pigmentation. For each individual gene we extracted related information including Symbol (e.g., TYR), Entrez Gene Name (e.g., tyrosinase), Location (e.g., Cytoplasm), Type (e.g., Enzyme), Biomarker Application (e.g., efficacy), drugs (e.g., hydroquinone, azelaic acid), Entrez Gene ID for Human (e.g., 7299), and Entrez Gene ID for Mouse (e.g., 22,173) and Rat (e.g., 30,880). Due to the page limitation, we could not show the complete list all data which we have used for analysis. A complete list could be sent upon request and/or in supplementary materials.

One of the goals of this study was to find the most frequent gene and gene type in all selected skin conditions. Term frequency-inverse document frequency (TF-IDF) was used to find the more common types and genes in all conditions based on all results. The document here is considered as all results. Term frequency (TF) is defined as the number of times a token appears in the report, which serves as an estimate of its importance. Inverse document frequency (IDF) is calculated based on the uniqueness of the token (e.g., type) in the overall corpus (all conditions). The product of the TF and IDF assigns a final, weighted importance to each token (e.g., type). Figure 3 shows the cloud for the most frequent types and genes in all conditions as a word cloud. The larger the size of the word in the cloud, the more frequent the term appeared in the document. The outcomes suggest the hypothesis and questions for further study to researchers including our team for next steps (e.g., what are the most common drugs to consider in all different conditions or why the type of enzymes is more than transporters in all conditions).
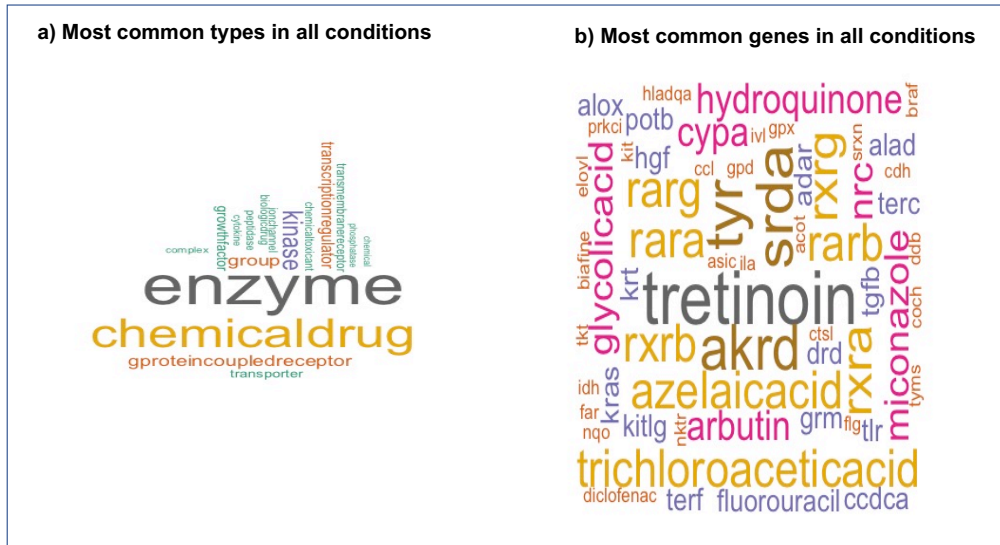
**Figure 3.** Word cloud shows the most common types and genes for all conditions based on TF/IDF.

Next, we sought to understand the phenotypes for each condition, such as relevant drugs. Figure 4 shows the word cloud for drugs that target each condition, and following that we show the four top drugs for each condition in Table 2 and discuss the overlap and responses.
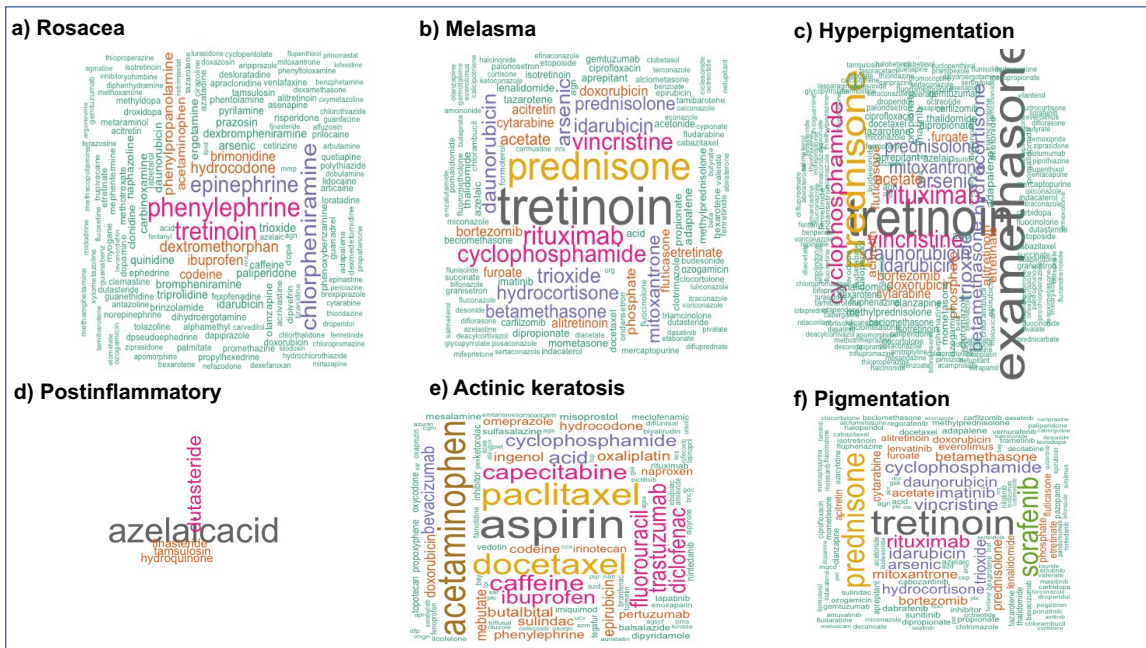


**Figure 4.** Word cloud shows the most common drugs for each conditions.

Our hypothesis is that by finding the overlap in genotype and phenotype between different skin conditions, we can suggest a drug that is used in one condition for treatment in another condition which has similar genes or other common phenotypes.

**Table 2.** Top drugs for different conditions.

| | | | | |
|---|---|---|---|---|
| **Actinic keratosis:** | aspririn | paclitaxel | docetaxel | acetaminophen |
| **Melasma:** | tretinoin | prednisone | rituximab | cyclophosphamide |
| **Post inflammatory:** | azelaic acid | dutasteride | finasteride | tamsulosin |
| **Pigmentation:** | tretinoin | prednisone | sorafenib | rituximab |
| **Hyperpigmentation:** | examethasone | tretinoin | prednisone | cyclophosphamide |
| **Rosacea:** | phenylephrine | tretinoin | chlorpheniramine | epinephrine |

We discuss other the common factors between different conditions. For example Azelaic Acid is a drug used to treat Rosacea and acne (even though it is not in our top drugs list), kills acne bacteria, and reduces production of keratin, reduces inflammation, reduces synthesis of melanin and is used for the treatment of skin pigmentation. Thioredoxin reductase and TYR genes are affected by Azelaic Acid and these two genes are presented as extracted genes in all these conditions. The overlap of these genes across conditions suggests the hypothesis of using this drug for treatment of all conditions and not just for Rosacea and/or Post inflammatory. AKR1D1 (aldo-keto reductase family 1 member D1) and SRD5A2 (steroid 5 alpha-reductase 2) as an enzyme also present in all these conditions, and are located in the cytoplasm, which is responded to the same drug.

Prednisone (drug) is used to decrease the immune system's to reduce symptoms such as swelling and allergic-type reactions [17], which could be used for the inflammatory problem as well. NR3C1 (nuclear receptor subfamily 3 group C member 1) [17] is common in all conditions as a coding protein, and is affected by this drug. We suggest the hypothesis of using this drug with a combination of other drugs for these skin conditions to reduce the itchy skin and inflammation.

What is interesting is that among the conditions the drug is effective against, some of the proteins are identified in all the conditions (for example for tretinoin, RXRA, RXRB and RXRG are identified in all the conditions; and prednisone, NR3C1, but then there are other genes affected by the drug that only show up for one of the conditions (for example KIT and RET identified in pigmentation for proteins affected by prednisone). These proteins that are common between conditions appear to be significant in the pigmentation process. Additionally, these proteins that are linked to the commonly effective drugs, but have not been identified across conditions serve as options to investigate further.

## 5. Conclusions and Future Work

We have shown that Advanced Data Analytics using Big Data is possible using an existing large knowledge base including literature and OMICS related data and more to find the common genes and phenotype such as use in drug repositioning in these conditions. Future works will focus on expanding this application to other phenotypes, as we already have results for the most common molecular signatures, biological process and more. Gathering all kind of phenotype in one paper was not possible due to page limitation and complexity. More comprehensive methods are needed to explore and show more results, as well as aggregate and integrate other data from the rest of our results.

We will extend our methods based on machine learning and text mining works in the biomedical field [18–22] from machine learning methods trained on textual features, to analyses of all results with consideration of other factors such as a biological process to find the common risk factors and outcomes in these conditions.

In summary, we have all the data we need to mine and explore more for future works. As an example, we have all related pathways for each condition as well and we need to use platforms such as Reactom [23], kegg [24], and/or Pathway Tools [25] to analyze them and find the common pathways between different conditions. We are also looking to find out the possibility of Housekeeping genes in selected genes and even the larger list of genes. Publishing this paper gives the community the opportunity to become familiar with our framework to investigate and collaborate, and will open the discussion for multidimensional similarity from genotype to phenotype and emphasize the advantages of large-scale data. Finding common drugs to treat different conditions helps reduce the side effects

of drugs on patients with multiple conditions, as well as reduce negative drug interactions that can worsen the skin condition or even cause a new skin condition. For example, a negative drug interaction that turns Hyperpigmentation into rosacea during the treatment or therapy for Hyperpigmentation caused by Melasma. Prediction and prevention of new skin conditions in patients also can be applied in similar conditions with a similar genotype or phenotype. As an example, knocking out specific biological processes can be used to discover different or similar results under different conditions. This can be used to both prospectively facilitate drug discovery and improve treatment for similar or related conditions, and to correct and augment existing data in OMICS databases. Since this is an informatics study and research so far is one of the possible ways to validate our hypothesis, we are planning to establish a collaboration with a biological lab to test our hypothesis.

## References

1. van Zuuren, E.J. Rosacea. *N. Engl. J. Med.* **2017**, *377*, 1754–1764. [CrossRef] [PubMed]
2. Gallo, R.; Granstein, R.; Kang, S.; Mannis, M.; Steinhoff, M.; Tan, J.; Thiboutot, D. Standard classification and pathophysiology of rosacea: The 2017 update by the National Rosacea Society Expert Committee. *J. Am. Acad. Dermatol.* **2018**, *78*, 148–155. [CrossRef] [PubMed]
3. Wu, I.; Lambert, C.; Lotti, T.M.; Hercogová, J.; Sintim-Damoa, A.; Schwartz, A.R. Melasma. *G. Ital. Dermatol. Venereol.* **2012**, *147*, 413–418. [PubMed]
4. Lee, B.W.; Schwart, R.A. Melasma. *G. Ital. Dermatol. Venereol.* **2017**, *152*, 36–45. [PubMed]
5. Panahiazar, M.; Taslimitehrani, V.; Jadhav, A.; Pathak, J. Empowering personalized medicine with big data and semantic web technology: Promises, challenges, and use cases. In Proceedings of the 2014 IEEE International Conference on Big Data, Washington, DC, USA, 27–30 October 2014.
6. Elston, D.M. Fitzpatrick's color atlas & synopsis of clinical dermatology. *J. Am. Acad. Dermatol.* **2010**, *62*, 166.
7. Ingenuity Pathway Analysis Webpage. Available online: https://www.qiagenbioinformatics.com/products/ingenuity-pathway-analysis/ (accessed on 17 October 2018).
8. Ortonne, J.P.; Passeron, T. Melanin pigmentary disorders: Treatment update. *Dermatol. Clin.* **2005**, *23*, 209–226. [CrossRef] [PubMed]
9. Choi, J.K.; Yu, U.; Kim, S.; Yoo, O.J. Combining multiple microarray studies and modeling interstudy variation. *Bioinformatics* **2003**, *19*, i80–i90. [CrossRef]
10. Yin, L.; Coelho, S.G.; Valencia, J.C.; Ebsen, D.; Mahns, A.; Smuda, C.; Miller, S.A.; Beer, J.Z.; Kolbe, L.; Hearing, V.J. Identification of Genes Expressed in Hyperpigmented Skin Using Meta-Analysis of Microarray Data Sets. *J. Investig. Dermatol.* **2015**, *135*, 2455–2463. [CrossRef] [PubMed]
11. Barrett, T.; Wilhite, S.E.; Ledoux, P.; Evangelista, C.; Kim, I.F.; Tomashevsky, M.; Marshall, K.A.; Phillippy, K.H.; Sherman, P.M.; Holko, M.; et al. NCBI GEO: Archive for functional genomics data sets—Update. *Nucleic Acids Res.* **2012**, *41*, D991–D995. [CrossRef] [PubMed]
12. Inkeles, M.S.; Scumpia, P.O.; Swindell, W.R.; Lopez, D.; Teles, R.M.B.; Graeber, T.G.; Meller, S.; Homey, B.; Elder, J.T.; Gilliet, M.; et al. Comparison of molecular signatures from multiple skin diseases identifies mechanisms of immunopathogenesis. *J. Investig. Dermatol.* **2015**, *135*, 151–159. [CrossRef] [PubMed]
13. National Human Genome Reserach Institute Webpage. Available online: https://www.genome.gov (accessed on 31 July 2018).
14. Kramer, A.; Green, J.; Pollard, J.J.; Tugendreich, S. Causal analysis approaches in Ingenuity Pathway Analysis. *Bioinformatics* **2013**, *30*, 523–530. [CrossRef] [PubMed]

15. IPA. QIAGEN Inc. Available online: https://www.qiagenbioinformatics.com/products/ingenuity-pathway-analysis (accessed on 31 July 2018).

16. Hadley, D.; Pan, J.; El-Sayed, O.; Aljabban, J.; Aljabban, I.; Azad, T.D.; Hadied, M.O.; Raza, S.; Rayikanti, B.A.; Chen, B.; et al. Precision annotation of digital samples in NCBI's gene expression omnibus. *Sci. Data* **2017**, *4*, 170125. [CrossRef] [PubMed]

17. Webmd Webpage. Available online: https://www.webmd.com (accessed on 31 July 2018).

18. Panahiazar, M.; Dumontier, M.; Gevaert, O. Context Aware Recommendation Engine for Metadata Submission. In Proceedings of the First International Workshop on Capturing Scientific Knowledge, Palisades, NY, USA, 7 October 2015; pp. 3–7.

19. Posch, L.; Panahiazar, M.; Dumontier, M.; Gevaert, O. Predicting structured metadata from unstructured metadata. *Database* **2016**, baw080. [CrossRef] [PubMed]

20. Taslimitehrani, V.; Dong, G.; Pereira, N.L.; Panahiazar, M.; Pathak, J. Developing EHR-driven heart failure risk prediction models using CPXR (Log) with the probabilistic loss function. *J. Biomed. Inform.* **2016**, *60*, 260–269. [CrossRef] [PubMed]

21. Panahiazar, M.; Taslimitehrani, V.; Pereira, N.; Pathak, J. Using EHRs and Machine Learning for Heart Failure Survival Analysis. *MedInfo* **2015**, *2015*, 40–44.

22. Panahiazar, M.; Dumontier, M.; Gevaert, O. Predicting biomedical metadata in CEDAR: A study of Gene Expression Omnibus (GEO). *J. Biomed. Inform.* **2017**, *72*, 132–139. [CrossRef] [PubMed]

23. Reactome Webpage. Available online: https://reactome.org (accessed on 31 July 2018).

24. Kegg Webpage. Available online: https://www.genome.jp/kegg/pathway.html (accessed on 31 July 2018).

25. Pathway Tools Webpage. Available online: http://bioinformatics.ai.sri.com/ptools/ (accessed on 31 July 2018).