The morbid cutaneous anatomy of the human genome revealed by a bioinformatic approach

Lilia Romdhane, Heni Bouhamed, Kais Ghedira, Cherif Ben Hamda, Amel Louhichi, Haifa Jmel, Safa Romdhane, Chérine Charfeddine, Mourad Mokni, Sonia Abdelhak, et al.

Title: The morbid cutaneous anatomy of the Human genome revealed by a bioinformatic approach

Authors:

Lilia Romdhane[1,2], Heni Bouhamed[3], Kais Ghedira[4], Cherif Ben Hamda[4], Amel Louhichi[3], Haifa Jmel[1], Safa Romdhane[1], Chérine Charfeddine[1,5], Mourad Mokni[6,7], Sonia Abdelhak[1], Ahmed Rebai[3]

Authors' affiliations:

1 Biomedical Genomics and Oncogenetics Laboratory LR11IPT05, LR11IPT16, Institut Pasteur de Tunis, Tunis, Université Tunis El Manar, Tunis, Tunisia.
2 Department of Biology, Faculty of Science of Bizerte, Jarzouna, Université Tunis Carthage, Tunis, Tunisia
3 Molecular and Cellular Screening Process Laboratory, Sfax, Centre of Biotechnology of Sfax, Sfax, Tunisia
4 Laboratory of Bioinformatics, Biomathematics and Biostatistics (LR16IPT09), Institut Pasteur de Tunis, Tunis, Université Tunis El Manar, Tunis, Tunisia.
5 High Institut of Biotechnology of Sidi Thabet, University of Manouba, BiotechPole of Sidi Thabet, Ariana, Tunisia.
6 Department of Dermatology, CHU La Rabta Tunis, Tunis, Tunisia.
7 Research Unit on Hereditary Keratinizaton Disorders UR12SP07, CHU La Rabta, Tunis, Tunisia.

Corresponding author's contact:

Lilia Romdhane
Biomedical Genomics and Oncogenetics Laboratory
Institut Pasteur de Tunis, BP 74, 13 Place Pasteur 1002 Tunis Belvédère Tunisia,
Department of Biology, Faculty of Science of Bizerte, Jarzouna, Université Tunis Carthage, Tunis, Tunisia
Tel: 216 71 849 110
Fax: 216 71 791 833, e-mail : lilia.romdhane@gmail.com ; lilia.romdhane@fsb.rnu.tn

Abstract:

Computational approaches have been developed to prioritize candidate genes in disease gene identification. They are based on different pieces of evidences associating each gene with the given disease. In this study, 648 genes underlying genodermatosis has been compared to 1808 genes involved in other genetic diseases using a bioinformatic approach. These genes were studied at the structural, evolutionary and functional levels. Results show that genes underlying genodermatosis present longer CDS and have more exons. Significant differences were observed in nucleotide motif and amino-acid compositions. Evolutionary conservation analysis revealed that genodermatoses genes have less paralogs, more orthologs in Mouse and Dog and are less conserved. Functional analysis revealed that genodermatosis genes involved in immune system and skin layers. The Bayesian network model returned a rate of good classification of around 80%. This computational approach could help investigators working in the field of dermatology by prioritizing positional candidate genes for mutation screening.

Keywords: Bayesian network, bioinformatics, classification, genetic diseases, genodermatosis, prioritization methods, skin

1. Introduction:

Genetic diseases represent a large burden on public health, especially in developing countries often featured by high rates of consanguinity, where more than 500 pathologies have been identified, namely in Tunisia (Mezzi N et al., unpublished). For many decades, linkage analysis is a successful procedure to associate diseases with genomic regions. This approach in combination with homozygosity mapping and linkage disequilibrium, has led to the identification of many genes involved in the molecular aetiology of genetic diseases mainly from North African families [1-6]. With the explosion of the molecular techniques of orphan genetic disease investigation, genomic and functional information around morbid genes have been accumulated and stored in specific databases. The genes known to be involved in human diseases as well as the underlying mutations are collected in centralized databases such as OMIM [7] and Human Gene Mutation Database (HGMD) [8].

The availability of the Human genome and other genome organisms', as well as functional data and large genetic databases provided the opportunity to the identification of intrinsic features of morbid genes traditionally by computational methods. Global analysis revealed distinct patterns between disease and non-disease genes [9]. Disease genes as well as their product and 3'UTR are longer and have more exons, a feature that could be correlated with the increase probability of deleterious mutation accumulation [9]. Furthermore, disease proteins seem to have signal peptides and preferentially enriched in alanine and glycine [10]. A study of the phylogenetic extent showed that morbid genes tend to be more conserved among species [11]. On the other hand, disease phenotype expression was also attributed to the fact that morbid genes have less paralogs [11]. Therefore, the altered protein function could not be rescued in the absence of a wild type paralog protein [11] .

Functional features of disease genes revealed also distinct patterns. They encode essentially for enzymes, transporters, transcription regulators, structural molecules, and protein function

77 modulators [12, 13]. With respect to biological processes, disease genes are involved in
78 metabolism, stress response, developmental process, cell communication and cell cycle [13].
79 Positional regions deduced from linkage analysis are often large containing hundreds of genes
80 making experimental methods for disease gene identification challenging. Computational
81 approaches have been developed as complementary methods to prioritize candidate genes [14-
82 16]. The proof of principle of such tools is "guilty by association": given a set of genes known
83 to be associated with a disease, computational tools for candidate gene prioritization rank
84 position genes according to their "similarity" profile to a set of reference genes. These
85 methods could integrate different data type ranging from sequence, phylogenic, functional and
86 interaction data and differ by their input and output as well as the classification algorithm
87 used to rank the candidate genes with diverse rates of accuracy [16, 17]. Recently, whole-
88 exome sequencing appears to be an efficient genomic technique to unravel the molecular
89 aetiology of Mendelian disorders [18]. Nevertheless, hundreds of *de novo* mutations can be
90 generated to be screened for a query disease [18]. The question that remains is how to infer
91 true causative genes from candidate genes that harbour such mutations. Thanks to
92 bioinformatics, appropriate tools could be used to help detecting genetic alterations and infer
93 their association with human diseases [19].
94 In general, these methods have considered that all disease genes are homogeneous thus
95 neglecting disease type specificities. In this context, we investigated a group of genes
96 implicated in a pathological group which is genetic disorders of the skin. Genetic diseases of
97 the skin or genodermatoses, such as *epidermolysis bullosa*, *Xeroderma pigmentosum* and
98 *ichthyosis,* are disabling and life-threatening [20]. In population where consanguinity is high,
99 genodermatoses are prevalent and most often show atypical phenotypes and could be
100 associated with other diseases thus hampering their diagnoses [20-24]. This prompted us to
101 investigated the features of genes involved in genodermatoses by comparing them to another

set of genes involved in other diseases. The analysis of the properties of genes and their products revealed interesting findings that allowed obtaining a global overview on genodermatosis genes at multiple levels.

2. Methods:

2.1 Morbid gene group delineation:

Disease genes were assigned into two groups. According to Feramisco et al., a genodermatosis is genetic disease with a primary cutaneous phenotype [25]. In order to identify genodermatosis genes, we queried the clinical synopsis section of the OMIM database using the key-word "skin" (Last accessed September 2015). Results were limited to entries with known molecular aetiology. This step led to the identification of genodermatoses with at least one symptom at the skin level. Genetic diseases except genodermatoses were identified using the same procedure but keeping blank the OMIM clinical synopsis search field [7]. Returned entries were manually curated and responsible genes were assigned to the corresponding groups. Furthermore, two other published databases for genetic diseases of the skin were also queried [25, 26]. Querying the "Entrez Gene" database defines a total of 648 genes (Skin disease group) and 1,808 genes underlying genetic diseases excepting genodermatoses referred to the "Other" disease group (Last accessed September 2015).

2.2 Structural analysis:

Genomic and functional sequences as well as structural features (Additional File 1) were downloaded from the "RefSeq" database (RefSeq Homo sapiens Build 37.2) [27] . DNA and protein motif frequencies were estimated using Wordcount and GeeCee softwares from the EMBOSS suite (http://emboss.sourceforge.net/). Protein structural features were downloaded using Ensembl Biomart [28]. Furthermore, we analysed coding sequences using the Z-curve, which is a bioinformatics algorithm for genome analysis describing a given DNA sequence at the three-dimensional level [29].

2.3 Evolutionary conservation study:

We also studied the phylogenetic extent (that is the extent to which a gene is conserved back through evolution based on homologs in other species) of each morbid gene examined. Therefore, orthologs in 6 Vertebrate species (*Pan troglotydes*, *Bos taurus*, *Canis lupus*, *Rattus norvegicus*, *Mus musculus* and *Gallus gallus*) were retrieved from the NCBI HomoloGene database [30]. For each ortholog, the percentage of identity at both nucleic and amino-acid levels were noted. To estimate the degree of paralogy, a BlastP (https://blast.ncbi.nlm.nih.gov/Blast.cgi) search was performed for each protein in the group against the human proteome. The best BlastP hit found in the human genome was kept for an "E-value" cuttof of 10^{-30}. For each paralog, an identity percentage is returned as well as an average of this measure for alternatively spliced isoforms. For each gene, the number of putative paralogs found is calculated in each disease gene group.

2.4 Calculation of nucleotide substitution rates at DNA level:

We estimated dN (nonsynonymous substitution rate), dS (synonymous substitution rate) and ω =dN/dS for each gene with the Codeml software of the PAML package (**Additional File 1**) [31]. Codeml uses the codon as the evolutionary unit in DNA alignments of coding regions and an explicit model incorporating the genetic code to estimate the amount of natural selection for or against protein-level changes. Total substitution rate (d) was estimated from the HomoloGene database. To test whether a difference exists between the disease groups, genes with no orthologs as well as those with dS= 0 were excluded.

To identify amino acids under positive selection in skin and non-skin genes, we used the Codeml program after aligning nucleic sequences of orthologs genes with MAFFT [32] and generating phylogenetic trees. The selective pressure is estimated with the ω ratio calculated with PAML. When $\omega > 1$, this is indicative of positive Darwinian selection [33]. A likelihood

ratio test (LRT) of positive selection comparing the nearly neutral and the positive selection models, as well as P-value were computed. Naïve empirical Bayes [33,34] and the Bayes empirical Bayes [35] were used for calculating the posterior probabilities for site classes and to identify sites under positive selection if the LRT and P-value are significant. Amino acids sites with posterior probabilities of Bayes Empirical Bayes analyses superior to 95 or 99% were considered as positively selected.

2.5 Functional analysis:

Enrichment in specific GO terms was assessed using FatiGO on the Babelomics server for gene expression, genome variation and functional analysis (http://babelomics.bioinfo.cipf.es/) [36], StringDB [37] and ClueGO cytoscape plugin [38]. Statistical significance in GO terms occurrence difference is considered using a Fisher's exact test and/or false discovery rate (FDR<=0.05).

2.6 Programming and statistic tests:

NCBI Databases were queried using in-house "Python" and "Biopython" scripts integrating "E-Utilities" [39] (www.r-project.org; www.python.org; www.biopython.org). Statistical analyses were performed using the "R" environment at significance of 5 %. Plots were carried out using the "ggplot2" R package [40]. To assess whether the distribution of a statistic feature differed between two groups of genes, we used a Kolmogorov-Smirnov test. To test whether the studied features differed between the two disease gene groups, a Wilcoxon test was performed.

2.7 Bayesian network

We proceeded to automatic learning from data of a Bayesian network that served later as a gene classifier (involved or not in genodermatosis). This classifier learns from training data the conditional probability of each attribute given the class label. The attributes used to build the model are all the studied features except the functional ones. We proceeded first to the

discretization of the continuous variables using the Colot algorithm sed as the model attributes [41]. For the construction and manipulation of the Bayesian network, we used the Bayesian network toolbox (BNT) [42] (http://www.ai.mit.edu/~murphyk/Software/BNT/bnt.html) on the Matlab software version 2010 (www.mathworks.com). We employed the « Augmented Naïve Bayes » [43] for the structure learning and the « Clique-tree propagation algorithm » [44] for the inference. During the learning phase, we used 548 genes (from the initial 648) involved in genodermatosis and 1,708 genes (from the initial 1,808) involved in other genetic diseases thus forming the learning set. The remaining 200 genes, the test set, were used during the test phase of the constructed classifier**.**

We used a ROC (Receiver Operating Curve) curve to evaluate the overall accuracy and predictive value of the method. The ROC analysis is a standard approach to evaluate the sensitivity and specificity of prediction methods. The ROC curve is obtained by plotting the true positive rate against the false positive rate for different values of the cut-off probability score. The 45° diagonal of the ROC space represents a random guess situation. The ROC curve was plotted using the ROCR package [45].

3. Results:

3.1 Structural features of genes and products:

3.1.1 Length features:

Length feature analysis reveals that significant differences exist between the two morbid gene sets concerning CDS and protein length as well as exon counts. CDS of genes of genodermatoses seem to be longer with more exons (**Figure 1**) and code for longer proteins than the other disease gene set (**Table 1**). Furthermore, exon count distribution is different (KS p-value = $4\ 10^{-3}$). Genes with exon count less than 20 are less prevalent among the genodermatosis gene group while those with exon count between 30 and 70 are more

prevalent among the genodermatosis group of gene (**Figure 1**). No difference has been noticed regarding remaining studied features (**Table 1**).

3.1.2 Mono, di and tri nucleotide motif frequencies:

Analysis of motif composition of CDS shows an enrichment of some motifs. Coding regions of genodermatosis genes are more enriched in **Guanine** (p-value = **0.03**). Differences in the composition of dinucleotide motifs as well as their distributions were revealed. Skin disease CDS are less enriched in CA, GG and TC motifs but more in TG. Trinucleotide motif frequency is significantly different for 11 motifs as highlighted by the Wilcoxon test but 8 of them show a different distribution (**Additional File 2**). Codon translation suggests that genodermatosis proteins could be composed essentially by hydrophobic amino-acids**.**

3.1.3 Z curve of CDS:

We aimed to study CDS composition according to the 3 components of the Z curve. Significant differences have been highlighted for one variable calculated from mononucleotide frequency, for 3 variables calculated from dinucleotide motif frequencies and for 13 variables estimated from trinucleotide motif frequencies. In order to present a global description of the 3 properties of these DNA sequences, we took the mean of each component (**Additional File 2**). According to the X component, CDS in both disease gene sets show a positive mean value meaning that all these CDS are more enriched in purine. Genodermatosis CDS are more enriched in purine (Wilcoxon p-value $< 10^{-6}$). Mean Y component is also positive in the 2 gene sets. This illustrates that CDS in both groups are more enriched in amine bases than in ketones. Nevertheless, genodermatoses CDS are less enriched in amine bases than the other gene group (Wilcoxon p-value = $1.13\ 10^{-6}$). Distributions of these 2 variables are also different. Similar frequencies and distributions of bases with weak/strong bounds among the CDS and corresponding to the Z component have been noticed.

3.1.4 Primary and secondary structure of proteins:

Proteins responsible for genodermatoses are more enriched in cysteine and glutamine (**Table 2**). Nevertheless, they are less enriched in isoleucine, methionine and phenylalanine. The distribution of these frequencies is also different in the two groups except for glutamine. In addition, we studied the prevalence of some protein structure such as the presence of signal peptide, transmembrane domains as well as alpha helices. Only 23.9 % of genodermatosis proteins versus 22.5 % of proteins of the second group have a secondary structure of helix type. This difference is not significant (Chi2 test p-value = 0.51). Similar frequencies of signal peptide (34.5% in genodermatoses vs 30.8% in the second group, Chi2 p-value = 0.36) and transmembrane domain (30.5 % in genodermatoses vs 32.6 % in the second group, p-value= 0.35) have also been noticed. Furthermore, genodermatosis proteins seem to be enriched in immunoglobulin C1-set domain (FDR=0.0048) while the second group of genes was enriched in Homeobox domain (FDR=0.00569) (**Additional File 3**).

3.2 Evolutionary conservation features:

3.2.1 Orthology:

Genes underlying skin genetic diseases have more orthologs than genes in the second morbid group especially in Dog (91 % vs 87 %, Chi2 p-value= $\mathbf{0.7\ 10^{-2}}$) and Mouse (97.22 % vs 94.8 %, Chi-square p-value= $\mathbf{1.52\ 10^{-2}}$). No significant difference has been reported in the other species **(Additional File 4)**. Nevertheless, orthologs of skin genetic diseases seem to be less conserved at the nucleic level in 3 species **(Table 3)**. At the protein level, chicken orthologs are less conserved than other species in the two gene groups. Distribution of conservation proportion of ortholog sequences is only different among the mouse orthologs at both DNA and protein levels (KS p-value= $1.8\ 10^{-2}$ and KS p-value =$2.2\ 10^{-2}$, respectively) (**Table 3**, Figure 2).

3.2.2 Paralogy:

Human protein alignment against the proteome using BlastP revealed that paralogs are less prevalent among genes of genodermatoses than do genes of the second group (63.88 % vs 69.85 %, Chi2 p-value = 5.9 10^{-3}). Nevertheless, genodermatosis proteins have **more** putative paralogs per gene, the median number of paralogs per gene being 5 versus 4 (Wilcox p-value = 1.15 10^{-3}). The paralog number per gene distribution is also significantly different (KS p-value = 3.2 10^{-2}). Thus, paralog count distribution shows that duplication is frequent among the non-skin disease genes having putative orthologs under 7 paralogs per gene and less frequent for those having more than 31 paralogs per gene. Furthermore, skin disease gene paralogs are less conserved (42.44 % vs 43 %, Wilcoxon p-value =3 10^{-2}). The conservation distribution is also significantly different (KS p-value = 4 10^{-2}). More conserved paralogs with values over 60 % are more represented in the non-skin disease gene group.

3.2.3 Selective pressure acting on morbid genes at the DNA level:

The degree of conservation at the DNA level of the 2 sets of morbid genes was analysed in order to shed light on the selective pressures acting on them. We examined total substitution (d), non-synonymous (dN), synonymous (dS) rates, as well as dN/dS ratio, from coding sequences of orthologs of human in 6 Vertebrate species (**Additional File 4**). The total substitution rate (d) is statistically higher only among Human-Chicken (**Wilcox** test p-values =10^{-2}, respectively), Human-Dog (**Wilcox** p-values = 3 10^{-2}) and Human-Mouse (**Wilcox** test p-values = 2 10^{-2}) orthologs in the skin disease gene group. The distribution is significantly different only for Human-Chiken and Human-Mouse orthologs (KS test p-values = **2.6 10^{-2}** and KS test p-values= **3.8 10^{-2}** respectively). The dN, which is indicative of the selective pressure acting on sites that involve an amino-acid change, is lower among the skin disease genes of Human-Primate orthologs (0.002 vs 0.003, MW p-value= 4.7 10^{-2}). Furthermore, the dN distribution is significantly different among all the studied orthologs except those of Cow

and Primate. Skin disease genes are thus evolving more slowly, in general, than the other disease genes, this indicates that they are subject to strong selective constraints.

The dS is more reflective of the background mutation level. Globally, skin disease genes show less dS values among the orthologs of the studied species (**Additional File 4**). The distribution is also statistically different (**Additional File 4**). This difference could be assigned to varying mutation rates in the genome. The analysis of dN/dS shows that the ratio is statistically lower only among the Human-Primate orthologs in the skin disease gene group (**Additional File 4)**. The distribution is also significantly different (**Additional File 4**). Similar dN/dS values were found among the orthologs of the remaining species in the two groups although the distribution of dN/dS values is different. Moreover, the investigation of the amino acid under positive selection between skin and non skin genes showed that the most enriched amino acids under positive selection within the skin protein coding genes compared to the non-skin coding genes includes W: Tryptophan, L: Leucine, S: Serine, P: Proline, A: Alanine and M: Methionine (**Figure 3**).

3.3 Function and process of morbid genes:

We have used GO terms to characterize protein function and elucidate trends in our morbid protein dataset (**Additional File 5**, Figure 4). Protein binding activity is the most significant molecular function that is over-represented in the skin genetic disease genes (FDR=4.43 10^{-41}) (**Additional File 5).** In terms of GO biological process, as expected, genodermatosis genes are highly involved in skin, tissue, ectoderm, immune system and epidermis development, protein localization, melanocyte differentiation, pigmentation, response to UV and regulation of inflammatory response (Figure 4). Genodermatoses are caused by mutations preferentially localized in the peroxysome, lytic vacuole, vacuole, intermediate filament cytoskeleton, cornified envelope and respiratory chain (**Additional File 5**).

3.4 Bayesian network

We developed and tested an automatic learning approach from data using a Bayesian network as a gene classifier allowing to classify genes as involved or not in genodermatosis based on gene sequence features. We used 90 % of the data for model learning and 10% for validation. To validate the model, we checked how well the model is able to rank genes using the remaining data fraction that has not been used before (10 %, the test set). Each gene of the test set was assigned a probability score of being involved in genodermatoses (**Additional File 6**). On average, the constructed Bayesian network ranked genes with a rate of 80 % of good classification. Taking a cut-off probability score of 0.5, we obtained 59 % sensitivity, 97 % specificity and 31 % accuracy. The accuracy of the method was evaluated using a ROC analysis (**Figure 5**).

4. Discussion:

As previous studies focused on comparing morbid genes and those not underlying genetic diseases [9,46,47], few works aiming to analyse features of genes and their products responsible for a particular pathogenic group such as cancers, deafness, cardiovascular, diseases of the eye or neuropsychiatric diseases have been performed [48-53]. In this context, we focused on establishing properties of skin diseases genes. Our study revealed that the genes involved in genodermatoses differ from the genes involved in other groups of genetic disorders.

Skin disease genes coding sequences and their proteins are longer than the remaining of the morbid genes with higher number of exons. A similar pattern has been noted in comparison to protein involved in cancer and hereditary diseases when comparing with non-disease genes [9-11,48,54-56]. Taking cancer genes as a set of morbid genes, length pattern differences have been revealed between mutated and translocated cancer genes showing that those muted have longer CDS than those translocated [48]. In genodermatoses, as in the other hereditary diseases, a longer CDS is more susceptible to accumulate point mutations as a consequence of

its length and therefore more likely to produce a dysfunctional protein. Therefore, this length pattern among morbid genes of distinct groups of diseases could be assigned to a different mutation process of these two groups of genes. This could be verified by the study of the variant pattern accumulation in the genodermatoses genomic sequences as well as their corresponding transcripts and CDS. In addition to the correlation gene length - high mutation accumulation probability, it was also suggested that gene overlap and multiple amino acid runs are also morbid disease gene features [57]. Similarly, motifs enriched in CDS of genodermatoses could be associated with mutational hotspots that increase the probability for a mutation to occur. With respect to over-representations of some amino acids in corresponding proteins, their implications in the morbid phenotype of the skin are not clear. However, it was established that protein regions lacking secondary or tertiary structures, known as intrinsically disordered region, could be involved in protein-protein interactions, regulation and signal transduction explaining their crucial role in disease development including cancers and cardiovascular diseases [51,58]. As at least 4 genodermatosis proteins are shown be known as intrinsically disordered charged protein involved in human diseases, further studies on protein intrinsic disorders in this kind of diseases are needed and could help gain insights on amino acid residues occurrence involved in genodermatoses [59].

Our results revealed also that genodermatoses genes are more enriched in Cysteine and Glutamine, two negatively charged amino-acids, and less enriched in Isoleucine, Methionine and Phenylalanine compared to the set of genes from other disorders. Moreover, this study showed that Tryptophan, Leucine, Serine, Proline, Alanine and Methionine are the most enriched amino acids under positive selection in skin proteins, thus suggesting conserved functional and structural properties. Indeed, it has been shown that bovine, rat and human skin keratins are all rich in the amino acids glycine, serine, leucine and glutamic acid [60]. Glutamine plays a key role in regulating the acid-base balance within the body and firms the

skin by contributing to its elasticity [61]. In addition, the Glutamine plays an important role in cross-linking of cornified envelop proteins, such as involucrin, to the lysine residue of the head domain of keratins molecules [62]. Intensive concentration of sulfur in the keratogeneous zone through the amino acids cysteine and methionine residues can be involved to form disulphide bonds, sulphur-sulphur covalent linkages that confer more folding and stability to proteins in the processes of hard cornification in mammalian skin modifications, such as hairs and nails [63]. In mammals, a high content of cysteine in the head and tail domains, as well as in their rod domain, is characteristic of the α-keratins [64, 65]. The intrinsic features of the keratin filaments and of the keratin filament-associated proteins that are cross-linked through dissulfide bonds influence the mechanical properties of keratinocytes and corneocytes [66]. In addition, dissulfide bonds cross-link the keratins and keratin filaments to the proteins of the cellular envelope, such as involucrin, locricrin and periplakin [67]. All of these intracellular factors contribute to the mechanical properties of each keratinized or keratinized and cornified cells.

In addition to cysteine, the acidic keratins in the follicular epidermis are characterized by a large number of proline residues in their head and tail domains [68]. The H1 subdomain of K1 of basic type II keratins contains many threonines and prolines [69] and may play an important role in the correct parallel alignment of keratin polypeptide chains during the formation of the coiled-coil heterodimer [70]. In addition, it was showed that Proline play a key role in protein binding and is the main component of the collagen, a high-tensile fiber found in connective tissue such as tendons and skin [71,72]. In all types of collagens, a sequence repeat occurs formed by the XaaYaaGly motif. The amino acids in Xaa and Yaa positions are proline (28%) and hydroxyproline (38%), with the ProHypGly being the most common triplet (10.5 %) [72]. Proper posttranslational modifications are critical for ultimate triple-helix formation of mature collagen and keratin assembly. These posttranslational

modifications include hydroxylation of proline residues in the procollagen. Consequently, the lack of a stable triple-helical collagen structure compromises the integrity of the skin, mucous membranes, blood vessels and bones [73]. In keratins, the serine residue of a head domain can become negatively charged by phosphorylation, resulting in the disassembly of the keratin filament. Therefore, this amino acid is instrumental in the disassembly and reassembly of keratin filaments [74].

Previous study reports the importance of Tryptophan and Methionine residues within polypeptides stabilizing structures/domains [75,76]. Tryptophan also plays a key role in promoting protein-protein, protein-peptide, or protein-biomolecule structural hydrophobic interactions [75] and playing a significant anchoring role and protein folding [75]. However, Isoleucine, Methionine and Phenylalanine are rarely directly involved in protein functions, and can be involved in substrate recognition, binding/recognition of hydrophobic ligands and stacking interactions with other aromatic side chains, respectively.

Furthermore, our results showed that genodermatosis proteins were enriched in immunoglobulin C1-set domain which are Ig-like domains involved in a variety of functions, including cell-cell recognition, cell-surface receptors, muscle structure and the immune system [77]. This is in concordance with our results revealing that genodermatosis genes are enriched in Immune system development and inflammatory response control.

Persistence of human genetic diseases in the population could be attributed to a combination of many factors including mutation, genetic drift and natural selection. Many studies have reported that morbid genes and their products seem to be more conserved mainly because they have been exposed to strong evolutionary constraints and therefore have not the opportunity to accumulate multiple variations [48,56,78]. Using the dN/dS ratio was a main approach to evaluate how selective pressure acts on coding gene sequences. Comparison results between disease and non-disease genes have been conflicting. Some studies found that morbid genes

have lower dN/dS values [10,79], others found higher values [56,78] and one found similar pattern [80]. The most plausible explanations could be that genetic disease genes have been considered as a homogeneous dataset thus neglecting that they could in fact be categorized into house-keeping and non-essential genes exhibiting different patterns of substitution rates [80]. Similarly, different values of substitution rate accumulations were found according to the transmission mode [48,82]. Correspondingly, our study revealed distinct patterns of substitution rates as well as dN/dS ratio in pathogenic genes according to tissue/organ type. Skin disease genes seem to evolve at slower rates by accumulating less non-synonymous variations and thus are under selective pressure constraints. In opposition, we have observed that frequent skin genes have greater dN/dS values indicating that less selective pressure have been able to change among species and to tolerate variations without leading to the expression of a morbid phenotype. In addition, we have also found that extremely conserved proteins are less frequently found to be involved in genodermatoses. One likely explanation is that variations in this group of extremely conserved genes are mostly lethal.

The functional analysis of the studied genodermatosis group of genes through GO enrichment, showed that these later are mainly involved in epithelium, organ, tissue, skin, cell development but also response to stress, immune response, pigmentation and the control of inflammatory response. These genes are mainly acting on peroxisome, vacuole, lysosome and cellular organelles. The most significant molecular function that is over-represented in these skin genetic disease genes is the binding activity. Function annotation pattern through GO terms could correlate with the conservation pattern. As assumed earlier that genes in skin diseases are under selective constraints, this seems to be obvious as the skin is the first natural barrier protection from the external environment. Skin mechanism protection acts mainly through its pigmentation, the melanin, and its variation influences both pigmentation and skin cancer risks [83]. Melanin provides a crucial filter for solar UV radiation and it is mainly

424 effective against the harmful effects of the shorter wavelengths of the electromagnetic
425 radiation that is the most damaging to DNA and proteins [84]. Sunlight is not only damaging
426 to human skin, but also to some essential nutrients in particular folate that is required for
427 DNA synthesis and repair. Complications during pregnancy, as spina bifida and anencephalus
428 can rise as a consequence of folate deficiency [85,86]. Prenatal and postnatal mortality in
429 some populations was caused by folate deficiency before preventive supplementation
430 introduction [87]. Moreover, it was suggested that photo degradation of folates could play a
431 key role in the increased tendency of populations of low melanin pigmentation living in areas
432 of high UV exposure to develop skin cancers [88].
433 In the present work, as we have focused only on the analysis of the sequences' features and
434 functions of genes involved in genodermatoses, limitations have to be highlighted. Although
435 OMIM is the most comprehensive database relating on human genetic diseases, its phenotypic
436 classification is incomplete and sometimes outdated. Therefore, a more reliable annotation
437 system of diseases should be used in order to classify each morbid gene in the correct
438 pathogenic group. Furthermore, it would be also interesting to explore features of
439 genodermatoses according to whether they manifest only cutaneous symptoms or in extra
440 dermatological manifestations. Differences in sequence patterns of morbid genes allowed us
441 to build a Bayesian network model able to differentiate these genes according to disease
442 category with a relatively good classification rate. In order to increase the accuracy of our
443 model, it would be interesting to integrate additional data type when available, for instance,
444 functional annotation, expression pattern or their genomic distribution. It will be also
445 interesting to explore the involvement of the proteins of this kind of disease in the context of
446 protein networks and gene regulatory networks to get further insights on genodermatosis
447 pathogenicity and prediction of genes involved in skin genetic diseases. Such models
448 integrating multiple data types were successfully used to assist disease gene discovery from a

candidate genomic region identified using a genetic linkage approach [89]. In the next-generation sequencing era, high throughput sequencing like a whole exome and whole genome sequence is becoming more and more accessible and constitutes a pivotal methodology for rapid and cost-effective detection of pathogenic variations in Mendelian disorders. Determining the causative mutation from a pool of variations that could not be involved in the disease aetiology is the principal challenge of such approach. Filtering strategies based on variation frequencies and functional annotations could be combined with computational tools of disease gene prioritization revealed to be an effective solution to reduce variation number on candidate genes and elucidation of the molecular aetiology of the disease in a single family [90]. Intersection of the results of variant analysis with an approach of prioritization of candidate genes according to their feature and relevance could be a key factor in exome prioritization algorithms [91]. Such approach will help elucidate aytpical phenotypic expression of rare genodermatoses unsolved by exome sequencing (Messaoud O, Personal communication)

5. Conclusion:

In summary, we have investigated the structural, evolutionary and functional properties of a group of genes known, when mutated, to be causative of genodermatoses compared to another group of genes not related to genetic diseases of the skin. We have detected clear trends in this group of genes in terms of biological process or cellular component as well as sequence and evolution properties. Based on these features, we have developed a Bayesian network classification model with which human genes have been scored for their likelihood of involvement in genetic diseases of the skin. These results could be useful for the future development of computational tools that allow prioritization of novel candidate genes for genodermatoses. Moreover, the *in silico* approach in this study could be expanded to investigate other disease-related genes. We believe that the division of disease genes by the

affected organ or tissue will enhance both understanding of the disease process, prediction and prioritization of candidate disease genes in the future.

ABBREVIATIONS:

CDS: coding sequence; **dN:** nonsynonymous substitution rate; **dS:** synonymous substitution rate; **d:** substitution rate; **GO:** gene ontology; **KS:** Kolmogorov-Smirnov; **OMIM:** Online Mendelian Inheritance in Man; **UTR:** untranslated region

DECLARATIONS:

Ethics approval and consent to participate:

This study did not involve human participants, human data, or human tissue requiring the need to include a statement on ethics approval and consent.

Consent to publish:

Not applicable.

Funding:

This work was supported by the Tunisian Ministry of Public Health, the Ministry of Higher Education and Scientific Research (LR11IPT05) and by the E.C. Grant agreement No. 295097 for FP7 project Genomedika (www.genomedika.org).

Acknowledgements:

The authors are also very grateful to Professor Fredj Tekaia for his help with comparative analysis and his careful and meticulous reading of the paper. The authors would also like to kindly acknowledge Dr Catherine Lethondal for her help with python programming.

Availability of data and materials:

The datasets generated, used and/or analysed during the current study are not currently publicly available as a database will be developed but are available from the corresponding author on reasonable request.

Competing interests:

The authors declare that they have no competing interests.

Authors' contributions:

LR: collected and generated data, performed bioinformatic and statistical analyses and wrote the manuscript

KG and **CBH**: performed functional data analysis, figures generation and contributed in the manuscript writing and revision

HB and **AL:** performed statistical and classification analysis

HJ and **SR:** managed the data

CC and **MM**: revised the manuscript

SA and **AR**: designed the study and revised the manuscript

REFERENCES:

1. Basit S, Lee K, Habib R, Chen L, Umm e K, Santos-Cortez RL, Azeem Z, Andrade P, Ansar M, Ahmad W *et al*: **DFNB89, a novel autosomal recessive nonsyndromic hearing impairment locus on chromosome 16q21-q23.2**. *Hum Genet* 2011, **129**(4):379-385.
2. Ben Hamida C, Doerflinger N, Belal S, Linder C, Reutenauer L, Dib C, Gyapay G, Vignal A, Le Paslier D, Cohen D *et al*: **Localization of Friedreich ataxia phenotype with selective vitamin E deficiency to chromosome 8q by homozygosity mapping**. *Nat Genet* 1993, **5**(2):195-200.
3. Gusella JF, Wexler NS, Conneally PM, Naylor SL, Anderson MA, Tanzi RE, Watkins PC, Ottina K, Wallace MR, Sakaguchi AY *et al*: **A polymorphic DNA marker genetically linked to Huntington's disease**. *Nature* 1983, **306**(5940):234-238.
4. Lander ES, Botstein D: **Homozygosity mapping: a way to map human recessive traits with the DNA of inbred children**. *Science* 1987, **236**(4808):1567-1570.
5. Romdhane L, Abdelhak S, Research Unit on Molecular Investigation of Genetic Orphan D, Collaborators: **Genetic diseases in the Tunisian population**. *Am J Med Genet A* 2011, **155A**(1):238-267.
6. Vona B, Nanda I, Hofrichter MA, Shehata-Dieler W, Haaf T: **Non-syndromic hearing loss gene identification: A brief history and glimpse into the future**. *Mol Cell Probes* 2015, **29**(5):260-270.
7. McKusick VA: **Mendelian Inheritance in Man and its online version, OMIM**. *Am J Hum Genet* 2007, **80**(4):588-604.
8. Cooper DN, Krawczak M: **Human Gene Mutation Database**. *Hum Genet* 1996, **98**(5):629.
9. Adie EA, Adams RR, Evans KL, Porteous DJ, Pickard BS: **Speeding disease gene discovery by sequence based candidate prioritization**. *BMC Bioinformatics* 2005, **6**:55.
10. Kondrashov FA, Ogurtsov AY, Kondrashov AS: **Bioinformatical assay of human gene morbidity**. *Nucleic Acids Res* 2004, **32**(5):1731-1737.
11. Lopez-Bigas N, Ouzounis CA: **Genome-wide identification of genes likely to be involved in human genetic disease**. *Nucleic Acids Res* 2004, **32**(10):3108-3114.
12. Jimenez-Sanchez G, Childs B, Valle D: **Human disease genes**. *Nature* 2001, **409**(6822):853-855.
13. Lopez-Bigas N, Blencowe BJ, Ouzounis CA: **Highly consistent patterns for inherited human diseases at the molecular level**. *Bioinformatics* 2006, **22**(3):269-277.
14. Oti M, Brunner HG: **The modular nature of genetic diseases**. *Clin Genet* 2007, **71**(1):1-11.
15. Perez-Iratxeta C, Bork P, Andrade MA: **Association of genes to genetically inherited diseases using data mining**. *Nat Genet* 2002, **31**(3):316-319.
16. Tranchevent LC, Capdevila FB, Nitsch D, De Moor B, De Causmaecker P, Moreau Y: **A guide to web tools to prioritize candidate genes**. *Brief Bioinform* 2011, **12**(1):22-32.
17. Moreau Y, Tranchevent LC: **Computational tools for prioritizing candidate genes: boosting disease gene discovery**. *Nat Rev Genet* 2012, **13**(8):523-536.
18. Bamshad MJ, Ng SB, Bigham AW, Tabor HK, Emond MJ, Nickerson DA, Shendure J: **Exome sequencing as a tool for Mendelian disease gene discovery**. *Nat Rev Genet* 2011, **12**(11):745-755.

19. Schwartz CE, Chen CF: **Progress in detecting genetic alterations and their association with human disease**. *J Mol Biol* 2013, **425**(21):3914-3918.
20. Mokni M, Charfeddine C, Abdelhak S: **Genomics and Health in the Developing World**. In: *Genetic Skin Diseases in the Arab World.* 2014.
21. Bchetnia M, Laroussi N, Youssef M, Charfeddine C, Ben Brick AS, Boubaker MS, Mokni M, Abdelhak S, Zili J, Benmously R: **Particular Mal de Meleda phenotypes in Tunisia and mutations founder effect in the Mediterranean region**. *Biomed Res Int* 2013, **2013**:206803.
22. Charfeddine C, Mokni M, Kassar S, Zribi H, Bouchlaka C, Boubaker S, Rebai A, Ben Osman A, Abdelhak S: **Further evidence of the clinical and genetic heterogeneity of recessive transgressive PPK in the Mediterranean region**. *J Hum Genet* 2006, **51**(10):841-845.
23. Messaoud O, Ben Rekaya M, Kefi R, Chebel S, Boughammoura-Bouatay A, Bel Hadj Ali H, Gouider-Khouja N, Zili J, Frih-Ayed M, Mokhtar I *et al*: **Identification of a primarily neurological phenotypic expression of xeroderma pigmentosum complementation group A in a Tunisian family**. *Br J Dermatol* 2010, **162**(4):883-886.
24. Romdhane L, Messaoud O, Bouyacoub Y, Kerkeni E, Naouali C, Cherif Ben Abdallah L, Tiar A, Charfeddine C, Monastiri K, Chabchoub I *et al*: **Comorbidity in the Tunisian population**. *Clin Genet* 2016, **89**(3):312-319.
25. Feramisco JD, Sadreyev RI, Murray ML, Grishin NV, Tsao H: **Phenotypic and genotypic analyses of genetic skin disease through the Online Mendelian Inheritance in Man (OMIM) database**. *J Invest Dermatol* 2009, **129**(11):2628-2636.
26. Leech SN, Moss C: **A current and online genodermatosis database**. *Br J Dermatol* 2007, **156**(6):1115-1148.
27. Pruitt KD, Tatusova T, Maglott DR: **NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins**. *Nucleic Acids Res* 2007, **35**(Database issue):D61-65.
28. Smedley D, Haider S, Ballester B, Holland R, London D, Thorisson G, Kasprzyk A: **BioMart--biological queries made easy**. *BMC Genomics* 2009, **10**:22.
29. Zhang R, Zhang CT: **A Brief Review: The Z-curve Theory and its Application in Genome Analysis**. *Curr Genomics* 2014, **15**(2):78-94.
30. Sayers EW, Barrett T, Benson DA, Bolton E, Bryant SH, Canese K, Chetvernin V, Church DM, Dicuccio M, Federhen S *et al*: **Database resources of the National Center for Biotechnology Information**. *Nucleic Acids Res* 2012, **40**(Database issue):D13-25.
31. Yang Z: **PAML: a program package for phylogenetic analysis by maximum likelihood**. *Comput Appl Biosci* 1997, **13**(5):555-556.
32. Katoh K, Toh H: **Recent developments in the MAFFT multiple sequence alignment program**. *Brief Bioinform* 2008, **9**(4):286-298.
33. Yang Z, Nielsen R: **Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models**. *Mol Biol Evol* 2000, **17**(1):32-43.
34. Yang Z, Nielsen R, Goldman N, Pedersen AM: **Codon-substitution models for heterogeneous selection pressure at amino acid sites**. *Genetics* 2000, **155**(1):431-449.
35. Yang Z, Wong WS, Nielsen R: **Bayes empirical bayes inference of amino acid sites under positive selection**. *Mol Biol Evol* 2005, **22**(4):1107-1118.
36. Al-Shahrour F, Diaz-Uriarte R, Dopazo J: **FatiGO: a web tool for finding significant associations of Gene Ontology terms with groups of genes**. *Bioinformatics* 2004, **20**(4):578-580.

37. Szklarczyk D, Morris JH, Cook H, Kuhn M, Wyder S, Simonovic M, Santos A, Doncheva NT, Roth A, Bork P *et al*: **The STRING database in 2017: quality-controlled protein-protein association networks, made broadly accessible**. *Nucleic Acids Res* 2017, **45**(D1):D362-D368.
38. Bindea G, Mlecnik B, Hackl H, Charoentong P, Tosolini M, Kirilovsky A, Fridman WH, Pages F, Trajanoski Z, Galon J: **ClueGO: a Cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks**. *Bioinformatics* 2009, **25**(8):1091-1093.
39. Sayers EW, Barrett T, Benson DA, Bryant SH, Canese K, Chetvernin V, Church DM, DiCuccio M, Edgar R, Federhen S *et al*: **Database resources of the National Center for Biotechnology Information**. *Nucleic Acids Res* 2009, **37**(Database issue):D5-15.
40. Wickham H. **ggplot2: elegant graphics for data analysis**. *Springer* 2016.
41. Colot OO, C.; Courtellemont, P.: **Information criteria and abrupt changes of probability laws**, vol. 7: Elsevier; 1994.
42. Murphy K: **The BayesNet Toolbox for Matlab**. *Computing Science and Statistics* 2001, **33**:2001.
43. Friedman N, Geiger D, Goldszmidt M: **Bayesian Network Classifiers**. *Machine Learning* 1997, **29**(2):131-163.
44. Lauritzen SL, Spiegelhalter DJ: **Local Computations with Probabilities on Graphical Structures and Their Application to Expert Systems**. *Journal of the Royal Statistical Society: Series B (Methodological)* 1988, **50**(2):157-194.
45. Sing T, Sander O, Beerenwinkel N, Lengauer T: **ROCR: visualizing classifier performance in R**. *Bioinformatics* 2005, **21**(20):3940-3941.
46. Franke L, van Bakel H, Fokkens L, de Jong ED, Egmont-Petersen M, Wijmenga C: **Reconstruction of a functional human gene network, with an application for prioritizing positional candidate genes**. *Am J Hum Genet* 2006, **78**(6):1011-1025.
47. van Driel MA, Cuelenaere K, Kemmeren PP, Leunissen JA, Brunner HG, Vriend G: **GeneSeeker: extraction and integration of human disease-related information from web-based genetic databases**. *Nucleic Acids Res* 2005, **33**(Web Server issue):W758-761.
48. Furney SJ, Higgins DG, Ouzounis CA, Lopez-Bigas N: **Structural and functional properties of genes involved in human cancer**. *BMC Genomics* 2006, **7**:3.
49. Accetturo M, Creanza TM, Santoro C, Tria G, Giordano A, Battagliero S, Vaccina A, Scioscia G, Leo P: **Finding new genes for non-syndromic hearing loss through an in silico prioritization study**. *PLoS One* 2010, **5**(9).
50. Jain P, Thukral N, Gahlot LK, Hasija Y: **CARDIO-PRED: an in silico tool for predicting cardiovascular-disorder associated proteins**. *Syst Synth Biol* 2015, **9**(1-2):55-66.
51. Cheng Y, LeGall T, Oldfield CJ, Dunker AK, Uversky VN: **Abundance of intrinsic disorder in protein associated with cardiovascular disease**. *Biochemistry* 2006, **45**(35):10448-10460.
52. Pandey P, Acharya M: **Disease-Phenotype Deconvolution in Genetic Eye Diseases Using Online Mendelian Inheritance in Man**. *Invest Ophthalmol Vis Sci* 2016, **57**(6):2895-2904.
53. Forero DA, Prada CF, Perry G: **Functional and Genomic Features of Human Genes Mutated in Neuropsychiatric Disorders**. *Open Neurol J* 2016, **10**:143-148.
54. Furney SJ, Calvo B, Larranaga P, Lozano JA, Lopez-Bigas N: **Prioritization of candidate cancer genes--an aid to oncogenomic studies**. *Nucleic Acids Res* 2008, **36**(18):e115.

55. Mushegian AR, Bassett DE, Jr., Boguski MS, Bork P, Koonin EV: **Positionally cloned human disease genes: patterns of evolutionary conservation and functional motifs**. *Proc Natl Acad Sci U S A* 1997, **94**(11):5831-5836.
56. Smith NG, Eyre-Walker A: **Human disease genes: patterns and predictions**. *Gene* 2003, **318**:169-175.
57. Karlin S, Chen C, Gentles AJ, Cleary M: **Associations between human disease genes and overlapping gene groups and multiple amino acid runs**. *Proc Natl Acad Sci U S A* 2002, **99**(26):17008-17013.
58. Babu MM, van der Lee R, de Groot NS, Gsponer J: **Intrinsically disordered proteins: regulation and disease**. *Curr Opin Struct Biol* 2011, **21**(3):432-440.
59. Choura M, Rebai A: **The disordered charged biased proteins in the human diseasome**. *Interdiscip Sci* 2019.
60. Fuchs E: **Evolution and complexity of the genes encoding the keratins of human epidermal cells**. *J Invest Dermatol* 1983, **81**(1 Suppl):141s-144s.
61. Welbourne TC, Mu X: **Extracellular glutamate flux regulates intracellular glutaminase activity in LLC-PK1-F+ cells**. *Am J Physiol* 1995, **268**(6 Pt 1):C1418-1424.
62. Steinert PM, Kartasova T, Marekov LN: **Biochemical evidence that small proline-rich proteins and trichohyalin function in epithelia by modulation of the biomechanical properties of their cornified cell envelopes**. *J Biol Chem* 1998, **273**(19):11758-11769.
63. Bern HA, Harkness DR, Blair SM: **Radioautographic Studies of Keratin Formation**. *Proc Natl Acad Sci U S A* 1955, **41**(1):55-60.
64. Rogers MA, Nischt R, Korge B, Krieg T, Fink TM, Lichter P, Winter H, Schweizer J: **Sequence data and chromosomal localization of human type I and type II hair keratin genes**. *Exp Cell Res* 1995, **220**(2):357-362.
65. Wang H, Parry DA, Jones LN, Idler WW, Marekov LN, Steinert PM: **In vitro assembly and structure of trichocyte keratin intermediate filaments: a novel role for stabilization by disulfide bonding**. *J Cell Biol* 2000, **151**(7):1459-1468.
66. Fudge DS, Gosline JM: **Molecular design of the alpha-keratin composite: insights from a matrix-free model, hagfish slime threads**. *Proc Biol Sci* 2004, **271**(1536):291-299.
67. Ruhrberg C, Hajibagheri MA, Parry DA, Watt FM: **Periplakin, a novel component of cornified envelopes and desmosomes that belongs to the plakin family and forms complexes with envoplakin**. *J Cell Biol* 1997, **139**(7):1835-1849.
68. Langbein L, Rogers MA, Praetzel-Wunder S, Bockler D, Schirmacher P, Schweizer J: **Novel type I hair keratins K39 and K40 are the last to be expressed in differentiation of the hair: completion of the human hair keratin catalog**. *J Invest Dermatol* 2007, **127**(6):1532-1535.
69. Steinert PM, Steven AC, Roop DR: **The molecular biology of intermediate filaments**. *Cell* 1985, **42**(2):411-420.
70. Hatzfeld M, Burba M: **Function of type I and type II keratin head domains: their role in dimer, tetramer and filament formation**. *J Cell Sci* 1994, **107 (Pt 7)**:1959-1972.
71. Williamson MP: **Nuclear magnetic resonance studies of peptides and their interactions with receptors**. *Biochem Soc Trans* 1994, **22**(1):140-144.
72. Ramshaw JA, Shah NK, Brodsky B: **Gly-X-Y tripeptide frequencies in collagen: a context for host-guest triple-helical peptides**. *J Struct Biol* 1998, **122**(1-2):86-91.

73. Pokidysheva E, Boudko S, Vranka J, Zientek K, Maddox K, Moser M, Fassler R, Ware J, Bachinger HP: **Biological role of prolyl 3-hydroxylation in type IV collagen**. *Proc Natl Acad Sci U S A* 2014, **111**(1):161-166.
74. Herrmann H, Aebi U: **Intermediate filaments: molecular structure, assembly mechanism, and integration into functionally distinct intracellular Scaffolds**. *Annu Rev Biochem* 2004, **73**:749-789.
75. de Jesus AJ, Allen TW: **The role of tryptophan side chains in membrane protein anchoring and hydrophobic mismatch**. *Biochim Biophys Acta* 2013, **1828**(2):864-876.
76. Valley CC, Cembran A, Perlmutter JD, Lewis AK, Labello NP, Gao J, Sachs JN: **The methionine-aromatic motif plays a unique role in stabilizing protein structure**. *J Biol Chem* 2012, **287**(42):34979-34991.
77. Teichmann SA, Chothia C: **Immunoglobulin superfamily proteins in Caenorhabditis elegans**. *J Mol Biol* 2000, **296**(5):1367-1383.
78. Huang H, Winter EE, Wang H, Weinstock KG, Xing H, Goodstadt L, Stenson PD, Cooper DN, Smith D, Alba MM *et al*: **Evolutionary conservation and selection of human disease gene orthologs in the rat and mouse genomes**. *Genome Biol* 2004, **5**(7):R47.
79. Bustamante CD, Fledel-Alon A, Williamson S, Nielsen R, Hubisz MT, Glanowski S, Tanenbaum DM, White TJ, Sninsky JJ, Hernandez RD *et al*: **Natural selection on protein-coding genes in the human genome**. *Nature* 2005, **437**(7062):1153-1157.
80. Thomas PD, Kejariwal A: **Coding single-nucleotide polymorphisms associated with complex vs. Mendelian disease: evolutionary evidence for differences in molecular effects**. *Proc Natl Acad Sci U S A* 2004, **101**(43):15398-15403.
81. Tu Z, Wang L, Xu M, Zhou X, Chen T, Sun F: **Further understanding human disease genes by comparing with housekeeping genes and other genes**. *BMC Genomics* 2006, **7**:31.
82. Blekhman R, Man O, Herrmann L, Boyko AR, Indap A, Kosiol C, Bustamante CD, Teshima KM, Przeworski M: **Natural selection on genes that underlie human disease susceptibility**. *Curr Biol* 2008, **18**(12):883-889.
83. Parra EJ: **Human pigmentation variation: evolution, genetic basis, and implications for public health**. *Am J Phys Anthropol* 2007, **Suppl 45**:85-105.
84. Rees JL: **Genetics of hair and skin color**. *Annu Rev Genet* 2003, **37**:67-90.
85. Lucock M: **Folic acid: nutritional biochemistry, molecular biology, and role in disease processes**. *Mol Genet Metab* 2000, **71**(1-2):121-138.
86. Off MK, Steindal AE, Porojnicu AC, Juzeniene A, Vorobey A, Johnsson A, Moan J: **Ultraviolet photodegradation of folic acid**. *J Photochem Photobiol B* 2005, **80**(1):47-55.
87. Jablonski NG, Chaplin G: **The evolution of human skin coloration**. *J Hum Evol* 2000, **39**(1):57-106.
88. Williams JD, Jacobson EL, Kim H, Kim M, Jacobson MK: **Folate in skin cancer prevention**. *Subcell Biochem* 2012, **56**:181-197.
89. Aerts S, Lambrechts D, Maity S, Van Loo P, Coessens B, De Smet F, Tranchevent LC, De Moor B, Marynen P, Hassan B *et al*: **Gene prioritization through genomic data fusion**. *Nat Biotechnol* 2006, **24**(5):537-544.
90. Erlich Y, Edvardson S, Hodges E, Zenvirt S, Thekkat P, Shaag A, Dor T, Hannon GJ, Elpeleg O: **Exome sequencing and disease-network analysis of a single family implicate a mutation in KIF1A in hereditary spastic paraparesis**. *Genome Res* 2011, **21**(5):658-664.

761 91. Smedley D, Kohler S, Czeschik JC, Amberger J, Bocchini C, Hamosh A, Veldboer J,
762 Zemojtel T, Robinson PN: **Walking the interactome for candidate prioritization in**
763 **exome sequencing studies of Mendelian diseases**. *Bioinformatics* 2014,
764 **30**(22):3215-3222.

765

766

FIGURES TITLES:

Figure 1: Exon count distribution within skin and non skin genes group

Figure 2: Average % identity of orthologs with human in the two group (skin and non skin protein coding genes)

Figure 3: Frequency of positive selected amino acids in the skin and non skin group of genes. (*) $P < 0.001$. Comparison between amino acids frequency was performed by means of chi-squared test

Figure 4: Enriched Biological processes in skin and non Skin diseases using Clue GO Cytoscape plugin

Figure 5: ROC curve for the prediction of genodermatosis genes.

TABLES TITLES:

Table 1: Median and p-values of Wilcox and KS tests for structural features

Table 2: Median and statistical analysis for peptide composition

Table 3: Median values and statistical analysis of ortholog conservation

ADDITIONAL FILES:

Additional-File1.xlsx:

Supplementary table 1: List of features analysed in this work with the corresponding tools

Additional-File2.xlsx:

Supplementary table 2: Median and statistical analysis for nucleotide motif properties

Supplementary table 3: Mean X, Y and Z components

Additional-File3.xlsx:

Supplementary table 4: PFAM domain Enrichment in skin and non-skin proteins using StringDB

Additional-File4.xlsx:

Supplementary table 5: Selective pressure study

Supplementary table 6: Selective pressure study

Additional-File5.xlsx:

Supplementary table 7: GO enrichment in skin and non-skin genes using FatiGO and StringDB

Additional-File6.xlsx

Supplementary table 8: Probability scores of genes in the test set

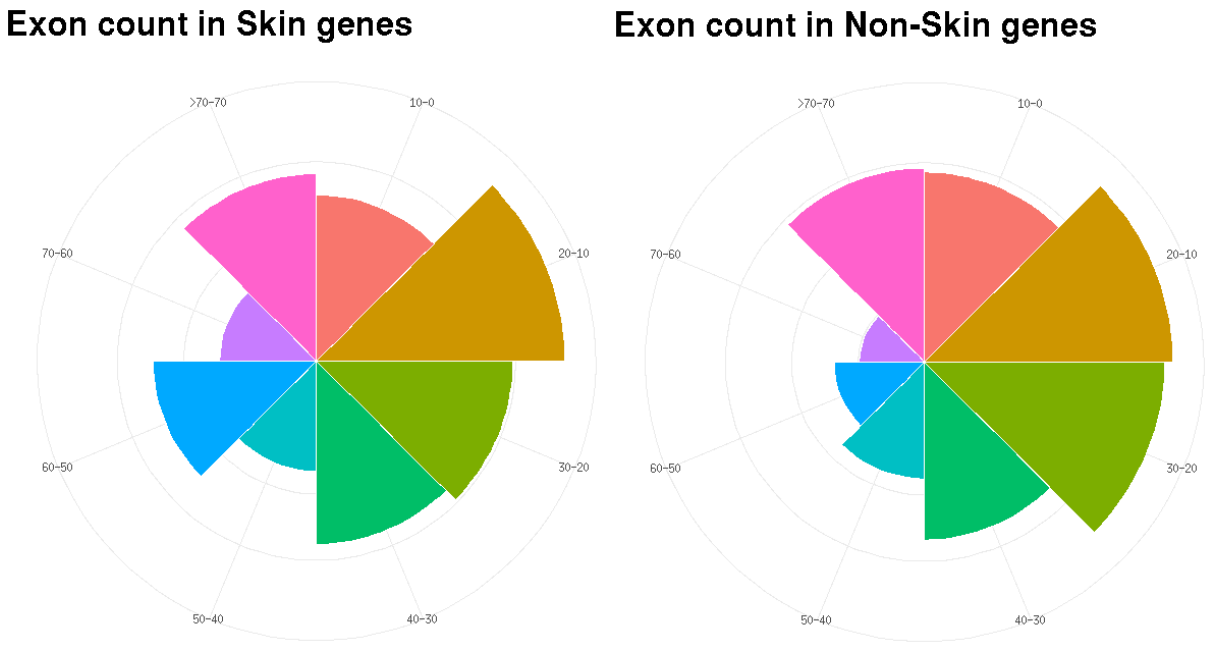

Figure 1: Exon count distribution within skin and non skin genes group

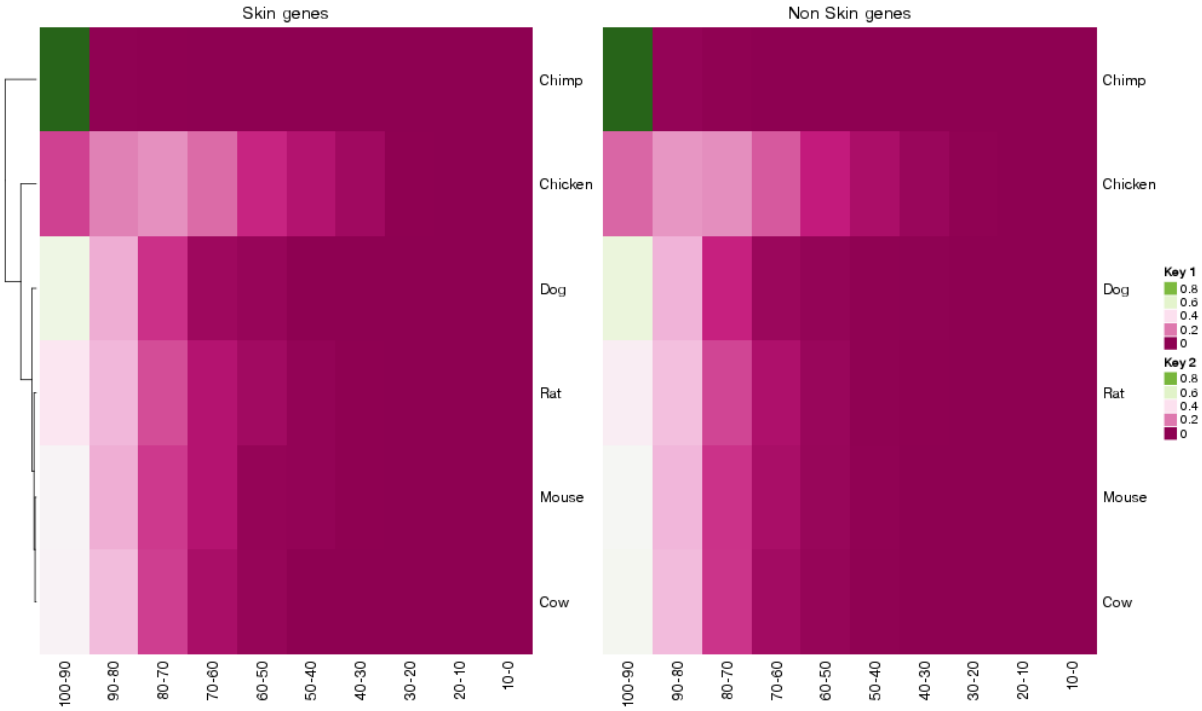

Figure 2: Average % identity of orthologs with Human genes in the two groups

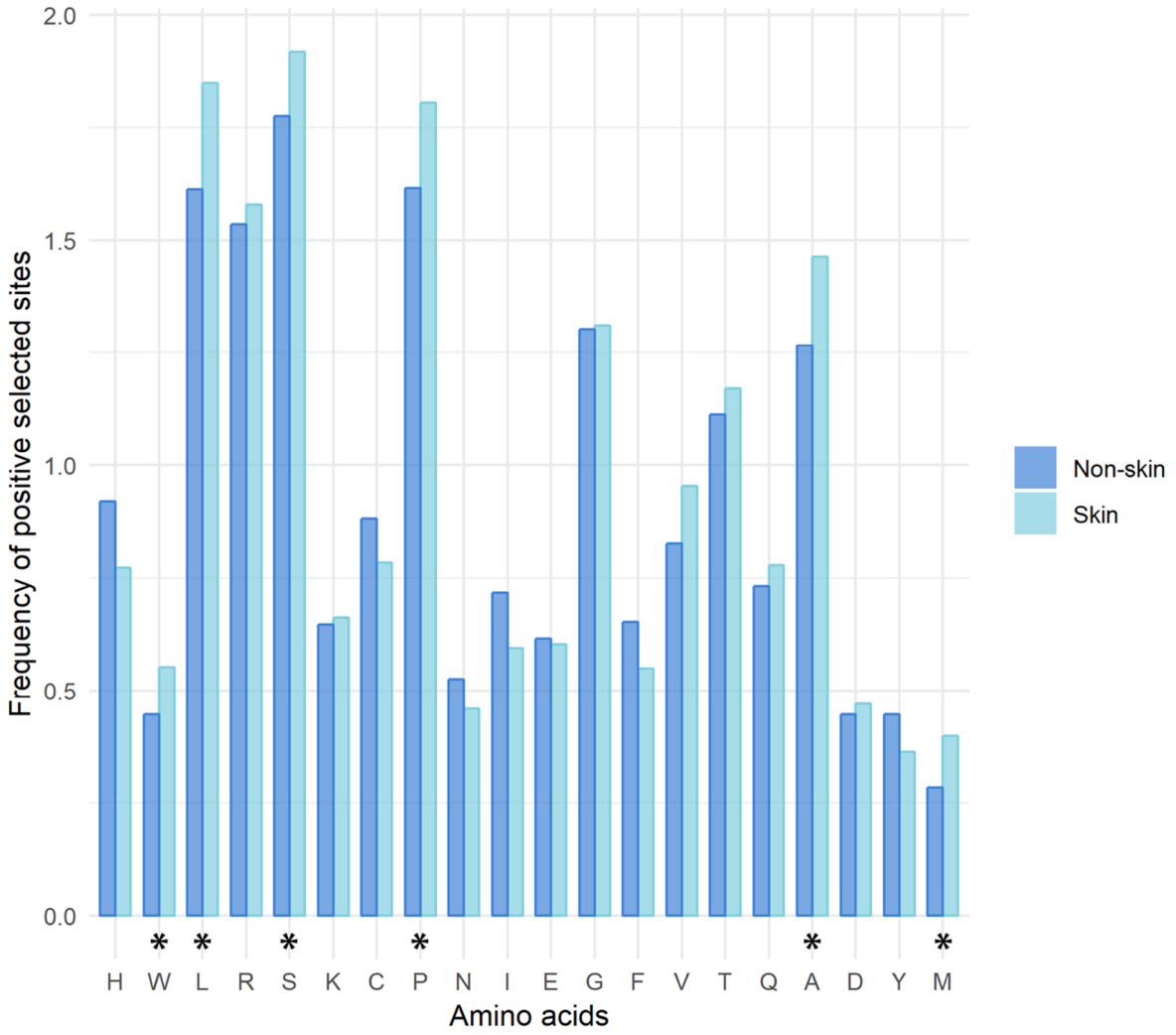

Figure 3: Frequency of positive selected amino acids in the skin and non skin group of genes. (*) P < 0.001. Comparison between amino acids frequency was performed by means of Chi-squared test

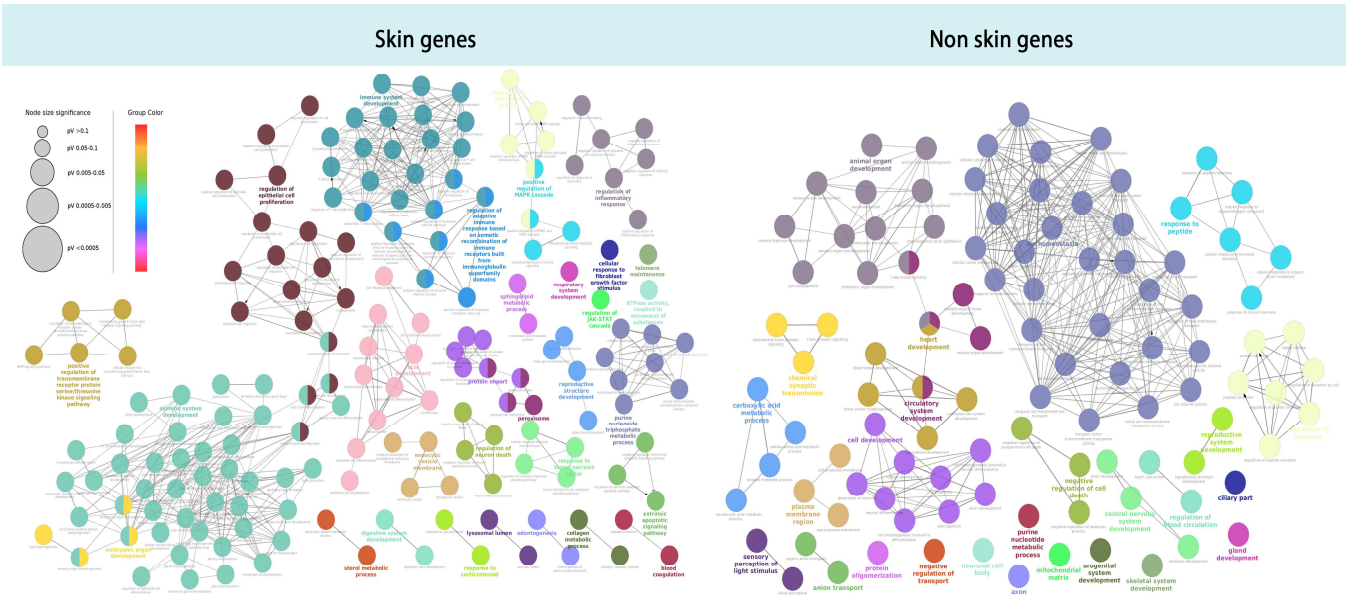

Figure 4: Enriched Biological processes in Skin and non Skin diseases using Clue GO Cytoscape plugin

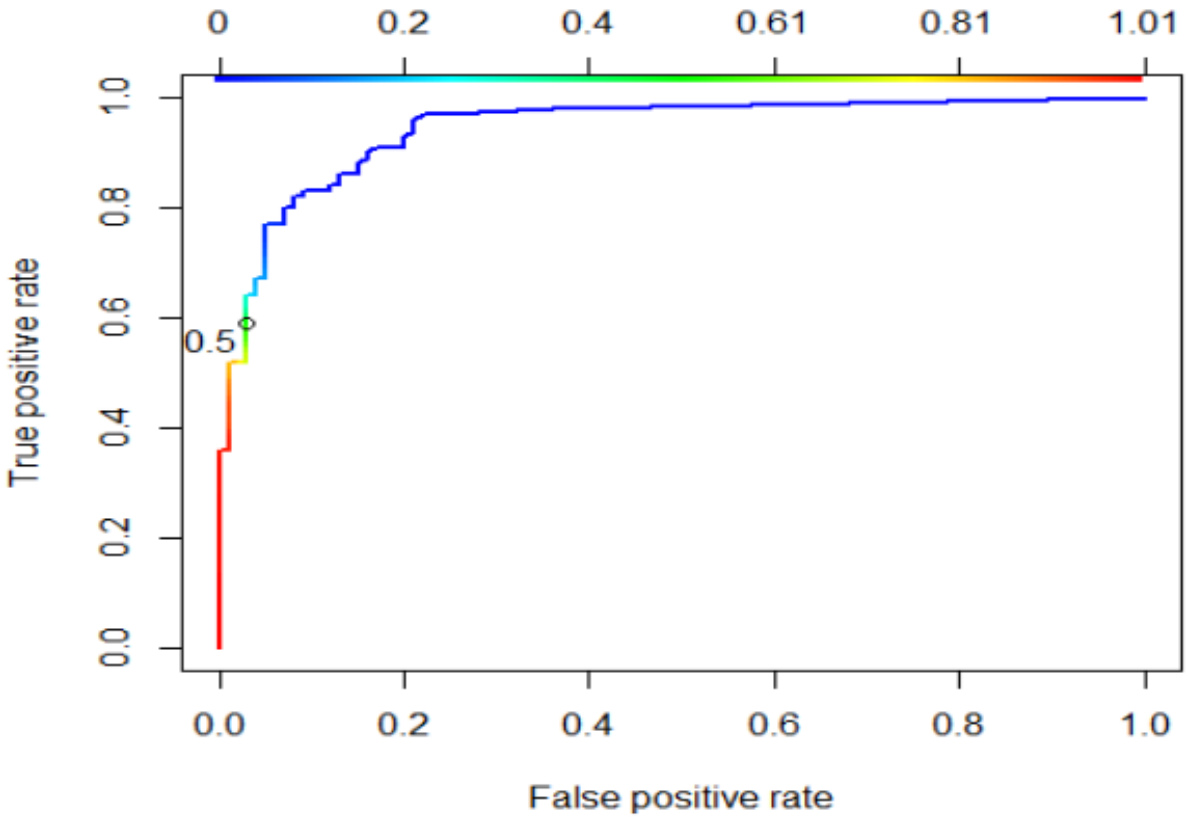

Figure 5: ROC curve for the prediction of genodermatosis genes.

Table 1: Median and p-values of Wilcox and KS tests for structural features

	Gene	CDS	cDNA	Exons	3'UTR	5'UTR	Isoform	Protein
Genodermatoses	29901	1530	2934	10.5	1061.5	176.9	1	509.5
Others	30025	1466	2889	9.5	1122.2	177.25	1	487.7
Wilcox p-value	0.36	**0.006**	0.27	**0.021**	0.47	0.81	0.47	**0.005**
KS p-value	0.8	0.091	0.52	**0.004**	0.2	0.87	0.88	0.086

Table 2: Median and statistical analysis for peptide composition

	Cysteine (%)	**Glutamine (%)**	**Isoleucine (%)**	**Methionine (%)**	**Phenylalanine (%)**
Genodermatoses	0.023	0.045	0.043	0.021	0.038
Others	0.022	0.044	0.045	0.023	0.039
Wilcox p-value	**0.039**	**0.032**	**0.046**	**$8.2\ 10^{-4}$**	**0.025**
KS p-value	**0.007**	0.094	**0.031**	**0.002**	**0.049**

Table 3: Median values and statistical analysis of ortholog conservation

Species		Genodermatoses (%)	Others (%)	Wilcox p-value	KS p-value
Gallus gallus (Chicken)	**DNA**	73.2	73.9	**10^{-3}**	0.05
	Protein	74.85	76.4	**$3.3\ 10^{-2}$**	0.19
Canis lupus (Dog)	**DNA**	89.2	89.6	**$3.1\ 10^{-2}$**	0.14
	Protein	91.3	91.8	0.05	0.09
Bos taurus (Cow)	**DNA**	88.7	89.1	0.09	0.29
	Protein	90.5	90.9	0.05	0.07
Ratus norvegicus (Rat)	**DNA**	85.3	85.6	0.06	0.14
	Protein	88.3	88.5	0.06	0.17
Mus musculus (Mouse)	**DNA**	85.5	85.8	**$1.8\ 10^{-2}$**	**$2.2\ 10^{-2}$**
	Protein	88.4	88.8	**$1.4\ 10^{-2}$**	**$1.8\ 10^{-2}$**
Pan troglotydes (Chimpanzee)	**DNA**	99.4	99.4	0.9458	0.99
	Protein	99.4	99.4	0.94	0.97