



**HAL**  
open science

## 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.

### ► To cite this version:

Lilia Romdhane, Heni Bouhamed, Kais Ghedira, Cherif Ben Hamda, Amel Louhichi, et al.. The morbid cutaneous anatomy of the human genome revealed by a bioinformatic approach. *Genomics*, 2020, 112 (6), pp.4232 - 4241. 10.1016/j.ygeno.2020.07.009 . hal-03491200

**HAL Id: hal-03491200**

**<https://hal.science/hal-03491200>**

Submitted on 22 Aug 2022

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution - NonCommercial 4.0 International License

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

3  
4 **Authors:**

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

8  
9  
10 **Authors' affiliations:**

11 1 Biomedical Genomics and Oncogenetics Laboratory LR11IPT05, LR11IPT16, Institut  
12 Pasteur de Tunis, Tunis, Université Tunis El Manar, Tunis, Tunisia.  
13 2 Department of Biology, Faculty of Science of Bizerte, Jarzouna, Université Tunis  
14 Carthage, Tunis, Tunisia  
15 3 Molecular and Cellular Screening Process Laboratory, Sfax, Centre of Biotechnology of  
16 Sfax, Sfax, Tunisia  
17 4 Laboratory of Bioinformatics, Biomathematics and Biostatistics (LR16IPT09), Institut  
18 Pasteur de Tunis, Tunis, Université Tunis El Manar, Tunis, Tunisia.  
19 5 High Institut of Biotechnology of Sidi Thabet, University of Manouba, BiotechPole of Sidi  
20 Thabet, Ariana, Tunisia.  
21 6 Department of Dermatology, CHU La Rabta Tunis, Tunis, Tunisia.  
22 7 Research Unit on Hereditary Keratinization Disorders UR12SP07, CHU La Rabta, Tunis,  
23 Tunisia.  
24

25 **Corresponding author's contact:**

26 **Lilia Romdhane**

27 Biomedical Genomics and Oncogenetics Laboratory  
28 Institut Pasteur de Tunis, BP 74, 13 Place Pasteur 1002 Tunis Belvédère Tunisia,  
29 Department of Biology, Faculty of Science of Bizerte, Jarzouna, Université Tunis Carthage,  
30 Tunis, Tunisia  
31 Tel: 216 71 849 110  
32 Fax: 216 71 791 833, e-mail : [lilia.romdhane@gmail.com](mailto:lilia.romdhane@gmail.com) ; [lilia.romdhane@fsb.rnu.tn](mailto:lilia.romdhane@fsb.rnu.tn)  
33

34 **Abstract:**

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

47

48

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

51

52 **1. Introduction:**

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

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

75 Functional features of disease genes revealed also distinct patterns. They encode essentially  
76 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 investigate the features of genes involved in genodermatoses by comparing them to another

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

## 105 **2. Methods:**

### 106 **2.1 Morbid gene group delineation:**

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

### 119 **2.2 Structural analysis:**

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

127

### 128 **2.3 Evolutionary conservation study:**

129 We also studied the phylogenetic extent (that is the extent to which a gene is conserved back  
130 through evolution based on homologs in other species) of each morbid gene examined.

131 Therefore, orthologs in 6 Vertebrate species (*Pan troglodydes*, *Bos taurus*, *Canis lupus*, *Rattus*  
132 *norvegicus*, *Mus musculus* and *Gallus gallus*) were retrieved from the NCBI HomoloGene  
133 database [30]. For each ortholog, the percentage of identity at both nucleic and amino-acid  
134 levels were noted. To estimate the degree of paralogy, a BlastP  
135 (<https://blast.ncbi.nlm.nih.gov/Blast.cgi>) search was performed for each protein in the group  
136 against the human proteome. The best BlastP hit found in the human genome was kept for an  
137 “E-value” cutoff of  $10^{-30}$ . For each paralog, an identity percentage is returned as well as an  
138 average of this measure for alternatively spliced isoforms. For each gene, the number of  
139 putative paralogs found is calculated in each disease gene group.

### 140 **2.4 Calculation of nucleotide substitution rates at DNA level:**

141 We estimated dN (nonsynonymous substitution rate), dS (synonymous substitution rate) and  
142  $\omega = dN/dS$  for each gene with the Codeml software of the PAML package (**Additional File 1**)  
143 [31]. Codeml uses the codon as the evolutionary unit in DNA alignments of coding regions  
144 and an explicit model incorporating the genetic code to estimate the amount of natural  
145 selection for or against protein-level changes. Total substitution rate (d) was estimated from  
146 the HomoloGene database. To test whether a difference exists between the disease groups,  
147 genes with no orthologs as well as those with dS= 0 were excluded.

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

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

## 158 **2.5 Functional analysis:**

159 Enrichment in specific GO terms was assessed using FatiGO on the Babelomics server for  
160 gene expression, genome variation and functional analysis (<http://babelomics.bioinfo.cipf.es/>)  
161 [36], StringDB [37] and ClueGO cytoscape plugin [38]. Statistical significance in GO terms  
162 occurrence difference is considered using a Fisher's exact test and/or false discovery rate  
163 (FDR $\leq$ 0.05).

## 164 **2.6 Programming and statistic tests:**

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

## 172 **2.7 Bayesian network**

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



177 discretization of the continuous variables using the Colot algorithm used as the model attributes  
178 [41]. For the construction and manipulation of the Bayesian network, we used the Bayesian  
179 network toolbox (BNT) [42] (<http://www.ai.mit.edu/~murphyk/Software/BNT/bnt.html>) on  
180 the Matlab software version 2010 ([www.mathworks.com](http://www.mathworks.com)). We employed the « Augmented  
181 Naïve Bayes » [43] for the structure learning and the « Clique-tree propagation algorithm »  
182 [44] for the inference. During the learning phase, we used 548 genes (from the initial 648)  
183 involved in genodermatosis and 1,708 genes (from the initial 1,808) involved in other genetic  
184 diseases thus forming the learning set. The remaining 200 genes, the test set, were used during  
185 the test phase of the constructed classifier.

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

### 192 **3. Results:**

#### 193 **3.1 Structural features of genes and products:**

##### 194 **3.1.1 Length features:**

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

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

### 203 **3.1.2 Mono, di and tri nucleotide motif frequencies:**

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

### 211 **3.1.3 Z curve of CDS:**

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

### 225 **3.1.4 Primary and secondary structure of proteins:**

226 Proteins responsible for genodermatoses are more enriched in cysteine and glutamine (**Table**  
227 **2**). Nevertheless, they are less enriched in isoleucine, methionine and phenylalanine. The  
228 distribution of these frequencies is also different in the two groups except for glutamine. In  
229 addition, we studied the prevalence of some protein structure such as the presence of signal  
230 peptide, transmembrane domains as well as alpha helices. Only 23.9 % of genodermatosis  
231 proteins versus 22.5 % of proteins of the second group have a secondary structure of helix  
232 type. This difference is not significant (Chi2 test p-value = 0.51). Similar frequencies of  
233 signal peptide (34.5% in genodermatoses vs 30.8% in the second group, Chi2 p-value = 0.36)  
234 and transmembrane domain (30.5 % in genodermatoses vs 32.6 % in the second group, p-  
235 value= 0.35) have also been noticed. Furthermore, genodermatosis proteins seem to be  
236 enriched in immunoglobulin C1-set domain (FDR=0.0048) while the second group of genes  
237 was enriched in Homeobox domain (FDR=0.00569) (**Additional File 3**).

### 238 **3.2 Evolutionary conservation features:**

#### 239 **3.2.1 Orthology:**

240 Genes underlying skin genetic diseases have more orthologs than genes in the second morbid  
241 group especially in Dog (91 % vs 87 %, Chi2 p-value= **0.7 10<sup>-2</sup>**) and Mouse (97.22 % vs 94.8  
242 %, Chi-square p-value= **1.52 10<sup>-2</sup>**). No significant difference has been reported in the other  
243 species (**Additional File 4**). Nevertheless, orthologs of skin genetic diseases seem to be less  
244 conserved at the nucleic level in 3 species (**Table 3**). At the protein level, chicken orthologs  
245 are less conserved than other species in the two gene groups. Distribution of conservation  
246 proportion of ortholog sequences is only different among the mouse orthologs at both DNA  
247 and protein levels (KS p-value= 1.8 10<sup>-2</sup> and KS p-value =2.2 10<sup>-2</sup>, respectively) (**Table 3**,  
248 Figure 2).

#### 249 **3.2.2 Paralogy:**

250 Human protein alignment against the proteome using BlastP revealed that paralogs are less  
251 prevalent among genes of genodermatoses than do genes of the second group (63.88 % vs  
252 69.85 %, Chi2 p-value =  $5.9 \cdot 10^{-3}$ ). Nevertheless, genodermatosis proteins have **more** putative  
253 paralogs per gene, the median number of paralogs per gene being 5 versus 4 (Wilcox p-value  
254 =  $1.15 \cdot 10^{-3}$ ). The paralog number per gene distribution is also significantly different (KS p-  
255 value =  $3.2 \cdot 10^{-2}$ ). Thus, paralog count distribution shows that duplication is frequent among  
256 the non-skin disease genes having putative orthologs under 7 paralogs per gene and less  
257 frequent for those having more than 31 paralogs per gene. Furthermore, skin disease gene  
258 paralogs are less conserved (42.44 % vs 43 %, Wilcoxon p-value =  $3 \cdot 10^{-2}$ ). The conservation  
259 distribution is also significantly different (KS p-value =  $4 \cdot 10^{-2}$ ). More conserved paralogs with  
260 values over 60 % are more represented in the non-skin disease gene group.

### 261 **3.2.3 Selective pressure acting on morbid genes at the DNA level:**

262 The degree of conservation at the DNA level of the 2 sets of morbid genes was analysed in  
263 order to shed light on the selective pressures acting on them. We examined total substitution  
264 (d), non-synonymous (dN), synonymous (dS) rates, as well as dN/dS ratio, from coding  
265 sequences of orthologs of human in 6 Vertebrate species (**Additional File 4**). The total  
266 substitution rate (d) is statistically higher only among Human-Chicken (**Wilcox** test p-values  
267 =  $10^{-2}$ , respectively), Human-Dog (**Wilcox** p-values =  $3 \cdot 10^{-2}$ ) and Human-Mouse (**Wilcox** test  
268 p-values =  $2 \cdot 10^{-2}$ ) orthologs in the skin disease gene group. The distribution is significantly  
269 different only for Human-Chicken and Human-Mouse orthologs (KS test p-values =  **$2.6 \cdot 10^{-2}$**   
270 and KS test p-values =  **$3.8 \cdot 10^{-2}$**  respectively). The dN, which is indicative of the selective  
271 pressure acting on sites that involve an amino-acid change, is lower among the skin disease  
272 genes of Human-Primate orthologs (0.002 vs 0.003, MW p-value =  $4.7 \cdot 10^{-2}$ ). Furthermore, the  
273 dN distribution is significantly different among all the studied orthologs except those of Cow

274 and Primate. Skin disease genes are thus evolving more slowly, in general, than the other  
275 disease genes, this indicates that they are subject to strong selective constraints.  
276 The dS is more reflective of the background mutation level. Globally, skin disease genes  
277 show less dS values among the orthologs of the studied species (**Additional File 4**). The  
278 distribution is also statistically different (**Additional File 4**). This difference could be  
279 assigned to varying mutation rates in the genome. The analysis of dN/dS shows that the ratio  
280 is statistically lower only among the Human-Primate orthologs in the skin disease gene group  
281 (**Additional File 4**). The distribution is also significantly different (**Additional File 4**).  
282 Similar dN/dS values were found among the orthologs of the remaining species in the two  
283 groups although the distribution of dN/dS values is different. Moreover, the investigation of  
284 the amino acid under positive selection between skin and non skin genes showed that the most  
285 enriched amino acids under positive selection within the skin protein coding genes compared  
286 to the non-skin coding genes includes W: Tryptophan, L: Leucine, S: Serine, P: Proline, A:  
287 Alanine and M: Methionine (**Figure 3**).

### 288 **3.3 Function and process of morbid genes:**

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

### 298 **3.4 Bayesian network**

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

#### 309 **4. Discussion:**

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

317 Skin disease genes coding sequences and their proteins are longer than the remaining of the  
318 morbid genes with higher number of exons. A similar pattern has been noted in comparison to  
319 protein involved in cancer and hereditary diseases when comparing with non-disease genes  
320 [9-11,48,54-56]. Taking cancer genes as a set of morbid genes, length pattern differences have  
321 been revealed between mutated and translocated cancer genes showing that those muted have  
322 longer CDS than those translocated [48]. In genodermatoses, as in the other hereditary  
323 diseases, a longer CDS is more susceptible to accumulate point mutations as a consequence of

324 its length and therefore more likely to produce a dysfunctional protein. Therefore, this length  
325 pattern among morbid genes of distinct groups of diseases could be assigned to a different  
326 mutation process of these two groups of genes. This could be verified by the study of the  
327 variant pattern accumulation in the genodermatoses genomic sequences as well as their  
328 corresponding transcripts and CDS. In addition to the correlation gene length - high mutation  
329 accumulation probability, it was also suggested that gene overlap and multiple amino acid  
330 runs are also morbid disease gene features [57]. Similarly, motifs enriched in CDS of  
331 genodermatoses could be associated with mutational hotspots that increase the probability for  
332 a mutation to occur. With respect to over-representations of some amino acids in  
333 corresponding proteins, their implications in the morbid phenotype of the skin are not clear.  
334 However, it was established that protein regions lacking secondary or tertiary structures,  
335 known as intrinsically disordered region, could be involved in protein-protein interactions,  
336 regulation and signal transduction explaining their crucial role in disease development  
337 including cancers and cardiovascular diseases [51,58]. As at least 4 genodermatosis proteins  
338 are shown be known as intrinsically disordered charged protein involved in human diseases,  
339 further studies on protein intrinsic disorders in this kind of diseases are needed and could help  
340 gain insights on amino acid residues occurrence involved in genodermatoses [59].  
341 Our results revealed also that genodermatoses genes are more enriched in Cysteine and  
342 Glutamine, two negatively charged amino-acids, and less enriched in Isoleucine, Methionine  
343 and Phenylalanine compared to the set of genes from other disorders. Moreover, this study  
344 showed that Tryptophan, Leucine, Serine, Proline, Alanine and Methionine are the most  
345 enriched amino acids under positive selection in skin proteins, thus suggesting conserved  
346 functional and structural properties. Indeed, it has been shown that bovine, rat and human skin  
347 keratins are all rich in the amino acids glycine, serine, leucine and glutamic acid [60].  
348 Glutamine plays a key role in regulating the acid-base balance within the body and firms the

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

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



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

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

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

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

399 have lower dN/dS values [10,79], others found higher values [56,78] and one found similar  
400 pattern [80]. The most plausible explanations could be that genetic disease genes have been  
401 considered as a homogeneous dataset thus neglecting that they could in fact be categorized  
402 into house-keeping and non-essential genes exhibiting different patterns of substitution rates  
403 [80]. Similarly, different values of substitution rate accumulations were found according to  
404 the transmission mode [48,82]. Correspondingly, our study revealed distinct patterns of  
405 substitution rates as well as dN/dS ratio in pathogenic genes according to tissue/organ type.  
406 Skin disease genes seem to evolve at slower rates by accumulating less non-synonymous  
407 variations and thus are under selective pressure constraints. In opposition, we have observed  
408 that frequent skin genes have greater dN/dS values indicating that less selective pressure have  
409 been able to change among species and to tolerate variations without leading to the expression  
410 of a morbid phenotype. In addition, we have also found that extremely conserved proteins are  
411 less frequently found to be involved in genodermatoses. One likely explanation is that  
412 variations in this group of extremely conserved genes are mostly lethal.

413 The functional analysis of the studied genodermatosis group of genes through GO enrichment,  
414 showed that these later are mainly involved in epithelium, organ, tissue, skin, cell  
415 development but also response to stress, immune response, pigmentation and the control of  
416 inflammatory response. These genes are mainly acting on peroxisome, vacuole, lysosome and  
417 cellular organelles. The most significant molecular function that is over-represented in these  
418 skin genetic disease genes is the binding activity. Function annotation pattern through GO  
419 terms could correlate with the conservation pattern. As assumed earlier that genes in skin  
420 diseases are under selective constraints, this seems to be obvious as the skin is the first natural  
421 barrier protection from the external environment. Skin mechanism protection acts mainly  
422 through its pigmentation, the melanin, and its variation influences both pigmentation and skin  
423 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

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

## 463 **5. Conclusion:**

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

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

476

477 **ABBREVIATIONS:**

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

481

482 **DECLARATIONS:**

483 **Ethics approval and consent to participate:**

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

486 **Consent to publish:**

487 Not applicable.

488

489 **Funding:**

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

493

494 **Acknowledgements:**

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

498

499 **Availability of data and materials:**

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

503

504 **Competing interests:**

505 The authors declare that they have no competing interests.

506

507 **Authors' contributions:**

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

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

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

513 **HJ and SR:** managed the data

514 **CC and MM:** revised the manuscript

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

516

517 **REFERENCES:**

- 518 1. Basit S, Lee K, Habib R, Chen L, Umm e K, Santos-Cortez RL, Azeem Z, Andrade P,  
519 Ansar M, Ahmad W *et al*: **DFNB89, a novel autosomal recessive nonsyndromic**  
520 **hearing impairment locus on chromosome 16q21-q23.2.** *Hum Genet* 2011,  
521 **129(4):379-385.**
- 522 2. Ben Hamida C, Doerflinger N, Belal S, Linder C, Reutenauer L, Dib C, Gyapay G,  
523 Vignal A, Le Paslier D, Cohen D *et al*: **Localization of Friedreich ataxia phenotype**  
524 **with selective vitamin E deficiency to chromosome 8q by homozygosity mapping.**  
525 *Nat Genet* 1993, **5(2):195-200.**
- 526 3. Gusella JF, Wexler NS, Conneally PM, Naylor SL, Anderson MA, Tanzi RE, Watkins  
527 PC, Ottina K, Wallace MR, Sakaguchi AY *et al*: **A polymorphic DNA marker**  
528 **genetically linked to Huntington's disease.** *Nature* 1983, **306(5940):234-238.**
- 529 4. Lander ES, Botstein D: **Homozygosity mapping: a way to map human recessive**  
530 **traits with the DNA of inbred children.** *Science* 1987, **236(4808):1567-1570.**
- 531 5. Romdhane L, Abdelhak S, Research Unit on Molecular Investigation of Genetic  
532 Orphan D, Collaborators: **Genetic diseases in the Tunisian population.** *Am J Med*  
533 *Genet A* 2011, **155A(1):238-267.**
- 534 6. Vona B, Nanda I, Hofrichter MA, Shehata-Dieler W, Haaf T: **Non-syndromic**  
535 **hearing loss gene identification: A brief history and glimpse into the future.** *Mol*  
536 *Cell Probes* 2015, **29(5):260-270.**
- 537 7. McKusick VA: **Mendelian Inheritance in Man and its online version, OMIM.** *Am*  
538 *J Hum Genet* 2007, **80(4):588-604.**
- 539 8. Cooper DN, Krawczak M: **Human Gene Mutation Database.** *Hum Genet* 1996,  
540 **98(5):629.**
- 541 9. Adie EA, Adams RR, Evans KL, Porteous DJ, Pickard BS: **Speeding disease gene**  
542 **discovery by sequence based candidate prioritization.** *BMC Bioinformatics* 2005,  
543 **6:55.**
- 544 10. Kondrashov FA, Ogurtsov AY, Kondrashov AS: **Bioinformatical assay of human**  
545 **gene morbidity.** *Nucleic Acids Res* 2004, **32(5):1731-1737.**
- 546 11. Lopez-Bigas N, Ouzounis CA: **Genome-wide identification of genes likely to be**  
547 **involved in human genetic disease.** *Nucleic Acids Res* 2004, **32(10):3108-3114.**
- 548 12. Jimenez-Sanchez G, Childs B, Valle D: **Human disease genes.** *Nature* 2001,  
549 **409(6822):853-855.**
- 550 13. Lopez-Bigas N, Blencowe BJ, Ouzounis CA: **Highly consistent patterns for**  
551 **inherited human diseases at the molecular level.** *Bioinformatics* 2006, **22(3):269-**  
552 **277.**
- 553 14. Oti M, Brunner HG: **The modular nature of genetic diseases.** *Clin Genet* 2007,  
554 **71(1):1-11.**
- 555 15. Perez-Iratxeta C, Bork P, Andrade MA: **Association of genes to genetically inherited**  
556 **diseases using data mining.** *Nat Genet* 2002, **31(3):316-319.**
- 557 16. Tranchevent LC, Capdevila FB, Nitsch D, De Moor B, De Causmaecker P, Moreau Y:  
558 **A guide to web tools to prioritize candidate genes.** *Brief Bioinform* 2011, **12(1):22-**  
559 **32.**
- 560 17. Moreau Y, Tranchevent LC: **Computational tools for prioritizing candidate genes:**  
561 **boosting disease gene discovery.** *Nat Rev Genet* 2012, **13(8):523-536.**
- 562 18. Bamshad MJ, Ng SB, Bigham AW, Tabor HK, Emond MJ, Nickerson DA, Shendure  
563 J: **Exome sequencing as a tool for Mendelian disease gene discovery.** *Nat Rev*  
564 *Genet* 2011, **12(11):745-755.**

- 565 19. Schwartz CE, Chen CF: **Progress in detecting genetic alterations and their**  
566 **association with human disease.** *J Mol Biol* 2013, **425**(21):3914-3918.
- 567 20. Mokni M, Charfeddine C, Abdelhak S: **Genomics and Health in the Developing**  
568 **World.** In: *Genetic Skin Diseases in the Arab World.* 2014.
- 569 21. Bchetnia M, Laroussi N, Youssef M, Charfeddine C, Ben Brick AS, Boubaker MS,  
570 Mokni M, Abdelhak S, Zili J, Benmously R: **Particular Mal de Meleda phenotypes**  
571 **in Tunisia and mutations founder effect in the Mediterranean region.** *Biomed Res*  
572 *Int* 2013, **2013**:206803.
- 573 22. Charfeddine C, Mokni M, Kassar S, Zribi H, Bouchlaka C, Boubaker S, Rebai A, Ben  
574 Osman A, Abdelhak S: **Further evidence of the clinical and genetic heterogeneity**  
575 **of recessive transgressive PPK in the Mediterranean region.** *J Hum Genet* 2006,  
576 **51**(10):841-845.
- 577 23. Messaoud O, Ben Rekaya M, Kefi R, Chebel S, Boughammoura-Bouatay A, Bel Hadj  
578 Ali H, Gouider-Khouja N, Zili J, Frih-Ayed M, Mokhtar I *et al*: **Identification of a**  
579 **primarily neurological phenotypic expression of xeroderma pigmentosum**  
580 **complementation group A in a Tunisian family.** *Br J Dermatol* 2010, **162**(4):883-  
581 886.
- 582 24. Romdhane L, Messaoud O, Bouyacoub Y, Kerkeni E, Naouali C, Cherif Ben Abdallah  
583 L, Tiar A, Charfeddine C, Monastiri K, Chabchoub I *et al*: **Comorbidity in the**  
584 **Tunisian population.** *Clin Genet* 2016, **89**(3):312-319.
- 585 25. Feramisco JD, Sadreyev RI, Murray ML, Grishin NV, Tsao H: **Phenotypic and**  
586 **genotypic analyses of genetic skin disease through the Online Mendelian**  
587 **Inheritance in Man (OMIM) database.** *J Invest Dermatol* 2009, **129**(11):2628-2636.
- 588 26. Leech SN, Moss C: **A current and online genodermatosis database.** *Br J Dermatol*  
589 2007, **156**(6):1115-1148.
- 590 27. Pruitt KD, Tatusova T, Maglott DR: **NCBI reference sequences (RefSeq): a curated**  
591 **non-redundant sequence database of genomes, transcripts and proteins.** *Nucleic*  
592 *Acids Res* 2007, **35**(Database issue):D61-65.
- 593 28. Smedley D, Haider S, Ballester B, Holland R, London D, Thorisson G, Kasprzyk A:  
594 **BioMart--biological queries made easy.** *BMC Genomics* 2009, **10**:22.
- 595 29. Zhang R, Zhang CT: **A Brief Review: The Z-curve Theory and its Application in**  
596 **Genome Analysis.** *Curr Genomics* 2014, **15**(2):78-94.
- 597 30. Sayers EW, Barrett T, Benson DA, Bolton E, Bryant SH, Canese K, Chetvernin V,  
598 Church DM, Dicuccio M, Federhen S *et al*: **Database resources of the National**  
599 **Center for Biotechnology Information.** *Nucleic Acids Res* 2012, **40**(Database  
600 issue):D13-25.
- 601 31. Yang Z: **PAML: a program package for phylogenetic analysis by maximum**  
602 **likelihood.** *Comput Appl Biosci* 1997, **13**(5):555-556.
- 603 32. Katoh K, Toh H: **Recent developments in the MAFFT multiple sequence**  
604 **alignment program.** *Brief Bioinform* 2008, **9**(4):286-298.
- 605 33. Yang Z, Nielsen R: **Estimating synonymous and nonsynonymous substitution**  
606 **rates under realistic evolutionary models.** *Mol Biol Evol* 2000, **17**(1):32-43.
- 607 34. Yang Z, Nielsen R, Goldman N, Pedersen AM: **Codon-substitution models for**  
608 **heterogeneous selection pressure at amino acid sites.** *Genetics* 2000, **155**(1):431-  
609 449.
- 610 35. Yang Z, Wong WS, Nielsen R: **Bayes empirical bayes inference of amino acid sites**  
611 **under positive selection.** *Mol Biol Evol* 2005, **22**(4):1107-1118.
- 612 36. Al-Shahrour F, Diaz-Uriarte R, Dopazo J: **FatiGO: a web tool for finding significant**  
613 **associations of Gene Ontology terms with groups of genes.** *Bioinformatics* 2004,  
614 **20**(4):578-580.



- 615 37. Szklarczyk D, Morris JH, Cook H, Kuhn M, Wyder S, Simonovic M, Santos A,  
616 Doncheva NT, Roth A, Bork P *et al*: **The STRING database in 2017: quality-**  
617 **controlled protein-protein association networks, made broadly accessible.** *Nucleic*  
618 *Acids Res* 2017, **45**(D1):D362-D368.
- 619 38. Bindea G, Mlecnik B, Hackl H, Charoentong P, Tosolini M, Kirilovsky A, Fridman  
620 WH, Pages F, Trajanoski Z, Galon J: **ClueGO: a Cytoscape plug-in to decipher**  
621 **functionally grouped gene ontology and pathway annotation networks.**  
622 *Bioinformatics* 2009, **25**(8):1091-1093.
- 623 39. Sayers EW, Barrett T, Benson DA, Bryant SH, Canese K, Chetvernin V, Church DM,  
624 DiCuccio M, Edgar R, Federhen S *et al*: **Database resources of the National Center**  
625 **for Biotechnology Information.** *Nucleic Acids Res* 2009, **37**(Database issue):D5-15.
- 626 40. Wickham H. **ggplot2: elegant graphics for data analysis.** *Springer* 2016.
- 627 41. Colot OO, C.; Courtellemont, P.: **Information criteria and abrupt changes of**  
628 **probability laws**, vol. 7: Elsevier; 1994.
- 629 42. Murphy K: **The BayesNet Toolbox for Matlab.** *Computing Science and Statistics*  
630 2001, **33**:2001.
- 631 43. Friedman N, Geiger D, Goldszmidt M: **Bayesian Network Classifiers.** *Machine*  
632 *Learning* 1997, **29**(2):131-163.
- 633 44. Lauritzen SL, Spiegelhalter DJ: **Local Computations with Probabilities on**  
634 **Graphical Structures and Their Application to Expert Systems.** *Journal of the*  
635 *Royal Statistical Society: Series B (Methodological)* 1988, **50**(2):157-194.
- 636 45. Sing T, Sander O, Beerenwinkel N, Lengauer T: **ROCR: visualizing classifier**  
637 **performance in R.** *Bioinformatics* 2005, **21**(20):3940-3941.
- 638 46. Franke L, van Bakel H, Fokkens L, de Jong ED, Egmont-Petersen M, Wijmenga C:  
639 **Reconstruction of a functional human gene network, with an application for**  
640 **prioritizing positional candidate genes.** *Am J Hum Genet* 2006, **78**(6):1011-1025.
- 641 47. van Driel MA, Cuelenaere K, Kemmeren PP, Leunissen JA, Brunner HG, Vriend G:  
642 **GeneSeeker: extraction and integration of human disease-related information**  
643 **from web-based genetic databases.** *Nucleic Acids Res* 2005, **33**(Web Server  
644 issue):W758-761.
- 645 48. Furney SJ, Higgins DG, Ouzounis CA, Lopez-Bigas N: **Structural and functional**  
646 **properties of genes involved in human cancer.** *BMC Genomics* 2006, **7**:3.
- 647 49. Accetturo M, Creanza TM, Santoro C, Tria G, Giordano A, Battagliero S, Vaccina A,  
648 Scioscia G, Leo P: **Finding new genes for non-syndromic hearing loss through an**  
649 **in silico prioritization study.** *PLoS One* 2010, **5**(9).
- 650 50. Jain P, Thukral N, Gahlot LK, Hasija Y: **CARDIO-PRED: an in silico tool for**  
651 **predicting cardiovascular-disorder associated proteins.** *Syst Synth Biol* 2015, **9**(1-  
652 2):55-66.
- 653 51. Cheng Y, LeGall T, Oldfield CJ, Dunker AK, Uversky VN: **Abundance of intrinsic**  
654 **disorder in protein associated with cardiovascular disease.** *Biochemistry* 2006,  
655 **45**(35):10448-10460.
- 656 52. Pandey P, Acharya M: **Disease-Phenotype Deconvolution in Genetic Eye Diseases**  
657 **Using Online Mendelian Inheritance in Man.** *Invest Ophthalmol Vis Sci* 2016,  
658 **57**(6):2895-2904.
- 659 53. Forero DA, Prada CF, Perry G: **Functional and Genomic Features of Human Genes**  
660 **Mutated in Neuropsychiatric Disorders.** *Open Neurol J* 2016, **10**:143-148.
- 661 54. Furney SJ, Calvo B, Larranaga P, Lozano JA, Lopez-Bigas N: **Prioritization of**  
662 **candidate cancer genes--an aid to oncogenomic studies.** *Nucleic Acids Res* 2008,  
663 **36**(18):e115.

- 664 55. Mushegian AR, Bassett DE, Jr., Boguski MS, Bork P, Koonin EV: **Positionally**  
665 **cloned human disease genes: patterns of evolutionary conservation and**  
666 **functional motifs.** *Proc Natl Acad Sci U S A* 1997, **94**(11):5831-5836.
- 667 56. Smith NG, Eyre-Walker A: **Human disease genes: patterns and predictions.** *Gene*  
668 2003, **318**:169-175.
- 669 57. Karlin S, Chen C, Gentles AJ, Cleary M: **Associations between human disease genes**  
670 **and overlapping gene groups and multiple amino acid runs.** *Proc Natl Acad Sci U*  
671 *S A* 2002, **99**(26):17008-17013.
- 672 58. Babu MM, van der Lee R, de Groot NS, Gsponer J: **Intrinsically disordered**  
673 **proteins: regulation and disease.** *Curr Opin Struct Biol* 2011, **21**(3):432-440.
- 674 59. Choura M, Rebai A: **The disordered charged biased proteins in the human**  
675 **diseasome.** *Interdiscip Sci* 2019.
- 676 60. Fuchs E: **Evolution and complexity of the genes encoding the keratins of human**  
677 **epidermal cells.** *J Invest Dermatol* 1983, **81**(1 Suppl):141s-144s.
- 678 61. Welbourne TC, Mu X: **Extracellular glutamate flux regulates intracellular**  
679 **glutaminase activity in LLC-PK1-F+ cells.** *Am J Physiol* 1995, **268**(6 Pt 1):C1418-  
680 1424.
- 681 62. Steinert PM, Kartasova T, Marekov LN: **Biochemical evidence that small proline-**  
682 **rich proteins and trichohyalin function in epithelia by modulation of the**  
683 **biomechanical properties of their cornified cell envelopes.** *J Biol Chem* 1998,  
684 **273**(19):11758-11769.
- 685 63. Bern HA, Harkness DR, Blair SM: **Radioautographic Studies of Keratin**  
686 **Formation.** *Proc Natl Acad Sci U S A* 1955, **41**(1):55-60.
- 687 64. Rogers MA, Nischt R, Korge B, Krieg T, Fink TM, Lichter P, Winter H, Schweizer J:  
688 **Sequence data and chromosomal localization of human type I and type II hair**  
689 **keratin genes.** *Exp Cell Res* 1995, **220**(2):357-362.
- 690 65. Wang H, Parry DA, Jones LN, Idler WW, Marekov LN, Steinert PM: **In vitro**  
691 **assembly and structure of trichocyte keratin intermediate filaments: a novel role**  
692 **for stabilization by disulfide bonding.** *J Cell Biol* 2000, **151**(7):1459-1468.
- 693 66. Fudge DS, Gosline JM: **Molecular design of the alpha-keratin composite: insights**  
694 **from a matrix-free model, hagfish slime threads.** *Proc Biol Sci* 2004,  
695 **271**(1536):291-299.
- 696 67. Ruhrberg C, Hajibagheri MA, Parry DA, Watt FM: **Periplakin, a novel component**  
697 **of cornified envelopes and desmosomes that belongs to the plakin family and**  
698 **forms complexes with envoplakin.** *J Cell Biol* 1997, **139**(7):1835-1849.
- 699 68. Langbein L, Rogers MA, Praetzel-Wunder S, Bockler D, Schirmacher P, Schweizer J:  
700 **Novel type I hair keratins K39 and K40 are the last to be expressed in**  
701 **differentiation of the hair: completion of the human hair keratin catalog.** *J Invest*  
702 *Dermatol* 2007, **127**(6):1532-1535.
- 703 69. Steinert PM, Steven AC, Roop DR: **The molecular biology of intermediate**  
704 **filaments.** *Cell* 1985, **42**(2):411-420.
- 705 70. Hatzfeld M, Burba M: **Function of type I and type II keratin head domains: their**  
706 **role in dimer, tetramer and filament formation.** *J Cell Sci* 1994, **107** ( Pt 7):1959-  
707 1972.
- 708 71. Williamson MP: **Nuclear magnetic resonance studies of peptides and their**  
709 **interactions with receptors.** *Biochem Soc Trans* 1994, **22**(1):140-144.
- 710 72. Ramshaw JA, Shah NK, Brodsky B: **Gly-X-Y tripeptide frequencies in collagen: a**  
711 **context for host-guest triple-helical peptides.** *J Struct Biol* 1998, **122**(1-2):86-91.

- 712 73. Pokidysheva E, Boudko S, Vranka J, Zientek K, Maddox K, Moser M, Fassler R,  
713 Ware J, Bachinger HP: **Biological role of prolyl 3-hydroxylation in type IV**  
714 **collagen.** *Proc Natl Acad Sci U S A* 2014, **111**(1):161-166.
- 715 74. Herrmann H, Aebi U: **Intermediate filaments: molecular structure, assembly**  
716 **mechanism, and integration into functionally distinct intracellular Scaffolds.**  
717 *Annu Rev Biochem* 2004, **73**:749-789.
- 718 75. de Jesus AJ, Allen TW: **The role of tryptophan side chains in membrane protein**  
719 **anchoring and hydrophobic mismatch.** *Biochim Biophys Acta* 2013, **1828**(2):864-  
720 876.
- 721 76. Valley CC, Cembran A, Perlmutter JD, Lewis AK, Labello NP, Gao J, Sachs JN: **The**  
722 **methionine-aromatic motif plays a unique role in stabilizing protein structure.** *J*  
723 *Biol Chem* 2012, **287**(42):34979-34991.
- 724 77. Teichmann SA, Chothia C: **Immunoglobulin superfamily proteins in**  
725 **Caenorhabditis elegans.** *J Mol Biol* 2000, **296**(5):1367-1383.
- 726 78. Huang H, Winter EE, Wang H, Weinstock KG, Xing H, Goodstadt L, Stenson PD,  
727 Cooper DN, Smith D, Alba MM *et al*: **Evolutionary conservation and selection of**  
728 **human disease gene orthologs in the rat and mouse genomes.** *Genome Biol* 2004,  
729 **5**(7):R47.
- 730 79. Bustamante CD, Fledel-Alon A, Williamson S, Nielsen R, Hubisz MT, Glanowski S,  
731 Tanenbaum DM, White TJ, Sninsky JJ, Hernandez RD *et al*: **Natural selection on**  
732 **protein-coding genes in the human genome.** *Nature* 2005, **437**(7062):1153-1157.
- 733 80. Thomas PD, Kejariwal A: **Coding single-nucleotide polymorphisms associated**  
734 **with complex vs. Mendelian disease: evolutionary evidence for differences in**  
735 **molecular effects.** *Proc Natl Acad Sci U S A* 2004, **101**(43):15398-15403.
- 736 81. Tu Z, Wang L, Xu M, Zhou X, Chen T, Sun F: **Further understanding human**  
737 **disease genes by comparing with housekeeping genes and other genes.** *BMC*  
738 *Genomics* 2006, **7**:31.
- 739 82. Blekhman R, Man O, Herrmann L, Boyko AR, Indap A, Kosiol C, Bustamante CD,  
740 Teshima KM, Przeworski M: **Natural selection on genes that underlie human**  
741 **disease susceptibility.** *Curr Biol* 2008, **18**(12):883-889.
- 742 83. Parra EJ: **Human pigmentation variation: evolution, genetic basis, and**  
743 **implications for public health.** *Am J Phys Anthropol* 2007, **Suppl 45**:85-105.
- 744 84. Rees JL: **Genetics of hair and skin color.** *Annu Rev Genet* 2003, **37**:67-90.
- 745 85. Lucock M: **Folic acid: nutritional biochemistry, molecular biology, and role in**  
746 **disease processes.** *Mol Genet Metab* 2000, **71**(1-2):121-138.
- 747 86. Off MK, Steindal AE, Porojnicu AC, Juzeniene A, Vorobey A, Johnsson A, Moan J:  
748 **Ultraviolet photodegradation of folic acid.** *J Photochem Photobiol B* 2005,  
749 **80**(1):47-55.
- 750 87. Jablonski NG, Chaplin G: **The evolution of human skin coloration.** *J Hum Evol*  
751 2000, **39**(1):57-106.
- 752 88. Williams JD, Jacobson EL, Kim H, Kim M, Jacobson MK: **Folate in skin cancer**  
753 **prevention.** *Subcell Biochem* 2012, **56**:181-197.
- 754 89. Aerts S, Lambrechts D, Maity S, Van Loo P, Coessens B, De Smet F, Tranchevent  
755 LC, De Moor B, Marynen P, Hassan B *et al*: **Gene prioritization through genomic**  
756 **data fusion.** *Nat Biotechnol* 2006, **24**(5):537-544.
- 757 90. Erlich Y, Edvardson S, Hodges E, Zenvirt S, Thekkat P, Shaag A, Dor T, Hannon GJ,  
758 Elpeleg O: **Exome sequencing and disease-network analysis of a single family**  
759 **implicate a mutation in KIF1A in hereditary spastic paraparesis.** *Genome Res*  
760 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

767 **FIGURES TITLES:**

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

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

771 Figure 3: Frequency of positive selected amino acids in the skin and non skin group of genes.

772 (\*)  $P < 0.001$ . Comparison between amino acids frequency was performed by means of chi-  
773 squared test

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

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

777

778 **TABLES TITLES:**

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

780 Table 2: Median and statistical analysis for peptide composition

781 Table 3: Median values and statistical analysis of ortholog conservation

782

783

784

785

786

787

788

789 **ADDITIONAL FILES:**

790 **Additional-File1.xlsx:**

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

792 **Additional-File2.xlsx:**

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

794 Supplementary table 3: Mean X, Y and Z components

795 **Additional-File3.xlsx:**

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

797 StringDB

798 **Additional-File4.xlsx:**

799 Supplementary table 5: Selective pressure study

800 Supplementary table 6: Selective pressure study

801 **Additional-File5.xlsx:**

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

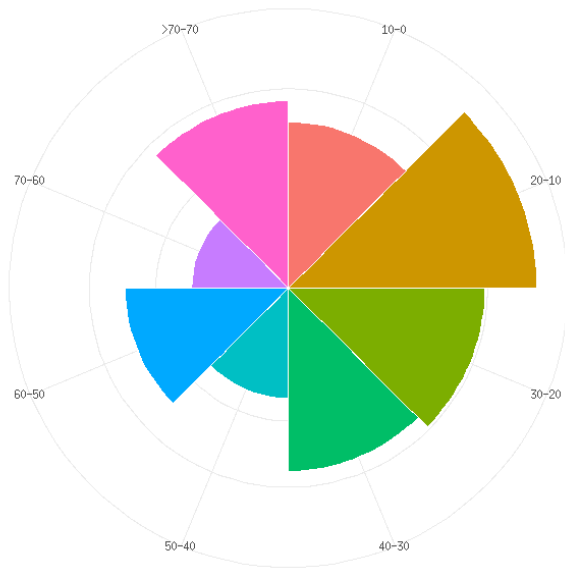
803 StringDB

804 **Additional-File6.xlsx**

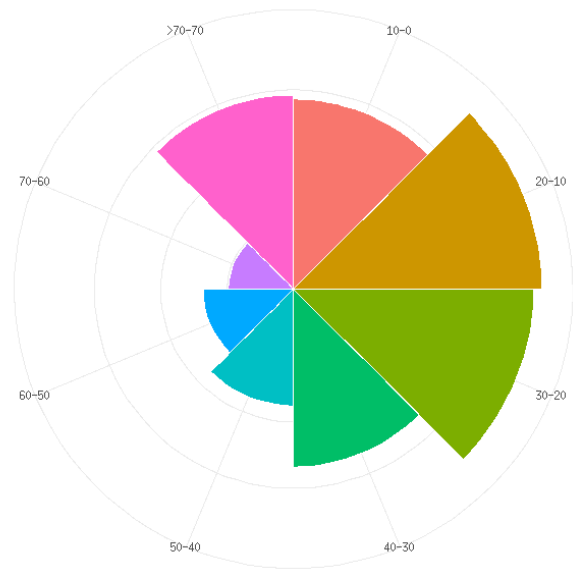
805

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

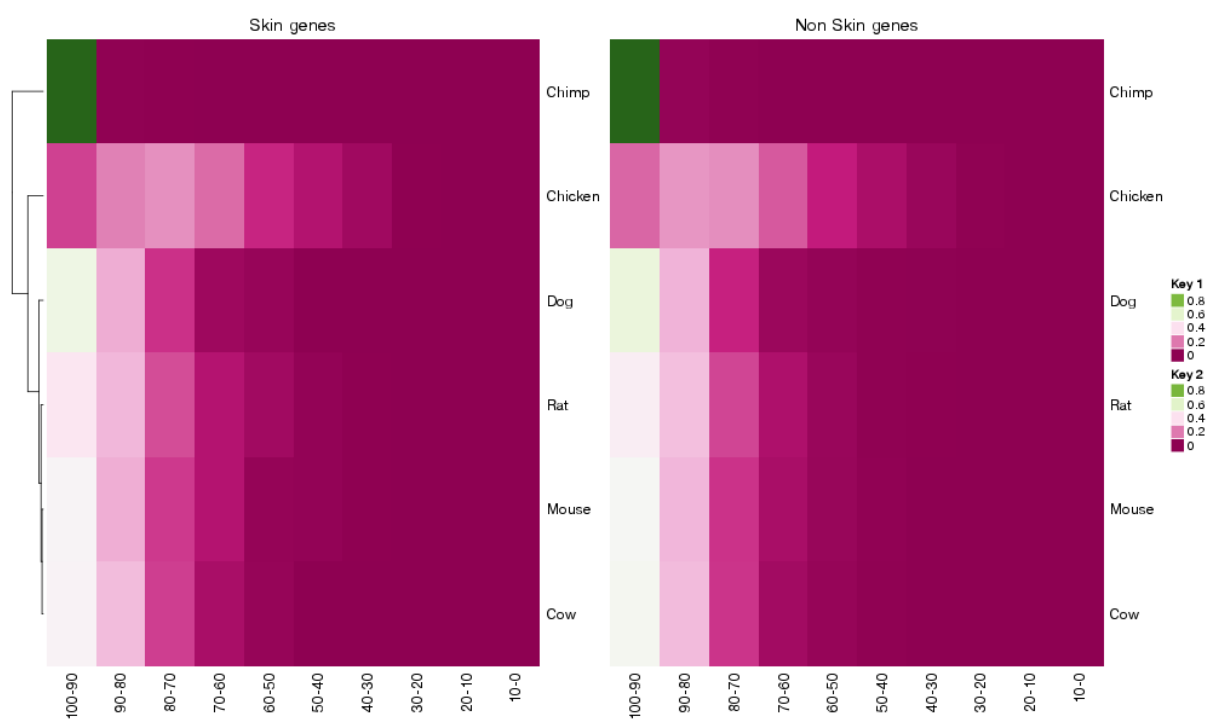
### Exon count in Skin genes



### Exon count in Non-Skin genes

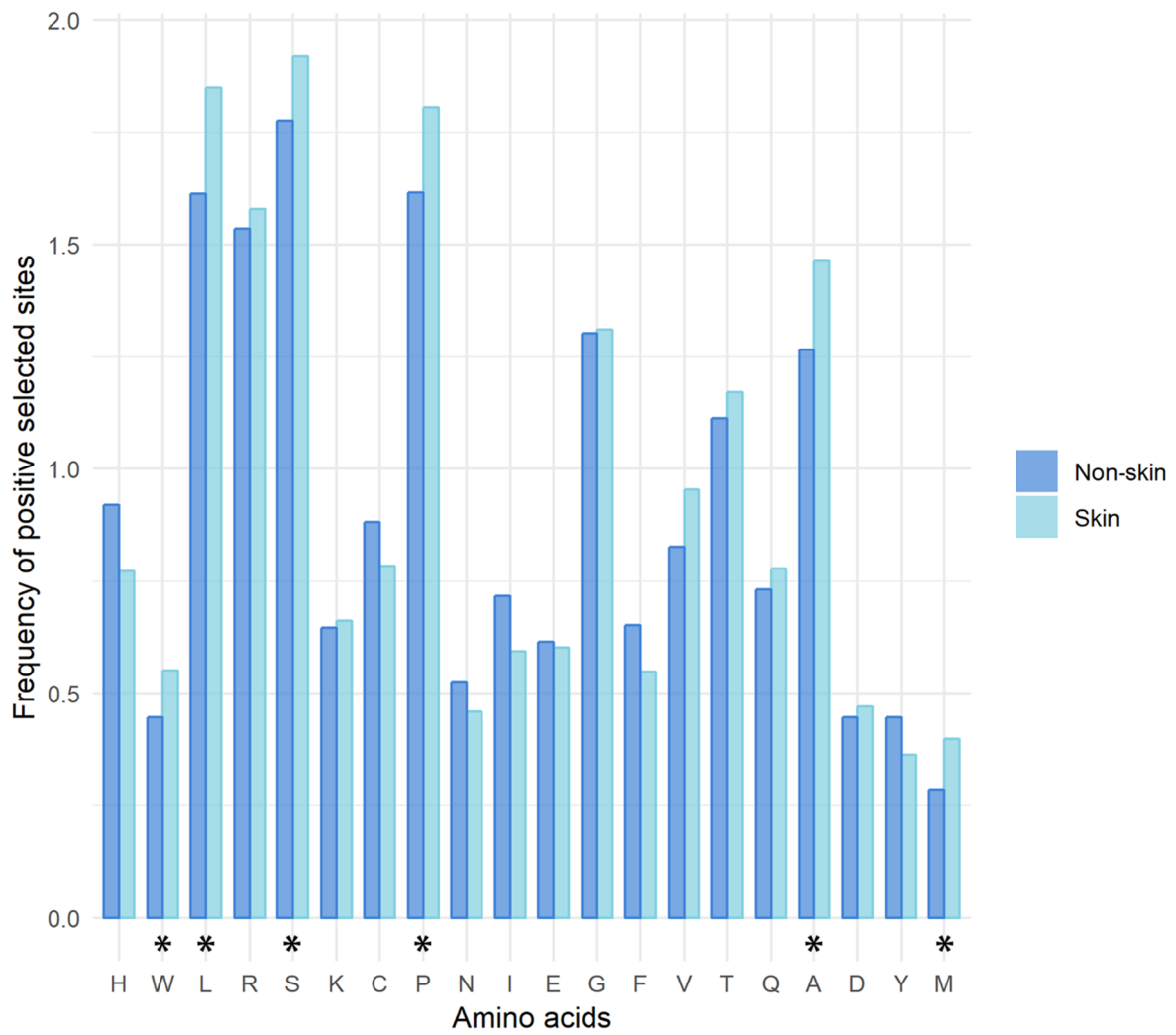


**Figure 1: Exon count distribution within skin and non skin genes group**



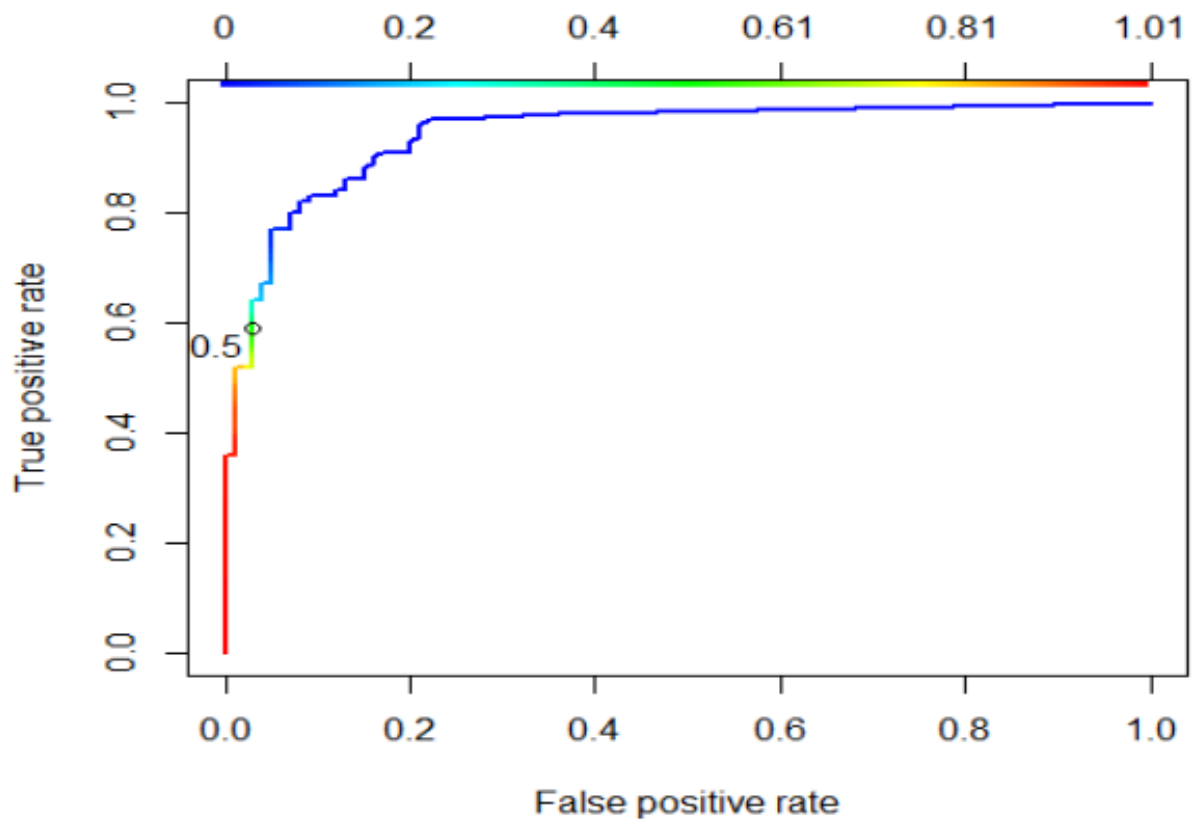
**Figure 2: Average % identity of orthologs with Human genes in the two groups**





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

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

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

**Table 2: Median and statistical analysis for peptide composition**

	<b>Cysteine</b> (%)	<b>Glutamine</b> (%)	<b>Isoleucine</b> (%)	<b>Methionine</b> (%)	<b>Phenylalanine</b> (%)
<b>Genodermatoses</b>	0.023	0.045	0.043	0.021	0.038
<b>Others</b>	0.022	0.044	0.045	0.023	0.039
<b>Wilcox p-value</b>	<b>0.039</b>	<b>0.032</b>	<b>0.046</b>	<b>8.2 10<sup>-4</sup></b>	<b>0.025</b>
<b>KS p-value</b>	<b>0.007</b>	0.094	<b>0.031</b>	<b>0.002</b>	<b>0.049</b>

**Table 3: Median values and statistical analysis of ortholog conservation**

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