Haplotype analysis of distantly related populations implicates corneodesmosin in psoriasis susceptibility


MATERIALS AND METHODS

Subjects
The UK dataset consisted of 171 parent-offspring trios of European descent, which have been described elsewhere. The Gujurati Indian cohort included 77 patients, 77 controls, and 30 parent-offspring trios, all ascertained as reported by Asumalaih et al. The unaffected control population typed for 96 unrelated UK subjects of European descent. For experimental confirmation of haplotypes, two generation pedigrees, composed of 29 and 40 subjects, were analysed.

All subjects provided their informed consent for participation in this study, which was undertaken following approval of the Guy’s and St Thomas’ Hospitals Ethics Committee of Kings College, London.

Key points

- Psoriasis is a multifactorial skin disorder. The major susceptibility locus maps to chromosome 6p21, within a genomic segment that contains three genes carrying disease associated alleles: HLA-C, α-helix coiled coil rod homologue (HCR), and corneodesmosin (CDSN).
- We have recently identified a rare susceptibility haplotype (cluster D), originating from a double recombination event that replaced HCR risk alleles, while preserving HLA-C and CDSN disease associated SNPs. This suggested that CDSN SNPs might be required to confer psoriasis susceptibility, in conjunction with HLA-C risk alleles.
- Here, we have detected cluster D as a risk chromosome, through the analysis of an Indian case/control dataset (77 patients, 77 controls; p=0.001). We have subsequently analysed six coding and 10 non-coding CDSN SNPs in 171 UK and 30 Indian parent-offspring trios. We have identified a 16 SNP chromosome showing significant disease association in both datasets (p<10^-10 and p=3.9 × 10^-6, respectively) and occurring on both sets of haplotypes (clusters D and E) known to confer psoriasis susceptibility.
- Altogether, our results support a disease model requiring the presence of CDSN susceptibility alleles on haplotypes bearing HLA-C related risk SNPs.

SNP identification

Four PCR primer pairs were designed based on GenBank sequence L28015, in order to amplify the CDSN coding region and the entire 3’UTR. A further primer pair was designed on the CDSN upstream region contained in genomic sequence AP000510, in order to amplify 785 bp of putative gene promoter. All primer sequences are available on our web site (www.leicester.ac.uk/ge/rt7/index2.html). PCR products were sequenced using Big Dye Terminators (Applied Biosystems) and electrophoresed on an ABI 377 automated sequencer (Applied Biosystems). SNPs were identified by visual inspection of chromatograms.

SNP typing

Samples were amplified using the primer pairs reported on our web site and the resulting PCR products were blotted onto Hybond N nylon membranes (Amersham). For each SNP, a pair of allele specific oligonucleotides (ASOs) was synthesised, end labelled, and used to probe dot blots of the corresponding PCR products, as described elsewhere. Synonym nucleotide changes and SNPs presenting a minor allele frequency <0.25 were excluded from the genotyping panel.
Statistical analyses

Family based association analysis of SNP haplotypes was performed by using the TRANSMIT program. Owing to computational limitations, overlapping segments of 7-9 SNPs were first determined. These were then pieced together in succession by identification of unique SNPs in overlapping regions until full length haplotypes were generated.

The PHASE program was used to derive haplotypes from the diploid genotypes of the Gujarati subjects. The frequencies of haplotypes predicted for cases and controls were compared using a chi-square test with 1 degree of freedom. Analysis of LD conservation among controls was performed using ad hoc software written in True Basic 4.1 by Alec Jeffreys. The haplotypes segregating in the three generation families were derived manually, by minimising recombination events.

Bioinformatic analyses

The sequence of the minimal promoter region was predicted using a Neural Network tool (available at http://www.fruitfly.org/seq_tools/promoter.html). The putative location of transcription factor binding sites was assessed by analysing 900 bp of upstream CDSN sequence with the TESS software (www.chb.upenn.edu/cgi-bin/tess/tess33). The sequence of the minimal promoter region was predicted using a chi-square test with 1 degree of freedom. The mouse sequence used for this analysis (nucleotides 117010-117910 in GenBank sequence AF111103) was identified by a BLAST search (www.ncbi.nlm.nih.gov/Blast) of the Mus musculus Cdsn gene was similarly analysed, and the results of human and mouse TESS analyses were compared, searching for conserved binding sites. The mouse sequence used for this analysis (nucleotides 117010-117910 in GenBank sequence AF111103) was identified by a BLAST search (www.ncbi.nlm.nih.gov/Blast) of the Mus musculus genome against transcript ENSMUST00000044804 from the Ensembl database (www.ensembl.org).

The P.IX tool (www.hgmp.mrc.ac.uk/registered/seqapp/pix/) was used to perform secondary structure analysis and to search the SBASE, Pfam, and PRODOM domain databases. Sites of N-glycosylation were predicted by the ScanProsite program (http://ca.expasy.org/tools/scanprosite/) and confirmed through the NetNGlyc web server (www.cbs.dtu.dk/services/NetNGlyc/).

RESULTS

PSORS1 haplotype analysis

We used the PHASE program to derive haplotypes from the SNP9, HCR*269T and CDSN*619, CDSN*1236, CDSN*1243 (defining CDSN*TTC) genotypes that we had previously generated in a case/control data set of Gujarati Indian descent. The results of the chi-square analysis implemented on PHASE showed that the diploid genotypes of the Gujurati subjects. The frequencies of haplotypes predicted for cases and controls were compared using a chi-square test with 1 degree of freedom. The mouse sequence used for this analysis (nucleotides 117010-117910 in GenBank sequence AF111103) was identified by a BLAST search (www.ncbi.nlm.nih.gov/Blast) of the Mus musculus genome against transcript ENSMUST00000044804 from the Ensembl database (www.ensembl.org).

The P.IX tool (www.hgmp.mrc.ac.uk/registered/seqapp/pix/) was used to perform secondary structure analysis and to search the SBASE, Pfam, and PRODOM domain databases. Sites of N-glycosylation were predicted by the ScanProsite program (http://ca.expasy.org/tools/scanprosite/) and confirmed through the NetNGlyc web server (www.cbs.dtu.dk/services/NetNGlyc/).

In order to exclude the possibility that the CDSN*TTC association might reflect LD with a coding variant located in a downstream gene, we typed two SNPs within the GTF2H4 gene lying 200 kb distal to CDSN. The analysis of our original 171 trios UK sample showed that these two marker alleles freely associate with all PSORS1 haplotypes, placing GTF2H4 outside the segment shared by clusters D and E, and defining an interval that encompasses the entire CDSN and STG genes, together with the first two exons of SEEK1 (fig 1). However, resequencing of STG has failed to identify any high frequency SNP while SEEK1 exons 1 and 2 lie within the gene 5'UTR (see GenBank sequence AP000510). Thus, CDSN appears to be the most likely positional candidate for the region shared by clusters D and E.

CDSN gene resequencing

Resequencing of eight unrelated patients carrying PSORS1 risk chromosomes identified 17 coding variants, all of which had been previously described. We also observed 10 polymorphisms in the putative promoter and the 3'UTR. Eight of these variants are novel, this study being the first systematic survey of CDSN non-coding regions. A complete list of the polymorphisms observed in resequenced patients is available on our web site.

LD conservation patterns

Fig 2 shows the results of diploid LD analysis between 14 SNP loci presenting a minor allele frequency >0.15. LD appears to be maintained along the upstream region and the proximal CDSN coding sequence (D' >0.8, columns -636 to 619 in fig 2), but it tends to decay with distance in the distal portion of the gene (D' < 0.4, columns 1243 to 1748).

Bioinformatic analyses

In order to generate a framework model allowing a preliminary assessment of SNP functional impact, we used bioinformatic tools to analyse the CDSN genomic sequence. Analysis by neural networks predicted a 41 bp polymerase II promoter site (nucleotides -21608-21648 in Genbank sequence AP000510; score = 0.96), containing a CATAAA box. This interval lies within a conserved 101 bp sequence showing 88% identity with the homologous mouse region. Analysis of human and murine CDSN upstream sequences highlighted conservation of potential binding sites for AP1 (-73/-67; -823/-817), Sp1 (-251/-246; -313/-308; -588/-593; -909/-915), CACCC box (-106/-101; -301/-296; -459/-454), and AP2-alpha (-851/-844) transcription factors, all of which have been implicated in keratinocyte differentiation.

Secondary structure analyses of the coding sequence identified an alternation of short beta-sheets and longer coil regions. Searches of domain database failed to identify any match to consensus sequences. An analysis with the ScanProsite program predicted glycosylation sites corresponding to Asn residues 134 and 156. These results were confirmed by the NetNGlyc server.

CDSN haplotype analysis

We typed a panel of 16 SNPs in our original UK data set and in 30 Gujurati parent-offspring trios. The resulting CDSN haplotypes are shown in table 2. Six chromosomes occurring with
a frequency >2% accounted for ~90% of haplotypes, both in the UK and the Gujurati population. Haplotype I was significantly over-transmitted to affected offspring in both data sets (p<10^{-10} and p=3.9×10^{-3}, respectively). Haplotype II, which is likely to have originated from I by ancestral mutation, was over-transmitted in the UK sample (p=3.9×10^{-3}), while showing a non-significant transmission increase in the smaller Gujurati data set. The cumulative frequency of haplotypes I and II was remarkably similar in the two populations, summing up to 40.1% in the UK and 41.9%

Figure 1 Occurrence of recombination on PSORS1 risk haplotypes. (A) Alignment of clusters D and E. Haplotypes are depicted as a succession of 59 squares, each corresponding to a SNP. Yellow and green fillings represent the two alleles at each locus, with risk variants labelled by black inserts. The Rec1 and Rec2 sites of recombination correspond to the boundaries of the genomic segment of haplotype divergence. (B) Closer view of the distal PSORS1 region, showing the location of positional candidates. The two arrows represent the opposite DNA strands genes are transcribed from.

Figure 2 Pairwise LD analysis of CDSN SNPs.
among the Gujurati. Haplotype V showed the highest number of substitutions with respect to I (11/16 variants), and was significantly under-transmitted to patients in the UK sample (p=4 × 10⁻⁴), while showing a lack of transmission of borderline significance (p=5 × 10⁻²) among the Gujurati.

TRANSMIT analysis of SNP9, HCR*269, and CDSN genotypes showed that haplotype I is found on disease bearing clusters D and E, together with SNP9C (table 3). This combination is unique to risk chromosomes as neutral clusters G1 and G2 respectively carry SNP9T + haplotype I and SNP9C + haplotype VI.

To validate experimentally the haplotypes derived by TRANSMIT, we typed SNP9, HCR*269, and the 16 CDSN variants in two extended three generation pedigrees, where we were able to confirm the segregation of all the haplotypes observed in the UK data set.

<table>
<thead>
<tr>
<th>Table 2A</th>
<th>CDSN chromosomes occurring with &gt;2% frequency; haplotype alignment</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>5' region</td>
</tr>
<tr>
<td></td>
<td>−712</td>
</tr>
<tr>
<td>I (Allele 1.11)*</td>
<td>c</td>
</tr>
<tr>
<td>II (Allele 1.21)</td>
<td>c</td>
</tr>
<tr>
<td>III (Allele 1.4)</td>
<td>c</td>
</tr>
<tr>
<td>IV (Allele 2.11)</td>
<td>t</td>
</tr>
<tr>
<td>V (Allele 2.2)</td>
<td>c</td>
</tr>
<tr>
<td>VI (Allele 1.5)</td>
<td>t</td>
</tr>
</tbody>
</table>

*Allele number according to the classification of CDSN coding haplotypes proposed by Guerrin et al.²⁷

| Table 2B | CDSN chromosomes occurring with >2% frequency; haplotype transmission to patients |
|---|---|---|---|---|---|---|---|---|---|
| | UK | Gujarati | UK | Gujarati | UK | Gujarati | UK | Gujarati |
| I | Over-transmitted | Over-transmitted | 31.8 | 35.8 | 42.8 | 8.3 |
| II | Over-transmitted | Neutral | 8.3 | 6.1 | 8.3 |
| III | Neutral | Neutral | 12.9 | 2.0 | 6.1 |
| IV | Under-transmitted | Neutral | 16.2 | 13.3 | 12.4 | 3.8 |
| V | Under-transmitted | Under-transmitted | 13.6 | 6.9 | 12.4 | 3.8 |
| VI | Under-transmitted | Neutral | 6.8 | 23.2 | 5.1 |

| Table 3 | CDSN haplotypes present on PSORS1 clusters |
|---|---|---|
| SNP9C | HCR*269T | CDSN haplotype |
| Cluster A | − | − | V |
| Cluster B | − | − | III, VI |
| Cluster C | − | + | IV |
| Cluster D | + | − | I |
| Cluster E | + | + | I, II |
| Cluster G1 | − | − | I |
| Cluster G2 | + | − | VI |

DISCUSSION

CDSN alleles have previously been implicated in psoriasis pathogenesis,¹⁴ ¹⁵ ¹⁷ ²⁹ but disease associations have generally been ascribed to LD with HLA-Cw*0602,¹⁵ ³⁰ which is currently recognised as the marker conferring the highest psoriasis risk.¹³ However, we recently showed that HLA-C and CDSN SNPs do not show linkage disequilibrium when non-disease bearing chromosomes are analysed.¹⁹ This suggested that the associations with HLA-C and CDSN SNPs might reflect a requirement of both sets of alleles on risk haplotypes. This hypothesis was also supported by the conservation of CDSN*TTC on cluster D.¹⁹ Here, we have validated cluster D as a risk chromosome, through the analysis of an Indian case-control data set. Studying this distantly related white population disclosed a distinctive repertoire of ancestral haplotypes. Thus, the identification of a common, neutral chromosome carrying SNP9C in the absence of CDSN*TTC (cluster G2) and the observation of significant cluster D increase among affected subjects concur to support a requirement of CDSN risk alleles on psoriasis susceptibility haplotypes.

The corneodesmosin gene being extremely polymorphic, we have derived SNP haplotypes spanning 5.3 kb of CDSN genomic sequence, in order to assess the contribution of additional variants that might be in LD with CDSN*TTC risk allele. Our analysis has identified a 16 SNP chromosome (haplotype I) that shows significant disease association both in the UK and the Gujurati data set. We also observed two further haplotypes (II and III) which are likely to have originated from I, by ancestral mutation of nucleotides 1215 (II) and 619 (III). Haplotype II shows increased transmission in both the UK and the Gujurati data sets, whereas chromosome III appears to be neutral in our sample, but encompasses a
shorter haplotype (allele 1.4), which shows disease association in the Japanese population. The presence of different CDSN*619 and CDSN*1215 alleles on the otherwise identical haplotypes I, II, and III suggests that these SNPs do not play a major pathogenic role.

Haplotypes IV, V, and VI display a higher degree of diversity, each carrying at least five substitutions with respect to the other two. Haplotype V shows the highest number of substitutions compared to I (11/16 variants), and is under-transmitted in both the UK and the Gujarati sample. Interestingly, the frequency of CDSN allele 2.21, which is contained in haplotype V, also appears to be increased among Japanese controls, if compared to patients. Altogether, these observations suggest that haplotype V may have a protective effect in a wide range of populations, CDSN being over-expressed in psoriatic lesions, it is of note that haplotypes I and V differ at 8/10 variants localised to potential regulatory regions. These include two SNPs (CDSN*–31 and CDSN*–22) lying within the predicted RNA polymerase II promoter and four variants from the 3′UTR. This latter group of alleles might alter gene expression levels by affecting mRNA stability, as shown for a number of non-coding SNPs involved in complex disease pathogenesis. As we have only resequenced 785 bp upstream of the CDSN coding region, we cannot exclude a contribution of additional regulatory variants lying in the distal genetic promoter. However, the pattern of LD decay observed in UK controls suggests that such SNPs would be unlikely to be in strong linkage disequilibrium with CDSN*TTT, which lies in the distal portion of the gene.

Haplotypes I and V also diverge at three (CDSN*442, CDSN*1243, and CDSN*1593) of the four coding variants that are common to chromosomes I, II, and III. CDSN*442 appears to introduce a potential site for N-glycosylation, a process which is known to take place in vivo, as N-linked oligosaccharides account for ~10% of CDSN molecular weight. The potential impact of CDSN*1243 and CDSN*1593 cannot be easily predicted, since both SNPs occur on open coil regions matching no conserved functional domain. However, a corneodesmosin 3D structural model, recently generated by our group using fold recognition methods, predicted that the amino acid affected by CDSN*1243 localises to an exposed region of the protein (Allen et al, manuscript in preparation), where a change in residue polarity might hamper the action of the proteases that digest CDSN during skin desquamation. Clearly, bioinformatic predictions have to be interpreted with caution; nonetheless these analyses have generated a number of working hypotheses that can now be assessed experimentally in order to dissect the functional contribution of regulatory and coding SNPs lying on CDSN risk haplotypes.

The localisation of haplotype I to clusters D and E, both of which also contain HLA-C related risk alleles, is consistent with our initial two locus model of the PSORS1 interval. Interestingly, haplotype III only occurs on cluster B, which lacks the second set of disease alleles. This may explain why haplotype III does not show psoriasis association among the white population, while conferring disease susceptibility in the Japanese, where it occurs on the background of an ancestral haplotype bearing an HLA-C risk allele.

Clustering of complex disease loci in regions defined by a single peak of linkage has previously been described, with accumulating evidence for multiple MHC genes conferring susceptibility to type 1 diabetes and multiple sclerosis. The occurrence of separate, closely linked disease susceptibility loci adds to the complexity of common disease analysis and warrants the study of divergent populations presenting with different patterns of linkage disequilibrium conservation.

ACKNOWLEDGEMENTS

The authors wish to thank all the psoriasis families whose participation made this project possible. The authors would also like to thank Professor Sir Alec Jeffreys for access to software for LD diploid analysis. This research was supported by a Wellcome Trust grant to RC and JNWNB (No 056713/Z/99/Z) and a Medical Research Council (MRC) UK Cooperative Group Grant. FC is a recipient of a Wellcome Trust Travelling Research Fellowship.

Supplementary tables 1 and 2 can be found on our web site, www.jmg.net/supplemental

Authors’ affiliations

F Copan, K T Tool, J C Evans, S Patel, R C Trembath, Division of Medical Genetics, University of Leicester, Leicester, UK
F Copan, Division of Human Genetics, “Tor Vergata” University of Rome, Italy
M H Allen, J N W N Barker, St John’s Institute of Dermatology, Kings College, London, UK
D Tillman, D Burden, Department of Dermatology, Western Infirmary, Glasgow, UK

Correspondence to: Professor R Trembath, Division of Medical Genetics, Department of Genetics and Medicine, Adrian Building, University Road, Leicester LE1 7RH, UK; rtremba@hgmp.mrc.ac.uk

REFERENCES

Browsing made easy

Collections

With a single click Collections allows you to find all articles that have been published in your chosen subject. Select from over 200 clinical and non-clinical topic collections and/or cross search other specialist journals, the BMJ and Cochrane Reviews

www.jclinpath.com

www.jmedgenet.com