Long-range comparison of human and mouse *Sprr* loci to identify conserved noncoding sequences involved in coordinate regulation

Natalia Martin, Satyakam Patel and Julia A. Segre

Genome Res. 2004 14: 2430-2438
Access the most recent version at doi:10.1101/gr.2709404

Letter

Long-range comparison of human and mouse *Sprr* loci to identify conserved noncoding sequences involved in coordinate regulation

Natalia Martin, Satyakam Patel, and Julia A. Segre[1]

National Human Genome Research Institute, National Institutes of Health, Bethesda, Maryland 20892, USA

Mammalian epidermis provides a permeability barrier between an organism and its environment. Under homeostatic conditions, epidermal cells produce structural proteins, which are cross-linked in an orderly fashion to form a cornified envelope (CE). However, under genetic or environmental stress, specific genes are induced to rapidly build a temporary barrier. Small proline-rich (SPRR) proteins are the primary constituents of the CE. Under stress the entire family of 14 *Sprr* genes is upregulated. The *Sprr* genes are clustered within the larger epidermal differentiation complex on mouse chromosome 3, human chromosome 1q21. The clustering of the *Sprr* genes and their upregulation under stress suggest that these genes may be coordinately regulated. To identify enhancer elements that regulate this stress response activation of the *Sprr* locus, we utilized bioinformatic tools and classical biochemical dissection. Long-range comparative sequence analysis identified conserved noncoding sequences (CNSs). Clusters of epidermal-specific DNaseI-hypersensitive sites (HSs) mapped to specific CNSs. Increased prevalence of these HSs in barrier-deficient epidermis provides in vivo evidence of the regulation of the *Sprr* locus by these conserved sequences. Individual components of these HSs were cloned, and one was shown to have strong enhancer activity specific to conditions when the *Sprr* genes are coordinately upregulated.

[Supplemental material is available online at www.genome.org.]

The complete sequencing of genomes has opened new avenues of research into the full interpretation of these codes. Not only will this information facilitate a comprehensive recognition of transcribed sequences, but it also enables the prediction of regulatory sequences responsible for the correct spatial and temporal gene expression. The identification of evolutionarily conserved noncoding sequences (CNSs) among orthologous genomic regions of different species has proven to be a powerful guide for the localization of *cis*-acting transcriptional regulatory elements. The elucidation of *cis*-regulatory elements in a genome is an important starting point for the identification of the specific binding factors and therefore for the subsequent discovery of the molecular pathways that control the mechanisms of cellular differentiation and physiology (for review, see Pennacchio and Rubin 2001; Frazer et al. 2003; Ureta-Vidal et al. 2003). With the release of the initial mouse and human sequences, there has been much interest in the use of comparative sequence analysis to identify gene regulatory elements. Between these two genomes, ~5% of the sequence is under positive selection. As this is much higher than the percent of the genomes predicted to be protein-coding sequences, the other regions are postulated to encode untranslated regions of genes, RNA genes, or regulatory elements under selection for biological function (Waterston et al. 2002).

A handful of studies in mammals have tested the hypothesis that CNSs are *cis*-regulatory modules that control the expression of nearby genes (Göttgens et al. 2001, 2002; Loots et al. 2000; Santagati et al. 2003; Santini et al. 2003; Fraser et al. 2004). A priori, noncoding sequences may be conserved because the sequences have not diverged in the evolutionary time between the organisms or because of strong selection to maintain *cis*-regulatory elements. After determining the local rate of sequence conservation, analysis focuses on regions with statistically lower rates of substitution. The full regulatory potential of sequences may not be appreciated since analysis is constrained by the assays to query functional significance.

The identification of long-range regulatory elements has been essential to understand the coordinate gene regulation of tandemly arrayed genes such as the β-globin, interleukins 4, 13, and 5, or Hox clusters (Grosveld et al. 1987; Loots et al. 2000; Santini et al. 2003). Still under investigation is how the sharing of common *cis*-regulatory elements is a selective force to maintain the physical proximity of genes tandemly arrayed. An open question is whether the mechanisms of coordinate regulation identified for these well studied gene clusters extend to other regions of the genome. The epidermal differentiation complex (EDC), a 2.2-Mb region of mouse chromosome 3 and human chromosome 1q21, consists of multiple genes that are expressed during keratinocyte differentiation (Cabral et al. 2001; Elder and Zhao 2002). Within the EDC are two distinct classes of genes. The central 1.2 Mb contains genes encoding structural proteins such as involucrin, loricrin, profilaggrin, the families of small proline-rich (SPRR) proteins, and the late envelope proteins. In response to the increased calcium levels of differentiating epidermal cells, transglutaminase cross-links these proteins. Flanking the genes encoding these structural proteins are genes of the *S100* family, which encode small calcium-binding proteins that mediate calcium signaling during epithelial cell differentiation (Fig. 1A; South et al. 1999).

Here we focus on the regulation of the family of *Sprr* genes, tandemly arrayed in 220 kb of the EDC. The SPRR proteins have

[1]**Corresponding author.**
E-mail jsegre@nhgri.nih.gov; fax (301) 402-4929.
Article and publication are at http://www.genome.org/cgi/doi/10.1101/gr.2709404.

14:2430–2438 ©2004 by Cold Spring Harbor Laboratory Press; ISSN 1088-9051/04; www.genome.org

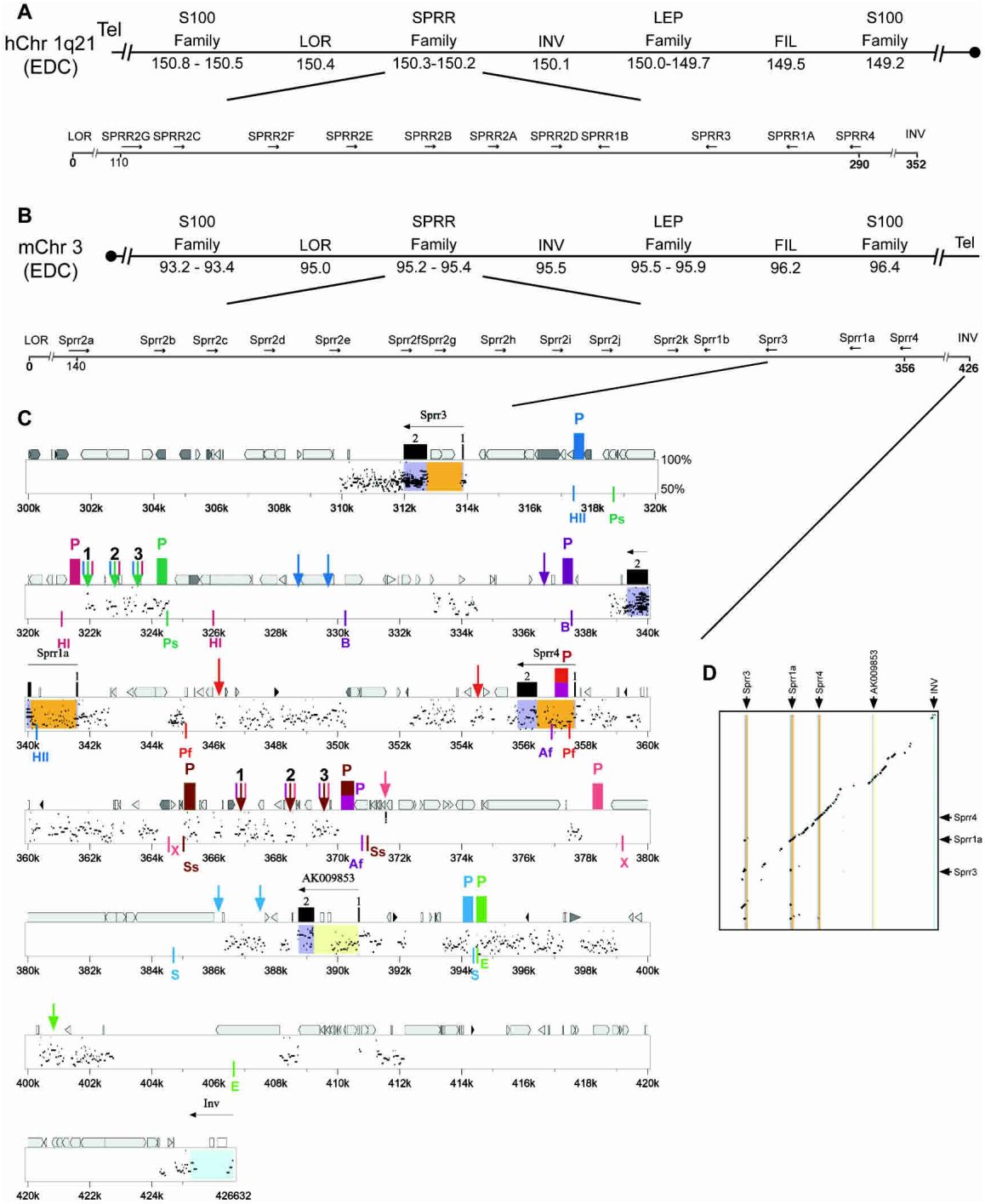


Figure 1. Physical map of comparative sequence analysis of the *Sprr* locus. (*A*) Representation of genes and families of genes found in the epidermal differentiation complex (EDC) on human chromosome 1q21. Sizes are in megabases from centromere. Gene order of *Sprr* locus are relative to flanking genes *Loricrin* and *Involucrin*. (*B*) Representation of genes and families of genes found in the EDC on mouse chromosome 3. Sizes are in megabases from centromere. (*C*) Percent identity plots (PIPs) comparing the mouse and human sequences of the *Sprr* locus. Analysis of the PipMaker program with the "all matches" option and the mouse sequence as the reference. The mouse genomic sequences is shown on the *x*-axis. Sizes are in kilobases relative to *Loricrin.* Percent sequence identity (50%–100%) is shown on the *y*-axis. In addition to the *Sprr* genes (introns, orange boxes; exons, blue boxes), predicted transcripts determined by BLAST are also shown (introns, yellow boxes; exons, blue boxes). The repeats of mouse sequence are depicted as follows: black pointed boxes, LINE2s; light gray pointed boxes, LINE1s; dark gray pointed boxes, LTRs; black triangles, MIRs; light gray triangles, SINEs other than MIRs; dark gray triangles, other repeats; white boxes, simple repeats. CpG islands are represented by short boxes. DNaseI HS results are also depicted in the figure: probes (P) are represented by colored boxes. Restriction enzymes used for each probe are indicated in the same color to the *left* and to the *right* of each probe below the plot (Ap, ApaI; HII, HpaII; HI, HpaI; Ps, PstI; B, BamHI; Ah, AhdI; Af, AflII; Pf, PflFI; Ss, SspI; X, XmaI; S, StuI; E, EcoRV). The DNaseI Hs are represented by arrows *above* the plot. To improve clarity of the figure, probes that did not reveal an HS are not displayed. (*D*) Dot matrix comparisons of human and mouse sequence in the *Sprr* region. Conservation in the dot blot is represented by black dots. The *x*-axis has a graphical representation of the mouse sequence in which exons are represented by coloration as in Figure 1C.

diversified their amino acid sequence and expression pattern, enabling the body to construct slightly different types of permeability barriers as needed for backskin, mouth, tongue, etc. Although the genes have unique expression patterns in normal tissues, under stressed conditions the entire family of genes can be upregulated. This upregulation has been documented in UV-irradiated cells and papilloma tissue (Song et al. 1999; Cabral et al. 2001) and in a genetic model of epidermal barrier dysfunction, the targeted ablation of the transcription factor Klf4, investigated in our laboratory (Patel et al. 2003). The clustering of the *Sprr* genes and their misregulation suggest that these genes may be coordinately regulated. Using a combination of comparative sequence analysis and classical biochemical approaches, we characterized an enhancer specific to the coordinate upregulation of the *Sprr* locus.

Results

Identification of candidates for *cis*-regulatory elements by comparative sequencing analysis

As stated above, the EDC consists of multiple genes that are expressed during keratinocyte differentiation. In the mouse and human, 14 and nine genes, respectively, encode SPRR proteins, which are tandemly arrayed within the EDC (Fig. 1A,B).

To identify regulatory elements of the *Sprr* locus, we first identified conserved sequence within this locus and the flanking sequences. We established *Loricrin* and *Involucrin* as the boundaries of our analysis because they are the closest known genes on either side of the *Sprr* locus and their expression is not affected in the epidermis of barrier-deficient mice that upregulate the entire *Sprr* locus (~426 kb in mouse and ~352 kb in human) (Patel et al. 2003). A priori, the larger size of the mouse interval compared to the human interval is unusual, since on average the mouse genome is 14% smaller than the human genome. However, lineage-specific expansion of gene families was noted with the initial sequencing and comparative analysis of the mouse and human genomes (Waterston et al. 2002). Our previously published analysis of the SPRR coding sequences suggested that this expansion occurred independently in mouse and human (Patel et al. 2003).

A comparison between the human and mouse sequences in the *Sprr* locus by PipMaker analysis revealed conservation of coding sequences, introns, and putative gene-specific regulatory regions, located within 2 kb of an exon. The 180-kb intergenic region from Sprr2a to Sprr3 contained only a few sparse bits of sequence conservation. The 40 kb between *Sprr3* and *Sprr4* contained large tracts of sequence conservation, similar to the sequences flanking the *Sprr* locus, which displayed a significant degree of similarity (Fig. 1C, Supplemental Fig. 1). These stretches of conserved sequence are conserved not only in sequence but also in relative position between mouse and human (Fig. 1D, Supplemental Fig. 2). (Note that due to the large size of the figure, only the region of extensive analysis from *Sprr3* to Inv is presented in Fig. 1C,D. The other panels are presented in Supplemental material.) Underscoring the significance of these alignments is the UCSC Mouse/Human Evolutionary Conservation Score, which predicts with high probability that these sequences are under selection rather than neutral evolution. We observed that 30.2% of the 352-kb human sequence aligns with the mouse sequence, significantly lower than the average of ~40%.

In contrast, the percent identity plot for the mouse-on-mouse sequence demonstrates multiple hits in the intergenic *Sprr2* region with little conservation elsewhere (Supplemental Figs. 3,4). This analysis aided us in designing unique primers and probes. Two regions (at 116 kb and 253 kb) had high sequence conservation with the coding exons of the other *Sprr* genes. The sequence at 116 kb contains a long open reading frame with the 5′ half of the gene corresponding to an *Sprr*-like coding sequence and the 3′ portion to non-*Sprr*-related sequence. The gene at 253 kb is an *Sprr*-pseudogene. These two blocks correspond to new *Sprr*-related sequences. However, as no spliced ESTs were found corresponding to these sequences for any tissue, they are not annotated as genes. Rather, the sequences at 116 and 253 kb along with *Sprr2c* are identified as pseudogenes.

A BLAST search was performed for all the putative mouse–human CNSs against genomic and EST sequences to determine whether they represent unique noncoding sequences. Three novel spliced genes were identified and are annotated in Supplemental Figure 1 with blue boxes representing the exons and yellow the intronic sequence. The genomic DNA corresponding to spliced ESTs was considered real exons and excluded from further investigations.

Two previous studies focused on CNSs of >100 bp >70% identity to identify regulatory elements (Loots et al. 2000; Santagati et al. 2003). Using BLASTZ and AVID programs to align the mouse and human DNA sequences in the noncoding regions, we identified 18 CNSs with >100 bp >70% identity, (allowing for gaps and mapping more than 2 kb from an exon): six blocks in the *Loricrin* to *Sprr2a* interval and 12 in the *Sprr3* to *Involucrin* interval. However, for certain gene clusters, for example the α-globin cluster, this criterion is too stringent to identify known regulatory elements (Flint et al. 2001). Instead of ranking regions solely on sequence identity and ungapped alignment length, our strategy was to prioritize aligned sequences (both ungapped and gapped) that contain keratinocyte-specific DNaseI-hypersensitive sites for further experimental analysis.

Keratinocyte-specific DNaseI-hypersensitive mapping of the *Sprr* locus

Based on the prediction that conserved noncoding sequences may represent regulatory elements, we searched for keratinocyte-specific chromatin alterations in these regions. Proteins typically bind DNA in open chromatin, which is preferentially cleaved by DNaseI. Mapping cell type-specific sites sensitive to digestion with this enzyme is a standard method to localize putative regulatory regions (Hardison et al. 1997; Sinha and Fuchs 2001). To identify keratinocyte-specific HSs in the *Sprr* locus, we utilized SP-1 cells, initiated mouse keratinocytes, which express all of the *Sprr* genes (Strickland et al. 1988). These cells were also selected because under differentiation conditions, they coordinately upregulate the entire *Sprr* locus, which will be important for experiments described below. As a control, we queried erythroid MEL cells, which do not express any *Sprr* gene (data not shown).

Nuclei from SP-1 and MEL cells were treated with increasing concentrations of DNaseI to visualize all possible HSs. To query the chromatin status of the regions with high sequence conservation, we used overlapping probes and restriction enzyme fragments encompassing the regions between *Loricrin* and *Sprr2a* and between *Sprr3* and *Involucrin*. Southern blot analysis of the DNA fragments hybridized with appropriate radiolabeled probes revealed the keratinocyte-specific DNaseI HSs (arrows in Fig. 1C). The Southern blot mapping of the HSs is shown in Figure 2. The

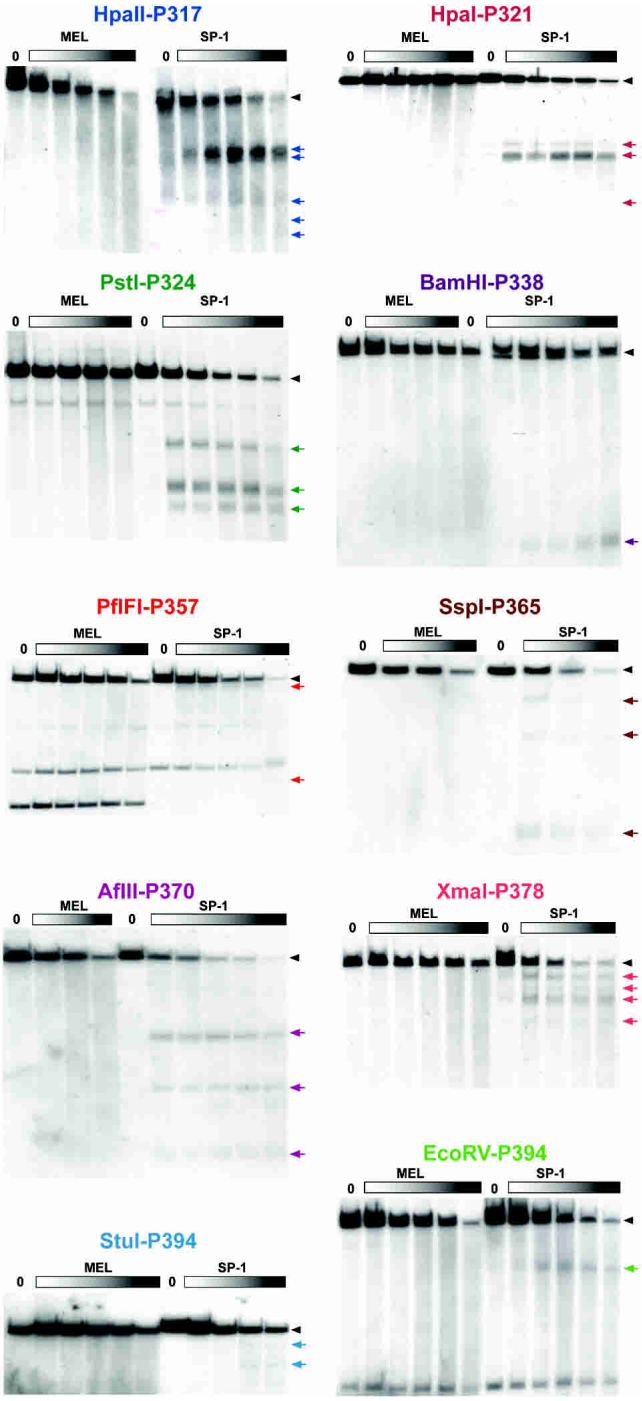


Figure 2. DNaseI HS mapping of the *Sprr* locus. Keratinocyte-specific patterns of DNaseI HSs at the conserved regions of the *Sprr* locus. Nuclei from SP-1 cells and MEL cells were treated with increasing amounts of DNaseI (closed rectangles), and isolated DNAs were digested with restriction enzymes and subjected to Southern blotting with a probe as indicated. The HSs bands are indicated on the *right* by colored arrows, and their localization in the mouse sequence is shown in Figure 1C. The exact locations of probes and restriction enzyme sites are given in Supplemental Table I.

locations of the probes (P), enzymes, and HSs are given in Figure 1C. In two cases, probes on both 5′ and 3′ sides of the restriction fragment were used to map the HSs: AflII with P357 and P370 and SspI with P365 and P370. The results generated with the two probes were identical, and thus only one is shown. Note that all of the data is color-coded so that a unique color is used for each enzyme, probe, and restriction site; for example, PstI in green cuts at 318.7 and 324.5 with a probe at 324.1 and HSs at 321.8, 322.7, and 323.1. The same colors are used consistently throughout the figures. The exact locations of each probe and restriction enzyme sites are given in Supplemental Table I. Since no HS mapped in the *Loricrin* to *Sprr2a* interval to a predicted CNS, we focused on the *Sprr3* to *Involucrin* interval.

Intriguingly, we detected two clusters of three HSs each in regions that overlapped with blocks of strong sequence identity that were more than 8 kb from any exon. We focused our efforts on these dense HSs from 322 to 324 and from 367 to 370 kb since clusters of HSs have traditionally been associated with control regions. The clusters of HSs contain alignments of 123 nucleotides (nt) with 77% identity (at 324.1), 155 nt at 71% identity (at 366.9), and 101 nt at 72% identity (at 369.3). Although no other HS mapped within 1 kb of a block of CNS with >70% identity over 100 bp, nine additional HSs, including the other HSs in these clusters at 324 and 370, mapped to regions of alignable sequence with lower conservation.

DNaseI-hypersensitive mapping of the *Sprr* locus in epidermis

To determine the biological significance of these two clustered HSs, we mapped these regions in neonatal epidermal cells, isolated from wild-type and barrier-deficient Klf4−/− epidermis, which we previously showed has a strong upregulation of the entire *Sprr* locus (Patel et al. 2003). To our knowledge, this type of in vivo verification has not previously been demonstrated for a keratinocyte-specific HS. We combined the method for isolating single cells from newborn epidermis with the method for isolating nuclei and mapped the HSs. Importantly, we discovered the identical pattern of HSs in the epidermis as the SP-1 cells; that is, no additional or missing HSs. Moreover, across all concentrations of DNaseI enzyme, the cluster of HSs from 322 to 324 is dramatically more prevalent in the Klf4-deficient than in the wild-type epidermal DNA. Normalizing the HSs to the undigested DNA, these HSs are ~three times more prevalent. Similarly, the clustered HSs from 367 to 370 are two- to sixfold more prevalent in the Klf4-deficient than in the wild-type epidermal DNA (Fig. 3).

Fine mapping and sequence analysis of individual hypersensitive sites

To enable testing of the HSs as enhancer elements, we need to fine-map the sites. Since the HS site may be generated anywhere within the nucleosome, we localized the HSs to 200–320 bp. For example, the HS at ~322 kb maps between the BglII and AciI sites, marking it to a 322-bp region. Each of the HSs was similarly localized (Fig. 4). Based on the fine mapping of the HSs, we now refer to them as 324HS1,2,3 and 370HS1,2,3 as shown on Figure 1C and Figure 4. With the fine localization of each HS, we returned to the concise alignment between the mouse and human sequences. We determined that the greatest block of sequence identity for 324 HS1,2,3 was 81 bp at 63%, 206 bp at 64%, and 201 bp at 63%, respectively. Similarly for 370HS1,2,3, the greatest block of sequence identity for HS1,2,3 was 300 bp at 69%, 293 bp at 60%, and 298 bp at 66%, respectively. These sequences are unique in the genome, and their localization relative to the *Sprr* genes is conserved in mouse and human.

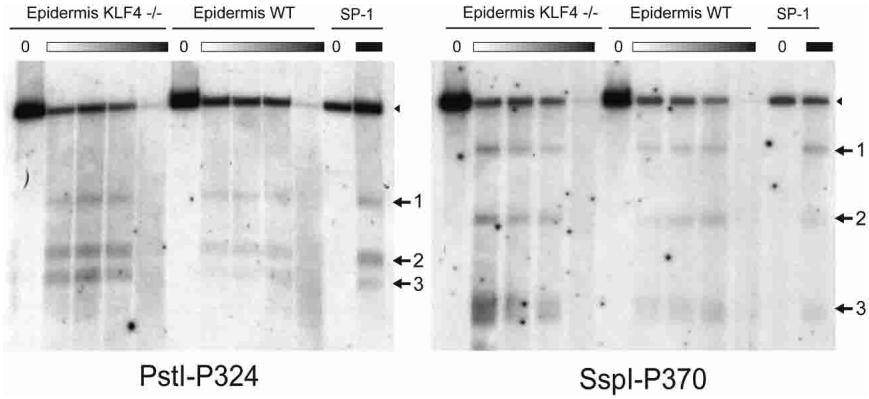


Figure 3. DNaseI HS mapping in skin epidermis. Nuclei were obtained from the epidermis of Klf4 −/− and wild-type newborn mice and treated with increasing amounts of DNaseI. The purified DNA was used for Southern blot analysis. The HSs detected by probe P324 on PstI-restricted DNaseI-treated DNA and probe P370 on SspI-restricted DNaseI-treated DNA in epidermis are the same as in the SP-1 cell line. Quantifying the intensity of the HSs bands in DNA digested to comparative levels indicates that the HSs are more prevalent in Klf4 −/− than in the wild-type DNA.

A PipMaker analysis between mouse and rat sequences in this region (Rat chr2: 184,984,342–185,508,416) revealed, as expected based on the shorter evolutionary distance between two rodents, longer blocks of conserved sequence with higher percent similarity. For the cluster of HSs at 370, we found that all three sites were conserved with 85.0%, 89.3%, and 87.3% identity over the 300 bp identified by fine mapping. In rat, these blocks of conserved sequence extended further than the mouse/human comparisons to 718, 339, and 1308 base pairs. Two surprising differences came with the PipMaker analysis of the 324 HS cluster. First, 324HS2 and 324HS3 are not conserved in sequence between mouse and rat. Second, 324HS1 is conserved at 85.6% identity over the 250 bp cloned. Between mouse and rat, this conserved sequence extends for 792 bp, including a gapless block of 386 bp.

Enhancer activity of individual hypersensitive sites

Our goal is to identify enhancer elements that are specific to the coordinate upregulation of the *Sprr* locus. First, we needed to design an in vitro system in which we could model the coordinate upregulation of the *Sprr* genes that is observed in vivo. The traditional way to differentiate epidermal cells is to shift the media from 0.05 to 1.3 mM Ca^{2+}. SP-1 cells were selected for the HS mapping because they express the *Sprr* genes, albeit at low levels, under growth conditions. Now, we queried the expression of the *Sprr* gene locus with the switch to high calcium. With gene-specific RT-PCR primers, we quantified the level of each transcript in the locus during growth and differentiation. Figure 5A shows the profile of *Sprr2g* as a typical example of the changes in *Sprr* gene expression when SP-1 cells are switched from growth to differentiation. Under low-calcium growth conditions, 32.9 cycles are required before the amplification of *Sprr2g* crosses the threshold. After 72 h in high Calcium, only 23.0 cycles are required to cross the threshold, indicating a significant upregulation (Fig. 5A). What is important, however, is not the expression level of any one gene in the family, but rather with high-calcium differentiation, SP-1 cells upregulate the entire family of *Sprr* genes (Fig. 5B). Only *Sprr2a* showed a modest increase with differentiation (~1 cycle increase in crossing the threshold with differentiation), and that may be due to the already high levels of *Sprr2a* that are expressed in the SP-1 cells during growth conditions. During differentiation, the levels of β-*actin*, *loricrin*, *involucrin*, and *Klf4* did not change substantially; that is, less than a 1-cycle difference in when they cross the threshold. The strength of this culture system is that in response to differentiation, SP-1 cells coordinately upregulate the entire *Sprr* locus.

To identify element(s) that coordinate the upregulation of the entire *Sprr* locus, we analyzed the enhancer activity of the three clustered HSs together (324HS1–3 and 370HS1–3) and each individual HS under both growth and differentiation conditions. The general design of the experiment was to clone the HS(s) upstream of a minimal *Sprr1a* promoter and query expression under the low-calcium (growth) conditions and the high-calcium (differentiation) conditions. The *Sprr1a* minimal promoter was selected because expression is unchanged under growth and differentiation conditions in the SP-1 cells.

When analyzed under growth conditions, none of the HSs showed any enhancer activity (data not shown). Since the SP-1 cells express the *Sprr* genes under these conditions, this suggests that these HSs are not functional in steady-state gene activation. Under differentiation conditions, both the cluster 324HS1–3 and 324HS1 showed increasing levels of activation: 4.0- and 4.9-fold

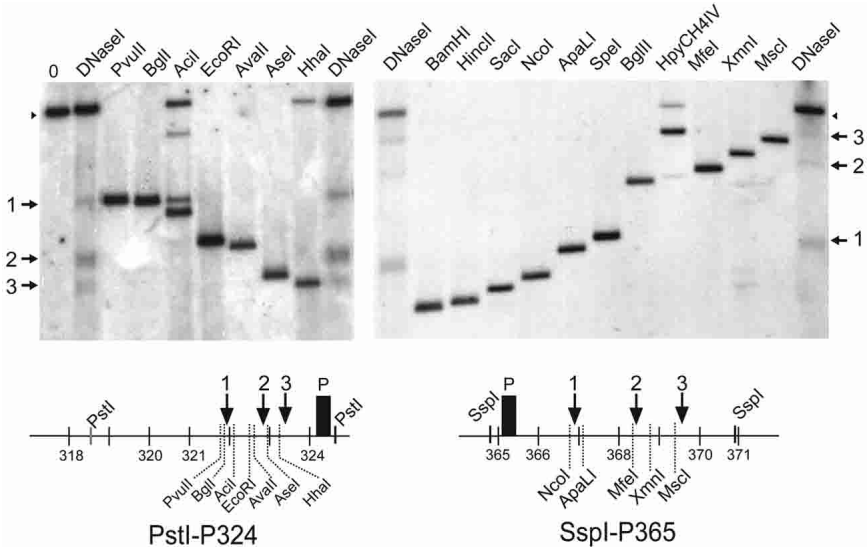
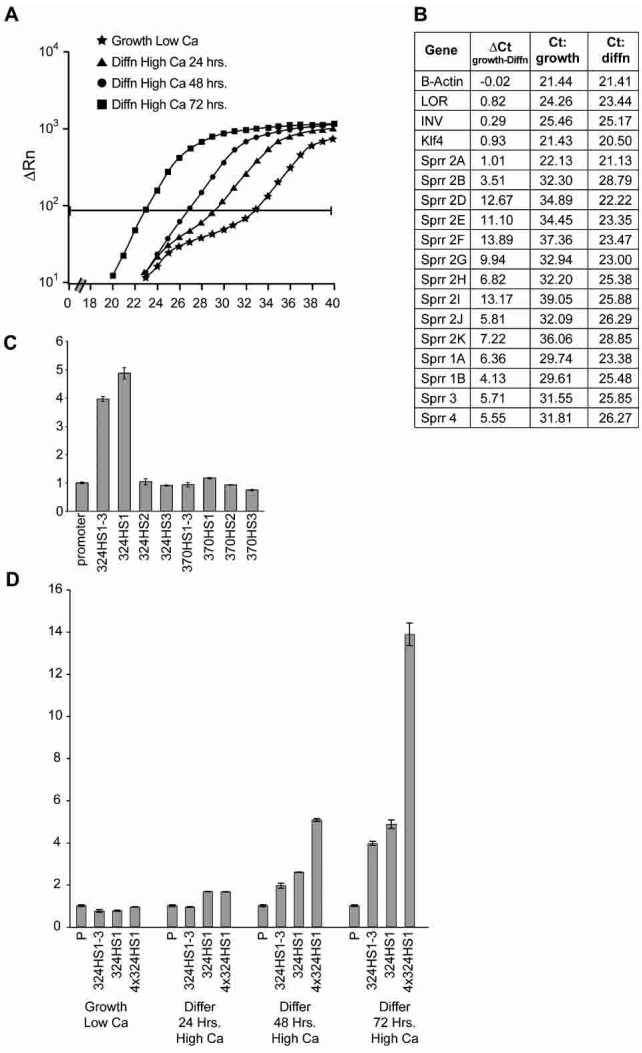


Figure 4. Fine mapping of clusters of HS. To more precisely localize the clusters of HSs, we mapped them with respect to other restriction sites in the genome. The restriction enzymes indicated on the figure were used to perform double digestions with either PstI (HS324) or SspI (P365) of SP-1 DNA, followed by Southern blotting and hybridization with the probe indicated.



Gene	ΔCt growth-Diffn	Ct: growth	Ct: diffn
B-Actin	-0.02	21.44	21.41
LOR	0.82	24.26	23.44
INV	0.29	25.46	25.17
Klf4	0.93	21.43	20.50
Sprr 2A	1.01	22.13	21.13
Sprr 2B	3.51	32.30	28.79
Sprr 2D	12.67	34.89	22.22
Sprr 2E	11.10	34.45	23.35
Sprr 2F	13.89	37.36	23.47
Sprr 2G	9.94	32.94	23.00
Sprr 2H	6.82	32.20	25.38
Sprr 2I	13.17	39.05	25.88
Sprr 2J	5.81	32.09	26.29
Sprr 2K	7.22	36.06	28.85
Sprr 1A	6.36	29.74	23.38
Sprr 1B	4.13	29.61	25.48
Sprr 3	5.71	31.55	25.85
Sprr 4	5.55	31.81	26.27

Figure 5. 324HS1 exhibits strong enhancer activity specific to coordinate upregulation of *Sprr* genes. (*A*) Expression of *Sprr2g* in SP-1 cells increases with differentiation conditions. Relative levels of mRNA transcript are determined by when the amplicon crosses the threshold. (*B*) Expression of the *Sprr* genes, *involucrin* (*Inv*), *Loricrin* (*Lor*), and *Klf4* in SP-1 cells under growth and differentiation conditions. Primer sequences and amplicon sizes are given in Supplemental Table III. (*C*) 324HS1–3 and 324HS1 have enhancer activity under differentiation conditions of SP-1 cells. All values are expressed relative to the activity of the minimal promoter. Activities represent the mean ± SEM of three experiments. (*D*) Multimerized 324HS1 (4×324HS1) has strong enhancer activity under differentiation conditions of SP-1 cells. The culture conditions are indicated on the *x*-axis. Activities represent the mean ± SEM of three experiments.

after 72 h, respectively. Each individual HS showed the same level of activation in either orientation relative to the promoter, and thus only one is shown (Fig. 5C). Since 324HS2 and 324HS3 did not show any enhancer activity and the activation of 324HS1–3 was either similar or slightly less than 324HS1, we focused on this individual HS. To test the specificity of this enhancer activity, we multimerized 324HS1 as four copies (4×324HS1). Due to the high sequence conservation (66% over 298 bp) and strong prevalence in the Klf4−/− epidermis, we simultaneously multimerized 370-HS3 (4×370-HS3). Multimerized 4×370HS3 showed ~1.6-fold increase in activity after 24 h in high calcium with no increased activity at 48 or 72 h (data not shown). 4×324HS1 did not show any activity under growth. However with terminal differentiation, 4×324HS1 conferred 13.9-fold increased enhancer activity at 72 h, indicating that this enhancer gives a dose-dependent activation (Fig. 5D).

To dissect the regulation of this sequence, we returned to the mouse–human identity in this region. 324HS1 is the hypersensitive site with the lowest conservation between these two species, with 63% conservation over 81 bp and a highly conserved core of 38 bp. Most transcription factors (TFs) have a small invariant core sequence (~4–6 bp), which generates a large number of false positives when a sequence is queried for multiple factor binding sites. Therefore, we utilized our local alignments in conjunction with known TF binding sites from the TRANSFAC database to predict binding sites only in the sequences that are conserved (Wingender et al. 2000). This search located binding sites for some TFs expressed in keratinocytes but no TFs specific to epidermal differentiation, such as *Klf4*. Because only a fraction of mammalian TFs and their binding sites are known and available for comparison, a large number of binding sites can be missed in such a TF database search.

Discussion

The current model suggests that a cornified envelope (CE) is assembled in a stepwise orderly process with tissue-specific *Sprr* genes forming the initial scaffold (Kalinin et al. 2002). However, under genetic or environmentally stressed circumstances, it has been hypothesized that substrates will be indiscriminately cross-linked to re-establish a temporary barrier (Michel et al. 1987; Robinson et al. 1997). Under stress, there is a specific upregulation of all 14 *Sprr* genes (Patel et al. 2003), encoding the initial CE scaffold proteins. Here, we explored the hypothesis that the *Sprr* genes are tandemly arrayed to enable coordinate regulation under stress.

For the sequence encompassing the *Sprr* locus, from *Loricrin* to *Involucrin*, we identified CNSs between mouse and human. Comparative analysis of mouse and human sequences is widely utilized because they are the first mammalian whole-genome sequences to be publicly available. As more genomes are selectively sequenced, multispecies DNA comparisons should be used to prioritize the CNS under the assumption that sequences conserved in multiple species are more likely to be functional. Powerful algorithms for identifying multispecies conserved sequences are being developed in anticipation of additional complete genome sequences (Margulies et al. 2003; Schwartz et al. 2003; Chapman et al. 2004). We were unable to compare with the chicken genome because the *Sprr* genes remain in small, unmapped contigs. We were unable to locate the *Sprr* genes or the flanking genes (*Loricrin* and *Involucrin*) in the zebrafish and pufferfish genome sequences. As expected, based on the shorter evolutionary distance between two rodents, comparing the mouse *Sprr* locus sequence with rat produced long blocks of conserved sequence with high percent similarity for the cluster of HSs at 370. However, for the cluster of HSs at 324, 324HS2 and 324HS3 are not conserved. In contrast, the functional enhancer 324HS1 is 85.6% identical with a block extending 792 bp between mouse and rat. In the future, analysis of additional vertebrate sequences may aid in understanding the conservation of 324HS1,2,3 sequences.

To identify potential enhancer regions, we mapped 15 keratinocyte-specific DNaseI-hypersensitive sites in the *Sprr3* to *In-*

volucrin interval. Of the 12 CNSs with >100 bp aligned at >70%, (allowing for gaps and mapping more than 2 kb from an exon) in this region, three map within 0.5 kb of an HS (resolution of the initial HS site mapping): CNSs at 324.1, 366.9, and 369.3. Two other CNSs (347.5 and 352.8) map within 2 kb of an HS, but precise overlap can be ruled out. Fine mapping of the HSs at 322–324 and 366–370 indicates a precise overlap between 370HS1=CNS366.9 and 370HS3=CNS369.3, but 324HS1,2,3 mapped more proximal than the strongest hit at 324.1. Under less stringent criteria, 12 of the 15 HSs map to regions of alignable sequence, albeit with lower percent conservation.

Under both growth and differentiation conditions, our assays were unable to detect any function for the cluster of HSs/CNSs at 370. Conversely, 324HS1 has strong enhancer activity specific to the upregulation of the *Sprr* locus. Ironically, this is the HS with the smallest alignable sequence of the HSs tested. This present analysis is constrained by the functional assays utilized. It is possible that these HSs could be enhancers for other promoters, cell types, or conditions than the ones queried in this analysis. Future studies will investigate the role of these HSs in the nucleosomal context.

The in vivo significance of the HSs is underscored by the analysis of epidermal nuclei from barrier-deficient Klf4−/− mice, which upregulate the *Sprr* locus. Not only are the same HSs identified in epidermis as in SP-1 cells, but also these HSs have increased prevalence in the Klf4−/− nuclei. To our knowledge, keratinocyte DNaseI HSs have not previously been verified in epidermal nuclei.

We sought here to query whether the *Sprr* genes are held tandemly arrayed to coordinate gene regulation under stressed conditions. With both in vitro and in vivo mapping, we identified clusters of hypersensitive sites in the distal region of the *Sprr* locus and identified one of these HSs as an enhancer specific to upregulation of the locus. Although our goal was to identify enhancer elements that regulate this stress response activation, future studies will address the role of these and other HSs as silencer, insulator, or boundary elements, the other properties of locus control regions. These studies demonstrate the power of using full-genome comparative sequence analysis to direct searches for regulatory regions.

Methods

Genomic sequence comparison

Complete murine (February 2003 freeze: chr3: 92963116–93389747) and human (April 2003 freeze: chr1: 149657368–150009693) sequences for the region from *Involucrin* to *Loricrin* were retrieved from the UCSC genome Web site (http://www.genome.ucsc.edu). All gaps were filled by long-range PCR with primers from the end sequences on bacterial artificial chromosome (BAC) DNA from this region. Alignment and search for homologous sequences between large genomic sequences were carried out with the PipMaker program with the "all matches" option of PipMaker (http://bio.cse.psu.edu/pipmaker) (Schwartz et al. 2000; Elnitski et al. 2002). The reference mouse sequence was masked of repetitive DNA with RepeatMasker (http://ftp.genome.washington.edu/RM/RepeatMasker.html).

Cell culture

The SP-1 cell line was obtained from Luowei Li and Stuart Yuspa (NCI, NIH) and cultured under the standard conditions of S-MEM media (GIBCO) with 8% Chelex-treated fetal bovine serum (Gemini) at 0.05 mM Ca^{2+}. The cells were maintained at 36°C in the presence of 7% CO_2 (Hennings et al. 1980; Strickland et al. 1988). For the switch to high calcium, $CaCl_2$ was added to the media to a final concentration of 1.3 mM. MEL cells were obtained from David Bodine (NHGRI, NIH) and cultured under the standard conditions of IMEM with 10% fetal calf serum (GIBCO).

Isolation of nuclei and DNaseI digestion

Approximately 10^8 SP-1 cells were scraped gently from the flask and washed in ice-cold PBS. The cells were centrifuged at 2100*g* at 4°C for 5 min. The cell pellet was resuspended in 14 mL of RSB buffer (10 mM Tris HCl, pH 7.5, 10 mM NaCl, 3 mM MgCl). The cell membrane was disrupted by adding 375 µL of 10% NP-40 drop-wise. Cell lysis was surveyed under the microscope by mixing 10 µL of cell suspension with 10 µL of Trypan blue. When ~50% of the cells had disrupted membranes (~10 min for SP-1 cells), the suspension was centrifuged at 1200*g* for 4 min at 4°C. Nuclei were resuspended in 1.5 mL RSB buffer. DNaseI (Amersham) was diluted in dilution buffer (10 mM Tris HCl, pH 7.5, 50 mM NaCl, 10 mM $CaCl_2$, 62.5 mM $MgCl_2$) to a final concentration range of 0, 0.5, 1, 2, 4, 6, 10 U/µL. Twenty µL of these DNaseI solutions were added to 200 µL of the resuspended nuclei. After incubation at 37°C for 10 min, reactions were quenched by adding 200 µL of stop buffer (50 mM Tris HCl, pH 7.5, 20 mM EDTA pH 8, 100 mM NaCl, 1% SDS). Forty µL of proteinase K (20 mg/mL) was added to each reaction, and nuclei were incubated overnight at 55°C to digest proteins. The DNA was isolated by phenol-chloroform extraction and ethanol precipitation. The final purified DNA pellet was dissolved in TE (pH 7.5).

Separation of newborn mouse skin epidermis by dispase and isolation of primary keratinocytes was done as described (Wang et al. 1997) . From this cell suspension (~10^6 cells per newborn mouse epidermis), nuclei were obtained by the same protocol described above for cultured cells, except that the initial amounts of RSB buffer (5 mL) and 10% NP-40 (80 µL) were modified.

Southern blot analysis

DNA (10 µg or 15 µg per sample) was digested with restriction enzymes; the DNA fragments were separated in 1% agarose gels by electrophoresis and transferred to nylon membranes (Zeta-Probe, Bio-Rad). The membranes were prehybridized at 65°C for at least 1 h in 7% SDS, 0.25 M NaPhosphate pH 7.5, and hybridized overnight at 65°C in a solution containing 7% SDS, 0.25 M NaPhosphate pH 7.5, 10 % dextran sulfate. The membranes were washed at 65°C with 2 × SSC, 0.1% SDS and with 0.4 × SSC, 0.1% SDS.

To generate the probes in targeted regions, unique sequences with no similarity in the *Sprr* locus or the mouse genome were selected. The probes were amplified by PCR from BACs in the region with primers given in Suppl. Table I, and amplicon sequences were confirmed by automated sequencing. The location of the probe is denoted by the sequence position shown on Figure 1C. The probes were radiolabeled by PCR with the primers used to amplify the original fragment.

DNA constructs and transfections

Recombinant plasmids for expression studies were constructed in the pGL3 firefly luciferase reporter vector (Promega). The *Sprr* promoter (−708 to +42 relative to the transcription start site of mouse *Sprr1a*) was cloned into pGL3 promoter vector with BglII and HindIII, which removes the endogenous SV40 promoter. Individual HSs were amplified by PCR from BAC DNA with primers containing XhoI sites at the 5′ ends and cloned 5′ of the *Sprr1a* promoter at the XhoI site. The sequence and size of the ampli-

cons are given in Supplemental Table II. Multimerized HS constructs were generated from PCR-amplified segments with KpnI and BamHI restriction sites at the 5′ end of the forward primer and BglII and NheI restriction sites at the 5′ end of the reverse primer. The amplicons were first cloned into the KpnI and NheI restriction sites of modified pBS (Stratagene and gift from Satrajit Sinha, SUNY, Buffalo), and a direct repeat was generated by re-inserting a BamHI/NheI fragment into the vector cut with BglII and NheI. The BamHI/BglII hybrid site is destroyed, and the process is repeated to generate four direct repeats. The multimerized enhancer was then cloned into a modified pGL3 containing the *Sprr1a* promoter at the KpnI and NheI restriction sites. All constructs were sequence verified.

Plasmid DNA was isolated using Maxiprep kits (QIAGEN). SP-1 cells were transfected with Lipofectamine Plus (Invitrogen) using conditions that were optimized according to the recommendations provided by the manufacturer. All transfections included a control *Renilla* luciferase plasmid (pRL-SV40) and were normalized to *Renilla* luciferase activity. Firefly and Renilla activities were measured using the dual-luciferase reporter assay system (Promega) on a Turner Designs 20/20 luminometer.

Quantitative real-time RT–PCR

RNA was prepared from SP-1 cells grown under low- or high-calcium conditions with Trizol (Invitrogen). RNA was treated with DNAse and first-strand cDNA was prepared according to the manufacturer's instructions (Invitrogen). Samples were queried with control primers to ensure that there was undetectable genomic DNA contamination. Quantification of cDNA synthesis was normalized to amplification of β-actin. Unique primers spanning intron boundaries specific to each *Sprr* gene were generated, and each resulting product was sequenced. All primer sequences are deposited as Supplemental Table III. Reactions were carried out with SyberGreen labeling using the Q-PCR mix (Invitrogen) and run on an ABI Prism 7700 sequence detector (PE Applied Biosystems). Each SyberGreen PCR reaction was run on an agarose gel to ensure that the correct-size product was generated.

Acknowledgments

We are very grateful for the excellent advice and reagents provided by David Bodine, Satrajit Sinha and Louwei Li. Michael Cichanowski assisted with the figures. Members of the lab provided useful advice and criticisms at all stages of the work. A huge thanks goes to Tyra Wolfsberg, Elliot Margulies and Tony Antonellis for many discussions.

References

Cabral, A., Voskamp, P., Cleton-Jansen, A.M., South, A., Nizetic, D., and Backendorf, C. 2001. Structural organization and regulation of the small proline-rich family of cornified envelope precursors suggest a role in adaptive barrier function. *J. Biol. Chem.* **276:** 19231–19237.

Chapman, M.A., Donaldson, I.J., Gilbert, J., Grafham, D., Rogers, J., Green, A.R., and Göttgens, B. 2004. Analysis of multiple genomic sequence alignments: A web resource, online tools, and lessons learned from analysis of mammalian SCL loci. *Genome Res.* **14:** 313–318.

Elder, J.T. and Zhao, X. 2002. Evidence for local control of gene expression in the epidermal differentiation complex. *Exp. Dermatol.* **11:** 406–412.

Elnitski, L., Riemer, C., Petrykowska, H., Florea, L., Schwartz, S., Miller, W., and Hardison, R. 2002. PipTools: A computational toolkit to annotate and analyze pairwise comparisons of genomic sequences. *Genomics* **80:** 681–690.

Flint, J., Tufarelli, C., Peden, J., Clark, K., Daniels, R.J., Hardison, R., Miller, W., Philipsen, S., Tan-Un, K.C., McMorrow, T., et al. 2001. Comparative genome analysis delimits a chromosomal domain and identifies key regulatory elements in the α-globin cluster. *Hum. Mol. Genet.* **10:** 371–382.

Frazer, K.A., Elnitski, L., Church, D.M., Dubchak, I., and Hardison, R.C. 2003. Cross-species sequence comparisons: A review of methods and available resources. *Genome Res.* **13:** 1–12.

Frazer, K.A., Tao, H., Osoegawa, K., de Jong, P.J., Chen, X., Doherty, M.F., and Cox, D.R. 2004. Noncoding sequences conserved in a limited number of mammals in the SIM2 interval are frequently functional. *Genome Res.* **14:** 367–372.

Göttgens, B., Gilbert, J.G., Barton, L.M., Grafham, D., Rogers, J., Bentley, D.R., and Green, A.R. 2001. Long-range comparison of human and mouse SCL loci: Localized regions of sensitivity to restriction endonucleases correspond precisely with peaks of conserved noncoding sequences. *Genome Res.* **11:** 87–97.

Göttgens, B., Barton, L.M., Chapman, M.A., Sinclair, A.M., Knudsen, B., Grafham, D., Gilbert, J.G., Rogers, J., Bentley, D.R., and Green, A.R. 2002. Transcriptional regulation of the stem cell leukemia gene (SCL)—Comparative analysis of five vertebrate SCL loci. *Genome Res.* **12:** 749–759.

Grosveld, F., van Assendelft, G.B., Greaves, D.R., and Kollias, G. 1987. Position-independent, high-level expression of the human β-globin gene in transgenic mice. *Cell* **51:** 975–985.

Hardison, R., Slightom, J.L., Gumucio, D.L., Goodman, M., Stojanovic, N., and Miller, W. 1997. Locus control regions of mammalian β-globin gene clusters: Combining phylogenetic analyses and experimental results to gain functional insights. *Gene* **205:** 73–94.

Hennings, H., Michael, D., Cheng, C., Steinert, P., Holbrook, K., and Yuspa, S.H. 1980. Calcium regulation of growth and differentiation of mouse epidermal cells in culture. *Cell* **19:** 245–254.

Kalinin, A.E., Kajava, A.V., and Steinert, P.M. 2002. Epithelial barrier function: Assembly and structural features of the cornified cell envelope. *Bioessays* **24:** 789–800.

Loots, G.G., Locksley, R.M., Blankespoor, C.M., Wang, Z.E., Miller, W., Rubin, E.M., and Frazer, K.A. 2000. Identification of a coordinate regulator of interleukins 4, 13, and 5 by cross-species sequence comparisons. *Science* **288:** 136–140.

Margulies, E.H., Blanchette, M., Haussler, D., and Green, E.D. 2003. Identification and characterization of multi-species conserved sequences. *Genome Res.* **13:** 2507–2518.

Michel, S., Schmidt, R., Robinson, S.M., Shroot, B., and Reichert, U. 1987. Identification and subcellular distribution of cornified envelope precursor proteins in the transformed human keratinocyte line SV-K14. *J. Invest. Dermatol.* **88:** 301–305.

Patel, S., Kartasova, T., and Segre, J.A. 2003. Mouse *Sprr* locus: A tandem array of coordinately regulated genes. *Mamm. Genome* **14:** 140–148.

Pennacchio, L.A. and Rubin, E.M. 2001. Genomic strategies to identify mammalian regulatory sequences. *Nat. Rev. Genet.* **2:** 100–109.

Robinson, N.A., Lapic, S., Welter, J.F., and Eckert, R.L. 1997. S100A11, S100A10, annexin I, desmosomal proteins, small proline-rich proteins, plasminogen activator inhibitor-2, and involucrin are components of the cornified envelope of cultured human epidermal keratinocytes. *J. Biol. Chem.* **272:** 12035–12046.

Santagati, F., Abe, K., Schmidt, V., Schmitt-John, T., Suzuki, M., Yamamura, K., and Imai, K. 2003. Identification of *cis*-regulatory elements in the mouse Pax9/Nkx2–9 genomic region: Implication for evolutionary conserved synteny. *Genetics* **165:** 235–242.

Santini, S., Boore, J.L., and Meyer, A. 2003. Evolutionary conservation of regulatory elements in vertebrate Hox gene clusters. *Genome Res.* **13:** 1111–1122.

Schwartz, S., Zhang, Z., Frazer, K.A., Smit, A., Riemer, C., Bouck, J., Gibbs, R., Hardison, R., and Miller, W. 2000. PipMaker—A web server for aligning two genomic DNA sequences. *Genome Res.* **10:** 577–586.

Schwartz, S., Elnitski, L., Li, M., Weirauch, M., Riemer, C., Smit, A., Green, E.D., Hardison, R.C., and Miller, W. 2003. MultiPipMaker and supporting tools: Alignments and analysis of multiple genomic DNA sequences. *Nucleic Acids Res.* **31:** 3518–3524.

Sinha, S. and Fuchs, E. 2001. Identification and dissection of an enhancer controlling epithelial gene expression in skin. *Proc. Natl. Acad. Sci.* **98:** 2455–2460.

Song, H.J., Poy, G., Darwiche, N., Lichti, U., Kuroki, T., Steinert, P.M., and Kartasova, T. 1999. Mouse *Sprr2* genes: A clustered family of genes showing differential expression in epithelial tissues. *Genomics* **55:** 28–42.

South, A.P., Cabral, A., Ives, J.H., James, C.H., Mirza, G., Marenholz, I., Mischke, D., Backendorf, C., Ragoussis, J., and Nizetic, D. 1999.

Human epidermal differentiation complex in a single 2.5 Mbp long continuum of overlapping DNA cloned in bacteria integrating physical and transcript maps. *J. Invest. Dermatol*. **112:** 910–918.
Strickland, J.E., Greenhalgh, D.A., Koceva-Chyla, A., Hennings, H., Restrepo, C., Balaschak, M., and Yuspa, S.H. 1988. Development of murine epidermal cell lines which contain an activated rasHa oncogene and form papillomas in skin grafts on athymic nude mouse hosts. *Cancer Res*. **48:** 165–169.
Ureta-Vidal, A., Ettwiller, L., and Birney, E. 2003. Comparative genomics: Genome-wide analysis in metazoan eukaryotes. *Nat. Rev. Genet*. **4:** 251–262.
Wang, X., Zinkel, S., Polonsky, K., and Fuchs, E. 1997. Transgenic studies with a keratin promoter-driven growth hormone transgene: Prospects for gene therapy. *Proc. Natl. Acad. Sci.* **94:** 219–226.
Waterston, R.H., Lindblad-Toh, K., Birney, E., Rogers, J., Abril, J.F., Agarwal, P., Agarwala, R., Ainscough, R., Alexandersson, M., An, P., et al. 2002. Initial sequencing and comparative analysis of the mouse genome. *Nature* **420:** 520–562.
Wingender, E., Chen, X., Hehl, R., Karas, H., Liebich, I., Matys, V., Meinhardt, T., Pruss, M., Reuter, I., and Schacherer, F. 2000. TRANSFAC: An integrated system for gene expression regulation. *Nucleic Acids Res*. **28:** 316–319.

Web site references

http://www.genome.ucsc.edu; UCSC genome Web site.
http://bio.cse.psu.edu/pipmaker; PipMaker.
http://ftp.genome.washington.edu/rm/repeatmasker.html; RepeatMasker.

Received April 27, 2004; accepted in revised form October 4, 2004.