ELSEVIER

Eur. J. Cell Biol. 83 (2004); 19–26
http://www.elsevier-deutschland.de/ejcb

EJCB

Comprehensive analysis of keratin gene clusters in humans and rodents

Michael Hesse[2)a], Alexander Zimek[2)b], Klaus Weber[b], Thomas M. Magin[1)a]

[a] Institut für Physiologische Chemie, Abteilung für Zellbiochemie, Bonner Forum Biomedizin and LIMES, Universitätsklinikum Bonn, Germany
[b] Max-Planck-Institute for Biophysical Chemistry, Department of Biochemistry, Göttingen, Germany

Received November 21, 2003
Received in revised version December 22, 2003
Accepted December 23, 2003

Gene clusters; Gene duplications; Keratins; Mammalian genomes; Synteny

Here, we present the comparative analysis of the two keratin (K) gene clusters in the genomes of man, mouse and rat. Overall, there is a remarkable but not perfect synteny among the clusters of the three mammalian species. The human type I keratin gene cluster consists of 27 genes and 4 pseudogenes, all in the same orientation. It is interrupted by a domain of multiple genes encoding keratin-associated proteins (KAPs). Cytokeratin, hair and inner root sheath keratin genes are grouped together in small subclusters, indicating that evolution occurred by duplication events. At the end of the rodent type I gene cluster, a novel gene related to K14 and K17 was identified, which is converted to a pseudogene in humans. The human type II cluster consists of 27 genes and 5 pseudogenes, most of which are arranged in the same orientation. Of the 26 type II murine keratin genes now known, the expression of two new genes was identified by RT-PCR. Kb20, the first gene in the cluster, was detected in lung tissue. Kb39, a new ortholog of K1, is expressed in certain stratified epithelia. It represents a candidate gene for those hyperkeratotic skin syndromes in which no K1 mutations were identified so far. Most remarkably, the human K3 gene which causes Meesmann's corneal dystrophy when mutated, lacks a counterpart in the mouse genome. While the human genome has 138 pseudogenes related to K8 and K18, the mouse and rat genomes contain only 4 and 6 such pseudogenes. Our results also provide the basis for a unified keratin nomenclature and for future functional studies.

Introduction

Keratins are structural proteins which form the intermediate filament cytoskeleton in all epithelia. They assemble into long filaments built from obligate heterodimeric double-stranded coiled coils which consist of a type I and a type II protein. Keratins are encoded by a large and well conserved gene family with more than 49 members in humans (Hesse et al., 2001). Their major function is to provide mechanical stability to epithelial cells against stress. Thus, mutations in at least 14 human epidermal keratin genes cause tissue fragility syndromes (Herrmann et al., 2003). At least 4 of the 11 *Caenorhabditis elegans* IF genes are essential for nematode development (Karabinos et al., 2001, 2003).

In human, all type I keratin genes, except for K18 (Waseem et al., 1990), are clustered on chromosome 17q21 and type II genes cluster on 12q13 (Int. Human Genome Sequencing Consortium, 2001). In the mouse, the orthologs are similarly clustered on chromosomes 11D and 15F, respectively (Waterston et al., 2002). In most mammalian species, there is an approximately equal number of type I and type II genes which give rise to so called expression pairs (Herrmann et al., 2003; Hesse et al., 2001). Unexpectedly, the preliminary gene complement of the teleost fish *Fugu rubripes* indicates a sizeable excess of type I over type II genes (Zimek et al., 2003). The functional significance of this imbalance is presently not clear.

Based on the work of Moll and Franke (Moll et al., 1982) and of Sun (Sun et al., 1983), a nomenclature based on the separation of keratins in high resolution 2D gels was established. The numbering system for type II keratins ranged from 1 to 8 and from 9 to 21 for type I keratins. Type II keratins identified subsequently had to be named with a letter following the number. Hair keratins were named in an analogous way with letters Ha and Hb indicating type I and II hair keratins, respectively (Langbein et al., 1999; Rogers et al., 2000). The more recently identified keratins expressed in the inner root sheath of the hair follicle were named in a similar fashion with an added "irs" referring to their major expression site. Other

[1)] **Corresponding author:** Prof. Dr. Thomas M. Magin, Institut für Physiologische Chemie, Abteilung für Zellbiochemie, Universitätsklinikum Bonn, Nussallee 11, D-53115 Bonn, Germany, e-mail: t.magin@uni-bonn.de, Fax: +49 228 734 558.
[2)] Both authors contributed equally to this work.

0171-9335/04/083/01-019 $30.00/0

terms e.g. KRT have been additionally used (HUGO Gene Nomenclature Committee (HGNC), under review).

The revised drafts of the human, mouse and rat genomes have now enabled us to conclude our analysis of the keratin gene clusters and have provided a few novel keratin genes not available previously (Hesse et al., 2001). Here, we present a comparative analysis of these 3 mammalian genomes, showing that the arrangement of keratin genes in humans and rodents is generally well conserved. However, we note several important changes. K3 for instance, the major type II protein of human corneal epithelium, is absent in the mouse genome. In view of the complexity of the keratin gene family and of the problems with a nomenclature partially based on the primary site of expression, we propose a new and unified nomenclature for all keratins which includes the principles established by Moll and Franke (Moll et al., 1982).

Materials and methods

Databases

For keratin gene annotation and discovery, the UCSC Genome Browser Database (Karolchik et al., 2003) was used which includes the sequenced genomes of human (Lander et al., 2001), mouse (Waterston et al., 2002), and rat. Genomic data used from GenBank included the latest human (hg16), mouse (mm3), and rat (rn3) assemblies. All accession numbers used were taken from GenBank.

Sequence analyses

Sequence similarity search was performed using the BLAT program (Kent, 2002). In addition, for more stringent similarity searches, BLAST (Altschul et al., 1990) was used for nucleotide sequences and PSI-BLAST (Altschul et al., 1997) for protein sequences, respectively. Newly detected and predicted keratin genes were checked for intron-exon accuracy by comparison to known keratin cDNA and protein sequences by BESTFIT and CLUSTAL programs (GCG Wisconsin package, HUSAR, Heidelberg).

Results and Discussion

Nomenclature of keratins

The present nomenclature of keratins is based on their position in 2D gels or their primary site of expression. This has resulted in a complex numbering system which in some cases does not reflect sequence relatedness despite identical numbers. Here, we propose a new nomenclature for all keratins based on the principles established by Moll and Franke (Moll et al., 1982). All type I keratin genes are named Ka9 to KaX and all type II keratin genes are named Kb1 to KbY. The numbers given by Moll et al. (1982) are maintained and all keratins described later, irrespective of their site of expression, receive numbers accordingly. The species is indicated by a prefix in parentheses (HUMAN, MOUSE, RAT). Table 1 lists known human and rodent keratins in the previous and the new nomenclature. The nomenclature proposed is an open system, allowing the addition of novel keratins in other species by additional numbers.

We followed the nomenclature principles for gene families proposed by the Hugo Gene Nomenclature Committee (HGNC). Pseudogenes that are not derived from a functional keratin gene have their own number and are indicated by a P. Pseudogenes derived from keratin genes by duplication or retrotransposition are numbered, starting with P1. Orthologs share the same numbers, but genes unique to a species are individually numbered (Table 1). For convenience of readers familiar with the keratin field, we provide the old gene names in parentheses throughout the text.

The mammalian keratin type I clusters

The human keratin I gene cluster located on chromosome 17q21.2 is completely sequenced (NCBI Build 34, July 2003) and its mouse counterpart on chromosome 11D is essentially known (release February 2003). The rat type I keratin gene cluster located on chromosome 10q31 still contains various gaps of different length (UCSC version rn3, June 2003), but essentially follows the murine cluster in gene arrangement (Fig. 1). Thus we have concentrated on the human and murine keratin type I and II gene clusters.

The human keratin type I gene cluster consists of 27 keratin I genes and 4 keratin I pseudogenes which are all arranged in the same orientation (Fig. 1). Inserted into this cluster is a domain of a large number of genes encoding the high and ultrahigh sulfur keratin-associated proteins (KAPs) (Rogers et al., 2001). These KAP genes show both orientations and divide the keratin cluster in two subclusters. For most but not all human and mouse type I keratin genes, cDNA accession numbers are available. The size of the human type I keratin cluster is 977 kb including the 362-kb KAP cluster. The gene density in the telomeric part of the keratin cluster is 14.4 kb (18 kb with pseudogenes excluded). The centromeric part has a gene density of 25.2 kb (29.8 kb with pseudogenes excluded). In the mouse the telomeric part has a size of 260 kb with a gene density of 16.2 kb (17.3 kb with pseudogenes excluded) and the centromeric part a size of 309 kb and a gene density of 25.8 kb (28.1 kb with pseudogenes excluded).

The first part of the mammalian type I cluster contains several genes encoding keratins of the inner root sheath of hair follicles (Bawden et al., 2001). The human pseudogene between genes *Ka24* (*K10B*) and *Ka38* (*K10C*) is absent from the rodent clusters. Past the gene for *Ka23* (*K23*) lie two newly defined hair keratin genes (*Ka35* and *Ka36*) and the subdomain of the genes encoding the high/ultrahigh sulfur KAPs. The KAP domain is flanked by the hair keratin gene *Ka27* (*KRTHA3A*), which is followed in the rodent clusters by 6 additional hair keratin genes. In man, this region contains two additional hair keratin I genes (*Ka32* (*KRTHA7*), *Ka33* (*KRTHA8*)) and the only hair keratin I pseudogene *Ka34P* (ψ*KRTHAA*). Together with the two newly discovered hair keratins (*Ka35* and *Ka36*), there are 9 and 11 hair type I keratin genes in rodents and man, respectively (Fig. 1).

The KAP subdomain comprises about 362 kb in human, 466 kb in mouse and 465 kb in rat, respectively. In man it contains 29 genes according to our own count of high/ultrahigh sulfur hair KAPs, which can be divided into 7 individual gene families. Their respective mRNAs are localized to the upper cortex of the hair shaft (Rogers et al., 2001). These KAPs have cysteine contents either below (high sulfur) or above (ultrahigh) 30%. The entire KAP region is inserted into the hair keratin domain (Fig. 1). Interestingly, 7 other human genes encoding high sulfur KAPs lie together with 17 genes for high glycine-tyrosine KAPs on human chromosome 21q22.1 (Rogers et al., 2002).

Table 1. Revised nomenclature of keratins.

Old name	New names					
Type I	HUMAN	Acc. No.	MOUSE	Acc. No.	RAT	Acc. No.
K9	Ka9	NM_000226	Ka9	AK028845	Ka9	NM_153476
K10	Ka10	J04029	Ka10	AK076508	Ka10	BK004032
Ka11*	~	~	Ka11P	BK004024	Ka11	BK004048
K12	Ka12	D78367	Ka12	NM_010661	Ka12	BK004033
K13	Ka13	X52426	Ka13	NM_010662	Ka13	BK004044
K14	Ka14	NM_000526	Ka14	BC011074	Ka14	BK004047
K15	Ka15	NM_002275	Ka15	NM_008469	Ka15	BK004045
K16	Ka16	NM_005557	Ka16	NM_008470	Ka16	BK004049
K17	Ka17	NM_000422	Ka17	NM_010663	Ka17	BK004050
K18	Ka18	NM_000224	Ka18	NM_010664	Ka18	gap
K19	Ka19	NM_002276	Ka19	NM_008471	Ka19	BK004046
K20	Ka20	BC031559	Ka20	AK018567	Ka20	M63665
Ka21P*	Ka21P	BK004052	Ka21P	AK031954	Ka21P	BK004026
Ka22P*	Ka22P	BK004056	Ka22	BK004025	Ka22	BK004051
K23	Ka23	BC028356	Ka23	AF102849	Ka23	BK004034
K24,K10B	Ka24	AK000268	Ka24	AK010165	Ka24	BK004027
K25A,K10C	Ka38	NM_181534	Ka38	BC018391	Ka38	BK004028
K25B,K10D	Ka39	NM_181539	Ka39	AK028591	Ka39	BK004029
K25C,K12b	Ka40	NM_181537	Ka40	AK077384	Ka40	BK004030
K25D	Ka41	NM_181535	Ka41	AK014642	Ka41	BK004031
KRTHA1	Ka25	X86570	Ka25	NM_010659	Ka25	BK004040
KRTHA2	Ka26	NM_002278	Ka26	NM_010665	Ka26	BK004041
KRTHA3A	Ka27	NM_004138	Ka27	BC029257	Ka27	BK004037
KRTHA3B	Ka28	X82634	Ka28	X75650	Ka28	BK004038
KRTHA4	Ka29	NM_021013	Ka29	NM_027563	Ka29	BK004039
KRTHA5	Ka30	NM_002280	Ka30	AF020790	Ka30	BK004042
KRTHA6	Ka31	NM_003771	Ka31	AK028845	Ka31	BK004043
KRTHA7	Ka32	NM_003770	~	~	~	~
KRTHA8	Ka33	NM_006771	~	~	~	~
ψKRTHAA	Ka34P	Y16795	~	~	~	~
Ka35*	Ka35	BK004054	Ka35	BK004022	Ka35	BK004035
Ka36*	Ka36	BK004055	Ka36	BK004023	Ka36	BK004036
Ka37P*	Ka37P	BK004053	~	~	~	~

Old name	New names					
Type II	HUMAN	Acc. No.	MOUSE	Acc. No.	RAT	Acc. No.
K1	Kb1	NM_006121	Kb1	M10937	Kb1	BK001580
K1b	Kb39	AJ564104	Kb39	BK003993	Kb39	BK003984
K2e	Kb2	AF019084	Kb2	X74784	Kb2	BK001581
K2p	Kb9	M99063	Kb9	AK009075	Kb9	BK003994
K3	Kb3	NM_057088	~	~	gap	
K4	Kb4	NM_002272	Kb4	X03491	Kb4	BK003985
K5	Kb5	NM_000424	Kb5	BC006780	gap	
K5b	Kb40	AK096419	Kb40	BK004080	Kb40	BK004081
K6a	Kb6	NM_005554	Kb6	NM_008476	gap	
K6b	Kb10	NM_005555	~	~	~	~
mK6b	~	~	Kb11	NM_010669	gap	
mK6d	~	~	Kb15	AK028791	Kb15	BK003995
K6e	Kb12	NM_173086	~	~	~	~
K6hf	Kb18	NM_004693	Kb18	NM_133357	Kb18	BK003980
K6irs1,K6i	Kb34	AJ308599	Kb34	NM_019956	gap	
K6irs2,K6k	Kb35	NM_080747	Kb35	BK001584	Kb35	BK003981
K6irs3	Kb36	AJ508776	Kb36	BK003988	Kb36	BK003982
K6irs4,K5c	Kb37	AJ508777	Kb37	BK003986	gap	
K6l	Kb38	BC039148	Kb38	BC031593	gap	
K7	Kb7	NM_005556	Kb7	NM_033073	Kb7	BK003973
K8	Kb8	X74929	Kb8	NM_031170	Kb8	M63482
Kb13*	~	~	Kb13	BC046626	gap	
Kb14*	~	~	Kb14	BK003987	gap	
mK6Ψ1	~	~	Kb16P	BK003992	gap	
mK6Ψ2	~	~	Kb17P	BK003991	Kb17P	BK003998
Kb19P*	Kb19P	BK003989	~	~	~	~
Kb20*	Kb20	BC047308	Kb20	AK004811	Kb20	BK001582
KRTHB1	Kb21	NM_002281	Kb21	AF312018	Kb21	BK003974
KRTHB2	Kb22	NM_033033	Kb22	AY028606	Kb22	BK003979
KRTHB3	Kb23	NM_002282	Kb23	M92088	Kb23	BK003976
KRTHB4	Kb24	NM_033045	Kb24	AY028607	Kb24	BK003978
KRTHB5	Kb25	NM_002283	Kb25	BK001583	Kb25	BK003976
KRTHB6	Kb26	NM_002284	Kb26	X99143	Kb26	BK003975
Kb27*	~	~	gap		Kb27	n.p.
ψHbA	Kb28P	Y19213	~	~	~	~
ψHbB	Kb29P	Y19214	~	~	~	~
ψHbC	Kb30P	Y19215	~	~	~	~
ψHbD	Kb31P	Y19216	~	~	~	~
Kb32P*	~	~	Kb32P	BK003990	Kb32P	BK003997
Kb33P*	~	~	Kb33P	NM_025487	Kb33P	BK003996

* newly identified in this work

m: mouse

h: human

n.p.: no reliable prediction

Past the hair keratin region, the 3 mammalian type I clusters remain very similar with very good conservation of synteny. We note a mouse pseudogene *Ka11P,* probably derived from *Ka16* (*K16*), which has no human counterpart. In rat, this gene is supposed to be functional (*Ka11*). The type I cluster ends with a novel rodent gene (*Ka22*) related to keratins 14 and 17. In man this gene was converted into a pseudogene (*Ka22P*). It is remarkable that cytokeratins (*Ka9-Ka24*), hair keratins (*Ka25-Ka36*) and inner root sheath keratins (*Ka38-Ka41*) are clustered together in small subdomains.

The gene for *K11*, proposed previously (Moll et al., 1982; Rieger and Franke, 1988) is undetectable in the human or rodent genomes and must therefore be considered a variant of an existing keratin, such as *Ka10* (*K10*) in agreement with former studies (Mischke, 1998).

The mammalian keratin type II clusters

The sequence of the human keratin II gene cluster located on chromosome 12q13.13 (size 783 kb) is now complete. The last release by the International Human Genome Sequencing Consortium (NCBI Build 34, July 2003) removed various violations of synteny versus the murine data for the last fifth of the cluster still present in previous releases (see Nov. 2002). The corresponding murine gene cluster, located on chromosome 15F3 (size 699 kb) still shows a variety of gaps (release Feb. 2003), but most of the genes in the cluster are unambiguously located versus their human orthologs. The rat keratin II gene cluster on chromosome 7Q35 (UCSC version rn3, June 2003) has a large central gap of 160 kbp, which may obscure 6 to 8 genes, and various smaller gaps as indicated in Figure 2. Nearly all type II genes have the same orientation. Notable exceptions are for instance *Kb7 (K7)* and the terminal type I *Ka18 (K18).*

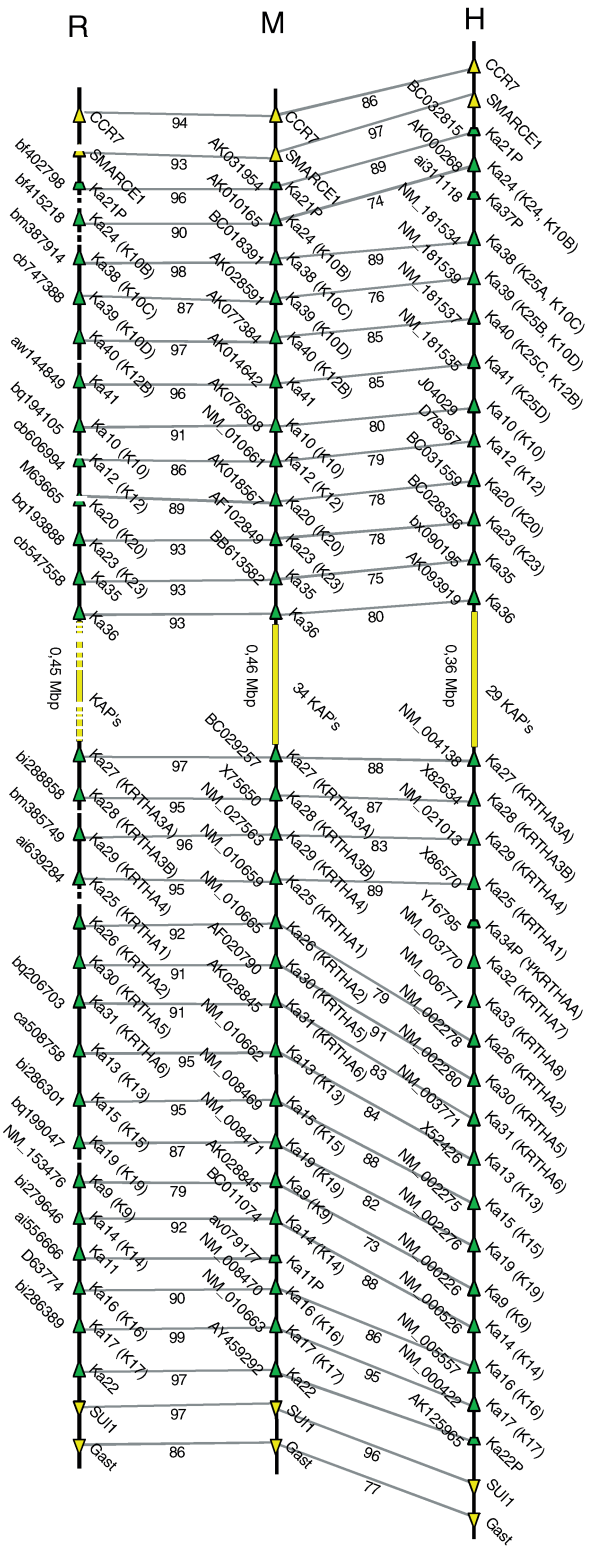

Fig. 1. The keratin I gene clusters of man (H), mouse (M) and rat (R). Remaining sequence gaps in the rat keratin type I cluster are indicated by gaps and breaks in the corresponding sequence line. Keratin genes are marked by green triangles with the tip corresponding to the 3′ end. Keratin genes are identified by the numbers used in the proposed nomenclature. For convenience, previously used numbers are added in parenthesis. Pseudogenes (blunted triangles) are marked by P. For details on the nomenclature see Table 1. If available, cDNA accession numbers are provided. Accession numbers of EST sequences start with small letters. The genes flanking the type I cluster are *CCR7* (C-C chemokine receptor type 7 precursor) and *SMARCE1* (SWI/SNF related, matrix associated, actin dependent regulator of chromatin, subfamily e, member 1) on one side and *SUI1* (translation initiation factor SUI1 isolog) and GAST (gastrin precursor) on the other side. These flanking genes are marked in yellow. The approximate position of the multiple KAP genes interrupting the type I cluster is given. For details see text. Conservation of synteny is indicated by the gray lines connecting corresponding human/mouse and mouse/rat genes respectively. Numbers at these lines give the percent sequence identity on the protein level.

The human keratin II cluster (size 783 kb) consists of 26 type II genes, 5 type II pseudogenes, 3 keratin-unrelated pseudogenes and ends with the type I keratin *Ka18 (K18)* gene. The gene density for the cluster is 22.4 kb (29 kb excluding pseudogenes). Figure 2 shows that cDNA accession numbers are available for all human type II genes. In the murine genome there are at least 27 functional type II genes, resulting in a gene density of 21.9 kb (25 kb excluding pseudogenes) and a further type II gene past *Kb7* (K7) present in the rat cluster may be obscured by a gap in the murine sequence (Fig. 2). Among the cDNA accession numbers for the murine genes there is also one for *Kb33P* (Fig. 2) indicating that this hair keratin pseudogene may be transcribed.

Previous work physically mapped the large region from the human keratin *Kb7* (K7) gene to the keratin *Kb2* (*K2e*) gene (Langbein et al., 2003), which harbors all type II hair keratin genes *Kb21* to *Kb26* (*Hb1* to *Hb6*), their 4 pseudogenes *Kb28P* to *Kb31P* (ψ*hHbA* to ψ*hHbD*), the genes for *Kb18* (K6hf), *Kb10* (*K6b*), *Kb12* (*K6e*), *Kb6* (*K6a, K6c*), *Kb5* (*K5*) and the four genes expressing the inner-root-sheath-specific type II keratins *Kb34* to *Kb37* (*K6irs1* to *K6irs4*). The region from *Kb7 (K7)* to *Kb2 (K2e)* is preceded by only a single type II gene *(Kb20)* in all 3 genomes. It seems difficult to relate this gene unambiguously to a certain subgroup of type II genes. Based on known cDNA sequences, *Kb20* is expressed in human cervix carcinoma, adult murine lung tissue and rat placenta. Using RT-PCR followed by DNA sequencing, we have verified the expression of this murine gene in lung tissue and oesophagus (data not shown). Past the human *Kb2* (K2e) gene are the genes for *Kb1* (*K1*), *Kb39* (*K1b*) and a pseudogene (*Kb19P)* followed by three keratin unrelated genes: a processed pseudogene of *prp28* (chromosome 12q13.11), which codes for a 100-kDa U5 snRNP-binding protein, a processed pseudogene of the potassium channel tetramerization protein *MGC 13007* (chromosome 11) and a processed pseudogene of *BART1* (chromosome 16), an ARL2 binding protein. These 3 non-keratin processed pseudogenes are not present in the rodent type II clusters. Indeed, the murine type II cluster seems to lack non-keratin genes entirely while the rat type II cluster harbors a processed pseudogene of ribosomal protein L4 (*RPL4*; chromosome 8). A short rat gene fragment predicting at 100% identity 370 nt (exon 21 and part of an intron) of alpha II spectrin (*Spna2*) is also present (Fig. 2).

The human gene for keratin *Kb39* (*K1b*) describes a novel keratin II gene (Hesse et al., 2001) with 66% identity to *Kb1* (K1) on the protein level. A human cDNA and an accession number of an EST sequence for the murine ortholog indicates that *Kb39* (*K1b*) is expressed in the skin and we have verified

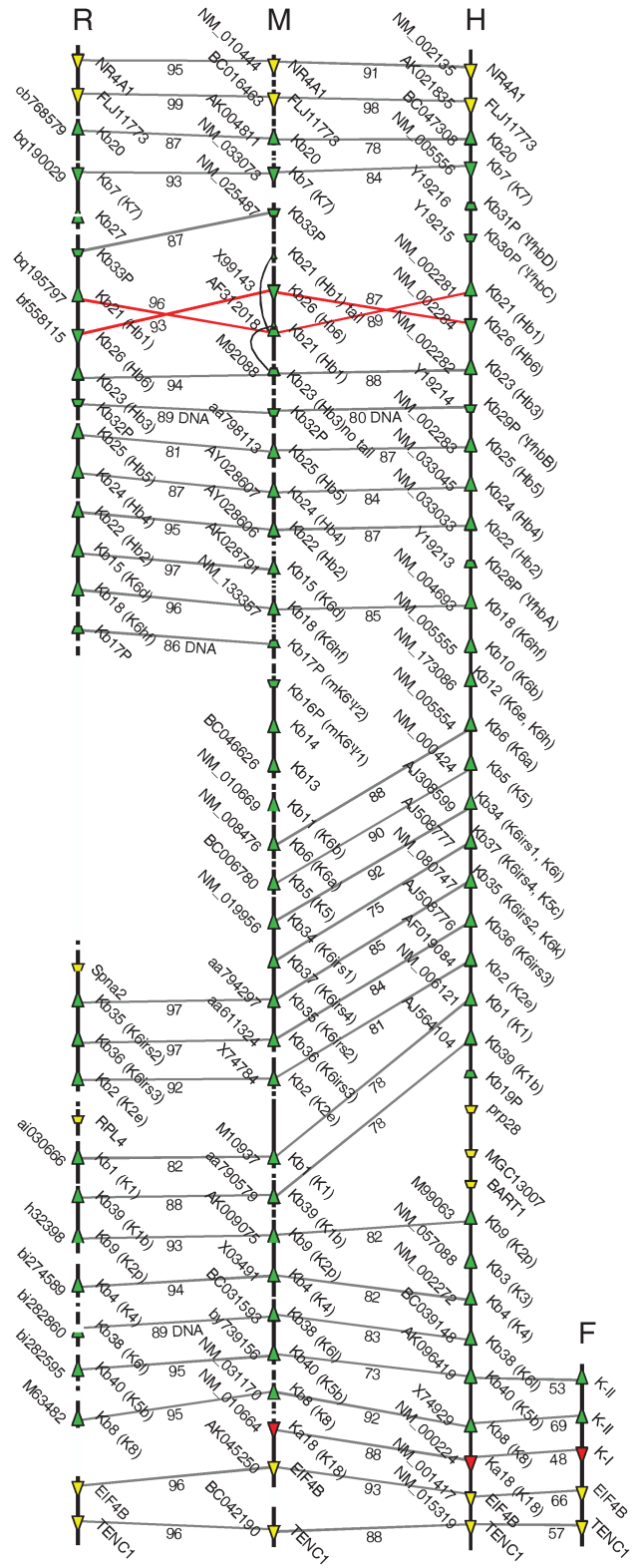

Fig. 2. The keratin II gene clusters of man (H), mouse (M) and rat (R). Remaining sequence gaps in the mouse and rat clusters are indicated by gaps and breaks in the corresponding sequence lines. Keratin genes are marked by green triangles with the tip corresponding to the 3′ end. Keratin genes are identified by the numbers used in the proposed nomenclature. For convenience previously used numbers are added in parenthesis. Pseudogenes (blunted triangles) are marked by P. For details on the nomenclature see Table 1. Flanking genes for the type II cluster are *NR4A1* (nuclear receptor subfamily 4, group A) and *FLJ11773* (hypothetical protein FLJ11773) on one side and *EIF4B* (eukaryotic translation factor 4A) and *TENC1* (tensin-like C1 domain containing phosphatase) at the other end. When available, cDNA accession numbers are provided. Accession numbers of EST sequences start with small letters. The type I keratin *Ka18 (K18)* gene at the end of the keratin II cluster is given in red. Yellow color is used to mark non-keratin genes. These are the non-keratin genes flanking the clusters and in the case of the human and rat type II gene cluster, 3 unrelated processed pseudogenes (human) and 2 processed pseudogenes (rat). Conservation of synteny is indicated by the gray lines connecting corresponding human/mouse and mouse/rat genes respectively. Numbers at these lines give the percent sequence identity on the protein level except for the cases marked 'DNA'. Note a potential case of violation of synteny (crossed red lines) early in the murine type II cluster which is discussed in the text. The right lower corner shows a DNA scaffold from the teleost fish *Fugu rubripes* (F). The five genes resemble in orientation and arrangement the end of the mammalian keratin II gene cluster (Zimek at al. 2003).

this assignment by RT-PCR (data not shown). Past the 3 non-keratin pseudogenes the human type II cluster continues with another six type II genes for *Kb9* (*K2p*), *Kb3 (K3)*, *Kb4 (K4)*, *Kb38 (K6l)*, *Kb40 (K5b)* and *Kb8 (K8)* before the end position, occupied by the type I keratin *Ka18 (K18)* gene, is reached.

Contrary to some previous releases the human keratin *Ka18 (K18)* gene is adjacent to the *Kb8 (K8)* gene. Keratins *Kb8 (K8)* and *Ka18 (K18)* are typical of interior epithelia and represent the earliest IF expression pair in embryogenesis. Within the human gene cluster 24 type II genes and 3 type II pseudogenes have the same orientation. The opposite orientation holds for only 2 type II genes, *Kb7 (K7)* and *Kb26* (*Hb6*), 2 type II pseudogenes, *Kb29P* and *Kb30P (ψhHbB* and ψ*hHbC*), the keratin *Ka18 (K18)* gene and the 3 non-keratin pseudogenes.

Except for a short region in the murine keratin type II gene cluster, there is no apparent violation of conservation of synteny. The murine hair keratin genes *Kb21* (*Hb1*) and *Kb26* (*Hb6*) differ in relative position and orientation from their human and rat counterparts, and the murine *Kb23* (*Hb3*) gene currently occurs in two parts adjacently located to *Kb21* (*Hb1*) and *Kb26* (*Hb6*), respectively. We expect that synteny will be established once the small murine sequence gaps in this region are closed (Fig. 2). In the region of keratin 6 gene isotypes, between *Kb18* (*K6hf*) and *Kb5* (*K5*), we have not drawn lines to indicate potential conservation of synteny because of the large gap in the rat cluster and the hypothesis that these murine genes arose independently from their human orthologs (Takahashi et al., 1998). The previously reported human *Kb6* (*K6*) genes *K6c*, *d*, and *f* (Takahashi et al., 1995) were not detectable in the human genome. In agreement with their high degree of sequence identity, *K6c* and *K6d* may be polymorphic protein variants of *Kb6* (*K6a*). *K6f* could be a variant of *Kb10* (*K6b*).

Interestingly, the human cluster contains between *Kb18 (K6hf)* and *Kb5 (K5)* only 3 related genes, *Kb10 (K6b)*, *Kb12 (K6e)* and *Kb6 (K6a)*, while the mouse shows 4 genes, *Kb14, Kb13, Kb11 (K6b), Kb6 (K6a)* and the two pseudogenes *Kb17P* and *K16P*.

There are a few changes in some gene/pseudogene relations between the human and the rodent type II clusters. While there are 4 human hair keratin II pseudogenes *Kb28P* to *Kb31P* (ψ*hHbA* to ψ*hHbD*), the rat and the mouse have only two

pseudogenes *Kb32P* and *Kb33P* (Fig. 2). The human hair keratin pseudogene *Kb28P* (ψ*hHbA*) corresponds in the mouse to the *Kb15* gene (80% identity on the DNA level) which is also present in the rat (Fig. 2). The accession number of a corresponding murine skin cDNA clone is compatible with an active keratin gene. Not always are there more pseudogenes in man than in rodents. Thus the murine type II cluster shows the two *Kb6 (K6)*-related pseudogenes *Kb16P* and *Kb17P* (*mK6ψ1*, *mK6ψ2*) reported earlier (Takahashi et al., 1998). They directly follow *Kb18* (*K6-hf*), but have no counterparts in the human keratin type II cluster. Furthermore, the human pseudogene *Kb19P* has no rodent counterparts.

Unexpectedly, the gene for human cornea *Kb3 (K3)* situated between genes *Kb9* (*K2p*) and *Kb4 (K4)* lacks a murine counterpart while a rat sequence gap obscures the possible presence of a gene located between *Kb9* (*K2p*) and *Kb4 (K4)*. Thus the human *Kb3 (K3)* gene may be the result of a recent gene duplication or an older mammalian *Kb3 (K3)* gene was lost on the lineage leading to mice. Using RT-PCR we found that *Kb5 (K5)* is the major type II keratin expressed in the murine cornea (data not shown).

For the teleost fish *Fugu rubripes* more than 12000 DNA scaffolds have been provided (Aparicio et al., 2002). Scaffold 173 contains 2 adjacent keratin type II genes in the same orientation. They are followed by a keratin type I gene, the gene *EIF-4B* (eukaryotic translation factor 4B) and the gene *TENC1* (tensin-like C1 domain containing phosphatase) all in opposite orientation (Zimek et al., 2003). Interestingly, this is precisely the gene arrangement at the *Ka18 (K18)* end of the mammalian keratin II gene cluster (see Fig. 2).

Processed keratin pseudogenes

With respect to keratins, the most significant difference between the human and the rodent genomes is the number of pseudogenes for *Kb8* (*K8*) and *Ka18* (*K18*). In the human genome there are at least 77 pseudogenes for *Ka18 (K18)* and 61 for *Kb8 (K8)* (Hesse et al., 2001). In contrast to our previous report, we now use the definition for pseudogenes of the Mouse Genome Sequencing Consortium, 2002, avoiding the term fragments. In mouse and rat the total number of pseudogenes for *Kb8 (K8)* and *Ka18 (K18)* is 4 and 6, respectively. In humans the 138 pseudogenes for *Kb8 (K8)/Ka18 (K18)* are dispersed throughout the genome and can be found on every chromosome (Hesse et al., 2001). Whether this difference is of any biological significance remains to be seen.

While the rodent genomes lack *Ka19 (K19)* pseudogenes, there are five processed pseudogenes in the human sequence. These are located on chromosomes 4, 6, 10 and 12 (two pseudogenes). Due to deletions and point mutations they are supposed to be inactive.

Gene duplications

During evolution, a region of the human type I cluster containing the genes for *Ka14 (K14), Ka16 (K16), Ka17 (K17)* and *Ka22P*, inserted four times in different regions of chromosome 17 (Fig. 3). Two of these duplications have been described previously (Troyanovsky et al., 1992). Remarkably, flanking regions with a size of up to 150 kb were subsequently duplicated together with the keratin genes. From the homology of the flanking regions and the size of the multiplied part of the type I cluster, we tentatively deduced the timely order of the duplication events shown in Figure 3. The insertions gave rise to three unprocessed pseudogenes for *Ka14 (K14)* and *Ka16 (K16)* each, and to four for *Ka17 (K17)* and *Ka22P*. Due to frameshift mutations all these genes are assumed to be non-functional. In contrast to human, similar duplications were not detected in the mouse or rat genomes.

Non-keratin IF genes and pseudogenes

Comparative analysis of the three genomes revealed one functional gene for each of the known non-keratin intermediate filament proteins (Hesse et al., 2001). A search for previously unknown IF genes using homology criteria revealed no further genes. Paranemin, a chicken IF gene, was not detected in all three mammalian genomes, despite a report on paranemin expression in the mouse based on histochemistry (Carlsson et al., 2000). Non-keratin IF genes do not cluster in the human genome (Hesse et al., 2001). The sole exception concerns the genes for the neurofilament proteins NF-L and NF-M, which are adjacent genes in all 3 mammalian genomes. Their distance is 33 kb in human, 32 kb in mouse and 58 kb in rat. While there are truncated, processed pseudogenes in man for NF-L (parts of exons 3 and 4 on the Y-chromosome) and NF-M (part of exon 4 on chromosome 10) and NF-H (middle part of exon 4 missing, on chromosomes 20 and 1), the mouse and the rat lack neurofilament pseudogenes.

There is no gene coding for the oocyte-specific lamin LIII of *Xenopus laevis* in either of the analysed mammalian genomes. It seems that this gene became redundant with the evolution of the mammalian reproductive tract. In contrast to humans, a pseudogene for lamin B2 could be detected in mice and rats. In rat, this pseudogene resides in intron 1 of a functional gene encoding a synaptic scaffolding molecule (Acc. No.: AF034863). The pseudogene for lamin B2 is processed in both species. It lacks part of exon 1, and is most likely non-functional.

Also the pseudogene for vimentin in humans consisting of parts of exons 3, 4, 6 and 9 is undetectable in the rodent genomes.

Perspective

Despite extensive analysis, the principles governing the pairwise and tissue-specific activity of keratin genes remain unknown. A preliminary search based on available algorithms has failed to reveal common regulatory sequences in promotors of coexpressed genes. The reported activity of *Kb8* (*K8*) and *Ka18* (*K18*) in several non-epithelial tissues (Bader et al., 1988) argues, that, whatever governs non-epithelial repression of the type II gene cluster, can be overcome by transcriptional activation of *Kb8* (*K8*) and *Ka18* (*K18*). Regarding the occurrence of 138 pseudogenes for *Kb8* (*K8*) and *Ka18* (*K18*) in humans but not in the mouse, keratin genes should present a useful object to study the mechanisms governing their evolution. Evidence from knockout mice and human genetics provides strong support for the vital function of at least 17 keratin genes (Coulombe and Omary, 2002; Herrmann et al., 2003). The notion, that in the mouse, *Kb5* (*K5*) is expressed instead of *Kb3* (*K3*) in the cornea, can be taken as an indicator that certain keratins replace others. While this is supported by *Ka18* (*K18*) and *Ka19* (*K19*) null mice (Magin et al., 1998; Tamai et al., 2000), it is unlikely as a general principle. The present report provides a rational basis to examine specific keratin functions by large scale genome engineering.

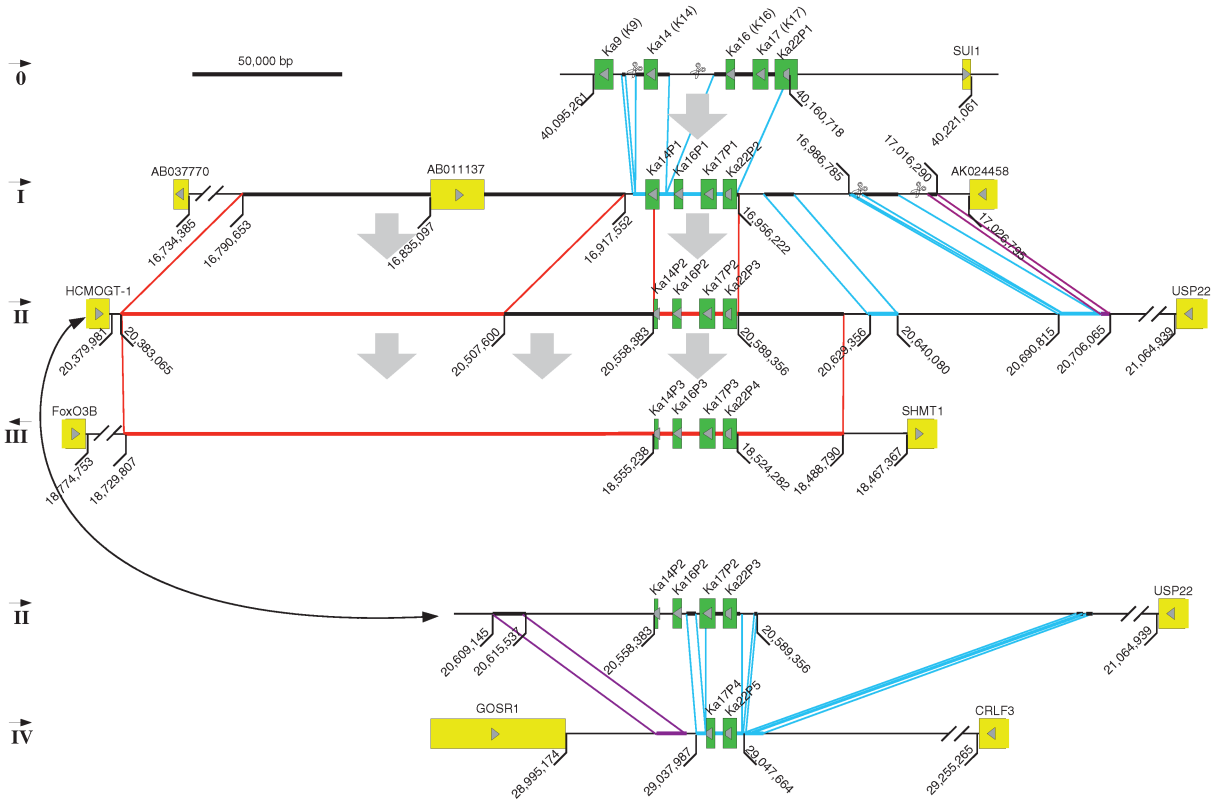

Fig. 3. Proposed events leading to duplication of a part of the keratin type I cluster on chromosome 17 in the human genome. Green boxes mark keratin genes and pseudogenes. Gray triangles indicate the orientation of the genes with the tips corresponding to the 3′-end of the gene. Big gray arrows mark regions duplicated with high homology, and indicate a "hierarchy" to explain, from which template the corresponding duplication was derived. Dashed lines border the duplicated regions. The scissors symbols indicate parts of the genomic DNA that were not duplicated. The numbers given correspond to the NCBI build 34, July 2003. Accession numbers were taken from GenBank (NCBI). The cytogenetic positions of the duplicated blocks 0-IV on chromosome 17 are as follows: 0:17q21; I–III: 17p11; IV: 17q11. *SUI1* protein translation factor SUI1 homolog (AF083441); *AK024458* FLJ00050 protein (AK024458); *AB037770* hypothetical zinc finger protein KIAA1349 (AB037770); *USP22* ubiquitin carboxyl-terminal hydrolase 22 (AB028986); *HCMOGT-1* sperm antigen HCMOGT-1 (BC021123); *FoxO3B* forkhead box O3B (AF041336); *SHMT1* serine hydroxymethyltransferase 1 (Y14485); *GOSR1* cis-Golgi SNARE p28 (AF073926); *CRLF3* cytokine receptor related protein 4 (AF046059). Red: 98–99.9% identity; blue: 94–97.9% identity, purple: 87–93.9% identity. Blunted arrowheads indicate fragments. The bar corresponds to 50 kb.

References

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., Lipman, D. J., 1990. Basic local alignment search tool. J. Mol. Biol. 215, 403–410.

Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W., Lipman, D. J., 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25, 3389–3402.

Aparicio, S., Chapman, J., Stupka, E., Putnam, N., Chia, J. M., Dehal, P., Christoffels, A., Rash, S., Hoon, S., Smit, A., Gelpke, M. D., Roach, J., Oh, T., Ho, I. Y., Wong, M., Detter, C., Verhoef, F., Predki, P., Tay, A., Lucas, S., Richardson, P., Smith, S. F., Clark, M. S., Edwards, Y. J., Doggett, N., Zharkikh, A., Tavtigian, S. V., Pruss, D., Barnstead, M., Evans, C., Baden, H., Powell, J., Glusman, G., Rowen, L., Hood, L., Tan, Y. H., Elgar, G., Hawkins, T., Venkatesh, B., Rokhsar, D., Brenner, S., 2002. Whole-genome shotgun assembly and analysis of the genome of *Fugu rubripes*. Science 297, 1301–1310.

Bader, B. L., Jahn, L., Franke, W. W., 1988. Low level expression of cytokeratins 8, 18 and 19 in vascular smooth muscle cells of human umbilical cord and in cultured cells derived therefrom, with an analysis of the chromosomal locus containing the cytokeratin 19 gene. Eur. J. Cell Biol. 47, 300–319.

Bawden, C. S., McLaughlan, C., Nesci, A., Rogers, G., 2001. A unique type I keratin intermediate filament gene family is abundantly expressed in the inner root sheaths of sheep and human hair follicles. J. Invest. Dermatol. 116, 157–166.

Carlsson, L., Li, Z. L., Paulin, D., Price, M. G., Breckler, J., Robson, R. M., Wiche, G., Thornell, L. E., 2000. Differences in the distribution of synemin, paranemin, and plectin in skeletal muscles of wild-type and desmin knock-out mice. Histochem. Cell Biol. 114, 39–47.

Coulombe, P. A., Omary, M. B., 2002. 'Hard' and 'soft' principles defining the structure, function and regulation of keratin intermediate filaments. Curr. Opin. Cell Biol. 14, 110–122.

Herrmann, H., Hesse, M., Reichenzeller, M., Aebi, U., Magin, T. M., 2003. Functional complexity of intermediate filament cytoskeletons: from structure to assembly to gene ablation. Int. Rev. Cytol. 223, 83–175.

Hesse, M., Magin, T. M., Weber, K., 2001. Genes for intermediate filament proteins and the draft sequence of the human genome: novel keratin genes and a surprisingly high number of pseudogenes related to keratin genes 8 and 18. J. Cell Sci. 114, 2569–2575.

Karabinos, A., Schmidt, H., Harborth, J., Schnabel, R., Weber, K., 2001. Essential roles for four cytoplasmic intermediate filament proteins in *Caenorhabditis elegans* development. Proc. Natl. Acad. Sci USA 98, 7863–7868.

Karabinos, A., Schulze, E., Schunemann, J., Parry, D. A., Weber, K., 2003. In vivo and in vitro evidence that the four essential intermediate filament (IF) proteins A1, A2, A3 and B1 of the nematode *Caenorhabditis elegans* form an obligate heteropolymeric IF system. J. Mol. Biol. 333, 307–319.

Karolchik, D., Baertsch, R., Diekhans, M., Furey, T. S., Hinrichs, A., Lu, Y. T., Roskin, K. M., Schwartz, M., Sugnet, C. W., Thomas, D. J., Weber, R. J., Haussler, D., Kent, W. J.; University of California Santa Cruz, 2003. The UCSC Genome Browser Database. Nucleic Acids Res. 31, 51–54.

Kent, W. J., 2002. BLAT – the BLAST-like alignment tool. Genome Res. 12, 656–664.

Lander, E. S. et al. (International Human Genome Sequencing Consortium), 2001. Initial sequencing and analysis of the human genome. Nature 409, 860–921.

Langbein, L., Rogers, M. A., Praetzel, S., Winter, H., Schweizer, J., 2003. K6irs1, K6irs2, K6irs3, and K6irs4 represent the inner-root-sheath-specific type II epithelial keratins of the human hair follicle. J. Invest. Dermatol. 120, 512–522.

Langbein, L., Rogers, M. A., Winter, H., Praetzel, S., Beckhaus, U., Rackwitz, H. R., Schweizer, J., 1999. The catalog of human hair keratins. I. Expression of the nine type I members in the hair follicle. J. Biol. Chem. 274, 19874–19884.

Magin, T. M., Schroder, R., Leitgeb, S., Wanninger, F., Zatloukal, K., Grund, C., Melton, D. W., 1998. Lessons from keratin 18 knockout mice: formation of novel keratin filaments, secondary loss of keratin 7 and accumulation of liver-specific keratin 8-positive aggregates. J. Cell Biol. 140, 1441–1451.

Mischke, D., 1998. The complexity of gene families involved in epithelial differentiation. Keratin genes and the epidermal differentiation complex. Subcell. Biochem. 31, 71–104.

Moll, R., Franke, W. W., Schiller, D. L., Geiger, B., Krepler, R., 1982. The catalog of human cytokeratins: patterns of expression in normal epithelia, tumors and cultured cells. Cell 31, 11–24.

Rieger, M., Franke, W. W., 1988. Identification of an orthologous mammalian cytokeratin gene. High degree of intron sequence conservation during evolution of human cytokeratin 10. J. Mol. Biol. 204, 841–856.

Rogers, M. A., Langbein, L., Winter, H., Ehmann, C., Praetzel, S., Korn, B., Schweizer, J., 2001. Characterization of a cluster of human high/ultrahigh sulfur keratin-associated protein (KAP) genes embedded in the type I keratin gene domain on chromosome 17q12–21. J. Biol. Chem. 276, 19440–19451.

Rogers, M. A., Langbein, L., Winter, H., Ehmann, C., Praetzel, S., Schweizer, J., 2002. Characterization of a first domain of human high glycine-tyrosine and high sulfur keratin-associated protein (KAP) genes on chromosome 21q22.1. J. Biol. Chem. 277, 48993–49002.

Rogers, M. A., Winter, H., Langbein, L., Wolf, C., Schweizer, J., 2000. Characterization of a 300 kbp region of human DNA containing the type II hair keratin gene domain. J. Invest. Dermatol. 114, 464–472.

Sun, T. T., Eichner, R., Nelson, W. G., Tseng, S. C., Weiss, R. A., Jarvinen, M., Woodcock-Mitchell, J., 1983. Keratin classes: molecular markers for different types of epithelial differentiation. J. Invest. Dermatol. 81, 109 s–115 s.

Takahashi, K., Paladini, R. D., Coulombe, P. A., 1995. Cloning and characterization of multiple human genes and cDNAs encoding highly related type II keratin 6 isoforms. J. Biol. Chem. 270, 18581–18592.

Takahashi, K., Yan, B., Yamanishi, K., Imamura, S., Coulombe, P. A., 1998. The two functional keratin 6 genes of mouse are differentially regulated and evolved independently from their human orthologs. Genomics 53, 170–183.

Tamai, Y., Ishikawa, T., Bosl, M. R., Mori, M., Nozaki, M., Baribault, H., Oshima, R. G., Taketo, M. M., 2000. Cytokeratins 8 and 19 in the mouse placental development. J. Cell Biol. 151, 563–572.

Troyanovsky, S. M., Leube, R. E., Franke, W. W., 1992. Characterization of the human gene encoding cytokeratin 17 and its expression pattern. Eur. J. Cell Biol 59, 127–137.

Waseem, A., Alexander, C. M., Steel, J. B., Lane, E. B., 1990. Embryonic simple epithelial keratins 8 and 18: chromosomal location emphasizes difference from other keratin pairs. New Biol. 2, 464–478.

Waterston, R. H. and others (2002. Initial sequencing and comparative analysis of the mouse genome. Nature 420, 520–562.

Zimek, A., Stick, R., Weber, K., 2003. Genes coding for intermediate filament proteins: common features and unexpected differences in the genomes of humans and the teleost fish *Fugu rubripes*. J. Cell Sci. 116, 2295–2302.