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

Michael Hesse[1,*], Thomas M. Magin[1,*] and Klaus Weber[2,*]

[1]Institute of Genetics, Division of Molecular Genetics and Bonner Forum Biomedizin, University of Bonn, 53117 Bonn, Germany
[2]Max-Planck-Institute for Biophysical Chemistry, Department of Biochemistry, 37070 Goettingen, Germany
*Authors for correspondence (e-mail: t.magin@uni-bonn.de; m.hesse@uni-bonn.de; r.longo@gwdg.de)

Accepted 23 May 2001
Journal of Cell Science 114, 2569-2575 (2001) © The Company of Biologists Ltd

SUMMARY

We screened the draft sequence of the human genome for genes that encode intermediate filament (IF) proteins in general, and keratins in particular. The draft covers nearly all previously established IF genes including the recent cDNA and gene additions, such as pancreatic keratin 23, synemin and the novel muscle protein syncoilin. In the draft, seven novel type II keratins were identified, presumably expressed in the hair follicle/epidermal appendages. In summary, 65 IF genes were detected, placing IF among the 100 largest gene families in humans. All functional keratin genes map to the two known keratin clusters on chromosomes 12 (type II plus keratin 18) and 17 (type I), whereas other IF genes are not clustered. Of the 208 keratin-related DNA sequences, only 49 reflect true keratin genes, whereas the majority describe inactive gene fragments and processed pseudogenes. Surprisingly, nearly 90% of these inactive genes relate specifically to the genes of keratins 8 and 18. Other keratin genes, as well as those that encode non-keratin IF proteins, lack either gene fragments/pseudogenes or have only a few derivatives. As parasitic derivatives of mature mRNAs, the processed pseudogenes of keratins 8 and 18 have invaded most chromosomes, often at several positions. We describe the limits of our analysis and discuss the striking unevenness of pseudogene derivation in the IF multigene family. Finally, we propose to extend the nomenclature of Moll and colleagues to any novel keratin.

Key words: Human genome, intermediate filament proteins, keratins, lamins, neurofilament proteins, pseudogenes, disease.

INTRODUCTION

The increase in specific cell types represents one hallmark of metazoan evolution. It is paralleled by the acquisition of multigene families, which often encode proteins of similar structure but distinct function. One such family is represented by the intermediate filament protein (IF) family. Its members form part of the cytoskeleton of most metazoan cells. Vertebrate IF are organised into five distinct gene families according to sequence identity and expression patterns (Fuchs and Weber, 1994; Herrmann and Aebi, 2000). These include keratins (K), which represent the type I and II homology groups encoded by more than 20 genes, and a further 15 hair keratin genes (Langbein et al., 1999; Rogers et al., 2000), the type III proteins desmin, vimentin, GFAP and peripherin, and the type IV homology group, which encompasses α-internexin, syncoilin (Newey et al., 2001), nestin, synemin and the neurofilament proteins NF-L, -M and -H. The nuclear lamins A/C, B1 and B2 form the type V IF, whereas the eye lens proteins phakinin and filensin constitute a separate group. All 16 known non-keratin IF proteins, including syncoilin (Newey et al., 2001) and synemin (Becker et al., 1995; M. Titeux et al., unpublished), were identified by biochemical, immunological and cDNA cloning methods. The power of the classical approach is best exemplified by the pioneering work of Moll and Franke, who in 1982 established the 'catalog of human cytokeratins' (Moll et al., 1982). They laid the groundwork for keratin expression profiles and provided a rational nomenclature. Their data were based on the isolation of keratins from microdissected normal and tumor tissues, as separated in high resolution 2D gels. The numbering system for type II keratins ranges from 1 to 8 with letters for later additions and from 9 to 21 for type I keratins. Hair keratins were named in an analogous way with letters Ha and Hb indicating type I and II hair keratins, respectively (Langbein et et al., 1999; Rogers et al., 2000). Subsequent work established that all IF proteins, with the exception of a few polymorphic variants (Mischke and Wild, 1987; Korge et al., 1992), are encoded by single copy genes (Fuchs and Weber, 1994). One difficulty of the classical biochemical and genetic approach is that potential minor keratins and other IF proteins, present in only a few cells of a tissue, or expressed transiently during embryonic development, may have escaped detection.

Gene mapping studies revealed that genes coding for non-keratin IF proteins are not clustered (International Human Genome Sequencing Consortium, 2001). All type I keratin genes (except K18; Waseem et al., 1990) are clustered on chromosome 17q21 and type II genes on 12q13 (International

IF Gene	Chromosome	# Pseudogenes	# Gene Fragments
Type I			
K9	17	-	-
K10	17	-	-
K10b‡	17	-	-
K10c‡	17	-	-
K10d‡	17	-	-
K12	17	-	-
K12b‡	17	-	-
K13	17	-	-
K14	17	1 (17)	1 (17)
K15	17	-	-
K16a	17	2 (17)	-
K17	17	2 (17)	2 (17)
K18	12	62	15
K19	17	3 (6,15,12)	
K20	17	-	-
K23	17	-	-
*	17	-	2
Hair Type I			
KRTHA1	17	-	-
KRTHA2	17	-	-
KRTHA3a	17	-	-
KRTHA3b	17	-	-
KRTHA4	17	-	-
KRTHA5	17	-	-
KRTHA6	17	-	-
KRTHA7	17	-	-
KRTHA8	17	-	-
ψKRTHaA	17	-	-
Type II			
K1	12	-	-
K2e	12	-	-
k2p	12	-	-
K3	12	-	-
K4	12	-	-
K5	12	-	-
K6a	12	-	-
K6b	12	-	-
K6hf	12	1 (12)	-
K7	12	-	-
K8	12	35	26
*	12	-	1

IF Gene	Chromosome	# Pseudogenes	# Gene Fragments
Hair Type II			
Hb1	12	-	-
Hb2	12	-	-
Hb3	12	-	-
Hb4	12	-	-
Hb5	12	-	-
Hb6	12	-	-
ψhHbA	12	-	-
ψhHbB	12	-	-
ψhHbC	12	-	-
ψhHbD	12	-	-
Type III			
vimentin	10	-	1 (6)
desmin	2	-	-
GFAP	17	-	-
peripherin	12	-	-
Type IV			
NF-L	8	-	2 (Y)
NF-M	8	-	1 (10)
NF-H	22	2 (20, 1)	-
α-Internexin	10	-	-
Syncoilin	1	-	-
nestin	1	-	-
synemin	15	-	-
Type V			
lamin A/C	1	-	-
laminB1	5	-	-
laminB2	19	-	-
Others			
Filensin	20	-	-
Phakinin	3	-	-
Novel Type II keratins			
K1b	12	-	-
K5b	12	-	-
K5c	12	-	-
K6h	12	-	-
K6i	12	-	-
K6k	12	-	-
K6l	12	-	-

Fig. 1. Classification and chromosomal localization of intermediate filament genes and pseudogenes. The table lists intermediate filament genes, pseudogenes and gene fragments identified in the draft of the human genome. Keratin genes 8 and 18, which gave rise to 62 and 35 processed pseudogenes, respectively, are marked with a red bar. Potential novel keratin genes/gene fragments in the type I and II clusters are indicated by an asterisk. Chromosomal localization of pseudogenes is indicated by numbers in brackets. Pseudogenes related to hair keratin genes are denoted by ψ; ‡ indicates type I keratin genes recently identified (Bawden et al., 2001). These are most closely related to K10. We propose to name them according to the Moll nomenclature as indicated in the text (Moll et al., 1982). The expression pattern of the newly identified keratin genes remains to be determined.

Human Genome Sequencing Consortium, 2001). Transcription analysis has demonstrated that the diversity of keratins is not increased further by alternative splicing.

Knowledge of IF genes and expression patterns stimulated the discovery of point mutations in a still growing number of IF genes, which has provided evidence for their pathogenic relevance in human disorders (Bonifas et al., 1991; Coulombe et al., 1991; Lane et al., 1992; reviewed by Irvine and McLean, 1999). Such 'experiments of nature' have demonstrated that mutations in at least 14 epidermal keratin genes cause fragility syndromes of epidermis and its appendages that seem to result from a collapse of a mutant keratin cytoskeleton. Formally, this was the genetic proof for a true cytoskeletal function of these proteins. Desmin mutations analogous to those in epidermal keratins were connected to myopathies of skeletal and heart muscle (Goldfarb et al., 1998), whereas point mutations in GFAP are now known to cause Alexander's disease (Brenner et al., 2001). At least two reports have linked NF-L mutations to Charcot-Marie-Tooth disease type 2E (Mersiyanova et al., 2000; De Jonghe et al., 2001). Finally, mutations in the genes coding for the nuclear lamins A/C give rise to several tissue-restricted disorders termed laminopathies (for a recent discussion, see Hutchison et al., 2001; Wilson et al., 2001). These data support the view that IF proteins also serve non-

cytoskeletal functions (Quinlan et al., 2001; Wilson et al., 2001).

Additional insight into IF protein function comes from genetically altered mice (H. Herrmann et al., unpublished). One common theme that emerges from such studies is that there are essential and nonessential IF protein functions depending on the tissue context. Ablation of keratins leads to extensive tissue fragility in the basal but not in the suprabasal epidermis (Lloyd et al., 1995; Peters et al., 2001; Reichelt et al., 2001). Moreover, knockout studies have demonstrated that certain IF proteins compensate each other (Magin et al., 2000). In addition, the phenotype of some IF gene knockout mice has shed light on new pathologies (Ku et al., 1999; Caulin et al., 2000; Hesse et al., 2000; Tamai et al., 2000).

The analysis of diseases with IF involvement as well as the understanding of IF function and evolution will be aided by the knowledge of the corresponding genes. Given that currently about 40 functional keratin genes had been identified, we were surprised by the large number of keratin genes in the recently published draft of the human genome. To clarify whether 111 keratin genes exist in the human genome (International Human Genome Sequencing Consortium, 2001), we have set out to analyze the data-set available in the public domain.

RESULTS

Number and organisation of keratin genes

We have used the NCBI and the Celera genome database for our search and included the most recently published keratins expressed in the inner root sheath (IRS) of hair follicles (Bawden et al., 2001). We found 208 keratin-related sequences in the draft (Fig. 1). Of these, 49 represent single copy genes for type I and II keratins. The type I keratin cluster contains at least 25 functional genes and 2 pseudogenes spread over nearly 1 Mb of DNA; the corresponding type II gene array harbours at least 24 functional genes and 5 pseudogenes distributed along 1.2 to 1.3 Mb.

The gene density in the two keratin clusters appears much higher than estimated for the overall genome and is approximately 35 kb per gene. There are 111 pseudogenes plus 47 gene fragments for all keratins. Intron-containing pseudogenes are mostly contained within the two keratin clusters, whereas those with features of processed pseudogenes have invaded most chromosomes, often at several positions (Fig. 2). A few earlier analyses have identified pseudogenes for keratins 8, 14, 16, 17, 18, 19 and hair keratins (Kulesh and Oshima, 1988; Rosenberg et al., 1988; Waseem et al., 1990; Troyanovsky et al., 1992; Ruud et al., 1999; Smith et al., 1999; Hut et al., 2000; Rogers et al., 2000; Winter et al., 2001). The peudogenes coding for K14, K16 and K17, which arose by gene duplication, are located outside the type I keratin cluster.

Unexpectedly, processed pseudogenes, which are cDNA derivatives, show a strikingly uneven gene relatedness. By far the highest number of processed pseudogenes relates to keratin genes 8 and 18, which map adjacently on chromosome 12q13 within the type II gene cluster. K8 and K18 are typical of internal epithelia and represent the earliest intermediate

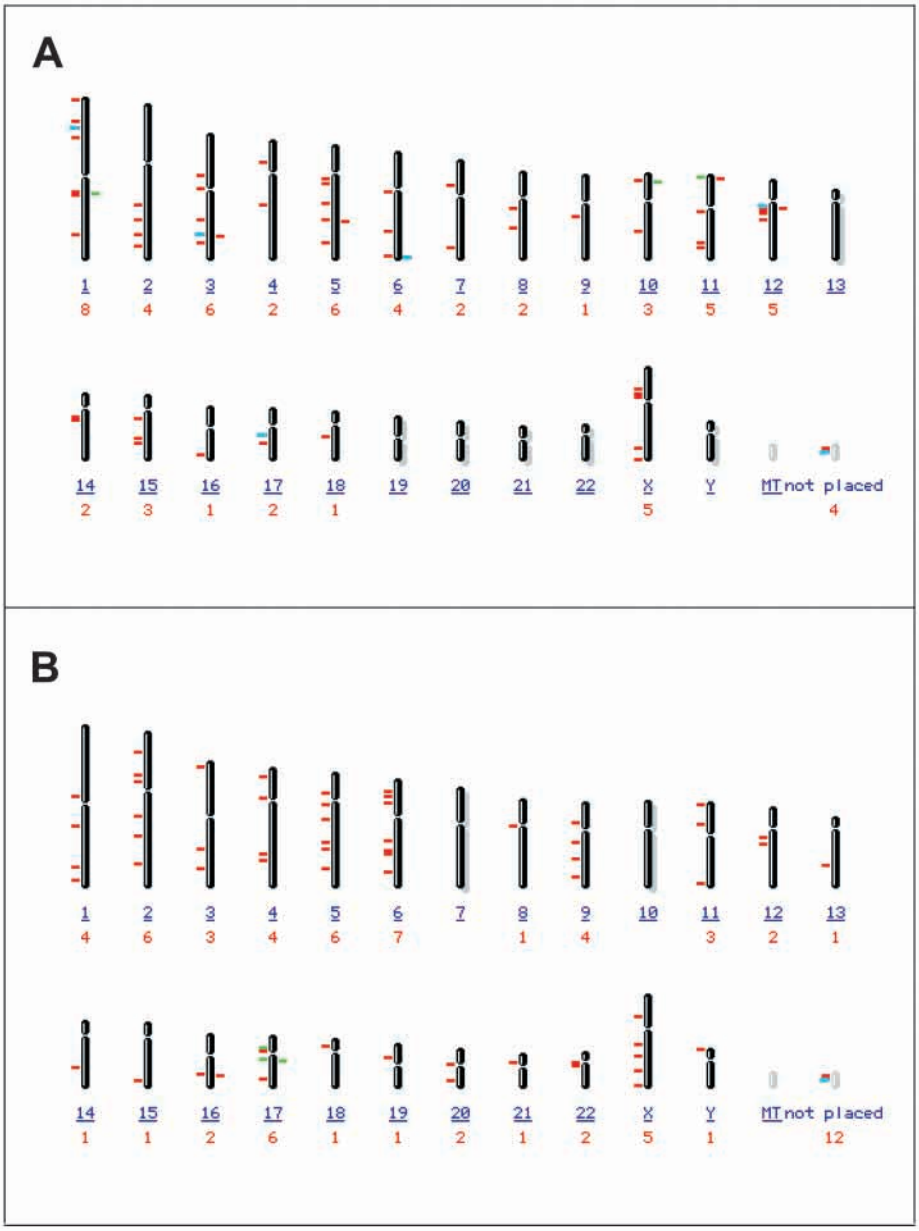

Fig. 2. Chromosomal localization of keratin 8 (A)and 18 pseudogenes (B). Chromosomes numbers are marked in blue. Integration sites per chromosome are marked in red. Coloured bars along chromosomes indicate the integration sites. The extent of sequence identity to K8 and 18 is indicated by red (alignment score >200), blue (alignment score 80-200) and green (alignment score 50-80) bars.

```
     1                                                  50
 K1  ~~SRQFSSGS GYRSGGGFSF GFAGIINYQR RTTSSSTRRS GGGGGRFSSS
K1b  ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~
 K5  ~MSRQSSVSF RSGGSRSFST ASAITPSVS. .RTSFTSVSR SGGGGGGGFG
K5b  ~~~~~~MSLS PCRAQRGFSA RSACSAR... ......SRGR SRGGFSS...
K5c  ~MSRQLNIKS .SGDKGNFSV HSAVVPRKA. ....VGSLAS YCAAGRG...
K6a  MASTSTTIRS HSSSRRGFSA SSARLPGVS. .RSGFSSISV SRSRGSG...
K6b  MASTSTTIRS HSSSRRGFSA NSARLPGVS. .RSGFSSISV SRSRGSG...
K6h  MASTSTTIRS HSSSRRGFSA NSARLPGVS. .RSGFSSISV SRSRGSG...
K6i  ~MSRQFTCKS GAAAKGGFSG CSAVLS.... ....GGSSSS FRAGSKG...
K6k  ~MSRQLTHFP .RGERLGFSG CSAVLSGGI. ....GSSSAS FRAR......
K6l  ~~MRSSVSRQ TYSTKGGFSS NSASGGSGSQ ARTSFSSVTV SRSSGSGG..

     51                                                 100
 K1  GGGGGSFGAG GGFGSRSLAG SGGSIASISG ARGGGGGSGF GGGYG.GGGF
K1b  ~~~~~~~~~~ ~~MG.RSTSG .........F CQGGGVG.GF GGGR....GF
 K5  RVSLAGACGV GGYGSRSLYN LGG..... ..SKRISIST RGGSF.RNRF
K5b  .........R GGFSSRSLNS FGR..... ...CLEG... SRGS......
K5c  ........AG AGFGSRSLYS LGG..... ..NRRISFNV AGGGVRAGGY
K6a  ..GLGGACGG AGFGSRSLYG LGG..... ..SKRISIGG GSCAI.SGGY
K6b  ..GLGGACGG AGFGSRSLYG LGG..... ..SKRISIGG GSCAI.SGGY
K6h  ..GLGGACGG AGFGSRSLYG LGG..... ..SKRISIGG GSCAI.SGGY
K6i  ........LS GGFGSRSLYS LGG..... ..VR..SLNV ASGSGKSGGY
K6k  .......... .......... .......... .......... ..........
K6l  ..GAHCGPGT GGFGSRSLYN LGG..... ..HKSISVSV AGGAL.SG..

     101                                                150
 K1  GGGGFGGGGF GGGGIGGGGF GGFGSGGGGF GGGGFGGGG. ...GYGGGYG
K1b  GVGSTGAGGF GGG..... .GFG..GAGF GTSNFGLG.. .......GFG
 K5  GAGAGGGYGF GGG.....AG SGFGFGGGAG GGFGLGGGAG FGGGFGGPGF
K5b  ......TWGS G......... ...GRLGVRF GEWSGGPG.. .........L
K5c  GFRPGSGYGG GRA..... ..SGFAGSMF GSVALGPAC. ....L.....
K6a  GSRAGAGYGF GG......AG SGFGFGGGAG IGFGLGGGP. ALLCFGGPGF
K6b  GSRAGGSYGF GG......AG SGFGFGGGAG IGFGLGGGAG LAGGFGGPGF
K6h  GSRAGGSYGF GG......AG SGFGFGGGAG IGFGLGGGAG LAGGFGGPGF
K6i  GFG......R GRA..... ..SGFAGSMF GSVALGPVC. ....P.....
K6k  .......... .......... .......... .......... ..........
K6l  .......RAL GG........ ..FGFGSRAF MGQGAGRQT. ....FG....

     151                                  Coil 1A       200
 K1  PVCSPGGIQE VTINQSLLQP LNVEIDPEIQ KVKSREREQI QSLNNQFASF
K1b  PYCPPGGIQE VTINQSLLEP LHLEVDPEIQ RIKTQEREQI MVLNNKFASF
 K5  PVCPPGGIQE VTVNQSLLTP LNLQIDPSIQ RVRTEEREQI KTLNNKFASF
K5b  SLCPPGGIQE VTINQNLLTP LKIEIDPQFQ VVRTQETQEI RTLNNQFASF
K5c  SVCPPGGIHQ VTVNKSLLAP LNVELDPEIQ KVRAQEREQI KVLNDKFASF
K6a  PVCPPGGIQE VTVNQSLLTP LNLQIDPAIQ RIGAEEREQI KTLNNKFASF
K6b  PVCPPGGIQE VTVNQSLLTP LNLQIDPAIQ RVRAEEREQI KTLNNKFASF
K6h  PVCPPGGIQE VTVNQSLLTP LNLQIDPAIQ RVRAEEREQI KTLNNKFASF
K6i  TVCPPGGIHQ VTVNESLLAP LNVELDPEIQ KVRAQEREQI KALNNKFASF
K6k  .......... VTVNKSLLAP LNVEMDPEIQ RVRAQEREQI KALNNKFASF
K6l  PACPPGGIQE VTVNQSLLTP LHVEIDPEIQ RVRTQEREQI KTLNNKFASF

     201                                                250
 K1  IDKVRFLEQQ NQVLQTKWEL LQQVD..TST RTHNLEPYFE SFINNLRRGV
K1b  IDKVQFLEQQ NQVLQTKWEL LQQVN..TST GTNNLEPLLE NYIGDLRRQV
 K5  IDKVRFLEQQ NKVLDTKWTL LQEQG..TKT VRQNLEPLFE QYINNLRRQL
K5b  IDKVRFLEQQ NKVLETKWHL LQQQG..LSG SQQGLEPVFE ACLDQLRKQL
K5c  IDKVRFLEQQ NQVLETKWEL LQQLD..LNN CKKNLEPILE GYISNLRKQL
K6a  IDKVRFLEQQ NKVLDTKWTL LQEQG..TKT VRQNLEPLFE QYINNLRRQL
K6b  IDKVRFLEQQ NKVLDTKWTL LQEQG..TKT VRQNLEPLFE QYINNLRRQL
K6h  IDKVRFLEQQ NKVLDTKWTL LQEQG..TKT VRQNLEPLFE QYINNLRRQL
K6i  IDKVRFLEQQ NQVLETKWEL LQQLD..LNN CKNNLEPILE GYISNLRKQL
K6k  IDKVRFLEQQ NQVLETKWNL LQQLD..LNN CRKNLEPIYE GYISNLQKQL
K6l  IDKVRFLEQQ NKVLETKWAL LQEQGQNLGV TRNNLEPLFE AYLGSMRSTL

     251                                      Coil 1B   300
 K1  DQLKSDQSRL DSELKNMQDM VEDYRNK... ..YEDEINKR TNAENEFVTI
K1b  DLLSAEQMRQ NAEVRSMQDV VEDYKSK... ..YEDEINKR TGSENDFVVL
 K5  DSIVGERGRL DSELRNMQDL VEDFKNK... ..YEDEINKR TTAENEFVML
K5b  EQLQGERGAL DAELKACRDQ EEEYKSK... ..YEEEAHRR ATLENDFVVL
K5c  ETLSGDRVRL DSELRSMRDL VEDYKKR... ..YEVEINRR TTAENEFVVL
K6a  DSIVGERGRL DSELRNMQDL VEDLKNK... ..YEVEINKR TAAENEFVTL
K6b  DNIVGERGRL DSELRNMQDL VEDLKNK... ..YEDEINKR TAAENEFVTL
K6h  DSIVGERGRL DSELRNMQDL VEDLKNK... ..YEDEINKR TAAENEFVTL
K6i  ETLSGDRVRL DSELRNVRDV VEDYKKR... ..YEEEINKR TAAENEFVLL
K6k  EMLSGDGVRL DSELRNMQDL VEDYKNKKKQ IWYEVEINRR TAAENEFVVL
K6l  DRLQSERGRL DSELRNVQDL VEDFKNK... ..YEDEINKH TAAENEFVVL

     301                                                350
 K1  KKDVDGAYMT KVDLQAKLDN LQQEIDFLTA LY...QAELS QMQTQISETN
K1b  KKDVDAAYVS KVDLESRVDT LTGEVNFLKY LF...LTELS QVQTHISDTN
 K5  KKDVDAAYMN KVELEAKVDA LMDEINFMKM FF...DAELS QMQTHVSDTS
K5b  KKDVDAVFLS KMELEGKLEA LREYLYFLKH LN...EEELG QLQTQASDTS
K5c  KKDADAAYAV KVELQAKVDS LDKDIKFLKC LY...DAEIA QIQTHASETS
K6a  KKDVDAAYMN KVELQAKADT LTDEINFLRA LY...DAELS QMQTHISDTS
K6b  KKDVDAAYMN KVELQAKADT LTDEINFLRA LY...DAELS QMQTHISDTS
K6h  KKDVDAAYMN KVELQAKADT LTDEINFLRA LY...DAELS QMQTHISDTS
K6i  KKDVDAAYAN KVELQAKVES MDQEIKFFRC LF...EAEIT QIQSHISDMS
K6k  KKDVDAAYMN KVELQAKVDS LTDEIKFFKC LY...EGEIT QIQSHISDTS
K6l  KKDVDAAYMG RMDLHGKVGT LTQEIDFLQQ LYEMHDAELS QVQTHVSNTN

     351                  Coil 2A                       400
 K1  VILSMDNNRQ FDLDSIIAEV KAQNEDIAQK SKAEAAESLYQ SKYEELQITA
K1b  VILFMDNNRS LDLDSIIDAV RTQYELIAQR SKDEAEALYQ TKYQELQITA
 K5  VVLSMDNNRN LDLDSIIAEV KAQYEEIANR SRTEAESWYQ TKYEELQQTA
K5b  VVLSMDNNRY LDFSSIITEV RARYEEIARS SKAEAEALYQ TKVQELQVSA
K5c  VILSMDNNRD LDLDSIIAEV RMHYEEIALK SKAEAEALYQ TKIQELQLAA
K6a  VVLSMDNNRN LDLDSIIAEV KAQYEEIAQR SRAEAESWYQ TKYEELQVTA
K6b  VVLSMDNNRN LDLDSIIAEV KAQYEEIAQR SRAEAESWYQ TKYEELQITA
K6h  VVLSMDNNRN LDLDSIIAEV KAQYEEIAQR SRAEAESWYQ TKYEELQVTA
K6i  VILSMDNNRN LDLDSIIDEV RTQYEEIALK SKAEAEALYQ TKFQELQLAA
K6k  IVLSMDNNRD LDLDSIIAEV RAQYEEIALK SKAEAETLYQ TKIQELQVTA
K6l  VVLSMDNNRN LDLDSIIAEV KAQYELIAQR SRAEAEAWYQ TKYEELQVTA


401 _                                                   450
 K1  GRHGDSVRNS KIEISELNRV IQRLRSEIDN VKKQISNLQQ SISDAEQRGE
K1b  GRHGDDLKNS KMEIAELNRT VQRLQAEISN VKKQIEQMQS LISDAEERGE
 K5  GRHGDDLRNT KHEITEMNRM IQRLRAEIDN VKKQCANLQN AIADAEQRGE
```

```
K5b  QLHGDRMQET KVQISQLHQE IQRLQSQTEN LKKQNASLQA AITDAEQRGE
K5c  SRHGDDLKHT RSEMVELNRL IQRIRCEIGN VKKQRASLET AIADAEQRGD
K6a  GRHGDDLRNT KQEIAEINRM IQRLRSEIDH VKKQCANLQA AIADAEQRGE
K6b  GRHGDDLRNT KQEIAEINRM IQRLRSEIDH VKKQCANLQA AIADAEQRGE
K6h  GRHGDDLRNT KQEIAEINRM IQRLRSEIDH VKKQCASLQA AIADAEQRGE
K6i  GRHGDDLKNT KNEISELTRL IQRIRSEIEN VKKQASNLET AIADAEQRGD
K6k  GQHGDDLKLT KAEISELNRL IQRIRSEIGN VKKQCADLET AIADAEQRGD
K6l  GKHGDNLRDT KNEIAELTRT IQRLQGEADA AKKQCQQLQT AIAEAEQRGE

     451             Coil 2B                            500
 K1  NALKDAKNKL NDLEDALQQA KEDLTRLLRD YQELMNTKLA LDLEIATYRT
K1b  QALQDAWQKL QDLEEALQQS KEELARLLRD YQAMLGVKLS LDVEIATYRQ
 K5  LALKDARNKL AELEEALQKA KQDMARLLRE YQELMNTKLA LDVEIATYRK
K5b  LALKDAQAKV DELEAALRMA KQNLARLLCE YQELTSTKLS LDVEIATYRR
K5c  NALKDAQAKL DELEGALHQA KEELARMLRE YQELMSLKLA LDMEIATYRK
K6a  MALKDAKNKL EGLEDALQKA KQDLARLLKE YQELMNVKLA LDVEIATYRK
K6b  MALKDAKNKL EGLEDALQKA KQDLARLLKE YQELMNVKLA LDVEIATYRK
K6h  MALKDAKNKL EGLEDALQKA KQDLARLLKE YQELMNVKLA LDVEIATYRK
K6i  NALKDARAKL DELEGALHQA KEELARMLRE YQELMSLKLA LDMEIATYRK
K6k  CALKDARAKL DELEGALHQA KEELARMLRE YQELVSLKLA LDMEIATYRK
K6l  LALKDAQKKL GDLDVALHQA KEDLTRLLRD YQELMNVKLA LDVEIATYRK

     501                                                550
 K1  LLEGEESRMS GECAPNVSVT VSTSHTSISG GGSRGGGGGG YGSGGSSYGS
K1b  LLEGEESRMS GELQSHVSIS VQNSQVSVNG G...AGGGGS YGSGG..YG.
 K5  LLEGEECRLS GEGVGPVNIS VVTSSVSSGY G.....SGSG YGGGLGGGLG
K5b  LLEGEECRMS GECTSQVTIS SVGGSAVMSG G...VGGGLG STCGLG....
K5c  LLEGEECRMS GENPSSVSIS VIS.SSSYSY H...HPSSAG VDLGAS....
K6a  LLEGEECRLN GEGVGQVNIS VVQSTVSSGY G.....GASG VGSGLG....
K6b  LLEGEECRLN GEGVGQVNIS VVQSTVSSGY G.....GASG VGSGLG....
K6h  LLEGEECRLN GEGVGQVNVS VVQSTISSGY G.....GASG VGSGLG....
K6i  LLESEECRSR .........S SG........ ....HPTGGC LQQSKP....
K6k  LLESEECRMS GEYPNSVSIS VIS.STN... A...GAGGAG FSMGFG....
K6l  LLESEESRMS GECPSAVSIS VTGNSTTVCG G.....GAAS FGGGIS....

     551                                                600
 K1  GGGSYGSGGG GGGGRGSYGS GGGSYGSGGS SYGSGGGGGG HGSYGSGSSS
K1b  .GG...SGGG YGGGR.SYR. GGGARGR... ...SGGG... ...YGSG...
 K5  GGLGGGLAGG SSGSY..YSS SSGGVGL... ......... .........
K5b  ......SGKG SPG..... .....S.... ......... .........
K5c  .....AVAGS SGS..... ......... ......... .........
K6a  ......LGGG SS..... ....YSY... ......... .........
K6b  ......LGGG SS..... ....YSY... ......... .........
K6h  ......LGGG SS..... ....YSY... ......... .........
K6i  .....DTAAQ RG..... ......... ......... .........
K6k  .....ASSSY SYK..... ......... ......... .........
K6l  ......LG.G SG..... ......... ......... .........

     601                                                650
 K1  GGYRGGSGGG GGGSSGGRGS GGGSSGGSSG GRGSSSGGVK SSGGSSSVKF
K1b  .......CGG GGGSYGG... .SGRS..... GRGSSR..VQ IIQTST....
 K5  .....G.... GGLSVGGSGF SASSGRGLGV GFGSGGGSSS SVKFVSTT..
K5b  ........CC TSIVTGGSNI ILGSGKDPVL DSCSVSGSSA GSSCHTILKK
K5c  .......... ..TQSGQTKT TEARGGDLKD TQGKSTPASI PARKATR~~~
K6a  .....G.... SGLGVGG.GF SSSSGRATGG GLSSVGGGSS TIKYTTTS..
K6b  .....G.... SGLGVGG.GF SSSSGRATGG GLSSVGGGSS TIKYTTTS..
K6h  .....G.... SGLGIGG.GF SSSSGRAIGG GLSSVGGGSS TIKYTTTS..
K6i  .......... ..NHQGLLGV KPAVDME~~~ ~~~~~~~~~~ ~~~~~~~~~~
K6k  .......... ..TAAADVKT KGSCGSELKD PLAKTSGSSC ATKKASR~~~
K6l  .......... .GATKGG..F STNVGYSTVK G.GPVSAGTS ILRKTTTV..

     651      662
 K1  VSTTYSGVTR ~~
K1b  .NTSHRRILE ~~
 K5  .SSSRKSFKS ~~
K5b  TVESSLKTSI TY
K5c  ~~~~~~~~~~ ~~
K6a  .SSSRKSYKH ~~
K6b  .SSSRKSYKH ~~
K6h  .SSSRKSYKH ~~
K6i  ~~~~~~~~~~ ~~
K6k  ~~~~~~~~~~ ~~
K6l  .KTSSQRY~~ ~~
```

Fig. 3. Comparison of type II keratins identified in this study. An alignment of the type II keratin sequences is given in the single letter code (residue numbers on top) with gaps introduced to maximize the amino acid alignment (dashes). Ends of the α-helical subdomains of the rod (1A, 1B, 2A and 2B) are indicated by solid arrowheads. For comparison, sequences of human keratins 1 and 5, the closest relatives, are co-aligned. For K6i, a different C-terminal sequence has been determined (M. Rogers, personal communication). Starting from position 443, it reads MSGEFPSPVS ISIISSTSGG SVYGFRPSMV SGGYVANS SNCISGVCSV RGGEGRSRGS ANDYKDTLGK GSSLSAPSKK TSR*. Asterisk indicates termination codon.

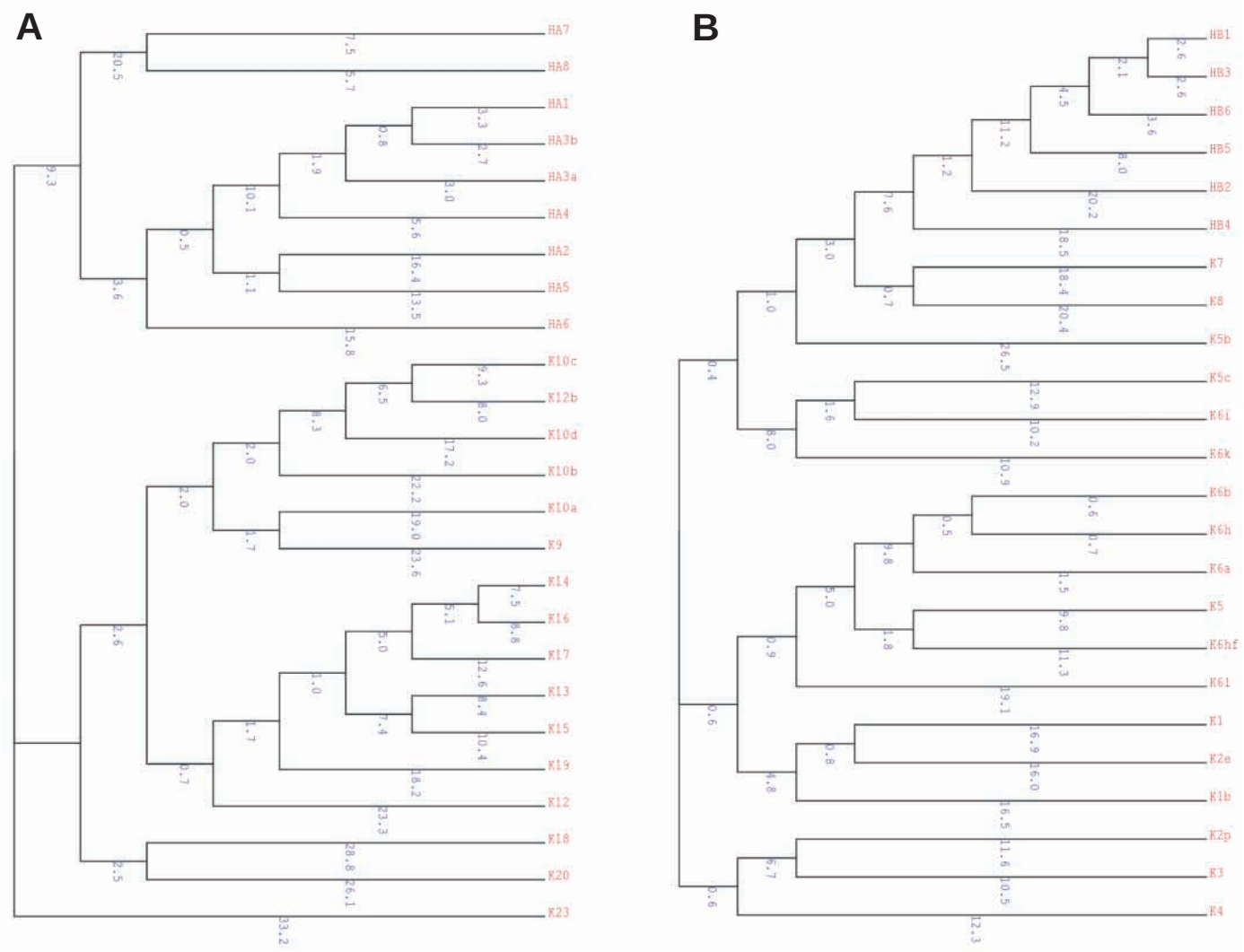

Fig. 4. Phylogenetic relationship of the human type I and II keratins. The phylogenetic tree shown was generated following the alignment of human type I (A) and type II keratins (B). Multiple sequence alignments were performed using the CLUSTAL program. Evolutionary tree construction was prepared using the CLUSTREE program. For the alignment, sequences published in the human genome draft were used (International Human Genome Sequencing Consortium, 2001).

filament expression pair in embryogenesis. There are 62 processed pseudogenes plus 15 gene fragments for the keratin 18 gene, and 35 processed pseudogenes plus 26 gene fragments for the keratin 8 gene (for a previous notion of pseudogenes, see Kulesh and Oshima, 1988; Waseem et al., 1990). These processed pseudogenes are dispersed over all chromosomes (see Fig. 2). None of these pseudogenes contained an intact open reading frame. Other keratin genes are either true single copy genes or are accompanied by one to four pseudogenes (Fig. 1).

In the present draft, no gene for keratin 11 (Moll et al., 1982), which may represent a polymorphic variant of K10 (Mischke and Wild, 1987; Korge et al., 1992) or for K6c-f (Takahashi et al., 1995) were found. The status of the latter may have to await the completion of the human genome.

Novel keratin genes and nomenclature

We discovered seven new type II keratins. Of these, five displayed homology to K6a, K6b and K5, one was most closely related to K1 and one was highly similar to K6b (Fig. 3). This new member of the K6 family has 99% protein sequence identity to K6b, but at the genomic level it contains a completely different intron 3. The evolutionary relationship of keratins is outlined in Fig. 4. Owing to the incomplete alignment of contigs, a few additional keratin genes and pseudogenes may exist.

The total number of keratin genes amounts to 49. Our survey of the current draft of the human genome conforms well with the view of 22 keratins expressed in various epithelia, 15 trichocyte-specific, 5 inner root sheath and 7 novel keratins described in this report. Together with the 13 genes for the non-keratin IF proteins, the number of genes encoding cytoplasmic IF proteins reaches 62. The three nuclear lamin genes bring the entire IF multigene family to 65.

Based on the numbering system introduced by Moll and colleagues (Moll et al., 1982), we propose to name novel type II keratins according to their sequence relationship with one of the existing eight type II genes, followed by a small letter. The

type II keratin genes reported in this study are therefore named K1b, K5b, K5c, K6h, K6i, K6k and K6l. Type I keratins should be named in the same way (see also Fig. 1). Novel genes not related to existing proteins should be given new numbers starting with K21.

Non-keratin IF genes

All 13 genes encoding the non-keratin cytoplasmic IF proteins are covered by the draft sequence (Fig. 1). Given the considerable sequence drift among these genes, the chicken sequence of synemin was non-informative for the identification of human synemin. The human orthologue was identified by D. Paulin (M. Titeux et al., unpublished). No additional functional IF gene was recognized in the current draft. Interestingly, pseudogenes are very rare among the non-keratin genes. Only the neurofilament NF-H gene is accompanied by two pseudogenes. Also, the genes for the three nuclear lamins (lamins A/C, B1 and B2) lack pseudogenes. If the completed version of the human genome lacks an additional lamin gene, the oocyte-specific lamin of certain amphibia (Döring and Stick, 1990) has no orthologue in the human genome.

CONCLUSIONS AND PERSPECTIVES

Our analysis is limited by two factors: (1) the alignment of contigs leading to the present draft is still incomplete; therefore, we cannot exclude the existence of a few more keratin genes. In light of the fidelity of the 'Moll catalog' and the concordant phenotypes of keratin-knockout mice (H. Herrmann et al., unpublished), we predict that any keratins yet to be discovered may be restricted to the hair follicle and/or other epidermal appendages. The existence of additional keratins specific for embryonic stages or specialized cells of internal epithelia appears unlikely. (2) Given the strong sequence drift among non-keratin IF genes, novel IF genes with yet unknown properties might exist. The prototype of such proteins could be represented by syncoilin, a constituent IF member of the dystrobrevin complex, which was proposed to link IF proteins to dystrobrevin at the neuromuscular junction (Newey et al., 2001). One task ahead will be to determine whether syncoilin does form copolymers with muscle-specific IF proteins or whether it serves different functions.

In view of the well-conserved structure of IF proteins and the common principles governing their assembly properties, a search for mutations in known and newly discovered IF protein genes is likely to reveal their involvement in additional disorders and to unravel new IF functions (see also Quinlan, 2001).

Most vertebrate gene families have pseudogenes, but these usually represent only a small minority of the total gene number (Mighell et al., 2000). Thus, the large number of pseudogenes for the keratin gene family is startling. Particularly striking is the finding that some 87% of these pseudogenes relate to keratin genes 8 and 18. An uneven distribution also holds for the human actin pseudogenes. There are 23 pseudogenes for β- and 6 for γ-cytoplasmic actin, while the four muscle actin genes lack pseudogenes (Pollard, 2001). The molecular mechanisms resulting in the generation of pseudogenes from some but not other genes are unknown. However, a future analysis of their integration sites may yield further information about the structural properties of human chromatin and the mechanisms of recombination.

We are grateful to D. Paulin (Paris) for providing the human synemin gene sequence, and to J. Schweizer and M. Rogers (Heidelberg) for helpful discussion and for providing sequence information on K6i. We also thank D. Siepe (Bonn) for advice on database searches. This work was supported by the DFG (SFB 284, C7) to T.M.M.

Note added in proof

While this manuscript was under review, Mizuno et al. characterized desmuslin, an IF protein that interacts with α-dystrobrevin and desmin (Mizuno et al., 2001). When we compared its sequence with that of human synemin, we found it to be nearly identical to the synemin α splice variant described by M. Titeux et al. (unpublished). Therefore, we propose to use the established name synemin.

REFERENCES

Bawden, C. S., McLaughlan, C., Nesci, A. and 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.

Becker, B., Bellin, R. M., Sernett, S. W., Huiatt, T. W. and Robson R. M. (1995). Synemin contains the rod domain of intermediate filaments. *Biochem. Biophys. Res. Commun.* **213**, 796-802.

Bonifas, J. M., Rothman, A. L. and Epstein, E. H. (1991). Epidermolysis bullosa simplex: evidence in two families for keratin gene abnormalities. *Science* **254**, 1202-1205.

Brenner, M., Johnson, A. B., Boespflug-Tanguy, O., Rodriguez, D., Goldman, J. E. and Messing, A. (2001). Mutations in GFAP, encoding glial fibrillary acidic protein, are associated with Alexander disease. *Nat. Genet.* **27**, 117-120.

Caulin, C., Ware, C. F., Magin, T. M. and Oshima. R. G. (2000). Keratin-dependent, epithelial resistance to tumor necrosis factor-induced apoptosis. *J. Cell Biol.* **149**, 17-22.

Coulombe, P. A., Hutton, M. E. Letai, A., Hebert, A., Paller, A. S. and Fuchs, E. (1991). Point mutations in human keratin 14 genes of epidermolysis bullosa simplex patients: genetic and functional analyses. *Cell* **66**, 1301-1311.

De Jonghe, P., Mersivanova, I. Nelis, E., Del Favero, J., Martin, J. J., Van Broeckhoven, C., Evgrafov, O. and Timmerman, V. (2001). Further evidence that neurofilament light chain gene mutations can cause Charcot-Marie-Tooth disease type 2E. *Ann. Neurol.* **49**, 245-249.

Doring, V. and Stick, R. (1990). Gene structure of nuclear lamin LIII of Xenopus laevis; a model for the evolution of IF proteins from a lamin-like ancestor. *EMBO J.* **9**, 4073-4081.

Fuchs, E. and Weber, K. (1994). Intermediate filaments: structure, dynamics, function, and disease. *Annu. Rev. Biochem* **63**, 345-382.

Goldfarb, L. G., Park, K. Y., Cervenakova, L., Gorokhova, S., Lee, H. S., Vasconcelos, O., Nagle, J. W., Semino-Mora, C., Sivakumar, K. and Dalakas, M. C. (1998). Missense mutations in desmin associated with familial cardiac and skeletal myopathy. *Nat. Genet.* **19**, 402-403.

Herrmann, H. and Aebi, U. (2000). Intermediate filaments and their associates: multi-talented structural elements specifying cytoarchitecture and cytodynamics. *Curr. Opin. Cell Biol.* **12**, 79-90.

Hesse, M., Franz, T., Tamai, Y., Taketo, M. M. and Magin, T. M. (2000). Targeted deletion of keratins 18 and 19 leads to trophoblast fragility and early embryonic lethality. *EMBO J.* **19**, 5060-5070.

Hut, P. H., Vlies, P., Jonkman, M. F., Verlind, E., Shimizu, H., Buys, C. H. and Scheffer, H. (2000). Exempting homologous pseudogene sequences from polymerase chain reaction amplification allows genomic keratin 14 hotspot mutation analysis. *J. Invest Dermatol.* **114**, 616-619.

Hutchison, C. J., Alvarez-Reyes, M. and Vaughan, O. A. (2001). Lamins in disease: why do ubiquitously expressed nuclear envelope proteins give rise to tissue-specific disease phenotypes? *J. Cell Sci.* **114**, 9-19.

International Human Genome Sequencing Consortium (2001). Initial sequencing and analysis of the human genome. *Nature* 409, 860-921.

Irvine, A. D. and McLean, W. H. (1999). Human keratin diseases: the increasing spectrum of disease and subtlety of the phenotype-genotype correlation. *Br. J. Dermatol.* **140**, 815-828.

Korge, B. P., Gan, S. Q., McBride, O. W., Mischke, D. and Steinert, P. M. (1992). Extensive size polymorphism of the human keratin 10 chain resides in the C-terminal V2 subdomain due to variable numbers and sizes of glycine loops. *Proc. Natl. Acad. Sci. USA* **89**, 910-914.

Ku, N. O., Zhou, X., Toivola, D. M. and Omary, M. B. (1999). The cytoskeleton of digestive epithelia in health and disease. *Am. J. Physiol.* **277**, G1108-G1137.

Kulesh, D. A. and Oshima, R. G. (1988). Cloning of the human keratin 18 gene and its expression in nonepithelial mouse cells. *Mol. Cell Biol.* **8**, 1540-1550.

Lane, E. B., Rugg, E. L., Navsaria, H., Leigh, I. M., Heagerty, A. H., Ishida, Y. A. and Eady, R. A. (1992). A mutation in the conserved helix termination peptide of keratin 5 in hereditary skin blistering. *Nature* **356**, 244-246.

Langbein, L., Rogers, M. A., Winter, H., Praetzel, S., Beckhaus, U., Rackwitz, H. R. and 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.

Lloyd, C., Yu, Q. C., Cheng, J., Turksen, K., Degenstein, Hutton, L. E. and Fuchs, E. (1995). The basal keratin network of stratified squamous epithelia: defining K15 function in the absence of K14. *J. Cell Biol.* **129**, 1329-1344.

Magin, T. M., Hesse, M. and Schroder, R. (2000). Novel insights into intermediate-filament function from studies of transgenic and knockout mice. *Protoplasma* **211**, 140-150.

Mersiyanova, I. V., Perepelov, A. V., Polyakov, A. V., Sitnikov, V. F., Dadali, E. L., Oparin, R. B., Petrin, A. N. and Evgrafov, O. V. (2000). A new variant of Charcot-Marie-Tooth disease type 2 is probably the result of a mutation in the neurofilament-light gene. *Am. J. Hum. Genet.* **67**, 37-46.

Mighell, A. J., Smith, N. R., Robinson, P. A. and Markham, A. F. (2000). Vertebrate pseudogenes. *FEBS Lett.* **468**, 109-114.

Mischke, D. and Wild, G. (1987). Polymorphic keratins in human epidermis. *J. Invest. Dermatol.* **88**, 191-197.

Mizuno, Y., Thompson, T. G., Guyon, J. R., Lidov, H. G., Brosius, M., Imamura, M., Ozawa, E., Watkins, S. C. and Kunkel, L. M. (2001). Desmuslin, an intermediate filament protein that interacts with alpha-dystrobrevin and desmin. *Proc. Natl. Acad. Sci. USA* **98**, 6156-6161.

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

Newey, S. E., Howman, E. V., Ponting, C. P., Benson, M. A., Nawrotzki, R., Loh, N. Y., Davies, K. E. and Blake D. J. (2001). Syncoilin, a novel member of the intermediate filament superfamily that interacts with alpha-dystrobrevin in skeletal muscle. *J. Biol. Chem.* **276**, 6645-6655.

Peters, B., Kirfel, J., Büssow, H., Vidal, M. and Magin, T. M. (2001) Complete cytolysis and neonatal lethality in keratin 5 knockout mice reveal its fundamental role in skin integrity and in EBS. *Mol. Biol. Cell* (in press).

Pollard, T. D. (2001). Genomics, the cytoskeleton and motility. *Nature* **409**, 842-843.

Quinlan, R. (2001). Cytoskeletal catastrophe causes brain degeneration. *Nat. Genet.* **27**, 10-11.

Reichelt, J., Büssow, H., Grund, C. and Magin, T. M. (2001). Formation of a normal epidermis supported by increased stability of keratins 5 and 14 in keratin 10 null mice. *Mol. Biol. Cell* (in press).

Rogers, M. A., Winter, H., Langbein, L., Wolf, C. and 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.

Rosenberg, M., RayChaudhury, A., Shows, T. B., Le, B. M. and Fuchs, E. (1988). A group of type I keratin genes on human chromosome 17: characterization and expression. *Mol. Cell Biol* **8**, 722-736.

Ruud, P., Fodstad, O. and Hovig, E. (1999). Identification of a novel cytokeratin 19 pseudogene that may interfere with reverse transcriptase-polymerase chain reaction assays used to detect micrometastatic tumor cells. *Int. J. Cancer* **80**, 119-125.

Smith, F. J., McKusick, V. A., Nielsen, K., Pfendner, E., Uitto, J. and McLean, W. H. (1999). Cloning of multiple keratin 16 genes facilitates prenatal diagnosis of pachyonychia congenita type 1. *Prenat. Diagn.* **19**, 941-946.

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

Takahashi, K., Paladini, R. D. and 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.

Troyanovsky, S. M., Leube, R. E. and 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., Gough, A. C., Spurr, N. K. and Lane, E. B. (1990). Localization of the gene for human simple epithelial keratin 18 to chromosome 12 using polymerase chain reaction. *Genomics* **7**, 188-194.

Winter, H., Langbein, L., Krawczak, M., Cooper, D. N., Jave-Suarez, L. F., Rogers, M. A., Praetzel, S., Heidt, P. J. and Schweizer, J. (2001). Human type I hair keratin pseudogene phihHaA has functional orthologs in the chimpanzee and gorilla: evidence for recent inactivation of the human gene after the Pan-Homo divergence. *Hum. Genet.* **108**, 37-42.

Wilson, K. L., Zastrow, M. S. and Lee, K. K. (2001). Lamins and disease: insights into nuclear infrastructure. *Cell* **104**, 647-650.