

- 7 Dobney, K. and Larson, G. (2006) Genetics and animal domestication: new windows on an elusive process. *J. Zool.* 269, 261–271
- 8 Cruz, F. *et al.* (2008) The legacy of domestication: accumulation of deleterious mutations in the dog genome. *Mol. Biol. Evol.* 25, 2331–2336
- 9 Machida, M. *et al.* (2008) Genomics of *Aspergillus oryzae*: learning from the history of Koji mold and exploration of its future. *DNA Res.* 15, 173–183
- 10 Baker, S.E. and Bennett, J.W. (2008) An overview of the genus *Aspergillus*. In *The Aspergilli: Genomics, Medical Applications, Biotechnology, and Research Methods* (Goldman, G.H. and Osmani, S.A., eds), pp. 3–13, CRC Press
- 11 Geiser, D.M. *et al.* (1998) Cryptic speciation and recombination in the aflatoxin-producing fungus *Aspergillus flavus*. *Proc. Natl. Acad. Sci. U. S. A.* 95, 388–393
- 12 Klich, M.A. (2002) Biogeography of *Aspergillus* species in soil and litter. *Mycologia* 94, 21–27
- 13 Payne, G.A. *et al.* (2006) Whole genome comparison of *Aspergillus flavus* and *A. oryzae*. *Med. Mycol.* 44 (Suppl), 9–11
- 14 Rokas, A. *et al.* (2007) What can comparative genomics tell us about species concepts in the genus *Aspergillus*? *Stud. Mycol.* 59, 11–17
- 15 Machida, M. *et al.* (2005) Genome sequencing and analysis of *Aspergillus oryzae*. *Nature* 438, 1157–1161
- 16 Abe, K. *et al.* (2006) Impact of *Aspergillus oryzae* genomics on industrial production of metabolites. *Mycopathologia* 162, 143–153
- 17 Koonin, E.V. (2005) Orthologs, paralogs, and evolutionary genomics. *Annu. Rev. Genet.* 39, 309–338
- 18 Thompson, J.D. *et al.* (1994) Clustal-W - improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* 22, 4673–4680
- 19 Bininda-Emonds, O.R. (2005) transAlign: using amino acids to facilitate the multiple alignment of protein-coding DNA sequences. *BMC Bioinformatics* 6, 156
- 20 Yang, Z. (1997) PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput. Appl. Biosci.* 13, 555–556
- 21 Yang, Z. (1998) Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. *Mol. Biol. Evol.* 15, 568–573
- 22 Rokas, A. and Galagan, J.E. (2008) The *Aspergillus nidulans* genome and a comparative analysis of genome evolution in *Aspergillus*. In *The Aspergilli: Genomics, Medical Applications, Biotechnology, and Research Methods* (Goldman, G.H. and Osmani, S.A., eds), pp. 43–55, CRC Press
- 23 Mathe, E. *et al.* (2006) Computational approaches for predicting the biological effect of p53 missense mutations: a comparison of three sequence analysis based methods. *Nucleic Acids Res.* 34, 1317–1325
- 24 Ruepp, A. *et al.* (2004) The FunCat, a functional annotation scheme for systematic classification of proteins from whole genomes. *Nucleic Acids Res.* 32, 5539–5545
- 25 Ronald, J. *et al.* (2006) Genomewide evolutionary rates in laboratory and wild yeast. *Genetics* 174, 541–544
- 26 Fay, J.C. and Benavides, J.A. (2005) Evidence for domesticated and wild populations of *Saccharomyces cerevisiae*. *PLoS Genet.* 1, 66–71
- 27 Legras, J.L. *et al.* (2007) Bread, beer and wine: *Saccharomyces cerevisiae* diversity reflects human history. *Mol. Ecol.* 16, 2091–2102
- 28 Taylor, J.W. *et al.* (2000) Phylogenetic species recognition and species concepts in fungi. *Fungal Genet. Biol.* 31, 21–32
- 29 Tsai, I.J. *et al.* (2008) Population genomics of the wild yeast *Saccharomyces paradoxus*: Quantifying the life cycle. *Proc. Natl. Acad. Sci. U. S. A.* 105, 4957–4962
- 30 Lynch, M. and Deng, H.W. (1994) Genetic slippage in response to sex. *Am. Nat.* 144, 242–261

0168-9525/\$ – see front matter © 2008 Elsevier Ltd. All rights reserved.
doi:10.1016/j.tig.2008.11.003 Available online 10 December 2008

Genome Analysis

Methylation and deamination of CpGs generate p53-binding sites on a genomic scale

Tomasz Zemojtel*, Szymon M. Kielbasa*, Peter F. Arndt, Ho-Ryun Chung and Martin Vingron

Department of Computational Molecular Biology, Max-Planck-Institute for Molecular Genetics, Ihnestrasse 73, D-14195 Berlin, Germany

The formation of transcription-factor-binding sites is an important evolutionary process. Here, we show that methylation and deamination of CpG dinucleotides generate *in vivo* p53-binding sites in numerous Alu elements and in non-repetitive DNA in a species-specific manner. In light of this, we propose that the deamination of methylated CpGs constitutes a universal mechanism for *de novo* generation of various transcription-factor-binding sites in Alus.

Methylated TEs as a source of transcription-factor-binding sites

The mobility of transposable elements (TEs) has been proposed to have an important role in spreading regulatory

elements throughout the genome [1]. In mammals, most TEs are rendered silent by DNA methylation of cytosines in the context of CpG dinucleotides. Methylated cytosines can easily be converted to thymine residues via deamination and this mutational process has the highest rate among all base substitutions [2]. Therefore, it becomes an attractive hypothesis that these silenced TEs are a source of transcription-factor-binding sites generated by means of cytosine deamination-driven mutagenesis.

In vivo p53-binding sites in Alus

Until recently, no efficient technique was available for identification of transcription-factor-binding sites residing in TEs on a genome-wide scale. However, a recently proposed approach that combines chromatin immunoprecipitation (ChIP) with paired end tag (PET) sequencing made it possible to detect p53-binding sites in the repetitive portion of the human genome [3]. This study has

Corresponding author: Zemojtel, T. (zemojtel@molgen.mpg.de).

* These authors contributed equally

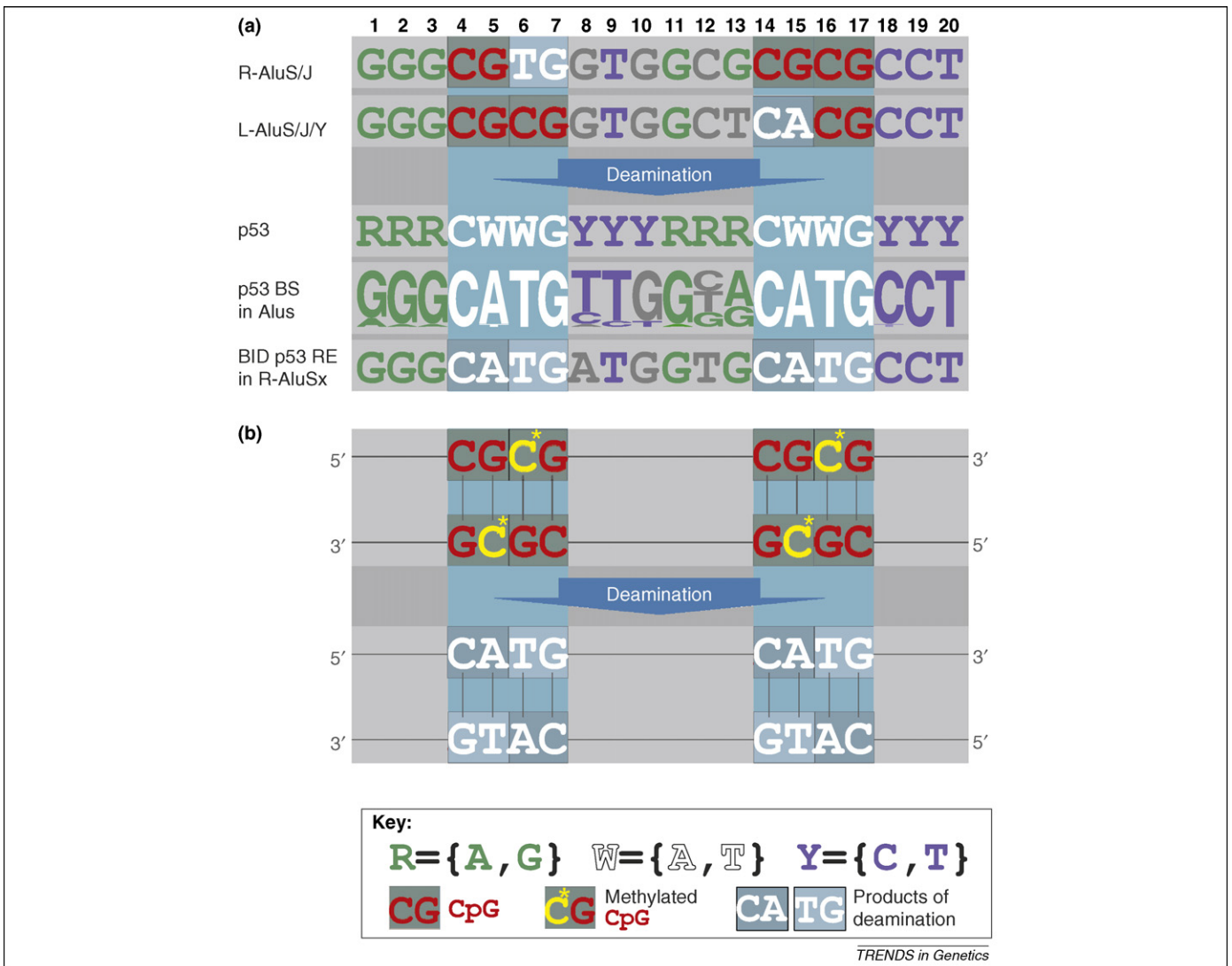


Figure 1. Methylation and deamination of pairs of CpG dinucleotides generates the preferred p53-binding sites. **(a)** Regions overlapping with the A-Box (Figure S2) in both arms of Alu consensus sequences containing CpG dinucleotides at the positions corresponding to the core of p53-binding motif are turned into functional motifs, as shown by CHIP-PET data. R-AluS/J and L-AluS/J/Y indicate the right and the left arm consensus sequences of the S, J and Y Alu subfamilies; p53 BS in Alus indicates the frequency logo of *in vivo* p53-binding sites in Alu elements, for sites see Tables S1, S2; BID p53 responsive element (RE) in R-AluSx indicates p53-responsive element in the right arm of AluSx element inserted within the *BID* gene loci. p53 indicates the consensus sequence of the p53-binding motif. **(b)** Pairs of CpG dinucleotides comprise perfect templates for deamination-driven formation of the preferred core motif in the p53-binding sequence. A proper pattern of cytosine deamination in both DNA strands generates the preferred p53 core motif.

revealed a potential link between tumor-suppressor protein p53 and transposable elements in the context of a human colorectal cancer cell line, HCT116. TEs belonging to the primate-specific long terminal repeat (LTR) class I endogenous retrovirus (ERV LTR) family have been shown to contain *in vivo* p53-binding sites. It has been suggested that the p53 site was present in the founder of the LTR subfamilies rather than it having arisen by mutations in the individual sequences [4]. Here, our analysis of the CHIP-PET data [3] revealed that as many as 106 of the 161 *in vivo* p53 sites reside within different classes of transposable elements, with the short interspersed nuclear element (SINE)-derived sites accounting for ~15% (Figure S1 and Table S1 in the supplementary material online). All p53-binding sites in the 19 primate-specific Alu elements (SINE) and three Alu-precursor free left arm monomer (FLAM) and free

right arm monomer (FRAM) sequences were composed of two directly repeated half-sites matching the p53-binding site consensus sequence RRRCWWGYYY [5] and contained at most three mismatches, located in the flanking RRR and YYY sequences, but not in the core CWWG sequence (Figure 1a and Table S2). The core sequences were invariantly found to be CATG, a sequence that has been shown to be bound preferably by p53 both *in vivo* and *in vitro* [6,7] (Figure 1a). We found that all these sites were located at the same position and overlapped with the A-Box of a bipartite RNA polymerase (Pol) III promoter in one of the two arms of Alu elements (Figure S2), with a very strong preference for the right arm (18 of 19 sites were located in the right arm; see Table S2). The reason for this is unclear at present, although it has been documented that the right arm possesses unique structural features [8], and one can speculate that this might result in higher

incidence of p53 binding to the right arm compared with the left arm.

Deamination of CpGs creates the preferred *in vivo* p53-binding sites in Alus

Are p53-binding sites present in the founders of the different Alu subfamilies, as proposed for the ERV LTR families or did they arise by mutation? To answer this question, we compared the sequences of the *in vivo* p53-binding sites in Alu elements to the reconstructed consensus sequences of different Alu subfamilies. Crucially, the comparison revealed that the consensus sequences contained CpGs at the locations corresponding to CpA and TpG in the core, and that p53-binding sites arose via mutation and were not present in the founding members of the subfamilies (Figure 1a, Table S2). Likewise, we detected four orthologous Alu loci in the rhesus and the chimp genomes that contained CpGs at the positions corresponding to the cores in the *in vivo* human p53 sites (Table S2), indicating that these sites arose in a species-specific manner, although we cannot rule out binding of p53 to these sites. Collectively, this implies that the CATG core sequence was generated by deamination of methylated cytosines in the context of CpGs (Figure 1b). In line with this interpretation, it has been established that methylation and deamination of CpGs, resulting in CpG exchange to either TpG or CpA, is one of the most prevalent mutational forces in Alu elements [9–11]. This process is time-dependent [12] and we found that the *in vivo* p53 site harboring AluJo sequences belonging to the oldest Alu subfamily, AluJ, contain at most one CpG per sequence.

Even though we detected *in vivo* p53 sites in sequences belonging to the oldest Alu subfamily, AluJ, and intermediate AluS subfamily, no single *in vivo* p53 site originated from the youngest AluY subfamily. An analysis of the right arm consensus sequence of the AluY subfamily revealed that the GG dinucleotide was present at the position corresponding to the CG dinucleotide in the core of the second half-site (positions corresponding to nucleotides [nts] 14–15 in the p53 dimer motif). This clearly reduced the capacity of the youngest Alu subfamily to be a template for p53-binding sites (see later).

Our findings indicate that Alu sequences can serve as templates for the generation of p53-binding-sites on a genome-wide scale. To confirm this idea, we compared the number of putative preferred p53-binding sites with the core motif CATG in both half-sites residing in Alu sequences with the total number in the human genome (Supplementary Methods). We found ~87 000 sites in the human genome matching our criteria (Table S3). Of these, ~80 000 resided within the repetitive portion of the genome, among which ~68 000 were harbored by Alu sequences, overlapped with the A-Box in both arms of different Alu subfamilies and were created via deamination of cytosines (Table S4). The majority of the detected sites (94%) originated from the sequences of the AluJ and AluS subfamilies. Only very few sites were present in the sequences of the AluY subfamily and a clear (> fivefold) enrichment of these was observed in the left arm compared with the right arm. This could be explained by the presence of the GG dinucleotide instead of the CG

dinucleotide in the core region, further indicating that the process of CpG deamination has a dominant role in generation of p53 sites in Alus (Table S4). Up to ~34% (11 417) of all (33 452) transcriptional units contained at least one Alu element with the putative preferred p53 site (Supplementary Table S5). Importantly, we found that one of these sites was located within the AluSx element residing in the first intron of the p53-responsive *BID* gene (Table S2), a member of the pro-apoptotic B-cell lymphoma 2 (Bcl-2) family. This Alu-derived site bound to p53 *in vivo* and mediated p53-dependent transactivation of a reporter gene, but its origin was not recognized [13]. Moreover, two preferred p53-binding sites detected by ChIP-PET residing in Alu elements inserted in the *NCK2* and *NAV3* genes, with potential roles in growth, differentiation and apoptosis, were found to be direct p53 targets by expression analyses [3]. Our analysis of a set of 209 genes identified as true p53 targets in colorectal cancer HCT116 cells [14] revealed that 112 of them are among the transcriptional units containing at least one Alu element with the putative preferred p53 site (Table S6; Fisher's exact test, $P < 10^{-8}$).

Thus, we conclude that methylation and deamination of cytosines generates a high number of preferred p53-binding sites in Alu elements, some of which were recruited to regulate target gene expression. Given that the region containing the A-Box of the RNA Pol III promoter is present in SINE elements of various species, these transposons might contribute to spreading p53-responsive elements. Indeed, we found numerous (~8000) putative preferred p53 sites in the mouse Alu-like B1 elements (Table S7).

Furthermore, we found evidence that formation of the p53 core motif was facilitated via methylation and deamination of CpGs in LTR/ERVs (Figure S3), indicating that our findings can be generalized to other TEs.

Deamination of CpGs creates the preferred *in vivo* p53-binding sites in non-repetitive DNA

Remarkably, >50% of ChIP-PET *in vivo* p53 sites (Table S1) and ~4600 of the annotated preferred p53-binding sites (Table S3) contained the core motif CATG in both half-sites and resided within the non-repetitive portion of the human genome. Among these, we identified one of the best studied p53-responsive elements located within the human *GADD45* [7] locus. Multiple sequence alignments across mammalian species indicated that the core of the p53-responsive element of human *GADD45* arose via methylation and deamination of CpGs (Figure S4a). Interestingly, an orthologous location in the mouse genome contains CpGs at the position corresponding to the core motif (Figure S4a) and was found to be inactive for p53-mediated transactivation [15]. Similarly, the earlier exemplified four orthologous Alus in the primates that contained CpGs at the positions corresponding to the cores in the human Alu-derived *in vivo* p53 sites are likely not to contain functional p53 sites. This is in line with the recent observation that many p53-responsive elements arose in a species-specific manner [16]. Similar data were collected for few other p53-binding sites in the non-repetitive portion (Figure S4).

Up to thousands of p53-binding sites can be formed via single deamination events

Finally, we used a statistical model (Supplementary Methods) to identify and characterize 20-mers in the human genome that lie on the fastest evolutionary trajectories of p53-site formation. Up to ~151 000 20-mers resided in the highest probability range and required only one cytosine deamination to become a p53 site. As expected, most of these (~119 000) were located in Alu sequences and ~10 000 resided in the non-repetitive portion of the genome (Figure S5). It has only recently been shown that noncanonical sites built of a single half-site or a 3/4-site can also function as p53 responsive elements [17]. Readily, such sites can evolve much faster than canonical sites in Alu sequences (i.e. a high number of noncanonical sites resides in the group of 119 000 Alus), however, the binding of p53 to noncanonical sites in Alus requires experimental conformation.

Conclusions

Our findings strongly indicate that the formation of p53-binding sites by CpG deamination, in particular in Alu repeats but also in non-repetitive DNA, is an important evolutionary process. Alu repeats, which amplified to over one million copies, harbor one-third of the total number of CpGs in the human genome [18], resulting from which, most Alus are transcriptionally silenced by methylation [19]. Because Alu elements are associated with gene-rich regions [20,21], the process of cytosine deamination is capable of transforming numerous silent Alus into functional regulatory elements. As we pointed out here, this process has assigned a role for Alus in spreading of p53-binding sites and in recruiting new target genes to the p53 regulatory network in a species-specific manner. In other work, it has been suggested that retinoic acid receptor (RAR)-binding motifs, DR2 s, which are present in many Alu sequences, likewise arose via deamination of a single methylated CpG dinucleotide [22]. Similarly, PAX6 homeodomain transcription-factor-binding motifs arose via deamination of methylated CpGs in Alus [23]. Taken together with our results, this implies that deamination of CpGs constitutes a universal mechanism for generation of different transcription-factor-binding sites in Alus.

Acknowledgements

T.Z. received funding from the European Commission within its FP6 Programme 'Biosapiens' (contract number LHSG-CT-2003-503265).

Supplementary data

Supplementary data associated with this article can be found at doi:10.1016/j.tig.2008.11.005.

References

- 1 Feschotte, C. (2008) Transposable elements and the evolution of regulatory networks. *Nat. Rev. Genet.* 9, 397–405
- 2 Arndt, P.F. *et al.* (2003) Distinct changes of genomic biases in nucleotide substitution at the time of Mammalian radiation. *Mol. Biol. Evol.* 20, 1887–1896
- 3 Wei, C.L. *et al.* (2006) A global map of p53 transcription-factor binding sites in the human genome. *Cell* 124, 207–219
- 4 Wang, T. *et al.* (2007) Species-specific endogenous retroviruses shape the transcriptional network of the human tumor suppressor protein p53. *Proc. Natl. Acad. Sci. U. S. A.* 104, 18613–18618
- 5 el-Deiry, W.S. *et al.* (1992) Definition of a consensus binding site for p53. *Nat. Genet.* 1, 45–49
- 6 Osada, M. *et al.* (2005) Differential recognition of response elements determines target gene specificity for p53 and p63. *Mol. Cell. Biol.* 25, 6077–6089
- 7 Inga, A. *et al.* (2002) Differential transactivation by the p53 transcription factor is highly dependent on p53 level and promoter target sequence. *Mol. Cell. Biol.* 22, 8612–8625
- 8 Mariner, P.D. *et al.* (2008) Human Alu RNA is a modular transacting repressor of mRNA transcription during heat shock. *Mol. Cell* 29, 499–509
- 9 Britten, R.J. *et al.* (1988) Sources and evolution of human Alu repeated sequences. *Proc. Natl. Acad. Sci. U. S. A.* 85, 4770–4774
- 10 Jurka, J. and Smith, T. (1988) A fundamental division in the Alu family of repeated sequences. *Proc. Natl. Acad. Sci. U. S. A.* 85, 4775–4778
- 11 Quentin, Y. (1988) The Alu family developed through successive waves of fixation closely connected with primate lineage history. *J. Mol. Evol.* 27, 194–202
- 12 Xing, J. *et al.* (2004) Alu element mutation spectra: molecular clocks and the effect of DNA methylation. *J. Mol. Biol.* 344, 675–682
- 13 Sax, J.K. *et al.* (2002) BID regulation by p53 contributes to chemosensitivity. *Nat. Cell Biol.* 4, 842–849
- 14 Kho, P.S. *et al.* (2004) p53-regulated transcriptional program associated with genotoxic stress-induced apoptosis. *J. Biol. Chem.* 279, 21183–21192
- 15 Jegga, A.G. *et al.* (2008) Functional evolution of the p53 regulatory network through its target response elements. *Proc. Natl. Acad. Sci. U. S. A.* 105, 944–949
- 16 Horvath, M.M. *et al.* (2007) Divergent evolution of human p53 binding sites: cell cycle versus apoptosis. *PLoS Genet.* 3, e127
- 17 Jordan, J.J. *et al.* (2008) Noncanonical DNA motifs as transactivation targets by wild type and mutant p53. *PLoS Genet.* 4, e1000104
- 18 Batzer, M.A. and Deininger, P.L. (2002) Alu repeats and human genomic diversity. *Nat. Rev. Genet.* 3, 370–379
- 19 Schmid, C.W. (1991) Human Alu subfamilies and their methylation revealed by blot hybridization. *Nucleic Acids Res.* 19, 5613–5617
- 20 Bailey, J.A. *et al.* (2003) An Alu transposition model for the origin and expansion of human segmental duplications. *Am. J. Hum. Genet.* 73, 823–834
- 21 Korenberg, J.R. and Rykowski, M.C. (1988) Human genome organization: Alu, lines, and the molecular structure of metaphase chromosome bands. *Cell* 53, 391–400
- 22 Laperriere, D. *et al.* (2007) Widespread Alu repeat-driven expansion of consensus DR2 retinoic acid response elements during primate evolution. *BMC Genomics* 8, 23
- 23 Zhou, Y.H. *et al.* (2002) Novel PAX6 binding sites in the human genome and the role of repetitive elements in the evolution of gene regulation. *Genome Res.* 12, 1716–1722

0168-9525/\$ – see front matter © 2008 Elsevier Ltd. All rights reserved.
doi:10.1016/j.tig.2008.11.005 Available online 26 December 2008