factors and the gene regulatory network in *Escherichia coli*. *Nucleic Acids Res.* 31, 1234–1244
15 Richter, G. *et al.* (1997) Biosynthesis of riboflavin: characterization of the bifunctional deaminase-reductase of *Escherichia coli* and *Bacillus subtilis*. *J. Bacteriol.* 179, 2022–2028
16 Moertl, S. *et al.* (1996) Biosynthesis of riboflavin. Lumazine synthase of *Escherichia coli*. *J. Biol. Chem.* 271, 33201–33207
17 Hecht, S. *et al.* (2001) Studies on the nonmevalonate pathway to terpenes: the role of the GcpE (IspG) protein. *Proc. Natl. Acad. Sci. U. S. A.* 98, 14837–14842
18 Gopalakrishnan, A.S. *et al.* (1986) Structure and expression of the gene locus encoding the phosphatidylglycerophosphate synthase of *Escherichia coli*. *J. Biol. Chem.* 261, 1329–1338
19 Schultz, J. *et al.* (1998) SMART, a simple modular architecture research tool: identification of signaling domains. *Proc. Natl. Acad. Sci. U. S. A.* 95, 5857–5864
20 Letunic, I. *et al.* (2002) Recent improvements to the SMART domain-based sequence annotation resource. *Nucleic Acids Res.* 30, 242–244
21 Bateman, A. *et al.* (2002) The Pfam protein families database. *Nucleic Acids Res.* 30, 276–280
22 Wolf, Y.I. *et al.* (2001) Genome alignment, evolution of prokaryotic genome organization, and prediction of gene function using genomic context. *Genome Res.* 11, 356–372

0168-9525/$ - see front matter
doi:10.1016/j.tig.2004.01.006

Mutational patterns correlate with genome organization in SARS and other coronaviruses

Andrei Grigoriev

GPC Biotech, Fraunhoferstr. 20, Martinsried 82152, Germany

Focused efforts by several international laboratories have resulted in the sequencing of the genome of the causative agent of severe acute respiratory syndrome (SARS), novel coronavirus SARS-CoV, in record time. Using cumulative skew diagrams, I found that mutational patterns in the SARS-CoV genome were strikingly different from other coronaviruses in terms of mutation rates, although they were in general agreement with the model of the coronavirus lifecycle. These findings might be relevant for the development of sequence-based diagnostics and the design of agents to treat SARS.

Previously, cumulative skew diagrams have been employed successfully to analyze mutational patterns in various viral genomes. They have been used to: (i) link the nucleotide content changes to the genome organization, replication and transcription of double-stranded DNA viruses [1]; (ii) correlate the transcriptional pattern of a bacteriophage T7 with its nucleotide content [2]; and (iii) associate the compositional biases with mutational pressures in retroviruses [3]. (See Box 1 on how to interpret cumulative diagrams.)

The severe acute respiratory syndrome coronavirus (SARS-CoV) plus-strand genomic RNA (plus-gRNA) consists of two distinct parts: one (comprising two thirds of the genome) encodes the replicase polyprotein and the other encodes structural and other proteins [4,5]. In this paper, these parts are referred to as the long and short arm, respectively. Strikingly, there is a change in behavior of the cumulative skew diagram at the border of the arms in all coronaviruses sequenced to date (six representatives are shown in Figure 1), indicating a lower GC skew on the short arm. This behavior suggests that biological processes that distinguish the two arms (Box 2) are responsible for the mutational pattern, rather than the fidelity of the replication machinery; the latter not would result in a constant slope of cumulative skew, as is the case in retroviruses [3]. The mutation rates (as indicated by the extent of the cumulative skew on the *y*-axis) do not appear to depend on a host organism: skews are similar in murine, avian and human 229E coronaviruses (Figure 1c,e,f) but substantially lower in SARS-CoV (Figure 1a, Table 1).

The skew diagrams support the current model of coronavirus replication and transcription (Box 2), and GC skew is particularly illustrative in this regard because in both of these processes one RNA strand is single stranded. Deamination of cytosine to uracil is >100 times faster in single-stranded DNA compared with double-stranded DNA [6], and this ratio is probably similar in

Table 1. Mean excess of guanines versus cytosines in coronavirus genomes

Virus genome[a]	Extra guanines compared with cytosines per 100 bp of genomic sequence[b]		
	L[c]	S[c]	L-S[c]
SARS-CoV	1.8	−1.7	3.5
BCoV	7.8	3.5	4.3
MHV	7.1	3.5	3.6
PEDV	4.4	1.4	3.0
HCoV	6.0	1.8	4.2
IBV	5.9	4.2	1.7

[a]Abbreviations: BCoV, enteric bovine coronavirus; IBV, avian infectious bronchitis virus; HCoV, human coronavirus (229E); PEDV, porcine epidemic diarrhea virus; SARS-CoV, severe acute respiratory syndrome coronavirus.
[b]These averages represent the trends depicted in Figure 1 but without taking into account G + C content (which ranges from 37% to 42% in *Coronaviridae*). GC content does not affect the trends observed in Figure 1.
[c]The change in number of guanines compared with cytosines is probably due to cytosine deamination in the minus strand on the short arm and reflects additional mutational pressure on that arm. Notably, this change is comparable with SARS-CoV and other coronaviruses, whereas the guanine excess on the long arm is much smaller. Definitions: L, long arm; S, short arm; L-S, change on short arm.

Corresponding author: Andrei Grigoriev (andrei.grigoriev@gpc-biotech.com).

Box 1. Interpreting cumulative skew diagrams

Cumulative skew diagrams [1,7,8] can simplify the interpretation of biases in nucleotide sequence. An example of such bias is GC skew, which is a measure of the relative excess of guanines against cytosines on one sequence strand. It is calculated as $([G] - [C])/([G] + [C])$, where [G] and [C] represent the occurrence of the guanines and cytosines within a specified sequence window.

Such biases have been reported for bacteria [7,20,21] and double-stranded (ds) DNA viruses [1,21], and interpreted as evidence of asymmetry in mutation pressure because skew changes the polarity at the replication origin. The GC skew has been linked to the time the DNA strand spends in a single-stranded state [7], for example, during replication or transcription because cytosine deamination is much faster in single-stranded (ss) DNA compared with dsDNA (see [22,23] for in-depth reviews of the underlying mechanisms).

Cumulative skew represents a numerical integration of the skew value across the genome and replaces the most significant changes in polarity by global maxima and minima. For example, a non-cumulative plot of GC skew is shown in Figure Ia for the genome of the virus SV40, where GC skew changes sign at a point near the 50% coordinate. It is unclear which of the multiple local polarity switches in the middle of the plot is actually the global switch. On the cumulative GC skew plot Figure Ib these polarity switches are seen to correspond to local minima and maxima on the GC diagram. The global maximum at 54% clearly separates two genome segments with the opposite deviations from the parity $[G] = [C]$, and the slopes of the opposite linear trends on the GC diagram correspond to the respective mean GC skews for the two genome segments. GC skew is positive for the leading (left-hand side of the GC diagram) and negative for the lagging strand, as is the case with microbial genomes.

The two segments of the GC diagram also correspond to the divergently transcribed coding sequences of SV40. Note that the slopes of the two halves of the GC diagram are different. The excess of G compared with C in the leading strand in the late mRNA region of SV40 is almost half of the excess of C compared with G in the lagging strand in the early mRNA region. This suggests a contribution of transcription to the overall picture.

Even more illustrative interplay of replication and transcription is a seen in a cumulative diagram of human papillomavirus [1] Figure Ic. Although the replication is bi-directional (from 0 or 100% on the diagram), transcription is unidirectional: all papillomavirus genes are transcribed from one strand. If there are separate biases induced by replication and transcription, they should act in the same direction in one half of a papillomavirus genome, and in the opposite directions in the other half. This model explains the observed behavior in Figure Ic such that the steeper slopes on the left-hand side reflect a sum of the net contributions of replication and transcription, and the right-hand side of the diagrams corresponds to their subtraction, where their effects almost cancel each other out (a near-horizontal cumulative plot corresponding to zero mean GC skew).

The same rules apply to the analysis of RNA viral genomes. For example, for plus-strand RNA viruses the events taking place on the minus strand can be taken into account in much the same way as is done for the second strand of dsDNA. Because GC skew measures the level of cytosine depletion on one strand relative to its complementary strand, changes in the diagram shape enable researchers to infer the contribution of processes occurring on both strands, even in taxonomical orders of single-stranded viruses.

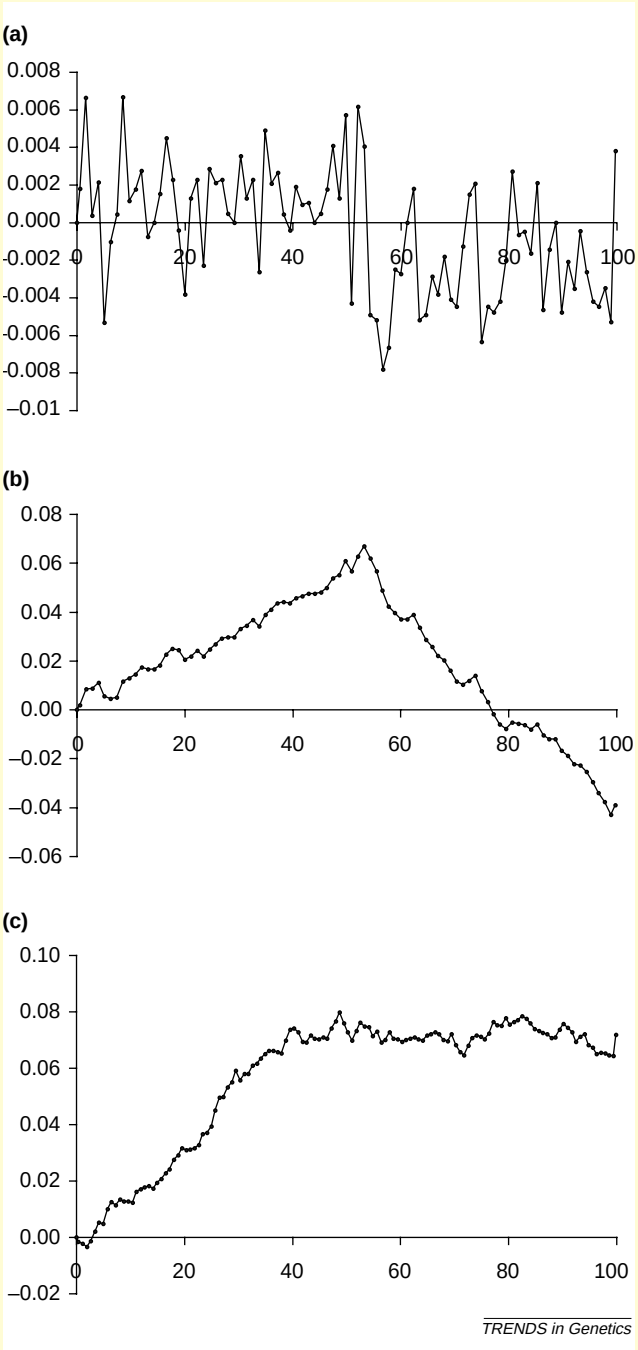

Figure I. **(a)** Non-cumulative and **(b)** cumulative GC-skew diagrams of the SV40 virus. **(c)** Cumulative GC skew of the human papillomavirus HPV-1A. For both viruses, the replication origin coordinate corresponds to 0% (or 100% because the genomes are circular). Reproduced with permission from Ref. [1].

RNA. Thus, cumulative GC skew can be interpreted as a measure of cytosine depletion on one strand relative to its complementary strand.

For most of the coronaviruses, there is almost a constant excess of G compared with C throughout the long arm (Figure 1), indicating an elevated C to U deamination in the plus strand. Similar to skews observed in DNA genomes [1,7,8], this probably results from the predominantly single-stranded nature of the plus-gRNA during replicase translation or minus-gRNA synthesis.

The skew is less pronounced on the short arm (although changes in the slope of the curve are sometimes small in Figure 1b–f, they are all significant; data not shown) and, remarkably, the cumulative diagram even reverses its trend in SARS-CoV (Figure 1a). Most probably, this reflects higher rates of cytosine deamination on the minus-strand related to subgenomic mRNA synthesis.

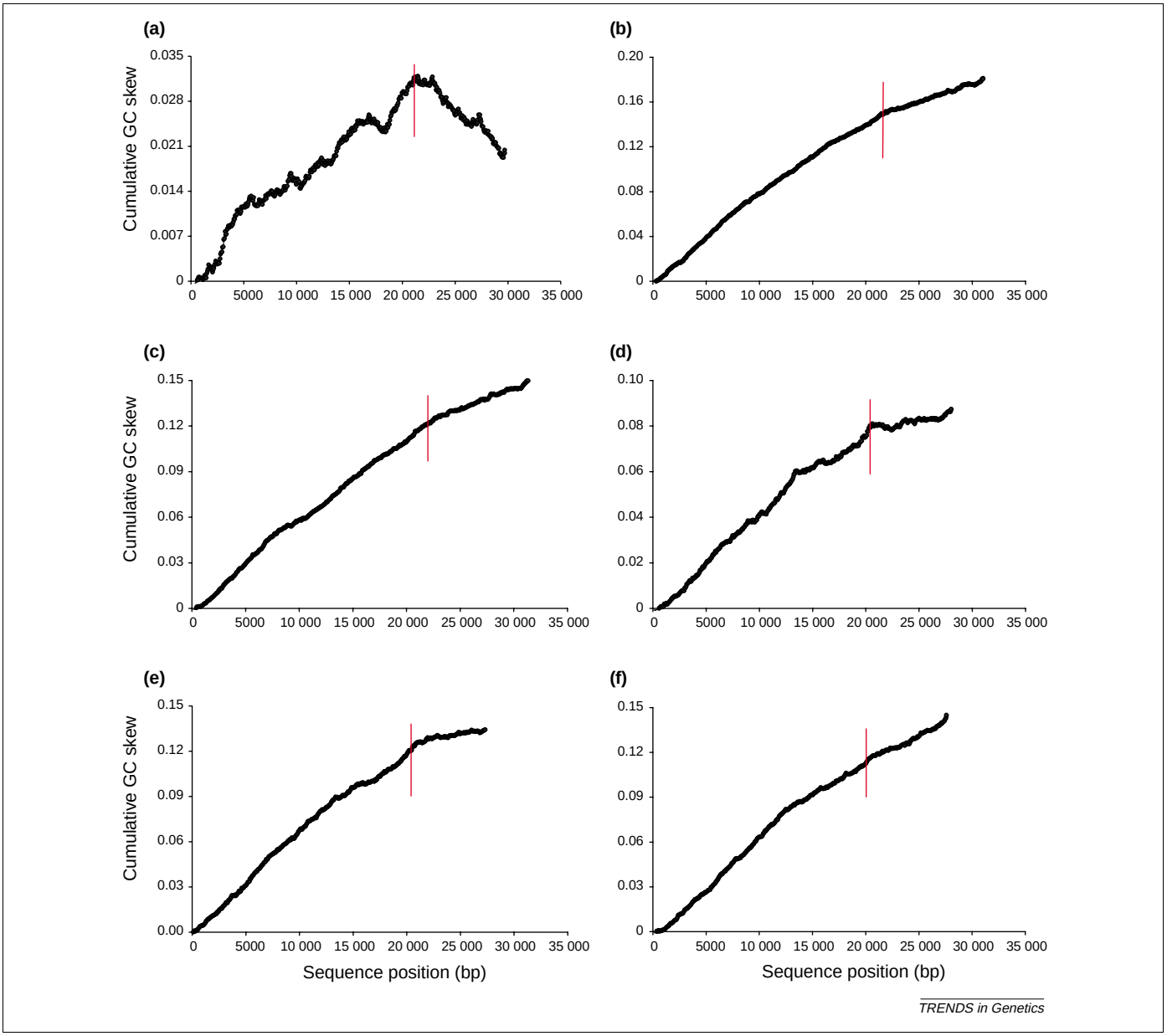

Figure 1. Cumulative GC skew diagrams of coronaviruses. RNA genomes of six representatives of the *Coronaviridae* family are shown: **(a)** severe acute respiratory syndrome coronavirus (SARS-CoV) [4,5], **(b)** enteric bovine coronavirus (BCoV) [13], **(c)** murine hepatitis virus (MHV) [14], **(d)** porcine epidemic diarrhea virus (PEDV) [17], **(e)** human coronavirus (229E) [18] and **(f)** avian infectious bronchitis virus (IBV) [19]. Diagrams with the window size of 60 bp were constructed as previously described [1,7]. Vertical bars mark the end of the replicase polyprotein gene in these genomes. Note the different slopes of the curves to the left and to the right of these vertical bars (which corresponds to the division points between the long and short arms) and the differences in vertical scales on different panels.

The intracellular duplex of minus-gRNA with plus-gRNA protects them from cytosine deamination. If the first stage of transcription, which involves subgenomic mRNA template synthesis from the plus-gRNA, leaves minus-gRNA on the short arm as a single strand (Box 2, Figure I), then cytosine deamination will lead to the accumulation of uracils on minus-gRNA. Subsequently, synthesis of the new viral plus-gRNA from minus-gRNA will propagate these mutations, depleting guanines and decreasing the overall GC skew on the short arm of the plus strand. This explanation concurs with the model of subgenomic mRNA synthesis from minus-strand subgenomic RNA templates [9,10], for which there is experimental evidence in arteviruses [11] and murine hepatitis virus (MHV) [12]. The rate of cytosine deamination that is related to sgmRNA synthesis is likely to be proportional to the difference between the slopes of the curves in the long and short arms (Table 1).

Such a combination of mutational pressures for the two RNA strands indicates a higher overall substitution rate for the short arm, compared with the long arm. The supporting evidence for this comes from the comparison of two bovine coronaviruses (respiratory and enteric) that have differences in 107 nucleotide positions [13]. More than 80% of these differences correspond to the third base of a codon, indicating mutational pressure. I analyzed the distribution of these 107 positions and found that 59 of them localized to the short arm, suggesting an ~2.5-fold increase in polymorphisms on that arm. Most of these polymorphisms (85, ~80%) correspond to a C to U

Box 2. Coronavirus replication and transcription in SARS-CoV

The genome of the severe acute respiratory syndrome coronavirus (SARS-CoV) is a plus-strand genome RNA (plus-gRNA) of ~30 Kbp in length. Translation of the replicase polyprotein on the long arm of the genome is followed by minus-gRNA synthesis and transcription from the short arm. The long and short arms of the SARS-CoV genome are shown, together with the transcriptional products [eight subgenomic mRNAs (sgmRNAs)], in Figure Ia [16].

Transcription on the short arm produces a nested set of 3′-coterminal sgmRNAs, containing at their 5′-end a short leader sequence derived from the 5′-end of the genome. A process for one of the subgenomic mRNAs is shown in Figure Ib [10–12]. After a minus-strand sgmRNA is synthesized on the short arm, a template switch enables the completion of the synthesis of the leader sequence (shown as open box on the left-hand side), skipping the long arm.

The relative levels of transcription and replication in coronaviruses mean that subgenomic mRNAs are by far the most abundant coronavirus RNAs in the cell, whereas the genome-length negative strand RNA (minus-gRNA) is the least abundant (it is ~10% of the level of plus-gRNA) [24]. These levels and localization of transcriptional activity are likely to be linked to the difference in mutation rates on the short and long arms (see main text).

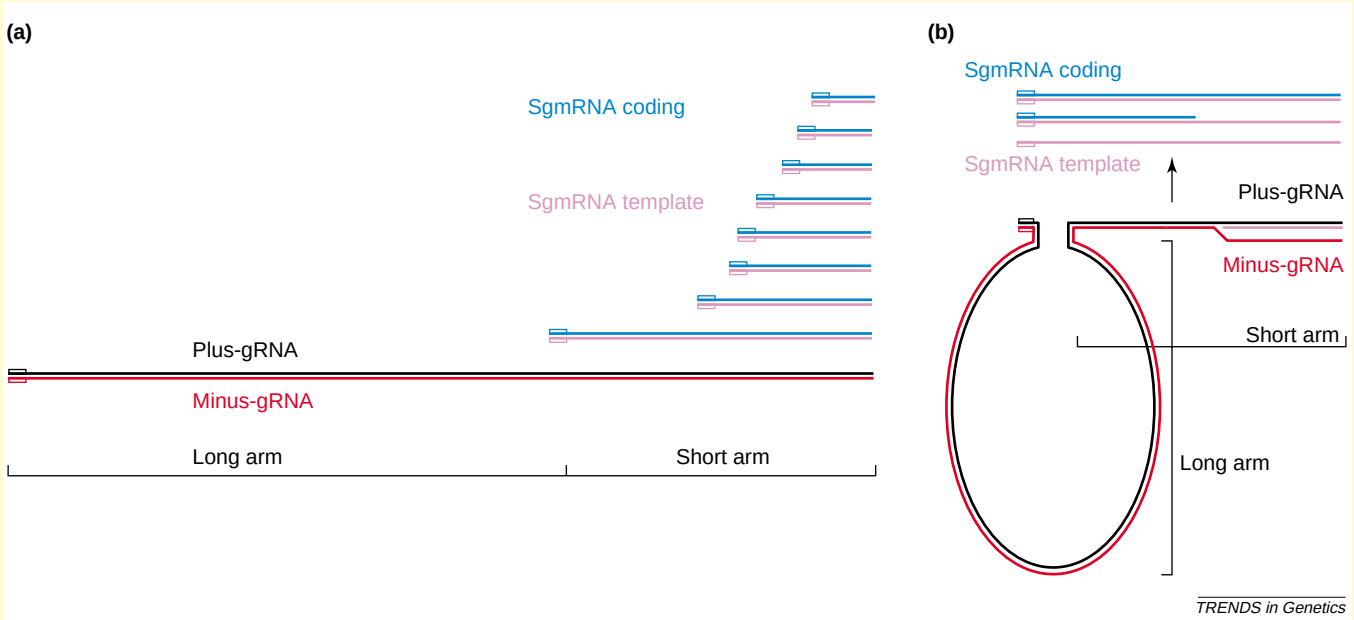

Figure I. **(a)** Genomic organization and **(b)** transcription process in the coronavirus genome. Different colors designate different types of RNA strands (i.e. coding and template strands). Plus-strand genome RNA (gRNA) and minus-gRNA are shown in black and red, respectively. Open box indicates the leader sequence (not drawn to scale).

substitution on one of the strands, further emphasizing the role of cytosine deamination as the primary mutational force in coronaviruses.

The rates of cytosine deamination in the SARS-CoV genome appear lower compared with other coronaviruses and this might explain the observation that the two sequenced strains diverged in genomic sequence by $<0.003\%$ [4,5]. Alternatively, if the epidemic came from a single clone, then only a short time span separates the two strains and that might explain the low divergence. Furthermore, the differences might be sequencing errors or PCR artifacts. However, it is worth pointing out that seven out of these eight polymorphisms also correspond to a C to U substitution on one of the strands.

Comparison of the skew diagrams places SARS apart from other groups of coronaviruses but does not provide any evidence of recent genomic recombination between members of those groups as the origin of SARS-CoV (such an event would have produced a skew diagram with fragments corresponding to the parent genomes). These observations are in agreement with the phylogenetic analyses of coronavirus-encoded proteins [4,5], which have also indicated lower conservation of the structural proteins, compared with replicase. This pattern appears to result from the mutational biases described above together with stronger selection on the replicase proteins and might influence the virulence and host-cell tropism of coronaviruses; examples of altered pathogenesis have been reported for murine coronavirus mutants [14].

Why are the mutational trends in the SARS-CoV genome so different from other coronaviruses? The cause is probably not in the host because another human coronavirus (229E) does not appear different from the other viruses examined (Figure 1e). Could the parameters of the virus-encoded RNA synthesis machinery, such as the speed of replication or transcription, or their relative turnover be responsible for this difference? The level of cytosine deamination, reflected in GC skew, has been hypothesized to depend on the time a DNA strand spends in a single-stranded state [1,7,8] (Box 1), and the same is probably true for RNA. Although the relative contribution of transcription in SARS-CoV is similar to that in other coronaviruses (Table 1, column L-S), the effect of replication is much lower (Table 1, column L). This suggests that either minus-strand synthesis is faster or plus-strand

replication is slower in SARS or their relative turnover is lower compared with the synthesis of subgenomic mRNA template RNA.

All these findings are relevant for sequence-based diagnostics and drug design against SARS-CoV and other coronaviruses because targeting the long arm with lower mutation rates should prove more robust against mutational changes in the target. This lends further support to a recent suggestion to design anti-SARS drugs based on the structure of the SARS 3C-like proteinase [15], which is encoded by genes on the long arm. These anti-SARS drugs will function as protease inhibitors that might block coronavirus replication. Another set of putative targets has been suggested in a recent publication [16] that has identified distant homologs of cellular RNA processing enzymes in the SARS genome. Notably, these are also encoded on the long arm as parts of the replicase polyprotein.

Acknowledgements

Thanks to D. Bancroft for help with improving the manuscript and interesting discussions, and to I. Ivanov and S. Meier-Ewert for critical reading of the manuscript and insightful comments.

References

1 Grigoriev, A. (1999) Strand-specific compositional asymmetries in double-stranded DNA viruses. *Virus Res.* 60, 1–19
2 Beletskii, A. *et al.* (2000) Mutations induced by bacteriophage T7 RNA polymerase and their effects on the composition of the T7 genome. *J. Mol. Biol.* 300, 1057–1065
3 Berkhout, B. *et al.* (2002) Codon and amino acid usage in retroviral genomes is consistent with virus-specific nucleotide pressure. *AIDS Res. Hum. Retroviruses* 18, 133–141
4 Rota, P.A. *et al.* (2003) Characterization of a novel coronavirus associated with severe acute respiratory syndrome. *Science* 300, 1394–1399
5 Marra, M.A. *et al.* (2003) The genome sequence of the SARS-associated coronavirus. *Science* 300, 1399–1404
6 Lindahl, T. and Nyberg, B. (1974) Heat-induced deamination of cytosine residues in deoxyribonucleic acid. *Biochemistry* 13, 3405–3410
7 Grigoriev, A. (1998) Analyzing genomes with cumulative skew diagrams. *Nucleic Acids Res.* 26, 2286–2290
8 Grigoriev, A. (1998) Genome arithmetic. *Science* 281, 1923a
9 Sawicki, S.G. and Sawicki, D.L. (1995) Coronaviruses use discontinuous extension for the synthesis of subgenome-length negative strands. *Adv. Exp. Med. Biol.* 380, 499–506
10 Sawicki, S.G. and Sawicki, D.L. (1998) A new model for coronavirus transcription. *Adv. Exp. Med. Biol.* 440, 215–219
11 Pasternak, A.O. *et al.* (2001) Sequence requirements for RNA strand transfer during nidovirus discontinuous subgenomic RNA synthesis. *EMBO J.* 20, 7220–7228
12 Sawicki, S.G. and Sawicki, D.L. (1990) Coronavirus transcription: subgenomic mouse hepatitis virus replicative intermediates function in RNA synthesis. *J. Virol.* 64, 1050–1056
13 Chouljenko, V.N. *et al.* (2001) Comparison of genomic and predicted amino acid sequences of respiratory and enteric bovine coronaviruses isolated from the same animal with fatal shipping pneumonia. *J. Gen. Virol.* 82, 2927–2933
14 Leparc-Goffart, I. *et al.* (1997) Altered pathogenesis of a mutant of the murine coronavirus MHV-A59 is associated with a Q159L amino acid substitution in the spike protein. *Virology* 239, 1–10
15 Anand, K. *et al.* (2003) Coronavirus main proteinase (3CLpro) structure: basis for design of anti-SARS drugs. *Science* 300, 1763–1787
16 Snijder, E.J. *et al.* (2003) Unique and conserved features of genome and proteome of SARS-coronavirus, an early split-off from the coronavirus group 2 lineage. *J. Mol. Biol.* 331, 991–1004
17 Kocherhans, R. *et al.* (2001) Completion of the porcine epidemic diarrhoea coronavirus (PEDV) genome sequence. *Virus Genes* 23, 137–144
18 Thiel, V. *et al.* (2001) Infectious RNA transcribed *in vitro* from a cDNA copy of the human coronavirus genome cloned in vaccinia virus. *J. Gen. Virol.* 82, 1273–1281
19 Boursnell, M.E. *et al.* (1987) Completion of the sequence of the genome of the coronavirus avian infectious bronchitis virus. *J. Gen. Virol.* 68, 57–77
20 Lobry, J.R. (1996) Asymmetric substitution patterns in the two DNA strands of bacteria. *Mol. Biol. Evol.* 13, 660–665
21 Mrazek, J. and Karlin, S. (1998) Strand compositional asymmetry in bacterial and large viral genomes. *Proc. Natl. Acad. Sci. U. S. A.* 95, 3720–3725
22 Francino, M.P. and Ochman, H. (1997) Strand asymmetries in DNA evolution. *Trends Genet.* 13, 240–245
23 Frank, A.C. and Lobry, J.R. (1999) Asymmetric substitution patterns: a review of possible underlying mutational or selective mechanisms. *Gene* 238, 65–77
24 Sethna, P.B. *et al.* (1989) Coronavirus subgenomic minus-strand RNAs and the potential for mRNA replicons. *Proc. Natl. Acad. Sci. U. S. A.* 86, 5626–5630

0168-9525/$ - see front matter
doi:10.1016/j.tig.2004.01.009