Avoidance of Protein Fold Disruption in Natural Virus Recombinants

Pierre Lefeuvre[1], Jean-Michel Lett[1], Bernard Reynaud[1], Darren P. Martin[2]*

1 CIRAD, UMR 53 PVBMT CIRAD-Université de la Réunion, Pôle de Protection des Plantes, Ligne Paradis, Saint Pierre, La Réunion, France, 2 Institute of Infectious Diseases and Molecular Medicine, University of Cape Town, Observatory, South Africa

With the development of reliable recombination detection tools and an increasing number of available genome sequences, many studies have reported evidence of recombination in a wide range of virus genera. Recombination is apparently a major mechanism in virus evolution, allowing viruses to evolve more quickly by providing immediate direct access to many more areas of a sequence space than are accessible by mutation alone. Recombination has been widely described amongst the insect-transmitted plant viruses in the genus *Begomovirus* (family Geminiviridae), with potential recombination hot- and cold-spots also having been identified. Nevertheless, because very little is understood about either the biochemical predispositions of different genomic regions to recombine or what makes some recombinants more viable than others, the sources of the evolutionary and biochemical forces shaping distinctive recombination patterns observed in nature remain obscure. Here we present a detailed analysis of unique recombination events detectable in the DNA-A and DNA-A-like genome components of bipartite and monopartite begomoviruses. We demonstrate both that recombination breakpoint hot- and cold-spots are conserved between the two groups of viruses, and that patterns of sequence exchange amongst the genomes are obviously non-random. Using a computational technique designed to predict structural perturbations in chimaeric proteins, we demonstrate that observed recombination events tend to be less disruptive than sets of simulated ones. Purifying selection acting against natural recombinants expressing improperly folded chimaeric proteins is therefore a major determinant of natural recombination patterns in begomoviruses.

Introduction

Besides its vital cellular role in maintaining and repairing broken DNA molecules [1,2], recombination is also evolutionarily significant in that it defends genomes against the otherwise unavoidable accumulation of deleterious mutations [3–5]. However, by enabling the creation of novel genetic combinations from existing genomes, recombination has the potential to do more than just reverse the mutational decay of genomes: it can also provide organisms with vastly more evolutionary options than are available through mutation alone [6,7].

In virology, two recombinational processes can be distinguished: genome reassortment and true recombination. Genome reassortment, also called pseudo-recombination, involves the exchange of intact genome components between viruses with multipartite genomes to yield viruses whose genomes are comprised of new combinations of components. True recombination, on the other hand, involves the exchange of genetic material between individual genomic molecules. The rearrangement of genetic information mediated by both true recombination and pseudo-recombination must yield fully functional and reasonably fit genomes for these processes to be easily detectable in nature. However, analysis of the functionality of recombinant genes [8,9] and the viability of recombinant genomes [10–12] has indicated that a large proportion (and possibly the vast majority) of recombination events between genomes sharing less than 90% nucleotide sequence identity yield progeny with decreased viability. Bacterial recombination [13] and DNA shuffling studies [8,9,14,15] have indicated that the evolu-

tionary value of recombination can vary depending on both the specific genes and sub-gene modules that are exchanged. A key factor determining the survival of recombinants is the degree to which recombination disrupts coevolved intra-genome interactions. At the whole genome scale, potentially disrupted interactions could include sequence-specific interactions between viral proteins, DNA, and RNA. At the scale of individual viral proteins, interactions include those occurring between amino acids required for proper folding.

While full accounts of experimentally verified intra-genome interactions are currently unavailable for any virus species, potential amino acid interactions within folded proteins can be inferred with reasonable accuracy given high resolution protein structural data. In the past five years, protein engineers have made substantial progress in the development of computational methods capable of accurately inferring degrees of recombination-induced fold disruption in experimentally generated chimaeras of proteins with known structures [8,14,15]. Although these methods

Abbreviations: CP, coat protein; hDNA, heterogeneous length linear dsDNA; IR, intergenic region; ORF, open reading frame; Rep, replication-associated protein; v-ori; virion strand origin of replication

* To whom correspondence should be addressed. E-mail: Darrin.Martin@uct.ac.za

Author Summary

The exchange of genetic material between different virus species, called inter-species recombination, has the potential to generate, within a single genome replication cycle, an almost unimaginable number of genetically distinct virus strains, including many that might cause deadly new human, animal, or plant diseases. Many fear that inter-species recombination could provide viruses with quick access to evolutionary innovations such as broader host ranges, altered tissue tropisms, or increased severities. However, mounting evidence suggests that recombination is not an unconstrained process and that most inter-species recombinants that occur in nature are probably defective. It is suspected that networks of coevolved interactions between different parts of virus genomes and their encoded proteins must be kept intact for newly formed inter-species recombinants to have any chance of out-competing their parents. One category of coevolved interaction is that between contacting amino acids within the 3-D structures of folded proteins. Here we examine the distributions of recombination events across the genomes of a group of rampantly recombining plant viruses and find very good evidence that this class of interaction tends to be preserved amongst recombinant sequences sampled from nature. This indicates that selection against misfolded proteins strongly influences the survival of natural recombinants.

have, to our knowledge, never been used to analyse any naturally generated chimaeric proteins, we realised they should also be useful for understanding breakpoint distribution patterns found within coding regions of recombining virus genomes.

Recently, Lefeuvre et al. [16] reported the first statistically supported evidence of recombination hot- and cold-spots in the genomes of begomoviruses, members of a highly recombinagenic family of single-stranded DNA viruses called the Geminiviridae. Importantly, they detected a substantial number of recombination events within a portion of the begomovirus replication–associated protein (*rep*) gene encoding a protein for which a high resolution crystal structure is available. In this paper we describe an expanded analysis of recombination amongst begomoviruses. We identify sets of unambiguously unique recombination events detectable in publicly available monopartite begomovirus DNA-A-like sequences and bipartite begomovirus DNA-A sequences. We then determine the distribution of recombination breakpoints across the analysed sequences and confirm the recombination hot- and cold-spots identified previously. We use a method called SCHEMA [8] to predict degrees of fold disruption in chimaeric begomovirus Rep and coat protein (CP) molecules (for which a reasonably high resolution structural model exists) expressed by viruses determined to have recombinant *rep* and *cp* genes. We then compare these predictions with those for an exhaustive set of all possible recombination breakpoint pairs within these genes and provide the first statistical evidence to our knowledge that avoidance of protein fold disruption is a major factor shaping the patterns of recombination that are detectable in natural virus populations.

Results/Discussion

We anticipated that general rules governing the evolutionary advancement of viruses through recombination should be most manifest in virus groups in which distinctive conserved patterns of recombination have emerged [16–21]. Given that begomoviruses are both highly recombinogenic [22] and display some evidence of recombination breakpoint hot- and cold-spots [16], we undertook a detailed analysis of recombination in this group.

Are Patterns of Recombination Conserved amongst All Begomoviruses?

We began by precisely mapping the distributions of recombination events across begomovirus DNA-A and DNA-A-like sequences sampled throughout the world. Using a battery of recombination signal detection tools and rigorous manual and automated evaluation of recombination signals, we identified sets of 120 and 164 non-ambiguous unique recombination events in the bipartite begomovirus DNA-A and monopartite DNA-A-like sequences, respectively (see Datasets S1 and S2, Figure S1, and Tables S1 and S2 for detailed descriptions of all the detected events).

These events were mapped onto "recombination count matrices" (Figure 1). These matrices represent the number of times that recombinational movement of sequence tracts within the analysed genomes has separated pairs of nucleotide sites. This representation of the characterised recombination events highlights the differential "exchangeability" of sequence tracts within begomovirus DNA-A and DNA-A-like sequences. Whereas highly exchangeable genome regions (i.e., those separated many times by recombination from their original genetic background) are represented by red/purple shades, the least exchangeable regions (i.e., those separated the fewest times by recombination) are represented by yellow/green shades. As can be seen from Figure 1, the region of the *rep* gene encoding the N-terminal portion of Rep and the adjacent intergenic region sequences up to the virion strand origin of replication (v-*ori*) are the regions of both monopartite and bipartite begomovirus genomes most frequently exchanged during recombination. As a result of this, the 5′ and 3′ portions of *rep* are very frequently inherited from different parents. This implies that *rep* must be comprised of highly modular subregions capable of proper functioning in diverse foreign genetic backgrounds. Conversely, the small numbers of detectable recombination events that separate fragments of the *cp* gene indicate that naturally occurring monopartite and bipartite begomovirus recombinants tend to inherit the portions of this gene encoding CP amino acids 75 through 220 from a single source.

To visualise the distribution of recombination breakpoints in monopartite and bipartite begomovirus genomes, all approximated recombination breakpoint locations were plotted on density maps and a permutation test was used to determine whether there were any statistically significant hot- or cold-spots in the breakpoint distribution. This test indicated that the distribution of breakpoints was significantly non-random, with clear recombination hot- and cold-spots being detectable (Figure 2). It is apparent that, as for the recombinant region count matrices (Figure 1), the recombination breakpoint distributions detected in the monopartite and bipartite datasets are very similar. The clusters of inferred breakpoint positions in the two datasets do not, however, have identical significance levels, probably due to differences in datasets with respect to both their sequence diversity, and the number of detectable recombination events they contain.
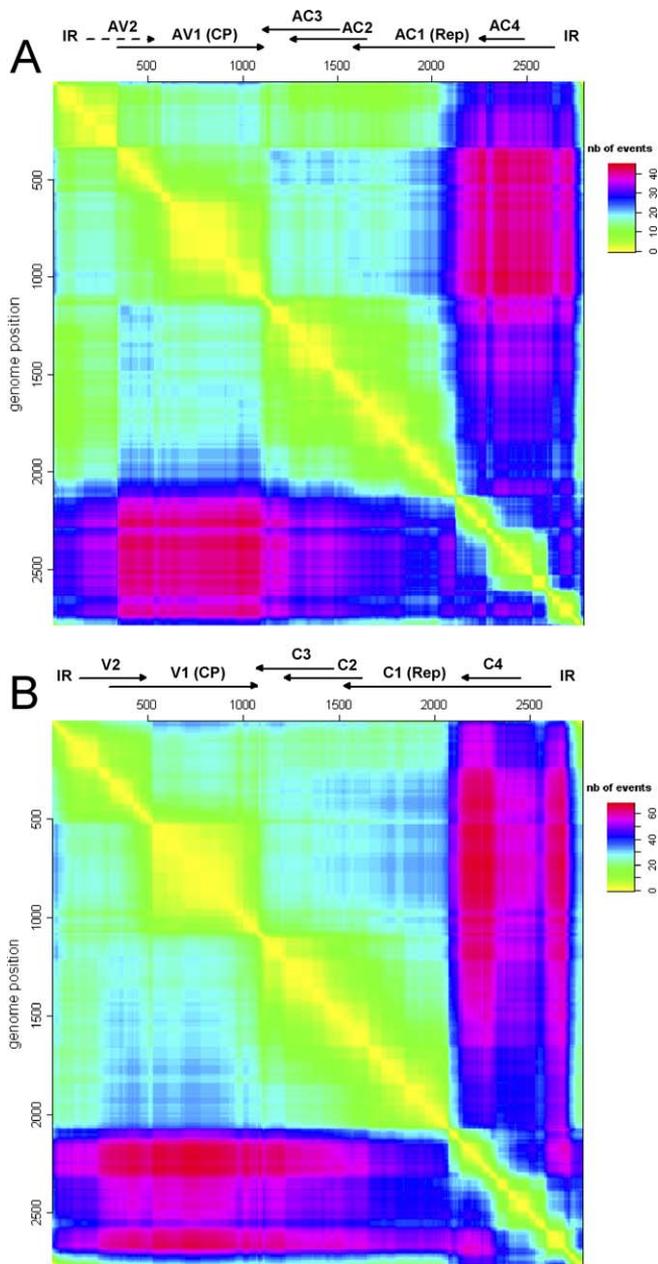
Figure 1. Recombination Region Count Matrix of Unique Recombination Events Detected amongst (A) DNA-A Sequences of Bipartite Begomoviruses and (B) DNA-A-Like Sequences of Monopartite Begomoviruses

Unique recombination events have been mapped onto the matrix based on their estimated breakpoint positions. The shades displayed are a function of the number of times pairs of nucleotides (plotted on the x- and y-axis) are separated during the observed set of unique recombination events. Diagrams indicating the positions of landmarks in begomovirus DNA-A/DNA-A-like sequences are shown on the top of the matrices. Positions were drawn in relative to EACMCV-[TZ] (AY795983) for bipartite sequences and ToLCYT-[Dem] (AJ865341) for monopartite sequences.

doi:10.1371/journal.ppat.0030181.g001

Figure 2. The Distribution of Recombination Breakpoints Detected within (A) DNA-A Sequences of Bipartite Begomovirus and (B) DNA-A-Like Sequences of Monopartite Begomoviruses

All estimated breakpoint positions are indicated by small vertical lines at the top of the graph. A 200-nucleotide window was moved along the alignment one nucleotide at a time and the number of breakpoints detected within the window region was counted and plotted (solid line). The horizontal lines at the top of each graph indicate 99% and 95% confidence thresholds for globally significant breakpoint clusters. Light and dark grey areas respectively indicate local 99% and 95% breakpoint clustering thresholds, taking into account local regional differences in sequence diversity that influence the ability of different recombination detection methods to identify recombination breakpoints. Red areas indicate recombination hot-spots, while blue areas represent recombination cold-spots. ORFs (horizontal arrows) and IR are represented on the top of the graph.

doi:10.1371/journal.ppat.0030181.g002

In both the monopartite and bipartite datasets, larger "globally" significant (global p-values <0.05 across its length) and smaller "locally" significant (local p-values <0.01) recombination hot-spots are apparent in the intergenic region (IR) and complementary strand ORFs. In the monopartite dataset one large globally significant ($p < 0.01$) hot-spot
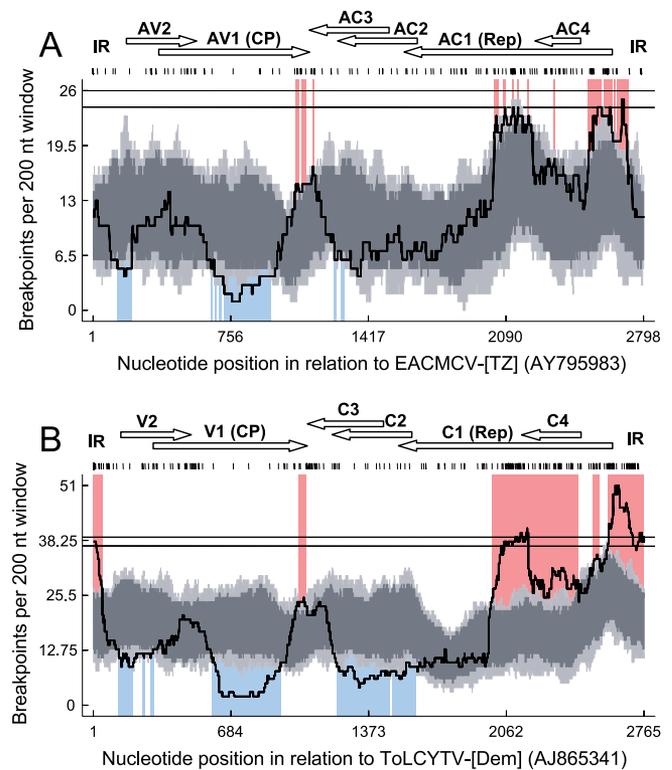
encompasses the entire IR 5′ of the v-ori, and another ($p < 0.01$) occurs near the centre of rep. Globally significant hotspots ($p < 0.05$ and $p = 0.05$) are also detected in these positions in the bipartite dataset, but the extent of the IR hot-spot is not as great. This is probably due in large part to the particularly low quality of nucleotide sequence alignment achievable amongst the highly divergent bipartite begomovirus IR sequences. Locally significant hot-spots occur at the interface of cp and the C3 ORFs in both bipartite and monopartite sequences. In addition to hot-spots, the analysis also revealed locally significant recombination cold-spots. These occurred in the first half of cp and in the third quarter of the V1 ORF for both datasets and in the overlapping region of the C2 and C3 ORFs of the bipartite dataset.

These results clearly indicate that recombination hot- and cold-spots previously identified amongst African and Mediterranean begomoviruses [16] are conserved amongst both monopartite and bipartite begomoviruses found worldwide.

Importantly, the distribution of these recombination hot- and cold-spots is largely consistent with observations made

during experimental analyses of geminivirus recombination [23–25] in which the V1/C3 ORF interface and the v-ori have been identified as potential recombination hot-spots. Also, analysis of replicating begomoviral DNA intermediates has revealed a wide distribution of so-called heterogeneous length linear dsDNA forms (hDNA). The ends of these hDNA molecules tend to map most frequently to the v-ori and either the AC2/AC3 transcription promoter at the hot-spot we detect in the centre of rep, or the C2/C3 terminator at the hot-spot we detect at the V1/C3 ORF interface. It has been convincingly demonstrated that these "broken" replicative intermediates are diverted into the recombination-dependent replication pathway of begomoviruses, which would neatly explain the recombination hot-spots detected in these regions [26].

Furthermore, population genetic analysis of recombination rates in large full genome datasets of very closely related groups of maize streak viruses (a geminivirus species in the genus *Mastrevirus*) and cassava-infecting geminiviruses has indicated that the base biochemical recombination rates in sequences encoding complementary sense genes are probably five to 12 times higher than they are in sequences encoding the virion sense genes [27]. Importantly, these studies also show that greatest changes in recombination rates occur near the V1/C3 ORF interface and the v-ori.

That all of these lines of evidence indicate the complementary sense ORFs of geminiviruses are biochemically more predisposed to recombination than their virion sense ORFs strongly suggests that something about the direction of transcription of these ORFs may be responsible for the recombination rate imbalance. For example, it has been proposed that complementary sense gene transcription, which occurs in the opposite direction to virion strand synthesis during rolling circle replication of geminivirus genomes, may be responsible for an increased rate of replication complex displacement during replication of the complementary sense ORFs [16,26]. Completion of replication from partially replicated virion strands would then proceed via the recombination-dependent replication pathway [26] which, in the presence of potential template DNAs with dissimilar sequences, could result in an increased prevalence of detectable recombination events across the complementary sense ORFs.

Do Recombination Breakpoints "Avoid" Disruption of Protein Folding?

While these mechanistic processes might account for both a general imbalance between recombination rates in the virion and complementary sense ORFs and hot-spots at the V1/C3 ORF interface, the centre of rep and the v-ori, they cannot completely explain, for example, the apparently conserved breakpoint clusters (albeit not hot-spots) at the 3′ end of the V2 ORF and breakpoint cool/cold-spots in the C2 ORF. We have noted previously [16] that, besides the hot-spot in rep, peaks in breakpoint density tend to occur either outside or near the ends of genes, whereas cold-spots tend to occur well within genes. As has been previously suggested in analyses of both virus recombination [28] and DNA shuffling experiments [8], this indicated to us that selection might preferentially favour the survival of recombinants that either do not express chimaeric proteins, or express chimaeric proteins in which recombination has not damaged amino

acid interactions required for proper protein folding. We therefore decided to test whether selection against disruption of protein folding might not also be at least partially responsible for some of the conserved breakpoint density peaks and troughs observable in Figure 2.

Recombination events with inferred breakpoint positions within the portions of rep and cp genes that encode protein fragments with available 3-D structural data were identified. These included 12 and five events in the cp genes of the monopartite and bipartite begomoviruses, respectively, and 29 and 19 events in the rep genes of the monopartite and bipartite begomoviruses, respectively.

We used the SCHEMA method to predict degrees of fold disruption in the chimaeric Rep and CP molecules expressed by the recombinant viruses we identified. This analysis indicated that the average degree of potential fold disruption in the chimaeric Rep molecules was higher than that in the chimaeric CP molecules (E-values in Table 1). Rather than indicating that recombination should be more tolerable in cp than it is in rep, this result simply reflects that CP molecules are much more conserved than Rep molecules and that there are consequently fewer potentially disruptive combinations of amino acids. It should also be pointed out that both CP and Rep have overlapping ORFs: V2/AV2 in the case of CP and C4/AC4 in the case of Rep. Whereas in the case of CP the overlap with V2/AV2 involves only approximately 3% (six codons) of the analysed CP region, the overlap of Rep and C4/AC4 involves approximately 58% (67 codons) of the analysed Rep region. It is possible that the evolutionary constraints of overlapping coding regions have made Rep more robust than CP with respect to potentially clashing amino acid interactions within its folded structure. Importantly, by increasing the number of tolerable Rep mutations, the only influence increased folding robustness might have had on our analyses would have been to decrease the power of our tests for preservation of intra-protein interactions.

We performed SCHEMA analyses on the two simulated datasets (described in Materials and Methods) to determine whether the degrees of Rep and CP fold disruption predicted for the observed recombinants were significantly lower than would be expected if selection did not act against recombinants expressing chimaeric proteins with high degrees of predicted fold disruption. Analysis of the exhaustive genome event dataset indicates that predicted degrees of fold disruption in real CP and Rep chimaeras were significantly lower (Table 1; p-values $= 5.1 \times 10^{-2}$ and $< 1.0 \times 10^{-4}$ for the bipartite and monopartite CP datasets, respectively, and 1.7×10^{-3} and 6.6×10^{-3} for the monopartite and bipartite Rep datasets, respectively) than would be expected in the absence of selection against fold disruption. However, the amino acid mutation levels (m-values in Table 1) of the real datasets indicate that the real recombination events also tended to involve transfers of significantly fewer non-synonymous mutations than the simulated datasets (Table1; p-values $= 1.0 \times 10^{-4}$ and 1.7×10^{-2} for the mono and bipartite CP datasets, respectively, and 1.0×10^{-4} and 2.4×10^{-2} for the monopartite and bipartite Rep datasets, respectively). This indicated that the significantly reduced degrees of predicted fold disruption in the real datasets relative to the simulated datasets are at least in part due to the real recombination events involving transfers of significantly fewer non-synonymous mutations than the simulated events. This implies that

Table 1. SCHEMA Derived Estimates of Recombination Induced Effective Mutation (m) and Protein Fold Disruption (E) for Monopartite and Bipartite Begomovirus Datasets Comparing Real and Simulated Recombination Events

Genomic Portion Analysed		V1 / AV1			C1 / AC1		
Position in Monopartite and Bipartite Sequences[a]		468-1055 / 502-1089			2249–2601 / 2282–2635		
Recombination Type		Real	Exhaustive Genome	Exhaustive Non-Synonymous	Real	Exhaustive Genome	Exhaustive Non-Synonymous
Monopartite	Number of events	12 (11[b])	3080	274	29 (28[b])	5281	835
	Mutation (m) mean ± sd (p-value[c])	1.7 ± 1.1	6.0 ± 3.2 (1.0e-4)	1.8 ± 1.0 (1.0)	6.5 ± 4.4	8.12 ± 3.5 (1.0e-4)	6.75 ± 4.3 (1.0)
	Disruption (E) mean ± sd (p-value[c])	0.33 ± 0.89	2.3 ± 2.4 (<1.0e-4)	1.3 ± 1.1 (3.0e-4)	6.0 ± 4.6	8.9 ± 6.6 (1.7e-3)	8.8 ± 6.9 (<1.0e-4)
Bipartite	Number of events	5 (5[b])	1543	188	19 (18[b])	3942	750
	Mutation (m) mean ± sd (p-value[c])	2.8 ± 1.6	9.1 ± 3.2 (1.7e-2)	2.8 ± 1.6 (1.0)	8.3 ± 5.0	10.8 ± 4.5 (2.4e-2)	8.7 ± 4.7 (1,0)
	Disruption (E) mean ± sd (p-value[c])	0.80 ± 0.84	4.0 ± 2.8 (5.1e-2)	2.2 ± 1.7 (2.9e-2)	8.7 ± 8.8	12.7 ± 8.7 (6.6e-3)	12.3 ± 8.3 (6.0e-4)

[a]Nucleotide positions within the ORFs analysed in relation to ToLCYTV-[Dem] (AJ865341) for monopartite sequences and EACMV-[TZ] (Z83256) for bipartite sequences.
[b]Number of events involving the transfer of non-synonymous polymorphisms.
[c]Probability that real events are not less mutative/disruptive than the simulated events (exhaustive genome events and exhaustive non-synonymous events). $p < 0.05$ was considered significant.
doi:10.1371/journal.ppat.0030181.t001

the real breakpoints tend to occur closer to the edges of the analysed regions than one would expect if breakpoints occurred randomly throughout the regions. Given the breakpoint density peaks on either side of the CP encoding V1 ORF (Figure 2), we had anticipated this result for the CP dataset. However, a similar result obtained for the Rep dataset implies that, despite the high density of breakpoints throughout the C1 region encoding the fragment of Rep that was analysed, there is still a significant tendency for breakpoints to occur closer to the edges of the analysed region than would be expected by chance.

We therefore decided to test whether avoidance of protein fold disruption is achieved only through avoidance of non-synonymous mutation mixing, or whether, controlling for unequal degrees of non-synonymous mutation mixing, it is also achieved through preferential mixing of non-disruptive non-synonymous mutations. Analysis of the simulated non-synonymous event dataset indicated that in the natural monopartite and bipartite *cp* and *rep* recombinants, there does indeed appear to have been preferential mixing of non-disruptive non-synonymous mutations (Table 1). This demonstrates, therefore, that whenever an interaction between a pair of polymorphic amino acid residues is predicted to be important for Rep or CP folding, there is a tendency for nucleotide sequences encoding these amino acids to be inherited from the same parental virus significantly more often than those encoding non-interacting pairs of polymorphic amino acids.

Despite the common conception that recombination is a highly efficient mechanism used by both microorganisms and protein engineers in the discovery of phenotypic novelty and/or improved fitness, rules constraining the evolutionary utility of recombination are beginning to emerge. It has been experimentally demonstrated that the viability of recombinant viruses and the activities of chimaeric proteins are strongly influenced by both the relatedness of their parents and the inherent "modularity" of the sequence tracts they inherit from them [8,10]. Put simply, fragments of sequence that do not interact extensively with other sequence fragments tend to function well when transferred into even

highly divergent foreign genetic backgrounds, whereas those that interact extensively with other sequence fragments tend to only work properly when transferred into foreign genetic backgrounds that are not very different from those in which they evolved [10,11]. It is therefore probable that high profile natural recombinant viruses, such as those that are responsible for disease outbreaks [29] or those that have novel host ranges and phenotypes [30,31], or even those that have simply emerged as prominent circulating members of virus populations [32–35], represent the exceptional, reasonably fit subset of a vastly greater but vastly more ephemeral group of defective "hopeful monsters" culled by purifying selection.

Our results provide clear supporting evidence for this notion that purifying selection is a major factor shaping at least part of the distinctive patterns of natural recombination found in begomoviruses. Within the genome regions analysed, we find strong statistical evidence that natural recombination events have tended to involve sequence exchanges that avoid the transfer of non-synonymous nucleotide polymorphisms (i.e., those encoding different amino acids in the different parents) between genomes. We also show that, when non-synonymous polymorphisms are transferred between genomes, there is a statistically significant tendency to avoid transfer of those non-synonymous polymorphisms that are predicted to disrupt the folding of expressed chimaeric proteins. While these results indicate that interaction networks required for proper protein folding are preserved in the natural recombinants, they imply that strong selective forces must operate against any novel recombinant in which these interaction networks are not preserved.

A major omission in our analysis of intra-protein amino acid interactions is our failure to consider all the other potential interactions that most likely occur within the genomes analysed. These include many sequence-specific inter-protein and protein–DNA interactions that might also constrain the viability of recombinants [10]. While experimental work towards obtaining high resolution genome-wide interaction maps has only begun for most virus taxa (including the begomoviruses), exciting new analysis methods

are being developed to identify both coevolving (or covarying) amino acids within protein sequence alignments [36–38] and epistatically interacting nucleotide sites within DNA sequence alignments [39,40]. Although these methods promise the mapping of interaction networks directly from naturally sampled viral genome sequences, it is currently unknown how well they will fare given datasets containing (1) large numbers of recombinant sequences or (2) obvious recombination hot-spots. It is very likely, for example, that many non-synonymous polymorphisms on sequence tracts between recombination hot-spots will be detectably "covariant" if they are frequently transferred amongst genomes (i.e., on an imposed phylogenetic tree it will appear as though the same sets of sites change simultaneously on multiple branches of the tree). If some of these or future related analytical methods prove robust to the influences of recombination, an obvious application of these would be to determine whether there is also a significant tendency for recombination to avoid disrupting these genome-wide protein–protein, protein–nucleic acid, and nucleic acid–nucleic acid interactions.

In fields as diverse as microbial evolution [41,42], protein engineering [8,9], and computer science [43], maintenance of interaction networks is emerging as a common theme unifying studies aimed at delimiting recombination's potential as an exploratory strategy. Many and complex interactions is a defining feature of living systems. When these interactions are encoded within genome sequences they form an epistatic architecture. It is really just common sense that for productive recombination to occur it must happen without damaging the integrity of these largely intangible network-like structures. Maintenance of these networks might in fact be directly responsible for the evolution of differential biochemical predispositions for recombination across genomes: If recombination events mostly occur in genome regions with low connectivity, a greater proportion of recombinants would be viable than if recombination events were randomly scattered across genome regions with low and high connectivity. Other evolutionary strategies to ensure maintenance of interaction networks in the face of continual recombination might be the evolution of network robustness, or an increased capacity to mutationally compensate for deleterious recombination events. Conversely, however, the network architectures themselves might also evolve over time to accommodate biochemically predisposed recombination hot-spots that have some biological importance. We have shown here that in the case of the begomoviruses at least, various biochemical and selective processes working in tandem most likely combine to produce the distinctive patterns of recombination seen in nature.

Materials and Methods

Sequence data. All available monopartite and bipartite begomovirus DNA-A and DNA-A-like sequences were obtained from public sequence databases using TaxBrowser (http://www.ncbi.nlm.nih.gov/) in May 2006. Multiple sequence alignments were constructed separately for monopartite and bipartite sequences using POA [44], the ClustalW [45] based sub-alignment tool available in MEGA 3.1 [46], and manual editing. While great care was taken to ensure the most accurate alignment possible, during subsequent recombination analyses additional alignment checks were performed in RDP3 (also using the ClustalW method) for every recombination signal detected to ensure that they were not misalignment artefacts [45]. To minimise

the number of tests performed during recombination analyses (and therefore increase the statistical power of the analyses) all but one sequence within groups of sequences sharing more than 98% nucleotide identity were discarded. The resulting monopartite DNA-A-like and bipartite DNA-A sequence alignments contained 123 and 116 sequences, respectively.

Structural data. The Rep protein structure (catalytic domain; residues 4–121) of TYLCSV has been determined by NMR spectroscopy. This Rep structure (PDB ID 1L2M) comprises five anti-parallel β sheets in the centre with a two-stranded β sheet, a β hairpin, and two α helices on the periphery [47]. The begomovirus capsid structures (196-aa core CP) has been modeled based on the crystal structure of *Satellite Tobacco Necrotic Virus* and fitted into an approximately 20-Å density map generated from cryoelectron microscopy reconstructions of *African cassava mosaic virus* particles [48]. The PDB file of this structure was kindly provided by B. Böttcher.

Recombination analysis. Identification of potential recombinants, parental sequences, and approximation of possible recombination breakpoint positions was carried out using the RDP [49], Geneconv [22], RecScan [50], Maximum Chi Square [51], Chimaera [52], and SisterScan [53] methods as implemented in RDP3 [52], which is available from http://darwin.uvigo.es/rdp/rdp.html (for full details of program settings, see Datasets S1 and S2). The analyses were performed with default settings for the detection methods, a Bonferroni-corrected p-value cutoff of 0.05, and a requirement that any potential event be detectable by two or more methods. It is important to point out that implementations of all these recombination detection methods in RDP3 were not severely constrained by the initial window size settings specified at the onset of the analyses. All of the methods used include an algorithm for dynamically optimising window sizes for the detection of recombination signals during an initial exploratory phase of recombination detection. Following this exploratory phase, RDP3 rechecks every detected recombination signal with all six methods with a starting window size seeded with that used by whatever method initially detected the signal.

The approximate breakpoint positions and recombinant sequence(s) inferred for every detected potential recombination event were manually checked and adjusted where necessary using the extensive phylogenetic and recombination signal analysis features available in RDP3. This process further reduced any possible influence that initial window size settings had on the final estimates of breakpoint positions. Once a set of unique recombination events was identified, a breakpoint map containing the positions of all clearly identifiable breakpoints was compiled. A breakpoint density plot was then constructed from this map as described in Heath et al. [17].

SCHEMA analysis. SCHEMA takes as input a PDB protein structure file and parental amino acid sequence files. It uses the protein structural information to properly fold the parental amino acid sequences and then identifies potentially interacting amino acid pairs based on their proximity (in this case within 4.5 Å) within the resulting folds. The amino acid contact map yielded by this process can be used to determine the degree of fold disruption expected in any conceivable chimaera of the parental amino acid sequences. The way this is done is relatively simple: For all the amino acid residues that are potentially interacting within a folded chimaeric protein, SCHEMA counts the number of instances where the interacting pairs are non-parental. Non-parental interacting amino acid pairs arise when the parental molecules differ from one another at two potentially interacting amino acid residues and the chimaera inherits one-half of the potentially interacting pair from one parent and the other half from the other parent. Counts of these non-interacting pairs in chimaeric proteins, called "E" values, have been shown to correlate directly with degrees of fold disruption experienced by the proteins. The value of E therefore corresponds with expected degree of fold disruption. SCHEMA also counts the number of amino acid substitutions that would be required to convert a chimaera into the parental sequence that it most closely resembles—this value is referred to as "m" [8].

We selected recombination events in the monopartite and bipartite sequence datasets for which (1) sequences closely resembling inferred parental sequences were identifiable and (2) recombination breakpoints occurred in genome regions encoding the portions of Rep and CP with known/approximated 3-D structures. These events constituted a "real event" dataset that we analysed using the SCHEMA method.

We devised a permutation test to determine whether predicted CP and Rep fold disruptions incurred by real events were less severe than those incurred by random recombination events with the same

parental sequences simulated throughout the *rep* and *cp* regions under consideration. The permutation test involved two different sets of simulated recombination events. The first set was derived from each real event by moving the breakpoints observed in the real event backwards and forwards along the entire nucleotide sequence alignment one polymorphic alignment position at a time until every possible unique recombination event involving the "exchange" of exactly the same number of polymorphic nucleotides as the real event were simulated within the parental sequences. We called the complete set of simulated events constructed from the entire real event dataset the "exhaustive genome event" dataset. Note that although these events involved exchanges of the same number of total polymorphic sites as was observed for the real events, they can involve a different number of polymorphic sites in the particular genomic regions analysed (i.e., those encoding portions of CP and Rep with available 3-D structure information). This set of simulated events was used to determine whether there was a significant tendency for the observed recombination breakpoints to occur on the edges of these analysed regions. The second set of simulated events was generated by considering only non-synonymous polymorphisms within the alignment regions encoding CP or Rep fragments with available structural data. A window containing the same number of non-synonymous mutations as a corresponding real event was moved along the analysed region, and all possible recombination events were simulated. The subsequent events share exchanges of exactly the same numbers of non-synonymous mutations as the real events and consequently all have the same SCHEMA m values. This set of events was called the "exhaustive non-synonymous event" dataset (see Figure S2 for simulation details). This set of simulated events was used to determine whether, given "exchanges" of the same numbers of polymorphic amino acids as were observed for real events, there was a significant tendency for the real recombination events to exchange less disruptive amino acid polymorphisms.

Quantification of potential fold disruption in real and simulated chimaeric CP and Rep molecules, respectively, expressed by real and simulated recombinants, was carried out using SCHEMA. For each of these chimaeras, amino acid sequences of inferred parents and chimeras were aligned with MUSCLE using default settings [54]. The python scripts SCHEMACONTACTS and SCHEMAENERGY [9] were used to compute m (mutational distance between chimaeras and their most closely related parent) and E (predicted fold disruption) scores for all simulated and real chimaeras. The analysis procedure was automated using some of the extensive functionalities available in the R package [55], APE [56], and seqinR (available on http://www.cran.r-project.org/). R scripts for these analyses are available on request. We grouped the E and m scores determined for the observed and simulated chimaeras and determined the sum of ranks for the observed chimaeras. We then repeated the entire process 10,000 times but with "real" events randomly chosen from amongst every subset of corresponding simulated events. We propose that the proportion of simulated events with a sum of ranks score lower than or equal to that of the observed event is equivalent to the probability that the breakpoint distributions observed in the real data have not tended to avoid disruption of protein folding. Put another way, we estimate a *p*-value from the proportion of permuted recombination event datasets that on the whole are predicted to be less disruptive to protein folding than the set of actual observed recombination events.

Supporting Information

Dataset S1. RDP3 Bipartite Sequences Analysis

RDP3 project file.

Found at doi:10.1371/journal.ppat.0030181.sd001 (379 KB ZIP).

Dataset S2. RDP3 Monopartite Sequences Analysis

RDP3 project file.

Found at doi:10.1371/journal.ppat.0030181.sd002 (412 KB ZIP).

Figure S1. Recombination Event Density and Parental Sequences Relatedness

Found at doi:10.1371/journal.ppat.0030181.sg001 (25 KB PDF).

Figure S2. Recombination Event Simulation Process

Figures describing how the simulated recombination events are created.

Found at doi:10.1371/journal.ppat.0030181.sg002 (108 KB PDF).

Table S1. Bipartite Sequences Recombination Events

List of recombination events detected with RDP3 in bipartite sequences.

Found at doi:10.1371/journal.ppat.0030181.st001 (132 KB XLS).

Table S2. Monopartite Sequences Recombination Events

List of recombination events detected with RDP3 in monopartite sequences.

Found at doi:10.1371/journal.ppat.0030181.st002 (140 KB XLS).

Reference

1. Michel B, Flores MJ, Viguera E, Grompone G, Seigneur M, et al. (2001) Rescue of arrested replication forks by homologous recombination. Proc Natl Acad Sci U S A 98: 8181–8188.
2. Cromie GA, Connelly JC, Leach DR (2001) Recombination at double-strand breaks and DNA ends: conserved mechanisms from phage to humans. Mol Cell 8: 1163–1174.
3. Felsenstein J (1974) The evolutionary advantage of recombination. Genetics 78: 737–756.
4. Keightley PD, Otto SP (2006) Interference among deleterious mutations favours sex and recombination in finite populations. Nature 443: 89–92.
5. Martin G, Otto SP, Lenormand T (2006) Selection for recombination in structured populations. Genetics 172: 593–609.
6. Crameri A, Raillard SA, Bermudez E, Stemmer WP (1998) DNA shuffling of a family of genes from diverse species accelerates directed evolution. Nature 391: 288–291.
7. Stemmer WP (1994) Rapid evolution of a protein in vitro by DNA shuffling. Nature 370: 389–391.
8. Voigt CA, Martinez C, Wang ZG, Mayo SL, Arnold FH (2002) Protein building blocks preserved by recombination. Nat Struct Biol 9: 553–558.
9. Meyer MM, Silberg JJ, Voigt CA, Endelman JB, Mayo SL, et al. (2003) Library analysis of SCHEMA-guided protein recombination. Protein Sci 12: 1686–1693.
10. Martin DP, van der Walt E, Posada D, Rybicki EP (2005) The evolutionary value of recombination is constrained by genome modularity. PLoS Genet 1: e51. doi:10.1371/journal.pgen.0010051
11. Escriu F, Fraile A, Garcia-Arenal F (2007) Constraints to genetic exchange support gene coadaptation in a tripartite RNA virus. PLoS Pathog 3: e8. doi:10.1371/journal.ppat.0030008
12. Moreno IM, Malpica JM, Diaz-Pendon JA, Moriones E, Fraile A, et al. (2004) Variability and genetic structure of the population of watermelon mosaic virus infecting melon in Spain. Virology 318: 451–460.
13. Jain R, Rivera MC, Moore JE, Lake JA (2003) Horizontal gene transfer accelerates genome innovation and evolution. Mol Biol Evol 20: 1598–1602.
14. Hayes RJ, Bentzien J, Ary ML, Hwang MY, Jacinto JM, et al. (2002) Combining computational and experimental screening for rapid optimization of protein properties. Proc Natl Acad Sci U S A 99: 15926–15931.
15. Saraf MC, Maranas CD (2003) Using a residue clash map to functionally characterize protein recombination hybrids. Protein Eng 16: 1025–1034.
16. Lefeuvre P, Martin DP, Hoareau M, Naze F, Delatte H, et al. (2007)

Begomovirus "melting pot" in the South West Indian Ocean Islands: molecular diversity and evolution through recombination. J Gen Virol 88: 3458–3468.

17. Heath L, van der Walt E, Varsani A, Martin DP (2006) Recombination patterns in aphthoviruses mirror those found in other picornaviruses. J Virol 80: 11827–11832.

18. Shapka N, Nagy PD (2004) The AU-rich RNA recombination hot spot sequence of Brome mosaic virus is functional in tombusviruses: implications for the mechanism of RNA recombination. J Virol 78: 2288–2300.

19. Ohshima K, Tomitaka Y, Wood JT, Minematsu Y, Kajiyama H, et al. (2007) Patterns of recombination in turnip mosaic virus genomic sequences indicate hotspots of recombination. J Gen Virol 88: 298–315.

20. Magiorkinis G, Paraskevis D, Vandamme AM, Magiorkinis E, Sypsa V, et al. (2003) In vivo characteristics of human immunodeficiency virus type 1 intersubtype recombination: determination of hot spots and correlation with sequence similarity. J Gen Virol 84: 2715–2722.

21. Chin MP, Rhodes TD, Chen J, Fu W, Hu WS (2005) Identification of a major restriction in HIV-1 intersubtype recombination. Proc Natl Acad Sci U S A 102: 9002–9007.

22. Padidam M, Sawyer S, Fauquet CM (1999) Possible emergence of new geminiviruses by frequent recombination. Virology 265: 218–225.

23. Stenger DC, Revington GN, Stevenson MC, Bisaro DM (1991) Replicational release of geminivirus genomes from tandemly repeated copies: evidence for rolling-circle replication of a plant viral DNA. Proc Natl Acad Sci U S A 88: 8029–8033.

24. Schnippenkoetter WH, Martin DP, Willment JA, Rybicki EP (2001) Forced recombination between distinct strains of Maize streak virus. J Gen Virol 82: 3081–3090.

25. Garcia-Andres S, Tomas DM, Sanchez-Campos S, Navas-Castillo J, Moriones E (2007) Frequent occurrence of recombinants in mixed infections of tomato yellow leaf curl disease-associated begomoviruses. Virology 365: 210–219.

26. Jeske H, Lutgemeier M, Preiss W (2001) DNA forms indicate rolling circle and recombination-dependent replication of Abutilon mosaic virus. EMBO J 20: 6158–6167.

27. Owor BE, Martin DP, Shepherd DN, Edema RE, Monjane AL, et al. (2007) Genetic analysis of maize streak virus (MSV) isolates from Uganda reveals widespread distribution of a recombinant MSV variant in Uganda. J Gen Virol 88: 3154–3165.

28. Bonnet J, Fraile A, Sacristan S, Malpica JM, Garcia-Arenal F (2005) Role of recombination in the evolution of natural populations of Cucumber mosaic virus, a tripartite RNA plant virus. Virology 332: 359–368.

29. Legg JP, Thresh JM (2000) Cassava mosaic virus disease in East Africa: a dynamic disease in a changing environment. Virus Res 71: 135–149.

30. Russell CJ, Webster RG (2005) The genesis of a pandemic influenza virus. Cell 123: 368–371.

31. Gibbs MJ, Weiller GF (1999) Evidence that a plant virus switched hosts to infect a vertebrate and then recombined with a vertebrate-infecting virus. Proc Natl Acad Sci U S A 96: 8022–8027.

32. Chare ER, Holmes EC (2006) A phylogenetic survey of recombination frequency in plant RNA viruses. Arch Virol 151: 933–946.

33. Rousseau CM, Learn GH, Bhattacharya T, Nickle DC, Heckerman D, et al. (2007) Extensive intrasubtype recombination in South african human immunodeficiency virus type 1 subtype C infections. J Virol 81: 4492–4500.

34. Varsani A, van der Walt E, Heath L, Rybicki EP, Williamson AL, et al. (2006) Evidence of ancient papillomavirus recombination. J Gen Virol 87: 2527–2531.

35. Tyler SD, Severini A (2006) The complete genome sequence of herpesvirus papio 2 (Cercopithecine herpesvirus 16) shows evidence of recombination events among various progenitor herpesviruses. J Virol 80: 1214–1221.

36. Fares MA, McNally D (2006) CAPS: coevolution analysis using protein sequences. Bioinformatics 22: 2821–2822.

37. Gloor GB, Martin LC, Wahl LM, Dunn SD (2005) Mutual information in protein multiple sequence alignments reveals two classes of coevolving positions. Biochemistry 44: 7156–7165.

38. Dutheil J, Pupko T, Jean-Marie A, Galtier N (2005) A model-based approach for detecting coevolving positions in a molecule. Mol Biol Evol 22: 1919–1928.

39. Shapiro B, Rambaut A, Pybus OG, Holmes EC (2006) A phylogenetic method for detecting positive epistasis in gene sequences and its application to RNA virus evolution. Mol Biol Evol 23: 1724–1730.

40. Zuker M, Jacobson AB (1998) Using reliability information to annotate RNA secondary structures. RNA 4: 669–679.

41. Jain R, Rivera MC, Lake JA (1999) Horizontal gene transfer among genomes: the complexity hypothesis. Proc Natl Acad Sci U S A 96: 3801–3806.

42. Fraser C, Hanage WP, Spratt BG (2007) Recombination and the nature of bacterial speciation. Science 315: 476–480.

43. Holland J (1975) Adaptation in natural and artificial systems. Ann Arbor (Michigan): University of Michigan Press.

44. Lee C, Grasso C, Sharlow MF (2002) Multiple sequence alignment using partial order graphs. Bioinformatics 18: 452–464.

45. Thompson JD, Higgins DG, Gibson TJ (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.

46. Kumar S, Tamura K, Nei M (2004) MEGA3: Integrated software for Molecular Evolutionary Genetics Analysis and sequence alignment. Brief Bioinform 5: 150–163.

47. Campos-Olivas R, Louis JM, Clerot D, Gronenborn B, Gronenborn AM (2002) The structure of a replication initiator unites diverse aspects of nucleic acid metabolism. Proc Natl Acad Sci U S A 99: 10310–10315.

48. Bottcher B, Unseld S, Ceulemans H, Russell RB, Jeske H (2004) Geminate structures of African cassava mosaic virus. J Virol 78: 6758–6765.

49. Martin D, Rybicki E (2000) RDP: detection of recombination amongst aligned sequences. Bioinformatics 16: 562–563.

50. Martin DP, Posada D, Crandall KA, Williamson C (2005) A modified bootscan algorithm for automated identification of recombinant sequences and recombination breakpoints. AIDS Res Hum Retroviruses 21: 98–102.

51. Maynard SJ (1992) Analyzing the mosaic structure of genes. J Mol Evol 34: 126–129.

52. Martin DP, Williamson C, Posada D (2005) RDP2: recombination detection and analysis from sequence alignments. Bioinformatics 21: 260–262.

53. Gibbs MJ, Armstrong JS, Gibbs AJ (2000) Sister-Scanning: a Monte Carlo procedure for assessing signals in recombinant sequences. Bioinformatics 16: 573–582.

54. Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 32: 1792–1797.

55. R Development Core Team (2006) R: A language and environment for statistical computing. Vienna: R Foundation for Statistical Computing. ISBN 3-900051-07-0. Available: http://www.R-project.org/. Accessed 23 October 2007.

56. Paradis E, Claude J, Strimmer K (2004) APE: analyses of phylogenetics and evolution in R language. Bioinformatics 20: 289–290.