Steele, E.J., Williamson, J.F., Lester, S., Stewart, B.J., Millman, J.A., Carnegie, P., Lindley, R.A., Pain, G.N. and Dawkins, R.L. (2011) Genesis of ancestral haplotypes: RNA modifications and reverse transcription-mediated polymorphisms. Human Immunology, 72 (3). 283-293.e1.

Accepted Dec 7 2010 *Human Immunology*

Genesis of ancestral haplotypes: RNA modifications and RT-mediated polymorphisms

Edward J Steele [a,b*], Joseph F Williamson [a,b], Susan Lester [a,c], Brent J Stewart [a,b], John A Millman [a], Pat Carnegie [a,b], Robyn A Lindley [a], Geoff N Pain [a] and Roger L Dawkins [a,b]

[a]C.Y.O'Connor ERADE Village Foundation
Canning Vale, Western Australia
AUSTRALIA

[b]School of Veterinary and Biomedical Sciences
Murdoch University, Western Australia
AUSTRALIA

[c]Rheumatology Department,
The Queen Elizabeth Hospital, Woodville,
South Australia
AUSTRALIA

Running head: *The RT-Mediated Long DNA Conversion (LDC) Hypothesis*

**Corresponding author:* Edward J Steele, CY O'Connor ERADE Village, PO Box 5100, Canning Vale South, Western Australia 6155, Australia Fax +618 9397 E-Mail: ejsteele@cyo.edu.au and e.j.steele@melvilleanalytics.com

Key words: polymorphic frozen block, somatic hypermutation, SNP clusters, RT-mediated gene conversion, transcription-coupled repair, transcription factories

Abstract

Understanding the genesis of the block haplotype structure of the genome is a major challenge. With the completion of the sequencing of the Human Genome and the initiation of the HapMap project the concept that the chromosomes of the mammalian genome are a mosaic, or patchwork, of conserved extended block haplotype sequences is now accepted by the mainstream genomics research community. Ancestral Haplotypes (AHs) can be viewed as a recombined string of smaller Polymorphic Frozen Blocks (PFBs). How have such variant extended DNA sequence tracts emerged in evolution? Here the relevant literature on the problem is reviewed from various fields of molecular and cell biology particularly molecular immunology and comparative and functional genomics. Based on our synthesis we then advance a testable molecular and cellular model. A critical part of the analysis concerns the origin of the strand biased mutation signatures in the transcribed regions of the human and higher primate genome, A-to-G *versus* T-to-C (ratio ~1.5 fold) and C-to-T *versus* G-to-A (≥1.5 fold). A comparison and evaluation of the current state of the fields of immunoglobulin Somatic Hypermutation (SHM) and Transcription-Coupled DNA Repair (TCR) focused on how mutations in newly synthesized RNA might be copied back to DNA thus accounting for some of the genome-wide strand biases (e.g. the A-to-G *vs* T-to-C component of the strand biased spectrum). We hypothesize that the genesis of PFBs and extended AHs occurs during mutagenic episodes in evolution (e.g. retroviral infections) and that many of the critical DNA sequence diversifying events occur first at the RNA level e.g. recombination between RNA strings resulting in tandem and dispersed RNA duplications (retroduplications), RNA mutations via adenosine-to-inosine pre-mRNA editing events as well as error prone RNA synthesis. These are then copied back into DNA by a cellular reverse transcription (RT) process (also likely to be error-prone) we have called "RT-mediated long DNA conversion" (RT-LDC). Finally we suggest that all these activities and others can be envisaged as being brought physically under the umbrella of special sites in the nucleus involved in transcription known as "Transcription Factories" (TF).

1. The Block Haplotype Structure of the Genome

The block haplotype structure of the human major histocompatibility complex (MHC) on chromosome 6 was first recognized in the early 1980s by the groups of Dawkins, McCluskey and associates in Perth, Australia [1,2] and Yunis, Alpers and associates in Boston, USA [3,4]. Spanning approximately 4Mb of DNA at chromosome band 6p 21.3 these strong linkage disequilibrium (LD) patterns among polymorphic alleles have been confirmed in many studies by these groups [5,6] and others [7,8]. The concept has been extended to the conserved polymorphic blocks of functionally related genes on other chromosomes, such as the genes encoding the proteins regulating complement activation in the RCA complex on chromosome 1 [9,10] as well as other mammalian species, such as the dog MHC [11]. In the 1990s these studies led to the development of the Genomic Matching Technique (GMT) to match donor and recipients at the MHC in bone marrow transplantation [12,13]. The GMT allowed successful matching of potential donors and recipients at the DNA sequence level in the MHC generating characteristic PCR fragment DNA profiles for polymorphic blocks within the MHC that do not undergo recombination. The blocks can be approximately 200-300 Kb in length spanning many genes and their combination is observed in a population as MHC haplotypes which have changed little and remained frozen. For example the beta block in the MHC spans 300 Kb and contains immunological relevant HLA-B, -C, genes as well as other non-HLA genes such as the natural killer cell receptor ligand PERB11 (MIC). In addition it contains two large segmental duplications containing PERB11.1 and PERB11.2 genes, and some other duplicated and polymorphic regions. The GMT is based on priming multiple sites within the block amplifying polymorphic complex sequences providing haplospecific and haplotypic signatures of the entire block rather than individual loci. Extended DNA sequences covering many exonic and intronic regions can be exactly matched by the technique without resorting to DNA sequencing [13].

These and other discoveries are not only relevant to establishing genetic markers associated with human disease but also to our understanding of recent human evolution [5]. The naïve expectation that recent human mutations may

spread rapidly by natural selection is misleading as most sequences are associated with conserved blocks. It was clear by the 1990s that the polymorphisms associated with these long DNA sequence blocks had been maintained with little mutational change for thousands of generations [5]. Thus blocks of frozen sequence containing many different (but possibly functionally related) genes are bounded by recombination hot spots. Such polymorphic blocks might therefore be shuffled by recombination and further modified by segmental duplications.

These early studies suggested that the wider genome may also be composed of a patch work of Ancestral Haplotypes (AHs) or conserved extended haplotypes (CEHs) each consisting of smaller stretches of conserved fragments or Polymorphic Frozen Blocks (PFBs,) bounded by short stretches of hyper recombination [5,6]. Indeed AHs could be considered as selected and thus preserved 'functional genetic units' encompassing many biochemically and physiologically related yet different genes in linkage disequilibrium and possibly regulated by both *cis*-, and in other cases, *trans*-epistasis [14]. They have been conserved over evolutionary time and the haplospecific sequences they contain are essentially the same in remote descendants of founding populations. Many of the common human MHC haplotypes may have therefore been handed down by inheritance for possibly 100,000 years or longer.

Haplospecific blocks range from 50 Kb up to 500 Kb in length and strings of such frozen blocks may extend into the megabase (Mb) range. Considerable evidence now supports this concept [15-17]. One practical motivation has been genome-wide association (GWA) studies [18] involving single nucleotide polymorphisms (SNP). An underlying assumption of many of the GWA studies is that if DNA sequences are inherited in large blocks, then SNPs diagnostic of one part of a DNA sequence should occur in perfect association (or Linkage Disequilibrium) with another more distal region where testing one SNP is exactly equivalent to testing the other such collections of SNPs [15-19]. The extensive DNA sequence SNP data of the HapMap project underpins the consensus of the haplotype block structure of the genome for DNA sequences far removed from the MHC region [15]. For example the Phase 1 SNP HapMap data show numerous

extended (1-4 Mb range) PFB sequences outside of conserved regions spanning centromeres on most autosomes (as well as numerous long PFBs on the X - chromosome). These are conservative estimates given the small size of the human populations examined in the HapMap Phase 1 study (see the Supplementary data in Tables 5 and 6 and Supplementary Fig 8b in ref [15]).

We have summarised the evidence supporting the block haplotype structure of the human genome beyond the MHC because of its relevance to wider genomic phenomena and recent human evolution. In our opinion the full implication of this genomic picture are still not fully appreciated by many biomedical scientists despite the common place application of block concepts to GWA analyses for disease or complex trait discovery [18-20], to forensics [21,22] and the recognition that the synteny of the haplotype block structure is conserved across mammals [5,23].

Many features of PFBs and AHs are therefore well understood. What is not clear is the mechanism of how they originated and why they have been conserved as extended block sequences for so long. Here we are primarily concerned with some new speculative ideas, some unconventional, on how long DNA AHs both within the MHC and beyond, may have been generated over evolutionary time. We review the literature from various fields of molecular and cell biology, molecular immunology as well as comparative and functional genomics leading to a testable molecular and cellular model.

2. Paradox of the Genesis of PFBs and AHs

Ancestral Haplotypes can be viewed as a recombined string of smaller PFBs. Despite their conservation such haplotype blocks also display considerable polymorphism. Variability is observed in terms of the alleles and single nucleotide polymorphisms (SNPs), insertions and deletions (indels) and other duplications, such as copy number variations (CNVs) they contain [5,24,25]. These PFB sequences also reveal an apparent paradox. In the MHC perhaps 4 Mb of contiguous stretches of DNA sequence are 'frozen' in the context of a given Ancestral Haplotype. In all individuals with the *same* AH there are *no* nucleotide

differences between them in this block of the chromosome (apart from a rare SNP). There appears to be a process that coordinates suppression of both SNP and indel generation in the entire block as a well as an apparent recombination suppression mechanism within a block. Recent work identifying recombination hot spot motif sequences e.g. the *PRDM9* gene, may be relevant to the mechanism of apparent recombination suppression within PFBs and AHs [26-29]. Indeed recombination hot spots of 1-2 Kb in length are found to border haplotype blocks [15,18]. The simple explanation is that recombination events are focused at these sites thus giving the impression of recombination suppression within a block. More work is required in identifying other recombination hot spot motifs and how their location on the chromosome is determined.

In comparing two closely related full length MHC ancestral haplotypes such as AH8.1 and AH7.1 they can be shown to differ by numerous SNPs and other sequence modifications [24]. How can we explain the genesis of these multipoint DNA sequence differences? If one haplotype arose by multiple mutations from the other it is axiomatic that the variant must first survive in germ cells and then pass through the "selection gates" of gametogenesis, embryogenesis, neonatal life and then survive to reproductive age.

There is also an apparent contradiction because the “SNP profile in MHC reveals extreme and interrupted levels of nucleotide diversity..between haplotypes” ([24,25]. How could this arise if SNPs are suppressed? A simple interpretation would consist of step-wise natural selection of point mutations in individual genes over aeons of evolutionary time *versus* some type of "big bang" mutational event, over the marked haplotypic region followed by bottle-neck selection, as discussed earlier [5,25]. The former would seem too slow given the known slow rate of mutational change estimated using "molecular clock" techniques. These measure evolutionary change over millions of years assuming spontaneously arising mutations at a constant rate which together with other dating information and the fossil record allow estimation of how long ago two related organisms diverged from a common ancestor.

As advanced earlier, we favour a "big bang" model. To be consistent in terminology we call this the 'mutational spray' model* followed by population

reduction, and thus bottle-neck selection, as the most likely genesis of a new PFB [5,25]. We do not restrict this explanation to blocks within the MHC but assume that may apply in other non-MHC regions of the genome (apart from some highly conserved polymorphic stretches spanning centromeres, which appear to be under a different form of structure-based selection pressure). Can a plausible molecular mechanism based on supporting evidence from a number of fields of molecular and cell biology be advanced to explain the emergence of PFBs and thus AHs both within and beyond the MHC?

(*Note on terminology : Terms such as "mutational/SNP cluster", "mutational/SNP burst" , "big bang burst model" , " quantal burst model", and "mutational/SNP shower" have all been used in the recent literature to describe these clustered mutation phenomena [5,24,25,30-34]. To simplify the terminology we use the terms "mutational spray " or "SNP spray" to describe the same apparent phenomena. The latter may be used interchangeably).

3. Changing views on the origin of SNPs

The number of SNPs constitute a major point of difference between any two closely related polymorphic regions [24,25]. SNP-based alleles are the foundation of the HapMap project. How SNPs might arise in the genome is now a question of general interest. The conventional *a priori* assumption would be that SNPs may originate as random point mutations during meiosis and thus derive from any one of a number of chemical or physical causes both internal and external to the body of the organism.

However recent analyses of the types of nucleotide substitution patterns seen in genomic SNP data sets reveals a very different picture: SNPs across the genome are now known to display a highly non-random pattern. In an important study, Polak and Arndt [35] examined intronic regions in approximately15,000 protein coding genes in a three way genome-wide comparison of human-chimpanzee-rhesus alignments. The results reveal some significant point mutation *strand bias* patterns such that:

a). A-to-G transitions significantly and systematically exceed T-to-C transitions by about 1.5 fold across the transcribed regions of the genome, whereas,

b). C-to-T transitions exceed G-to-A transitions by about 1.5 fold in a smaller 1-2 kb window around transcription start sites (TSS). (The Appendix contains an outline of how strand biased mutation signatures are detected in mutation data sets and see [36,37]).

The first component of the genome-wide SNP signature, A-to-G >> T-to-C, is also the defining strand bias of somatic hypermutation (SHM) of immunoglobulin (Ig) genes [36,37]. It is best now understood as the signature of RNA editing events copied back into DNA by cellular reverse transcription. In our view such a specific mutation ratio is indicative of strand-biased deamination events at the RNA level mediated by the transcription-coupled pre-mRNA editor mediating adenosine-to-inosine (A-to-I) modifications, ADAR1. This is discussed further below in the context of the origin of the strand-biased A-to-G component of the Ig SHM spectrum.

The second component of the genomic SNP signature, C-to-T >> G-to-A, is most likely via a DNA-based mechanism as discussed by Polak and Arndt [35]. C-to-T transitions exceeding G-to-A transitions are indicative of strand-biased deamination events at the DNA level around promoter regions and CpG islands, targeting ssDNA in the displaced non-transcribed strand of the transcription bubble (^{5me}C-to-T and C-to-U).

4. Implications of the work of Mattick and associates in the light of the Polak and Arndt genome-wide SNP analyses

Before we proceed to a more formal analysis relating somatic gene diversification processes focused on rearranged Ig genes (1-2 Kb scale) to the Mb scale of the genome, it is important to discuss the work of Mattick and associates [38,39]. They established two important facts about the higher mammalian genome:

(a). Only ~ 2% of the genomic DNA is transcribed into classical mRNAs

which translate into proteins ie. protein coding genes, and,

(b). Up to 98% or more of the entire genomic DNA is transcribed into short and long non-coding RNAs. This is suggestive of a multi-layered over-lapping mosaic pattern of transcription, interpreted by Mattick et al as indicative of a yet to be defined universe of RNA regulatory networks regulating expression of protein coding genes.

It is possible to draw an important implication from this work: If >98% of the genome is transcribed and if the major genome wide strand-biased SNP signature is of A-to-G exceeding T-to-C in transcribed regions [35] this could suggest reverse transcriptase-mediated fixation of RNA editing mutations [36,37 and below] across the genome. A controversial implication from these propositions therefore is that perhaps ≥98% of the human genomic DNA sequence may have passed through an RNA intermediate at some point during evolution.

We realise this conclusion will remain controversial in the absence of direct evidence. However it is one *likely* inference and a key point of departure from conventional DNA-based thinking about the origins of genomic DNA sequence diversity.

5. "Microgenomic Diversification": 0rigin of somatic mutations in rearranged immunoglobulin variable genes

Mutational spray events do in fact occur in real-time in one important biological system, namely, during somatic diversification of Ig genes during an immune response. Antigen-driven somatic hypermutation (SHM) of rearranged immunoglobulin variable genes (so called VDJs) causes somatic point mutations to accrue at frequencies from 1-10 point mutations per 100 bp over a short time period (5-10 days). These mutations are focused on a 1-2 Kb region targeting coding VDJ genes and intronic flanks in Germinal Center B lymphocytes (short indels comprise maybe 2-3% of all somatic mutational events in VDJ genes even in non-coding intronic regions). This is coupled to antigen-binding selection to ensure that mutated B cells bearing surface Ig antigen receptors with similar or

better binding affinity for antigen survive, proliferate and then become part of the memory B cell pool [40]. The overall process is beneficial to the organism; it is based on intense Darwinian selection involving receptor-ligand binding [40].

In the case of SHM we have shown that there are now several lines of independent evidence pointing to a role for the Ig mRNA template acting as an intermediate to guide the genesis of the two main strand biased somatic mutation signatures at A:T and G:C base pairs [36]. From the data analysed these nucleotide substitutions are interpreted to appear first as RNA modifications which must then become fixed in the B lymphocyte DNA by a cellular (i.e. non-viral) reverse transcription step [41].

The first is the prominent strand bias at A:T base pairs whereby there is a significant excess of mutations from A compared with mutations from T, in particular A-to-G exceeding T-to-C, by up to three fold. In the case of the A-to-G *vs* T-to-C strand bias there exists a strong and specific Pearson correlation ($P<0.002$) modeled on the molecular requirements for adenosine-to-inosine (A-to-I) pre-mRNA editing mediated by the transcription-coupled ADAR1 deaminase acting on W<u>A</u>-sites in the context of a dsRNA stem loop [37] (where W = A or T/U).

The second is the strand bias whereby mutations from G exceed mutations from C by at least 1.7 fold ($P<0.001$). This is a newly identified SHM strand bias which has hitherto gone undetected because it has been masked by the presence of strand bias-suppressing PCR-hybrids, or recombinant DNA molecules, which have contaminated many SHM data sets [36]. When allowance is made for such artifacts and analyses performed only on data sets either completely free of them or where their level is minimized, the G>>C strand bias is apparent [36]. It is consistent with the misincorporation signature of RNA polymerase II copying the template DNA strand carrying lesions such as uracils (U) or abasic sites [42] generating biases G-to-C/C-to-G = 2.4x and G-to-A/C-to-T = 1.5x as RNA Pol II inserts C opposite an abasic site and A opposite template U [42]. Uracil and abasic site DNA lesions are the hallmarks of the activation-induced cytosine deaminase (AID) converting C-to-U in single stranded regions of DNA and thus activating the sequelae of aberrant ('error-prone') DNA repair enzymes which

triggers both somatic hypermutation and Ig class switch recombination [43-45].

In somatic hypermutation therefore we have a picture whereby a mutational spray of point mutations are introduced into a DNA region of perhaps 1-2 Kb at the 5' end of rearranged VDJ genes, distributed from the TSS, peaking over the VDJ and tailing off into the J-C intron region downstream of the VDJ [36]. If this mutational spray happens to improve the antigen-binding affinity of the mutated antibody protein then the B cell will be selected for survival for both antibody production in the periphery and sequestered into the memory B lymphocyte compartment [40].

In a recent analysis we have shown that the somatic mutation patterns of some well characterised non-lymphoid cancer genomes (lung carcinomas, breast carcinomas and squamous cell carcinomas) strongly resemble *in toto* or in part the strand biased spectra of somatic point mutations observed in normal physiological SHM in antibody VDJ genes [46]. Once again, as already discussed, these striking strand-biased mutation spectra are best understood as occurring first in RNA molecules which are then copied back into DNA. It is most likely that this occurs by a cellular reverse transcription (RT) process [36,37,41] carried out by the sole error-prone DNA polymerase known to be involved in SHM, DNA polymerase–η (eta) [47,48] which has been shown by *in vitro* experiments to be an efficient reverse transcriptase [49]. The significance of these findings is that SHM-induced strand biased mutation signatures can be potentially generated in non-Ig loci across the genome in many different genes expressed in different tissue types.

A summary of how the major strand biased mutation signatures are most likely generated in SHM is shown in Figure 1. By extension similar RNA-based mutator processes could occur in many other non-Ig protein coding genes during aberrant regulation of the SHM machinery [46].

Is it conceivable that these mutation processes may also occur across the wider genome over evolutionary time? At present there is no evidence for such processes. However given the biological precedence of the somatic mutation processes for Ig and non-Ig somatically expressed genes in non-lymphoid cancer tissues just described, we advance the possibility that such RNA-based

diversification processes *may* take place under certain conditions coincident with meiosis in mammalian germ cells, for example during the genesis of a new block haplotype sequence (below)

6. RT-mediated long DNA conversion, Transcription Factories and the genesis of Polymorphic Frozen Blocks

Here we offer an outline of a plausible hypothesis which has been arrived at by merging two different molecular approaches to genetics: (1) The work of Dawkins and associates [5,12,13,24,25] at the genomic level concerning the particular genetic features, and thus questions on the origin of, Polymorphic Frozen Block haplotypes, as just discussed, and (2) The accumulated knowledge gained at the 'microgenomic' level from the work of Steele and associates [36,37,41,46,49] and others [43-45,47,48] on the molecular mechanism of SHM of rearranged immunoglobulin V(D)J genes expressed in Germinal Center B lymphocytes.

It is our considered view that the spray of point mutations observed during somatic hypermutation of antibody V genes over 1-2 Kb represents a highly specialized, regulated and adapted process typical of wider SNP generation in the genome. Indeed the analysis of Polak and Arndt [35] is consistent with this view. We therefore propose that a similar spray of SNPs, and maybe short/long indels also takes place via RNA intermediates during the genesis of a polymorphic frozen ancestral block haplotype. The main implication of this as a new hypothesis is that all the really significant genetic mutations and recombination events, *do not occur first* at the DNA level: *they occur first* as multiple recombination events between base-modified RNA molecules which are then copied back into the genomic DNA.

This model is also consistent with the hypothesis advanced by Mattick and associates that most important genetic regulatory action in higher cells does not necessarily occur by specific interactions by proteins with DNA or RNA, or DNA molecules interacting with DNA, but at the level of long and short RNA regulatory molecules communicating with other RNA molecules as well as with other DNA

and protein assemblies [38,39,50].

To be more specific, a large RNA recombinant string is formed which is believed to be A-to-I edited by ADAR enzymes in the nucleus [51,52] and may carry other RNA base modifications (and indels as well as RNA duplications resulting in CNVs). This step is followed by a highly processive reverse transcriptase step to copy the inosine-containing long recombinant RNA (Figure 2). The RT-priming step would be as envisaged for SHM (Figure 1 and ref. [36]) such that the nicked transcribed strand (TS) DNA with a free 3'-OH end anneals to the long modified RNA thus allowing extension of the cDNA to produce a long newly synthesized transcribed strand with all the RNA mutations now embodied within the DNA strand as SNPs (or indels etc). The last steps would involve strand invasion, endonuclease action to remove the displaced resident strand 'flap' and then integration to seal the gap on the TS (via ligation). These events could happen during gametogenesis and meiosis and manifest as a biased or directional DNA conversion tract from one parental chromosome to another. The processes involved may also alter the structure and position of recombination hotspot motifs, such as the *PRDM9* gene and relocate them to the boundaries of the PFB and thus minimizing recombination within the newly formed PFB [26-30]. That is, the donor strand low in PRDM9 motifs would invade and convert the target strand to create a tract of low density PRDM9 motifs.

Thus, given a stress-induced mutagenic episode in evolution (eg. retroviral infection [5]) it is suggested that many of the critical DNA sequence diversifying events occur first at the RNA level which are then copied back into DNA by reverse transcription in a process we term "RT-mediated Long DNA Conversion", RT-LDC (Figure 2).

It would be a distinct advantage for the hypothesis if one could point to known molecular processes within the nucleus that brought distant regions of the genome physically close together in one sub-nuclear location. It turns out that current cell biology studies reveal highly ordered chromosomal structures at the level of transcription. These are identified as "Transcription Factories" (TFs) by Cook and associates who first discovered them [53-56]. In Table 1 we list questions suggesting experiments at the well studied MHC locus to test

predictions our hypothesis.

One possibility is that all the RNA-linked processive activities discussed above may be linked to a special Transcription Factory (TF) that has a wide functional agenda e.g. the MHC locus. In one version of the hypothesis the specialized TF coordinates the synthesis of different RNAs responsible for co-ordinating a "functional interacting cascade" of mutational pathways involving RNA intermediates. The TF also regulates quality control (DNA repair) and context (recombination, duplication) of the manufacturing program (transcription/RT cycle). At one extreme genes being expressed in a specific PFB would associate within their own TF.

This operating concept can be tested experimentally eg. by observing if functionally related genes which are often under *cis*-regulation are expressed in their own highly specialised TF. That is, our model predicts that a process requiring extreme somatic diversity, as in immunoglobulin somatic hypermutation, will involve its own specialised TF. It would not need to always be a *cis*-interaction, as there is compelling evidence that the expressed *Myc* proto-oncogene on Chromosome 15 preferentially relocates to the *same* Transcription Factory as the highly transcribed IgH gene located on Chromosome 12 [57]. It is conceivable therefore that oncogenic *cMyc* translocations at expressed rearranged IgH loci may involve an RNA intermediate and not a straight forward DNA-DNA recombination interaction as implied in the current translocation paradigm. Further, Cook's Transcription Factory model for genome organization suggests a role for specialized TFs in homologous chromosome pairing in mitosis and meiosis: the physical lining up of homologues prior to the formation of a base paired-mediated DNA crossover is brought about by binding interactions between transcription factors, promoters and RNA polymerases in the factories which mediate the pairing [58].

In our hypothesis a simple prediction is that a single PFB associating with a specific TF will not always apply to *trans*-regulation. Many functionally related genes are not located near each other and can therefore not be under *cis*-regulatory control in the conventional sense of a bacterial polycistronic array. Good examples are the key complement proteins controlled by the C4 gene

(within MHC on Chr 6), the C3 gene (Chr 19), the C5 gene (Chr 9) and the Regulators of Complement Activation (RCA, Chr 1). In these examples it is conceivable that the chromosomal loops from multiple different chromosomes may be located in the same TF for coordinated expression eg. complement component activation and control. Such conditions may favor DNA-DNA or RNA-RNA duplication and/or recombination events which could result in *trans* rearrangement chromosome events.

Within the environs of the TF, or nearby, the following molecular events are envisaged to take place: synthesis of pre-mRNA by RNA Pol II; errors in RNA due to copying C-to-U deaminations or abasic sites in template DNA; A-to-I pre-mRNA editing; pre-mRNA splicing as well as aberrant RNA splicing resulting in RNA-RNA recombinations; conventional DNA-DNA recombinations events since widely disparate chromosomal looped regions can be brought into close proximity [56], and Figure 3; conventional transcription-coupled (TCR) DNA repair processes; and finally, processive non-viral cellular reverse transcription (below). Retroduplication and retrotransposition events could be a molecular outcome of such events [5,59].

What conditions favor such long read-through transcripts and thus reverse transcription events? (Figure 3) Conserved long RNAs have been established by Lander and associates for long intergenic non-coding RNAs (lincRNAs, of length 2-17 Kb) implying that such long transcripts are the norm and thus possible [60]. However we also know that certain very large poly-exon/intron genes such as the major muscle protein *bungy* encoded by the *Titin* gene is about 250 Kb in length, so long transcription events are also possible. We might therefore assume that long reverse transcripts are also possible, and, with the aid of accessory 'clamp' proteins, are highly processive *in vivo*.

There is another problem in creating and conserving a PFB: how could a cluster of different genes often with opposing transcriptional polarities be *cis*-regulated or *cis*-inherited as a block? A good example is the gamma block of the MHC (e.g. Fig 7 ref [5]). *Cis*-regulation in this situation may involve the hypothesized cluster specific promoter controlling the synthesis of a long cluster specific transcript (Figure 3) for the entire block structure (e.g. ≥500 Kb). This

long transcript would then be reverse transcribed to lock in the DNA conversion tract (Figure 2).

This leads to a further prediction. The RT-LDC process will preserve pre-existing SNP strand asymmetries as established by Polak and Arndt [35]. But more importantly, if the new spray of SNPs is significant it will also superimpose this new strand asymmetry as an overlay on the target sequence. This means comparative sequence analyses may reveal superimposed strand biases, particularly in the case of multiple opposing transcription polarities (Figure 3). It may also reveal local inversions of the A-to-G >> T-to-C SNP ratio.

Expanding the reasoning to a specific example of gene duplication within a frozen block e.g C4B and C4A within the gamma block of MHC (Fig. 7 ref [5]) a comparative SNP analysis may be able to be used to reconstruct the temporal order of sprays of SNPs laid down between closely related duplicates.

7. Literature reports consistent with the RT-LDC hypothesis

Recent analyses of *GC*-biased gene conversion (gBGC) tracts in many eukaryotic genomes including mammals and humans [61] are consistent with our hypothesis. These DNA sequence tracts are indicative of the A-to-G >> T-to-C strand biased genomic SNP signatures [35] and the gBGC tracts correlate strongly with recombination frequency [61]. This suggests, that such tracts could involve both error-prone processive polynucleotide polymerisation followed by strand invasion and recombination (Figure 2).

More recently it has been shown that SNP sprays, even small ones, are non-randomly distributed in the genome [30] which is also consistent with our hypothesis. Indeed the *Lac1* Big Blue mouse transgenic model of spontaneous mutation shows that nonrandom sprays of mutations are the rule not the exception [31,32]. And there are recent other reports of mutational sprays in the human genome involving copy number variations (CNVs) [33,34]. We posit that such clustering of mutations is at least consistent with a *processive* polynucleotide synthetic process.

A plausible explanation for non-random SNP sprays put forward by Amos and

not inconsistent with the present argument, invokes the idea that pre-existing polymorphic sites, particularly in heterozygotes, act as foci targeting error-prone gene conversion events causing sprays of SNPs to occur near pre-existing SNPs and indels [30].

An interesting example in this regard concerns the emergence of the chondrodysplastic (short legs) haplotype in domestic dogs [62]. Most of the 50 SNPs in the breed-defining homozygous 24 Kb tract on Chr18 (Figure. 4) are "wild-type' and appear to have existed in the dog genome before the 24 Kb haplotype was created. Seven SNPs are new, in order they are T-to-G, A-to-T, G-to-A , G-to-A, G-to-A, G-to-A and A-to-G. The data suggest a SNP spray size of 7 in 24,000 nucleotides (1 in 3429). This may reflect the SNP burst size expected in the creation of a new long haplotype (although the length over which the main SNP spray has occurred suggests the real burst size frequency may be higher than this). However clustering is evident (Figure 4) around the insertion site of the translocated fully processed retrogene *Fgf4* into the 3' end of the LINE element. Amongst these 7 SNPs the mutations are mainly off A or off G. Most are therefore consistent with the idea of being "RNA mutations" by the criteria already developed for base substitutions in immunoglobulin somatic hypermutation: the G-to-As could have arisen via RNA Pol II synthesis off a DNA template carrying Uracil lesions; the A-to-G could have arisen by A-to-I RNA editing by ADAR1, and the A-to-T could have arisen at the reverse transcriptase step ([36] and Figure 1).

Thus the emergence of this newly formed long 24 Kb haplotype could be interpreted as being consistent with significant RNA recombination taking place between short (*Fgf4* retrogene insert) and long haplotype-specific RNAs regulating, and thus physically marking, this block of genes on Chr18. Such a long read-through RNA, encompassing coding and non-coding regions would mark the boundaries of the haplotypic region. Under the hypothesis advanced here such a long mutated RNA transcript would then be reverse transcribed and integrated into the genomic DNA. Intense human-directed breeding selection for short legged dogs has ensured homozygosity at this locus.

8. Candidate cellular reverse transcriptases

There are several possible sources of cellular reverse transcriptases that may mediate the long processive RT-step. These have been identified in three different genetic situations. We discuss each type as a possible source of cellular reverse transcription. At this stage we do not favour any particular cellular RT although some can be ruled out.

1. LINE reverse transcriptases. The retrotransposon encoded RTs identified by Spadafora and associates provide strong evidence for their involvement in the genesis of some completely new DNA sequences. In a series of innovative studies they have established sperm-mediated uptake of foreign DNAs and RNAs and identified LINE-1 encoded reverse transcriptases as being involved in converting the absorbed RNAs back to cDNA. Some is integrated into the genomic DNA, with the majority being transmitted to (and potentially expressed) in progeny as extrachromosal episomal DNA elements. Spadafora concludes that "RT-mediated machinery operates in sperm cells and is responsible for the genesis and non-Mendelian propagation of new genetic information" [63]. Indeed the role of RT activity from endogenous and exogeneous transposable elements (TEs) shaping genomic diversity has recently been reviewed in the context of RNA-based gene duplications [59] as well as episodic surges in TE activity that could be an explanation for punctuated equilibrium as observed in the paleontological record [64].

2. DNA Polymerse-γ. The high fidelity mitochondrial-associated RT, DNA Polymerase-γ, was identified by Anderson and associates. The implications are not yet widely appreciated. They have shown that biologically significant RT activity of DNA Polymerse-γ, a high fidelity proof-reading RT encoded in the nuclear genome and used to replicate the circular mitochondrial genome which is synthesized via an RNA intermediate [65]. Aberrant control of the activity or substrate specificity of this DNA polymerase may cause it to be deployed in non-mitochondrial reverse transcription as envisaged in the hypothesized genomic RT-LDC process.

3. Y Family Translesion DNA Polymerases. The significant RT activity identified in the Y family of DNA translesion repair polymerases by Steele and associates, DNA Pol-η, DNA Pol-k, and DNA Pol-i [49]. DNA polymerase eta (-η) is the sole error-prone DNA polymerase involved in somatic hypermutation [48] and is thus the most likely cellular RT involved in the fixing of the RNA mutation patterns in the DNA of hypermutating B lymphocytes [36,37,49]. In isolated systems *in vitro* the processivity of DNA Pol-η is thought to be low but the role of sliding clamps (e.g. PCNA, proliferating cell nuclear antigen) in its RT mode [49] has not been investigated either *in vitro* or *in vivo*. In anycase, as already discussed,"processivity" is a biochemical concept describing the affinity of a DNA polymerase for its template determined in isolated *in vitro* systems. These conditions are far removed from the supramolecular protein complexes mediating complex DNA and RNA interactions and synthetic events *in vivo* which would be expected within the environs of a "Polynucleotide Synthetic Factory". Cook repeatedly reminds us that DNA copying templates *in vivo* are *reeled past* the fixed polymerases and not *vice versa* (the polymerase is popularly thought of as tracking along the template as expected to occur in an *in vitro* PCR reaction [56].

It should be noted also that Pol-η has the strand invasion and homologous recombination properties [66-68] necessary both for SHM (Figure 1) and our RT-LDC hypothesis for the genesis of PFBs (Figure 2).

In another context the concepts of A-to-I RNA editing, known to be widespread in the human transcriptome [51,52] and reverse transcription (DNA re-coding) has been deemed necessary for a better understanding of brain mechanisms of neural transmission and long-term memory and higher-order cognition [50].

4. Telomerases and origin of short repeat DNA sequences We include discussion of Telomerases in this section because they were the first non-viral cellular reverse transcriptases to be discovered by Blackburn, Grieder and associates [69]. Indeed the telomerase RNA moiety provides a short template sequence on which the telomerase enzyme has evolved the ability to copy in a

reiterative fashion [70-72]. The repeat sequence (TTAGGG) is generated by successive cycles of polymerisation to the end of the template and then a translocation and repositioning of the telomerase complex over the RNA template to synthesise a new cDNA copy contiguously joined to the previous repeat.

Recent evolutionary analysis now suggest that centromere repeat structures have been derived from telomeres during the evolution of eukaryotic chromosomes [73]. Telomerases however are also known to be highly specialised reverse transcriptases (TERTs) which function within the context of a ribonucleic protein particle. It seems unlikely that there is a primary role for TERTS in the RT-LDC hypothesis. However disregulated or aberrant TERT particles may generate tandem repeats (microsatellites) or even dispersed repeats akin to retrotransposition events at sites of single or double strand DNA breaks within chromosome arms and thus causing indel [25] or CNV sprays [33] directly in genomic DNA during mutagenic episodes in evolution.

9. Potential criticism of the RT-LDC hypothesis

The analysis thus far has depended on our interpretation of the genomic strand biased SNP signature in transcribed regions. The process involves A-to-G substitutions exceeding T-to-C [35], and it involves wide spread A-to-I editing of the transcriptome [51,52] with intermittent DNA re-coding or reverse transcriptase-mediated fixing of the SNPs into the DNA.

Authors such as Polak and Arndt [35] rely on a more conservative interpretation of this type of strand bias. In their view it is mediated as a consequence of 'normal strand-asymmetrical processes' of transcription-coupled repair (TCR) at the transcription bubble [74]. This is the current mainstream view in the TCR field. We have critically re-evaluated the recent TCR literature beginning with the authoritative review of the field by Hanawalt and Spivak [75]. In this review a key paper is cited on the clearance of lesions from the transcribed strand (TS) in mutations in the Chinese hamster HPRT gene [76]. We also evaluated papers cited by Polak and Arndt [35] such as Green et al [74] and the work on TCR-mediated asymmetry in the A-to-G/T-to-C ratios they cite viz.

Jiricny 1998 [77].

Whilst the mechanism for TCR in bacteria is agreed this is not the case with eukaryotic systems. Indeed ".. eukaryotic cell-free systems have failed to fully validate *in vivo* TCR observations … and this has hindered detailed biochemical analysis." [75].

In Fig 3 in Hanawalt and Spivak [75] various optional mechanisms are described as possible outcomes following RNA Pol II arrest at a lesion on the transcribed strand (TS). Most informative, for the present analysis, is option b in their Fig 3 " .. for some lesions, translesion transcription is possible but might result in transcriptional mutagenesis". This tacitly could imply DNA re-coding of RNA mutations, of the type observed when RNA Pol II copies a DNA template with AID-type lesions, Uracils and Abasic sites [36].

Further, in the view of Hanawalt and Spivak … " The most important function of (conventional) TCR is probably to remove obstructions to RNAP translocation rather than simply to repair expressed genes more rapidly". They then go on to say that the operation of TCR can result in strand bias of mutagenesis and they cite Vrieling et al [76] on the clearance of photoproduct lesions from the TS in mutations in the Chinese hamster HPRT gene (following exposure of CHO cells to moderate to low doses of UV). In this study cyclobutane pyrimidine dimers are removed from the TS following 4-8 hrs incubation and most point mutations in recovered mutant HPRT cells are found to be located on the non-transcribed strand (NTS). In our opinion the number of point mutations in this study are low (n ~ 20). Vrieling et al conclude "There did not seem to be a preference for a specific type of change, although transversions of GC base pairs were underrepresented". This is precisely the outcome to be expected if a lesion-free TS is re-synthesized. One would expect *all* types of mutations to be seen on the NTS (not just, for example, A-to-G transitions). This is an important point about conventional TCR outcomes removing bulky lesions or adducts (see Fig 3 options a, c and d in ref [75]). However, for an *unconventional TCR outcome* we need to consider option b in Fig 3. of Hanawalt and Spivak [75] as this fits the A-to-I RNA editing model coupled to a reverse transcription step fixing the strand biased RNA mutation pattern in the newly synthesized DNA of the TS. It is

consistent with the work of Kuraoka et al [42] which clearly shows RNA Pol II does not stall at minor lesions such as Uracils and Abasic sites. It simply copies over them incorporating signature mutations in the newly synthesized RNA.

In Polak an Arndt [35] an important reference is made to Green et al [74] as providing support for a conventional TCR explanation of the genomic strand-specific A-to-G/T-to-C ratio. Thus Green et al [74] bolster their case for a conventional TCR explanation by claiming the following: " Moreover, the fact that the strongest asymmetry occurs for A-to-G transitions, which in this model would result from the resolution of G-T mispairs arising from misinserted G, is consistent with the observation that MutSα is particularly efficient at recognizing G-T mispair (Jiricny, 1998)".

In contrast what Jiricny [77] *actually concluded* : "..Thus the take-home message from the binding studies is that affinity of the protein for a particular mispair or a DNA modification *in vitro* cannot be taken as an indication of repair efficiency *in vivo*." In other words, Jiricny backs away from the conclusions drawn by Green et al. We believe that this imisrepresentation by Green et al may be replicated by others in the TCR field. The Hanawalt and Spivak review in 2008 [75] is more comprehensive and they have incorporated option b in their Fig 3 as part of a plausible explanation of some types of "TCR-like" strand-biased data.

We conclude from our detailed review of the TCR literature that the data of the TCR field are entirely consistent with our RT-LDC hypothesis.

10. Concluding remarks on RNA intermediates and the preservation of frozen blocks

The molecular processes discussed in this paper have been limited to *how* initial genome-wide single nucleotide diversity *may* be generated via RNA template intermediates. It does not take into account more recent sophisticated population genetic theories on the types of selection forces that could maintain polymorphic haplotype blocks within the MHC and beyond [78]. Nor does it consider other known more complex mechanisms involved in genome-wide re-arrangement involving RNA intermediates (the existence of which suggests we

may not be being bold enough in our speculations in the current analyses cf. ref [79]). The process of genome-wide DNA re-arrangement is known to occur in many higher animals (as well as single cell animals) during mitosis and meiosis - with a whole range of quite perplexing mechanisms now being discovered. A key point is that the process of updating and reassembly of the genome in such wholesale re-arrangements relies on precisely the same locations in chromosomes. Initial evidence suggests that RNAs are probably responsible for caching and guiding the reassembly-and conserving the integrity of large stable regions of the genome during re-arrangement. RNA-mediated epigenetic re-programming is also involved in some complex genome re-arrangement pathways. These ideas are underlined by the work of Nowacki and associates on the ciliate genome *Oxytricha trifallax* [80]. During development of the somatic macronucleus, 95% of its germline DNA is fragmented and the organism then unscrambles hundreds of thousands of fragments by permutation or inversion in the chromosome reassembly process.

With respect to frozen blocks a possible conventional mechanism of haplotype conservation could involve normal DNA repair mechanisms associated with DNA-DNA recombination and may play an important role in PFB conservation. Thus recombination repair may more easily, and thus more frequently, occur between sister chromatids (from the same PFB) which of course will result in sequence conservation within the PFB for that chromosome.

The other paradox of the genomic diversity field is the clear conservation of an Ancestral Haplotype over hundreds if not thousands of breeding generations. eg. MHC haplotype 8.1 occurs in many human populations at very respectable frequencies ≥1% [5]. It is possible that most of the time the original long RNA delineating the haplotype is not mutated (or the "mutators' such as ADAR deaminases are switched off or quarantined) leading to any long RNA-mediated DNA conversion event replacing unmutated with unmutated i.e. this is part of the molecular maintenance mechanism.

However the genetic mechanisms responsible for conserving a PFB structure once generated remains a mystery. It may appear to be “non-Darwinian” in the simple sense that many genes are conserved as a block irrespective of whether

some genes predispose to significant life threatening disease [5,6]. Such conservation is of course "Darwinian" if selection forces preserve the "bloc" as a functional unit. In this regard one recent model on the evolution of the MHC is that deleterious recessive mutations could accumulate as a "sheltered load" near MHC genes ('hitch-hikers') and they become common as they are rarely expressed as homozygotes; this could be coupled with inefficient purifying selection and low recombination rates [78]. As recognized many years ago presumably such a block is a “genetic compromise” as the beneficial nature of the sum total of the genetic expressions within a block is compatible with life (Hill-Robertson Effect, [81]).

Thus in acknowledging that mechanisms be responsible for maintaining the integrity of the structure of PFBs over evolutionary time frames, we also posit a testable RT-LDC hypothesis to explain the generation of polymorphisms associated with mutational sprays. Future research will no doubt shed more light on the complex regulatory mechanisms involved in both cases.

References

[1] McCluskey J, Kay PH, Stuckey M, Christiansen FT, Dawkins RL, Wilson G MHC "supratype" predicting heterozygous 21-hydroxylase deficiency. Lancet. 1983 Apr 2;1(8327):764-5.

[2] Dawkins RL, Christiansen FT, Kay PH, Garlepp M, McCluskey J, Hollingsworth PN, Zilko PJ. Disease associations with complotypes, supratypes and haplotypes. Immunol Rev. 1983;70:1-22.

[3] Awdeh ZI, Raum D, Yunis EY, Alper CA Extended HLA/complement allele haplotypes. Proc. Natl. Acad. Sci. USA 1983; 80: 259-263.

[4] Alper CA, Awdeh ZL, Yunis EJ. Conserved extended MHC haplotypes. Exp.Clin Immunogenetics 1992; 9:58-71.

[5] Dawkins R, Leelayuwat C, Gaudieri S, Tay G, Hui J, Cattley S, Martinez, Kulski J. Genomics of the major histocompatibility complex: haplotypes, duplication, retroviruses and disease. Immunol Rev 1999; 167: 275-304.

[6] Yunis EJ, Larsen CE, Ferndez-Vina M, Awdeh ZL, Romero T, Hansen JA, Alper CA. Inheritable variable sizes of DNA stretches in the human MHC : conserved extended haplotypes and their fragments or blocks. Tissue Antigens 2003; 62:1-20.

[7] Smith WP, Vu Q, Li SS, Hansen JA, Zhao LP, Geraghty DE. Toward understanding MHC disease associations: partial resequencing of 46 distinct HLA haplotypes. Genomics 2006; 87:561-571.

[8] Jeffreys AJ, Kauppi L, Neumann R. Intensely punctate meiotic recombination in the class II region of the major histocompatibility complex. Nat. Genet. 2001; 29: 217-222.

[9] McLure CA, Williamson JF, Smyth LA, Agrawal S, Lester S, Millman JA, Keating PJ, Stewart BJ, Dawkins RL. Extensive genomic and functional polymorphism of the complement control proteins. Immunogenetics 2005; 57:805-815.

[10] Williamson JF, McLure CA, Baird PN, Male D, Millman J, Lawley B, Ashdown ML, Keating PJ, Dawkins RL. Novel sequence elements define ancestral haplotypes of the region encompassing complement factor H. Human Immunology 2008; 69:207-219.

[11] McLure CA, Kesners, PW, Lester S, male D, Amadou C, Dawkins JR, Stewart BJ, Williamson JF, Dawkins RL. Haplotyping of the canine MHC without the need for DLA typing. J Immunogenetics 2005; 32: 407-411.

[12] Tay GK, Witt CS, Christiansen FT, Charron D, Baker D, Herrmann R, Smith LK, Diepeveen D, Mallal S, McCluskey J, Lester S, Loiseau P, Teisserenc H, Chapman J, Tait B, Dawkins RL. Matching for MHC haplotypes results in improved survival following unrelated bone marrow transplantation. Bone Marrow Transplant. 1995; 15:381-385.

[13] Gaudieri S, Longman-Jacobsen N, Tay GK, Dawkins RL. Sequence analysis of the MHC Class I region reveals the basis of the Genomic Matching Technique. Human Immunology 2001; 62: 279-283.

[14] Lester S, McLure C, Williamson J, Bardy P, Rischmueller M, Dawkins RL. Epistasis between the MHC and the RCAα block in primary Sjogren syndrome. Ann. Rheum. Dis. 2008; 67: 849-854.

[15] Altshuler D, Brooks LD, Chakravarti A, Collins FS, Daly MJ, Donnelly P (Members of writing group: The International HapMap Consortium.) A haplotype map of the human genome. Nature 2005; 437: 1299-1320.

[16] Gabriel SB, Schaffner SF, Nguyen H, Moore JM, Roy J, Blumenstiel B, Higgins J, DeFelice M, Lochner A, Faggart M, Liu-Cordero SN, Rotimi C, Adeyemo A, Cooper R, Ward R, Lander ES, Daly MJ, Altshuler D. The structure of haplotype blocks in the human genome. Science 2002; 296: 2225-2229.

[17] Stumpf MPH. Haplotype diversity and the block structure of linkage disequilibrium. Trends in Genetics 2002; 18: 226-228.

[18] Wall JD, Pritchard JK. Haplotype blocks and linkage disequilibrium in the human genome. Nat. Rev. Genetics 2003; 4: 587-597.

[19] Goddard ME, Hayes BJ. Mapping genes for complex traits in domestic animals and their use in breeding programmes. Nat. Rev. Genetics 2009; 10: 381- 391.

[20] Hauser E, Cremer N, Hein R, Deshmukh H. Haplotype-based analysis : A summary of GAW16 group 4 analysis. Genetic Epidemiology 2009 ; 33 (Supp 1) : S24-S28.

[21] Laird R, Dawkins RL, Gaudieri S. Use of the genomic matching technique to complement multiplex STR profiling reduces DNA profiling costs in the high volume crimes and intellige3nec led screens. Forensic Sci Int. 2005; 151: 249- 257.

[22] Ge J, Budowle B, Planz JV, Chakraborty R. haplotype block: a new type of forensic DNA markers. Int J Legal Med. 2009; In press

[23] Guryev V, Smits BMG, van de Belt J, Verheul M, Hubner N, Cuppen E. Haplotype clock structure is conserved across mammals. PLoS Genetics 2006; July vol 2 e121: 1111-1118

[24] Gaudieri S, Dawkins RL, Habara K, Kulski JK, Gojobori T. SNP profile within the human major histocompatibility complex reveals and extreme

and interrupted level of nucleotide diversity. Genome Research 2000; 10: 1579-1586.

[25] Longman-Jacobsen N, Williamson JF, Dawkins RL, Gaudieri S. In polymorphic genomic regions indels cluster with nucleotide polymorphism: Quantum Genomics. Gene 2003; 312:257-261.

[26] Cheung VG, Sherman SL, Feingold E. Genetic control of hotspots. 2010 Science; 327:79-792.

[27] Baudat F, Buard J, Grey C, Fledel-Alon A, Ober C, Przeworski M, Coop G, de Massy B. PRDM9 is a major determinant of meiotic recombination hotspots in humans and mice. Science 2010; 327;836-840.

[28] Myers S, Bowden R, Tumian A, Bontrop RE, Freeman C, MacFie TS, McVean G, Donnelly P. Drive against hotspot motifs in primates implicates the PRDM9 gene in meiotic recombination. Science 2010; 327:876-877.

[29] Parvanov E, Petkov PM, Paigen K. Prdm9 controls activation of mammalian recombination hotspots. Science 2010;327:835.

[30] Amos W. Even small SNP clusters are non-randomly distributed: is this evidence of mutational non-independence? Proc Roy Soc Ser B 2010; 277:1443-1449.

[31] Drake JW. Mutations in clusters and showers. Proc. Natl Acad. Sci USA 2007; 104:8203-8204.

[32] Wang J, Gonzalez KD, Scaringe WA, Tsai K, et al. Evidence for mutation showers. Proc. Natl. Acad. Sci USA 2007; 104:8403- 8408.

[33] Singh SM, Castellani CA, O'Reilly RL. Copy number variation showers in schizophrenia: an emerging hypothesis. Mol. Psych. 2009; 14:356-358.

[34] Chen J-M, Ferec C. Cooper DN. Closely spaced multiple mutations as potential signatures of transient hypermutability in human genes. Human Mutation 2009; 30:1435-1448.

[35] Polak P, Arndt PF. Transcription induces strand-specific mutations at the 5' end of human genes. Genome Research 2008; 18: 1216-1223.

[36] Steele EJ. Mechanism of somatic hypermutation: critical analysis of strand biased mutation signatures at A:T and G:C base pairs. Molecular Immunology 2009; 46:305-320.

[37] Steele EJ, Lindley RA, Wen J, Weiller G. Computational analyses show A-to-G mutations correlate with nascent mRNA hairpins at somatic hypermutation hotspots. DNA Repair 2006; 5:1346-1363.

[38] Mattick JS. A new paradigm for developmental biology. J Exp. Biology 2007; 210: 1526-1547.

[39] Mattick JS, Maral PP, Dinger ME, Mercer TR, Mehler MF. RNA regulation of epigenetic processes. Bioessays 2009: 31: 51-59.

[40] MacLennan ICM. Germinal centers. Annu. Rev. Immunol. 1994; 12:117-139

[41] Steele EJ, Pollard JW. Hypothesis: somatic hypermutation by gene conversion via the error prone DNA-to-RNA-to-DNA information loop. Molecular Immunology 1987; 24:667-673.

[42] Kuraoka I, Endou M, Yamaguchi Y, Wada Y, Handa H, Tanaka K. Effects of endogenous DNA base lesions on transcription elongation by mammalian RNA polymerase II. J Biol. Chem. 2003; 278:7294-7299.

[43] Di Noia JM, Neuberger MS. Molecular mechanisms of antibody somatic hypermutation. Annu. Rev. Biochem. 2007; 76:1-22.

[44] Muramatsu M, Kinoshita K, Fagarasan S, Yamada S, Shinkai Y, Honjo T. Class switch recombination and hypermutation require activation-induced cytidine deaminase (AID), a potential RNA editing enzyme. Cell 2000; 102: 553-563.

[45] Wilson TM, Vaisman A, Martomo SA, Sullivan P, Lan L, Hanaoka F, Yasui A, Woodgate R, Gearhart PJ. MSH2-MSH6 stimulates DNA polymerase eta, suggesting a role for A:T mutations in antibody genes. J. Exp. Med. 2005; 201: 637-645.

[46] Steele EJ, Lindley RA. Somatic mutation patterns in non-lymphoid cancers resemble the strand biased somatic hypermutation spectra of antibody genes. DNA Repair 2010 ; 9:600-603.

[47]. Zeng, X, Winter DB, Kasmer C, Kraemer KH, Lehmann AR, Gearhart PJ. DNA polymerase η is an A-T mutator in somatic hypermutation of immunoglobulin variable genes, Nat. Immunol. 2001; 2 : 537-541.

[48] Delbos F, Aoufouchi S, Faili A, Weill J-C, reynaud C-A. DNA polymerase η is the sole contributor of A/T modifications during immunoglobulin hypermutation in the mouse.J Exp.Med,2007;204: 17-23.

[49] Franklin A, Milburn PJ, Blanden RV, Steele EJ. Human DNA polymerase-η, an A-T mutator in somatic hypermutation of rearranged immunoglobulin genes, is a reverse transcriptase. Immunol. Cell Biol. 2004; 82: 219-225.

[50] Mattick JS, Mehler MF. RNA editing, DNA recoding and the evolution of human cognition. Trends in Neuroscience 2008; 31:227-233.

[51] Levanon EY, Eisenberg E, Yelin R, Nemzer S, Halegger M, Shemesh R, Fligelman ZY, Shoshan A, Pollock SR, Sztybel D, Olshansky M, Rechavi, G, Jantsch, MF. Systematic identification of abundant A-to-I editing sites in the human transcriptome. Nature Biotechnology 2004: 22: 1001-1005.

[52] Athanasiadis A, Rich A, Mass S. Widespread A-to-I RNA editing of Alu-containing mRNAs in the human transcriptome. PLoS Biology 2004; 2: e391.

[53] Carter DRF, Eskiw C, Cook PR. Transcription Factories Biochem. Soc. Trans. 2008;36:585-589.

[54] Eskiw CH, Rapp A, Carter DRF, Cook PR. RNA polymerase II activity is located on the surface of protein-rich transcription factories. J Cell Sci 2008; 121:1999-2007.

[55] Xu M, Cook PR. Similar active genes cluster in specialized transcription factories. J Cell Biol. 2008; 181: 615-623.

[56] Cook PR. A model for all genomes: the role of transcription factories. J Mol Biol 2010; 395:1-10.

[57] Osborne CS, Chakalova L, Mitchell JA, Horton A, Wood AL, Bolland DJ, Corcoran AE, Fraser P. *Myc* dynamically and preferentially relocates to a Transcription Factory occupied by *Igh*. PloS Biology 2007; 5: issue 8 e192.

[58] Xu M, Cook PR. The role of specialized transcription factories in chromosome pairing. Biochim Biophys. Acta 2008; 1783: 2155- 2160.

[59] Kaessmann H, Vinckenbosch N, Long M. RNA-based gene duplication: mechanisms and evolutionary insights. Nat. Rev. Genet. 2009; 10:19- 31.

[60] Guttman M, Amit I, Garber M, French C, Lin MF, Feldser D, Huarte M, Zuk O, Caret BW, Cassady JP, Cabili MN, Jaenisch R, Mikkelsen TS, Jacks T, Hacohen N, Bernstein BF, Kellis M, Regev A, Rinn JL, Lander ES. Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. Nature 2009; 458: 223-227.

[61] Duret L, Galtier N. Biased gene conversion and the evolution of mammalian genomic landscapes. Annu. Rev. Genomics Hum.Genet. 2009; 10: 285-311.

[62] Parker HG, VonHoldt BM, Quignon P, Margulie EH, Shao, S, Mosher DS, Spady TC, Elkahloun A, Cargill M, Jones PG, Maslen CL, Acland GM, Sutter NB, Kuroki K, Bustamanti CD, Wayne RK, Ostrander EA. An expressed Fgf4 retrogene is associated with the breed-defining Chondrodysplasia in domestic dogs. Science 2009; 325:995-998.

[63] Spadafora C. Sperm-mediated 'reverse' gene transfer: a role of reverse transcriptase in the generation of new genetic information. Human Reproduction 2008; 23: 735-740.

[64] Oliver KR, Greene WK. Transposable elements: powerful facilitators of evolution. BioEssays 2009; 31:703-714.

[65] Murakami E, Feng JY, Lee H, Hanes J, Johnson KA, Anderson KS. Characterization of novel reverse transcriptase and other RNA-associated catalytic activities by human DNA polymerase- γ.J Biol. Chem. 2003; 278: 364013-36409.

[66] Rattray AJ, Strathern JN. Homologous recombination is promoted by translesion polymerase Pol η. Mol. Cell 2005; 20: 658-659.

[67] McIlwraith MJ, Vaisman A, Liu Y, Fanning E, et al. Human DNA polymerase eta promotes DNA synthesis from strand invasion intermediates of homologous recombination. Mol. Cell 2005; 20: 783-792.

[68] Kawamoto T, Araki K, Sonoda E, Yamashita YM, Harada K, et al. Dual roles for DNA polymerase eta in homologous DNA recombination and translesion DNA synthesis. Mol. Cell 2005; 20: 793-799.

[69] Blackburn EH. Telomerases. Annu. Rev. Biochem. 1992; 61: 113-129.

[70] Drosopoulos WC, DiRenzo R. Prasad VR. Human telomerase RNA template sequence is a determinant of telomere repeat extension rate. J. Biol. Chem. 2005; 280: 32801-32810.

[71] Collins K. Ciliate telomerase biochemistry. Annu. Rev. Biochem. 1999; 68: 187-218.

[72] Kelleher C, Teixeira MT, Forstemann K, Lingner J. Telomerase:biochemical considerations for enzyme and substrate. Trends Biochem. Sci 2002; 27:572-579.

[73] Villasante A, Abad JP, Mendez-Lago M. Centromeres were derived from telomeres during the evolution of the eukaryotic chromosome. Proc. Natl. Acad. Sc. USA 2007; 104:10542-10547.

[74] Green P, Ewing B, Miller W, Thomas PJ. Transcription-associated mutational asymmetry in mammalian evolution. Nature Genetics 2003;33:514-517.

[75] Hanawalt PC, Spivak G. Transcription-coupled DNA repair: two decades of progress and surprises. Nat Rev. Mol. Cell Biol 2008; 9:958-970.

[76] Vrieling H, Venema J, van Rooyen M-L, van Hoffen A, Menichini P, Zdzienicka MZ, Simons WIM, Mullenders LHF, van Zeeland AA. Strand specificity for UV-induced DNA repair and mutations in the Chinese hampster HPRT gene. Nucl. Acids. Res. 1991; 19:2411-2415.

[77] Jiricny J. Replication errors : cha(lle)nging the genome. EMBO J 1998; 17: 6427-6436.

[78] van Oosterhout C A new theory of MHC evolution: beyond selection on the immune genes. Proc R. Soc B 2009; 276 : 657-665.

[79] Steele EJ. Lamarck and immunity: somatic and germline evolution of antibody genes. J Roy Soc Western Australia 2009; 92:437-446.

[80] Nowacki M, Vijayan V, Zhou Y, Schotanus K, Doak TG, Landweber LF. RNA mediated epigenetic programming of a genome re-arrangement pathway. Nature 2007; 451 :153-158.

[81] Hill WG, Robertson A. The effect of linkage on limits to artificial selection. Genet. Res. 1966; 8:269-294.

Figure Legends

Figure 1. Explanations for strand biased mutation signatures in the antigen-driven somatic hypermutation of antibody genes. Adapted from Table 1a and Figure 5 in Steele 2009 [36]. All mutations are read from the non-transcribed 5' to 3' strand (NTS). Proportion of all mutations expressed as a percent of total and each value is the mean of 12 independent studies (standard error of the mean is in brackets) where the incidence of strand-biased blunting PCR hybrid artefacts are either non-existent or significantly minimized [36]. Highlighted in the base substitution table are the strand biases noted for mutation from A versus mutations from T (A>>T) and mutations from G versus mutations from C (G>>C). Thus for A >>T, adenosine-to-inosine (A-to-I) pre-mRNA editing by ADAR1 deaminase (for A-to-G) and then error prone reverse transcription (via DNA Pol-η) to generate A-to-T and A-to-C. For G>>C, the copying of DNA template carrying uracil and abasic site lesions (typical of AID deaminase) by RNA Pol II inserting G-to-A opposite template U and G-to-C opposite abasic sites [42] and then reverse transcription via DNA Pol-η. Thin black lines are DNA strands, thick black lines are mRNA, hatched thick lines are cDNA strands copied off mRNA. AID, activation induced cytidine deaminase, causes C-to-U deaminations in ssDNA regions. The question marks at the last steps indicate an unknown and indeterminant number of steps involving strand invasion, heteroduplex formation and/or resolution of heteroduplex and full length copying of newly synthesized transcribed strand. See Steele 2009 [36] for further details.

Figure 2. RT-Mediated Long DNA Conversion. See text and Figure 1 for more explanations and details. Thin black lines are DNA strands, thick black lines are mRNA, hatched thick lines are cDNA strands copied off mRNA.

Figure 3. A cluster specific promoter drives the synthesis of a long transcript in a transcription factory. Adapted from Cook and associates [53-56]. For details on the structure of Transcription Factories, see particularly references for the ~ 100nm dimensions of a protein rich factory [54,56]. Chromosomes are looped and anchored at the sites of RNA synthesis. The arrows shows the direction of transcription for that gene. The large hooked arrow denotes a cluster specific promoter driving the synthesis of a very long transcript (≥500 Kb).

Figure 4 The origin of the chondrodysplasia (short legged) haplotype in domestic dogs. From Parker et al 2009 [62]. SNP positions drawn approximately to scale. See text for further details.

Appendix : Detection of Strand Biased Mutation Signatures. In a data set containing a large number of somatic mutations or single nucleotide germ line polymorphisms (SNPs) strand biased base substitution signatures are revealed by comparing the base exchange frequencies of Watson-Crick complements on the *same* strand. By convention nucleotide substitutions are read from the non-transcribed strand (NTS). However the known direction of transcription in a region of genomic DNA allows identification of the strands. Thus, in the example, if A-to-G mutations occur with equal frequency on both strands, then its Watson-Crick complement, T-to-C will occur with equivalent frequency when scored off the same strand. However if there is a bias in the mutations favouring the NTS then A-to-G mutations will exceed T-to-C mutations. If there are systematic strand biases involving excessive mutations off A or G (e.g. as seen in Figure 1) then the sum total of mutations off A will exceed the sum total of mutations off T (at A:T base pairs where A>>T) and the sum total of mutations off G will exceed the sum total of mutations off C (at G:C base pairs where G>>C).

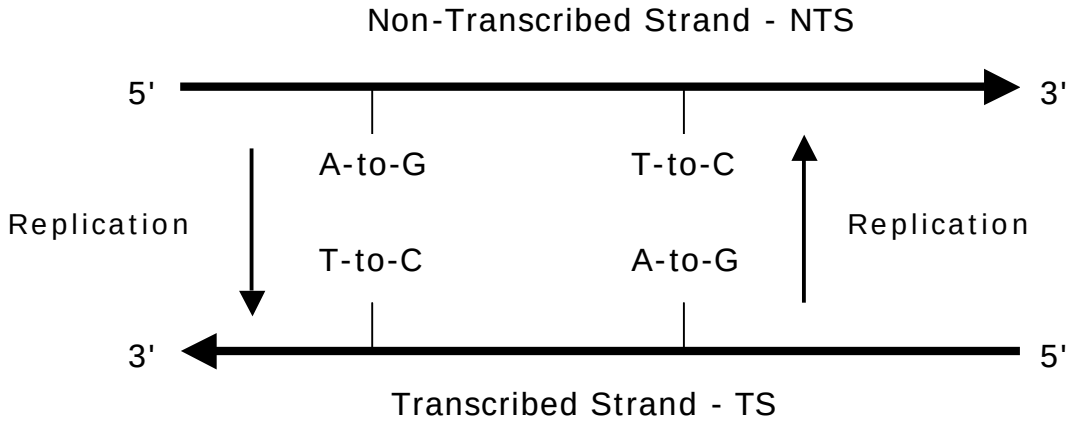

Table 1 Questions and Suggested Experiments on Transcription Factories (TF) and Polymorphic Frozen Blocks (PFB) in the MHC

• How many transcription factories are associated with the ~4Mb of the human MHC?

• How many TFs are associated with each PFB eg. the γ-block of the MHC?

• Do identifiable TFs correlate with known polymorphic frozen blocks?

• Do different MHC haplotypes have different TFs?

• Is the pattern of TFs, for say MHC, the same in somatic cells as in a germ cell? (e.g. male spermatogonia mother cell?)

• Where does genetic cross-over (recombination) occur in relation to MHC associated transcription factories - inside or outside block-specific TFs?

• Is there a difference between recombination sites (in relation to TFs) in "normal" versus "aberrant" physiological situations such as cancer?

• During an aberrant stress episode (eg retroviral infection), do TFs release their control of "transcriptional quality"? (ie. that would normally suppress RNA/DNA recombination, RNA editing, reverse transcription, aberrant RNA splicing, aberrant DNA repair etc).

• Given that TFs will exert "quality control of RNA transcripts", and possibly also regulate DNA recombination and DNA repair - what is the pattern of DNA repair activity in TFs in health and disease?

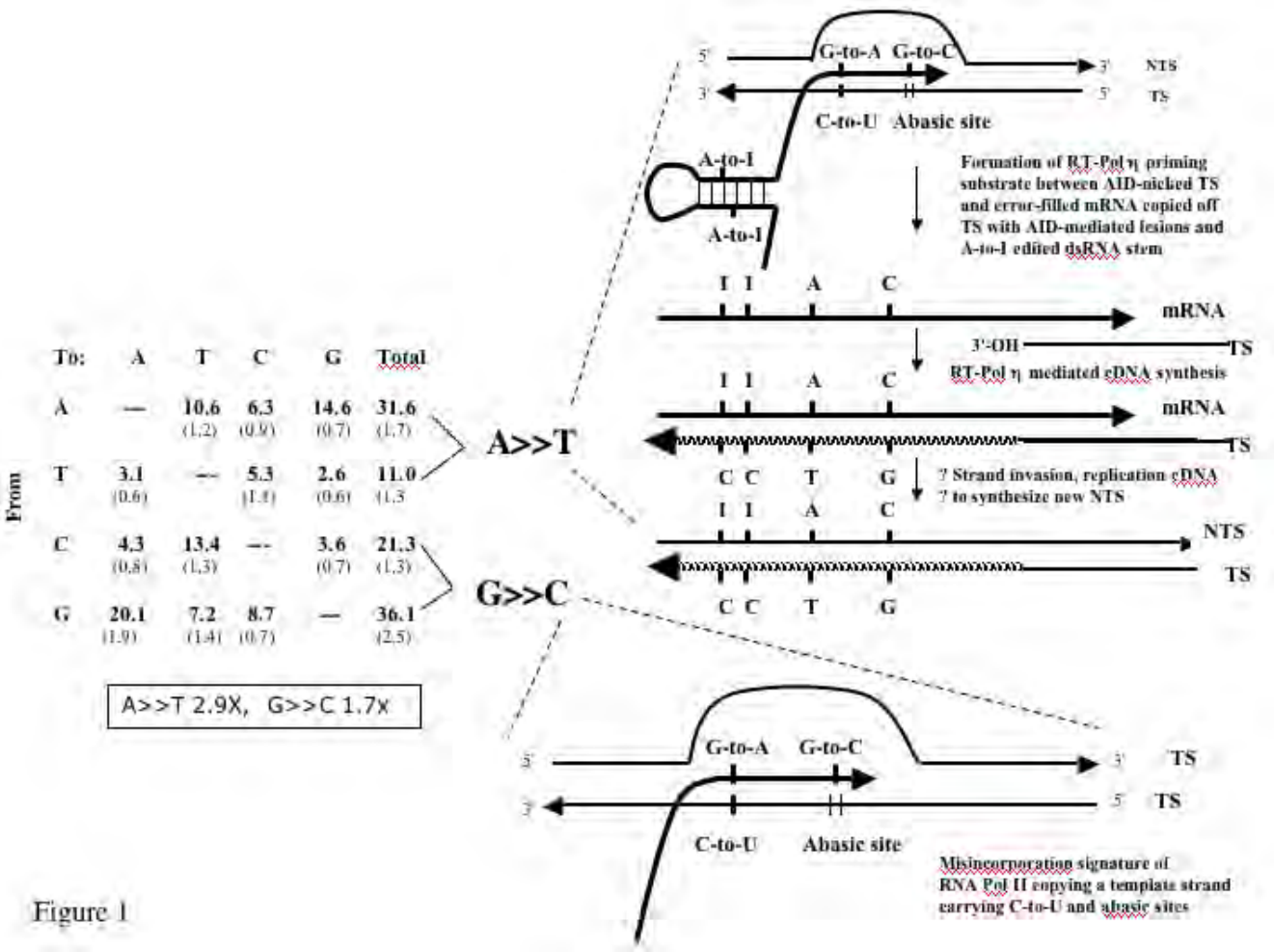
To:
A
T
C
G
Total
From
A
—
10.6 (1.2)
6.3 (0.9)
14.6 (0.7)
31.6 (1.7)
T
3.1 (0.6)
—
5.3 (1.1)
2.6 (0.6)
11.0 (1.3
C
4.3 (0.8)
13.4 (1.3)
—
3.6 (0.7)
21.3 (1.3)
G
20.1 (1.9)
7.2 (1.4)
8.7 (0.7)
—
36.1 (2.5)
A>>T
G>>C
A>>T 2.9X, G>>C 1.7X
5'
3'
G-to-A
G-to-C
NTS
TS
C-to-U
Abasic site
A-to-I
A-to-I
Formation of RT-Pol η priming substrate between AID-nicked TS and error-filled mRNA copied off TS with AID-mediated lesions and A-to-I edited dsRNA stem
I I A C
mRNA
3'-OH
TS
RT-Pol η mediated cDNA synthesis
C C T G
? Strand invasion, replication cDNA ? to synthesize new NTS
Misincorporation signature of RNA Pol II copying a template strand carrying C-to-U and abasic sites

Figure 1

.

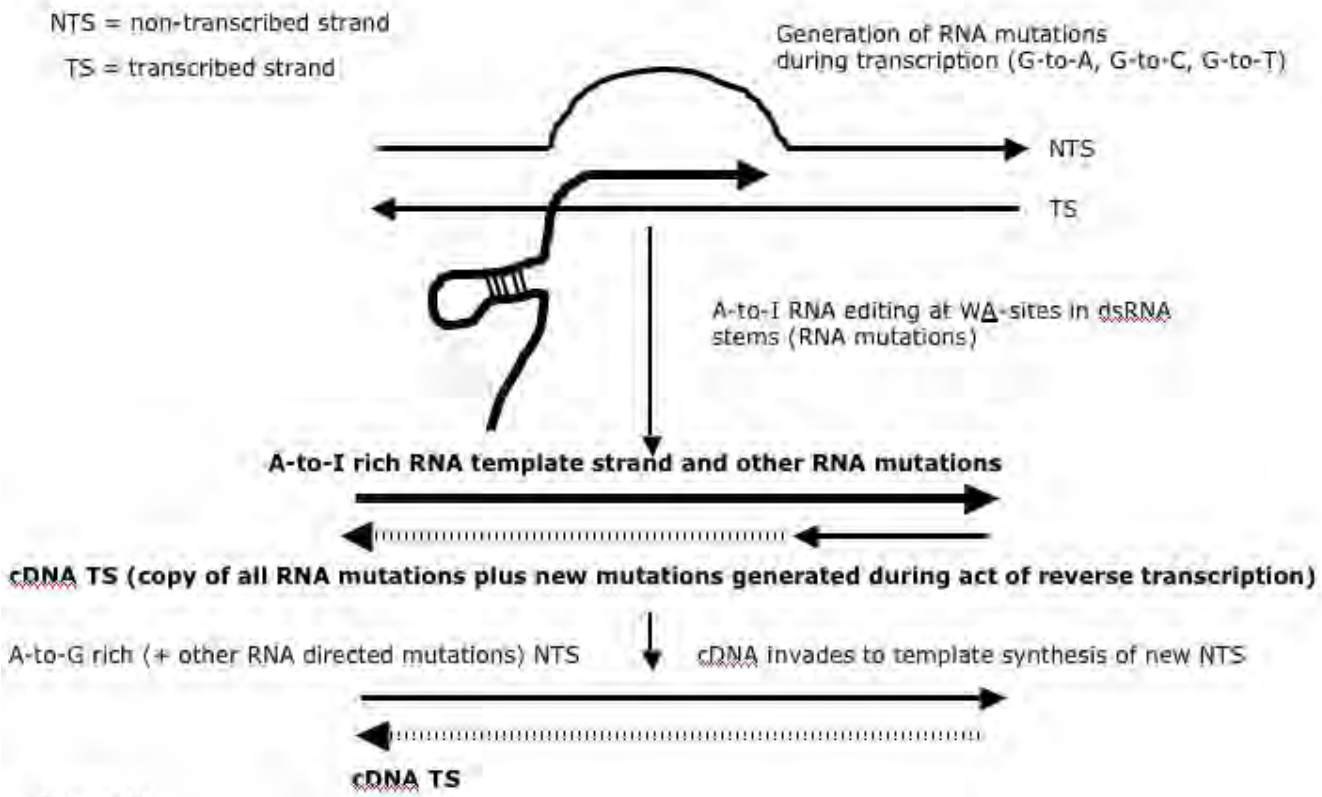
NTS = non-transcribed strand
TS = transcribed strand
Generation of RNA mutations
during transcription (G-to-A, G-to-C, G-to-T)
NTS
TS
A-to-I RNA editing at WA-sites in dsRNA
stems (RNA mutations)
A-to-I rich RNA template strand and other RNA mutations
cDNA TS (copy of all RNA mutations plus new mutations generated during act of reverse transcription)
A-to-G rich (+ other RNA directed mutations) NTS
cDNA invades to template synthesis of new NTS
cDNA TS

Figure 2

A cluster specific promoter could be engaged driving the synthesis of a long transcript within the Transcription Factory

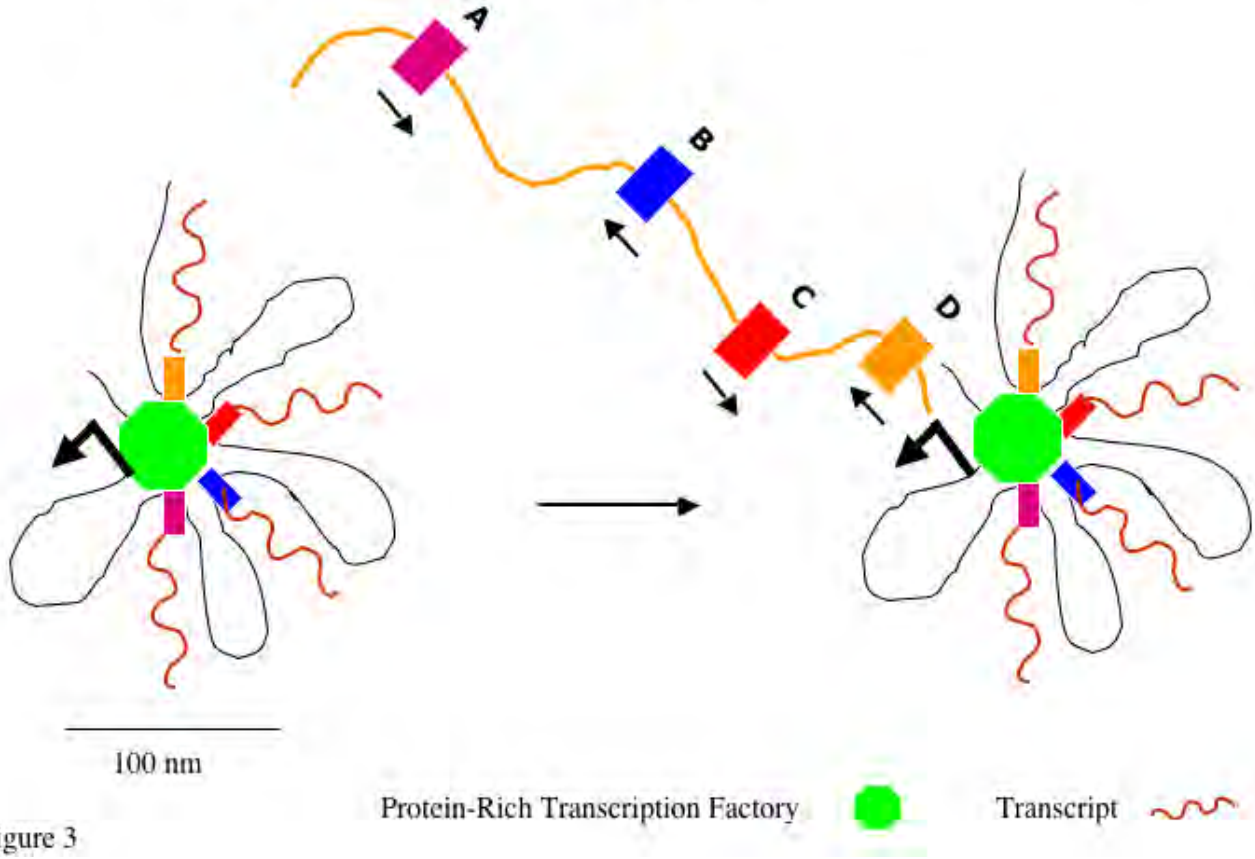

Figure 3

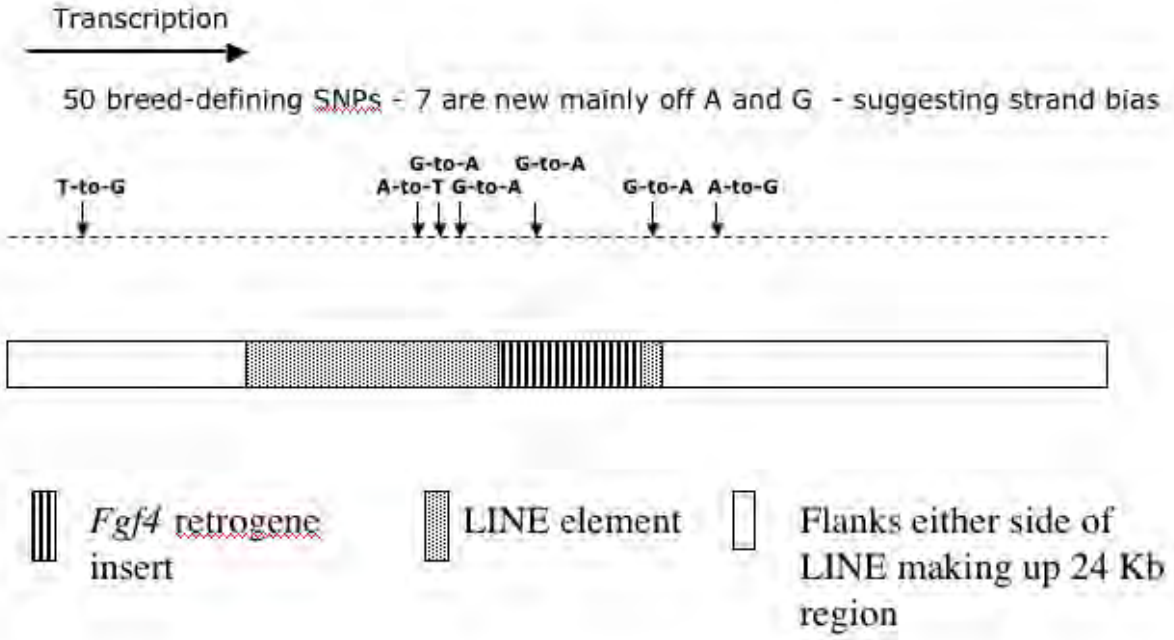
Transcription
50 breed-defining SNPs - 7 are new mainly off A and G - suggesting strand bias
T-to-G
G-to-A
A-to-T
G-to-A
G-to-A
G-to-A
A-to-G
Fgf4 retrogene insert
LINE element
Flanks either side of LINE making up 24 Kb region

Figure 4

.

.