FREQUENT INTRA-FAMILY RECOMBINATION IN THE LARGEST REPOSITORY OF ANTIGEN VARIANTS IN THE PROTOZOAN PATHOGEN *TRYPANOSOMA CRUZI*

by

DUO PENG

(Under the Direction of RICK TARLETON)

ABSTRACT

The genome of the protozoan parasite *Trypanosoma cruzi*, the causative agent of Chagas disease, encodes a family of genes consisting of 3209 trans-sialidase (TcTS) and TcTS-like genes. Simultaneous expression of many members of the TcTS/TcTS-like family (henceforth TcTS gene family) presents variant peptide antigens to the host immune system. Recombination is a major force in generating and spreading genetic variation in gene families and such a process has been documented in *Trypanosoma brucei* and *Plasmodium* to contribute to antigenic variation. To investigate the extent to which recombination creates genetic variation in TcTS gene family, we have developed a computational pipeline capable of analyzing recombination events in the entire TcTS gene family. Using this computational pipeline, we demonstrate that TcTS gene family members are undergoing frequent recombination, generating new variants from the thousands of functional and non-functional gene segments.

INDEX WORDS: trans-sialidase, recombination, gene conversion, *T. cruzi*

FREQUENT INTRA-FAMILY RECOMBINATION IN THE LARGEST REPOSITORY OF ANTIGEN VARIANTS IN THE PROTOZOAN PATHOGEN *TRYPANOSOMA CRUZI*

by

DUO PENG

A Thesis Submitted to the Graduate Faculty of The University of Georgia in Partial Fulfillment of the Requirements for the Degree

MASTER OF SCIENCE

ATHENS, GEORGIA

2016

FREQUENT INTRA-FAMILY RECOMBINATION IN THE LARGEST REPOSITORY OF ANTIGEN VARIANTS IN THE PROTOZOAN PATHOGEN *TRYPANOSOMA CRUZI*

by

Duo Peng

Major Professor:	Rick Tarleton
Committee:	Jessica C. Kissinger
	Robert Sabatini

Electronic Version Approved:

Suzanne Barbour
Dean of the Graduate School
The University of Georgia
May 2016

ACKNOWLEDGEMENTS

First and foremost I would like to acknowledge my advisor, Dr. Rick Tarleton, who has been a tremendous mentor for me, allowing me to grow as an independent researcher. I extend my thanks to my committee members Dr. Jessica Kissinger and Dr. Robert Sabatini for their advice and assistance. I would also like to thank my parents and friends for their support and as great source of information.

TABLE OF CONTENTS

Page

LIST OF TABLES

Page

LIST OF FIGURES

Page

CHAPTER 1

INTRODUCTION AND LITERATURE REVIEW

Trypanosoma cruzi is a protozoan parasite capable of causing zoonotic infection of mammals and it infects 10-20 million people in the Americas. The parasite cycles through 4 distinct life cycle stages between insect vectors and mammalian hosts. Epimastigotes replicate in the insect midgut, and differentiate into metacyclic trypomastigotes, which are the infective form capable of entering mammalian cells following parasite defecation. Inside mammalian cells, *T. cruzi* alternates between the infective trypomastigote form and the replicative rounded amastigote form.

Sialic acids, a family of 9-carbon carboyxlated sugars usually found as terminal monosaccharides of animal glycoproteins, are exploited by *T. cruzi* to interact with host cells in multiple ways. *T. cruzi* cannot synthesize sialic acids (Confalonieri, Martin et al. 1983, Schauer, Reuter et al. 1983) and it expresses an enzyme named trans-sialidase (TcTS), in mammalian life cycle stages, to catalyze the transfer of sialic acids from host cell glycoconjugates to its own surface glycoproteins (namely mucins) (Previato, Andrade et al. 1985, Scudder, Doom et al. 1993). The TcTS protein's N-terminal domain harbors the catalytic site responsible for trans-sialidase and neuraminidase activity. A C-terminal lectin-like domain (Trk binding domain) often found in TcTS proteins, is required for enzyme oligomerization, and is capable of binding to sialic acid. The C-terminus domain may also contain an additional site for glycophosphatidylinositol (GPI), which anchors TcTS protein to the plasma membrane of *T. cruzi* (Cross and Takle 1993).

Cleavage of the GPI anchor can lead to shedding of TcTS into host cell cytoplasm or extracellular space. A 12 amino acid repeat sequence named Shed Acute Phase Antigen (SAPA) motif is often found near the C-terminal of TcTS protein, helping to increase TcTS protein half-life when it is released from the parasite surface (Buscaglia, Alfonso et al. 1999).

The importance of TcTS in the *T. cruzi* invasion process has been demonstrated by several studies (Schenkman, Jiang et al. 1991, Ming, Chuenkova et al. 1993). Inhibiting catalytic activity of TcTS or reducing available host sialic acids negatively impacts adhesion to and infection of host cells (Schenkman, Vandekerckhove et al. 1993, Agusti, Paris et al. 2004, Carvalho, Sola-Penna et al. 2010, Lieke, Grobe et al. 2011). TcTS is also linked to promoting escape from the host cell vacuole into the cytosol, likely via interaction with lysosome integral membrane proteins during vacuole formation (fusion of lysosome and plasma membrane) when parasite enters non-professional phagocytic cells (Rubin-de-Celis, Uemura et al. 2006, Albertti, Macedo et al. 2010). Furthermore, over-expression of TcTS results in earlier parasite escape from vacuoles (Rubin-de-Celis, Uemura et al. 2006, Albertti, Macedo et al. 2010).

TcTS can modulate mammalian host innate and acquired immune responses (Dc-Rubin and Schenkman 2012). Sialyation of *T. cruzi* surface proteins (by TcTS) protects parasites from lysis by the complement system (Kipnis, David et al. 1981). The sialylated surface of *T. cruzi* interacts with dendritic cells and suppresses the production of the pro-inflammatory cytokine IL-12 (Erdmann, Steeg et al. 2009), a key cytokine in the activation of immune response. TcTS shed into the infected cell cytosol can activate serine-threonine kinase (Akt) signaling through its Trk binding domain, resulting in

increased protection against oxidative stress and pro-inflammatory cytokines, and prevents host cell apoptosis during infection (Chuenkova, Furnari et al. 2001, Chuenkova and PereiraPerrin 2009). Trypomastigote-derived TcTS in the blood can induce cell apoptosis and prevent proliferation of lymphocytes in the thymus and peripheral ganglia by sialyation of lymphocyte surface molecules (Vercelli, Hidalgo et al. 2005, Mucci, Risso et al. 2006). TcTS can also modulate T helper responses, by delaying and limiting Th1 responses (Ruiz Diaz, Mucci et al. 2015), which are crucial for controlling intracellular pathogens such as *T. cruzi*. Thymocyte and T cell apoptosis are also observed in acute *T. cruzi* infection and TcTS has been linked to the pro-apoptotic effects on thymocytes and mature T cells (Leguizamon, Mocetti et al. 1999, Mucci, Hidalgo et al. 2002). Both desialylation and sialylation activity by TcTS's neuraminidase and trans-sialidase activity, respectively, have been linked to hematological alternations of immature and mature T cells (de Titto and Araujo 1988, Tribulatti, Mucci et al. 2005, Mucci, Risso et al. 2006).

TcTS is encoded by a super family of TcTS and TcTS-like genes. The first genome sequencing effort of *T. cruzi* CL Brener strain identified 1430 TcTS genes, which represents one of the largest phenomena of gene expansion, though only 15 TcTS genes are predicted to be enzymatically active (El-Sayed, Myler et al. 2005). Weatherly et al. (unpublished) recently identified an additional 1779 TcTS genes in the *T. cruzi* CL Brener strain using partially assembled sequences in the original genome sequence dataset, raising the total number of predicted TcTS genes to 3209. Of this set, 810 TcTS genes are annotated as to produce truncated products. The coding sequence of these TcTS contains premature stop codons and/or frame-shifts after a leading signal peptide. There

are 2516 TcTS genes annotated as partial genes, which lack coding sequence for a significant proportion of N- and/or C-terminal domain. An archetype TcTS protein contains: sialidase FRIP motif, (multiple copies of) sialidase motif SXDXGXTW (Asp-box), conserved VTVxNVfLYNR (FLY) motif in the N-terminal domain; SAPA repeats and GPI addition site in the C-terminal domain. The TcTS super family is further divided into 5 families; each family with a distinct combination and representation of motifs and sometimes a complete N- or C-terminus domain.

Concomitant expression in *T. cruzi* of many TcTS was reported by several studies (Kahn, Colbert et al. 1991, Takle and Cross 1991). Proteomic studies have identified at least 246 expressed TcTS genes (Atwood, Weatherly et al. 2005). Furthermore, there are 530 expressible TcTS genes which harbor signal peptide and encode TcTS or TcTS-like peptides of 200 amino acids or longer. TcTS proteins expressed by mammalian life cycle stages of *T. cruzi* are released into the cytosol of infected cells as well as extracellular space (Pereira 1983, Rosenberg, Prioli et al. 1991, Schenkman, Pontes de Carvalho et al. 1992). TcTS proteins in the host cell cytosol and extracellular space are subject to be processed by the host immune system. Although the host immune system generates a robust CD8+ T cell response against TcTS proteins, only a small subset of TcTS peptides are responsible for the majority of CD8+ T cell response (Martin, Weatherly et al. 2006). Furthermore, the generation of CD8+ T cell response is delayed (Padilla, Simpson et al. 2009) and insufficient to cure infection (Martin and Tarleton 2004). Two mechanisms have been proposed to explain the sub-optimal T cell response to *T. cruzi*. (1) The diverse proteins released by *T. cruzi* can overload the host antigen processing pathway and prevent it from inducing a potent focused response. (2) The concomitant presentation of

homologous peptides may antagonize naïve or effector T cells. Both mechanisms require co-expression of a large number of variant peptides.

T. cruzi has disproportionally expanded its TcTS gene family in sharp contrast to the majority of enzyme-coding genes in its genome. Furthermore, maintenance of 3209 TcTS and TcTS-like genes in its genome and co-expression of hundreds of enzymatically inactive TcTS proteins is not justified by the catalytic, lectin-like binding functions of the TcTS proteins. The fact that TcTS gene family encodes a large number of variant peptides for immune evasion could potentially justify the purpose of maintaining this sizable gene family and co-expression of hundreds of its members. It is postulated that the TcTS gene family as well as other large gene families in *T. cruzi* that encode for surface proteins, have expanded in response to immune pressure, presumably by duplication, recombination and mutation of the ancestral founding family members.

Gene conversion-style recombination serves as one of the major mechanisms to generate and propagate variations in gene families consisting of large number of homologous sequences (Roth, Jacquemot et al. 1991, Ohta 1992, Thompson 1992, Parham and Ohta 1996, Gjini, Haydon et al. 2012). Azuaje et al. used computational simulation to show that gene conversion can act as an effective variation generation mechanism in TcTS gene family, maximizing variation at the amino acid level with relatively low mutation rates (Azuaje, Ramirez et al. 2007). Gene conversion can also deliver variations from silent pseudogenes to expressed counterparts (Thon, Baltz et al. 1989, Taylor and Rudenko 2006), effectively utilizing the whole gene family repertoire as genetic variation material pool, which allow for a potential justification of maintaining thousands of TcTS pseudogenes in the *T. cruzi* genome. Direct observation of a gene

conversion event between TcTS genes was made by Ruef et al., who cloned 3 TcTS sequences that, when aligned, demonstrated a clear switch in homology between two pairings (Ruef, Dawson et al. 1994). This patchy homology is indicative of a gene conversion event. Taken together, it is reasonable to suspect that gene conversion events were integral for generating the variation observed in the TcTS gene family.

Classical analysis of recombination typically relies on highly accurate sequence alignment and evidence of local disruption of sequence signatures or phylogenetic signals as markers of recombination. Rigorous statistical analysis can then be used to confirm the significance of the identified recombination breakpoint. However for gene families containing thousands of members such as the TcTS gene family, such methods are not feasible. Alternatively, BLAST algorithms can be used to find local sequence matches between different family members (Weirather, Wilson et al. 2012) or phylogenetic incompatibility distribution can be used as an index to estimate recombination activity, and such methods have been applied to document recombination events in the *T. brucei* VSG gene family (Jackson, Berry et al. 2012). However, such alternative methods, unlike the classical recombination analysis, cannot ensure both statistical support and the search-comprehensiveness of recombination events. Here I propose a recombination analysis pipeline that can be used on large and diverse genes families while also providing statistical validation of the results. I use this tool to thoroughly analyze the TcTS genes to understand the extent of gene conversion as a variation-generating mechanism in the TcTS gene family.

CHAPTER 2

ASSESSMENT OF RECOMBINATION EVENTS IN THE TCTS GENE FAMILY

Codon substitution analysis indicates more frequent protein sequence mutation in TcTS genes compared to core metabolic genes

To explore the rates of accumulation of variationsin TcTS genes, we performed codon substitution analysis in TcTS genes and *T. cruzi* core metabolic genes. The majority of core metabolic genes have low synonymous codon substitution rates (<0.06, Figure 1), and very low nonsynonymous codon substitution rates (<0.04), suggesting strong conservation of amino acid sequence. In contrast, both synonymous and nonsynonymous codon substitution rate of majority of TcTS genes are much higher, falling in a range of 0.04-0.5. Furthermore, the average nonsynonymous codon substitution rate of TcTS genes is 5 times higher than that of core metabolic genes, indicating a higher frequency of amino acid substitutions in the protein sequence of TcTS genes. These new peptide sequence variations are a prime candidate source for antigenic variation.

Phylogenetic analysis of related TcTS sequences suggests frequent recombination

The distribution throughout the *T. cruzi* genome of TcTS sequences of varying sizes but with a conserved structure (Weatherly et al. unpublished) strongly suggests a mechanism of gene duplication followed by recombination and mutation in the evolution

of this large and diverse gene family. To explore this possibility further, we first examined phylogenetic trees of TcTS sequences for evidence of recombination. Due to the substantial variation in gene length (from less than 500bp to over 4000bp) and nucleotide polymorphism among the >3000 TcTS sequences (which resulted in poor sequence alignment), we performed phylogenetic tree inference on small sets of more closely related TcTS sequences identified by MEGABLAST with custom thresholds (see Materials and Methods for details). Although the potential mosaic nature of TcTS sequences could hinder accurate phylogeny assessment, we were able to estimate a well-supported Bayesian tree (Figure 2) by a lengthy and aggressive Markov chain Monte Carlo exploration in parameter space. This Bayesian phylogenetic tree has long terminal branches that indicate loss of phylogenetic signal and is suggestive of recombination. Notably, a similar tree pattern is observed for the variant surface glycoprotein (VSG) genes in *T. brucei* and *T. congolense* (Jackson, Berry et al. 2012).

A computational pipeline to assess recombination events in TcTS gene family

In order to more thoroughly evaluate the role of recombination in the evolution of the TcTS family, a computational pipeline automated using PERL scripts was developed for identification of recombination events in TcTS (Figure 3). First, all TcTS (nucleotide) sequences were grouped into similarity sets by scanning each TcTS gene for closest-matching non-self sequences within the TcTS gene family (Figure 3a). To identify these closest-matching non-self sequences, each trans-sialidase sequence (termed the "group identifier gene") was first split into non-overlapping 50-mers and each 50-mer MEGABLASTed (Zhang, Schwartz et al. 2000) against a database containing all TcTS

sequences using default parameter settings. For each 50-mer, the closest matching TcTS gene with 100% coverage was identified; the similarity group for each TcTS gene consisted of the closest-matching non-self-hits for all 50-mers from that gene. Conserved regions identified by having multiple identical closest-matching hits, were excluded. Trials varying the size of the split window and overlap between windows showed the 50-mer, 0 overlap setting to provide adequate sensitivity without being computationally burdensome.

To identify possible recombinant regions, recombination breakpoints, and the major and minor sequence donor(s) (see glossary for definition) among the trans-sialidase sequence groups, each group of sequences was aligned by Clustalw2 (v2.1) (Larkin, Blackshields et al. 2007) and each alignment was then analyzed for recombination by the RDP (Recombination Detection Program) v4.17, (Martin, Lemey et al. 2010) using the following parameters: Linear sequence, Highest acceptable P-Value=0.01, Bonferroni correction, no permutations, check alignment consistency (Figure 3b). The settings for the individual algorithms are provided in Table 2. RDP-predicted recombination events were further filtered to meet 3 criteria: (1) the recombination event is detected by at least 2 algorithms, (2) the detected mosaic gene must be the group identifier gene, and (3) redundant recombination events, characterized by recombination events with the same recombinant region in same mosaic gene with distinct, but highly similar donors, are removed.

Since RDP relies on accurate alignment as input for analysis of recombination events, alignment artifacts introduced by distantly related sequences in each sequence group might create false recombination signals (as we observed), despite the fact that

RDP checks for alignment consistency before analyzing for recombination. As a final step, the sequences predicted to contribute to each recombination event were realigned using only the predicted mosaic gene, the major donor gene and the minor donor gene for that event, and the alignment resubmitted to recombination analysis and resulting events again filtered as described above (Figure 3c).

Evaluate specificity and sensitivity of recombination detection pipeline

To evaluate the specificity and sensitivity of our recombination detection pipeline, simulated trans-sialidase gene family datasets with and without recombination were generated as negative and positive datasets, respectively. Simulated datasets were subjected to analysis by the recombination detection pipeline. Results were compared to documented recombination events in the control dataset.

To generate negative and positive datasets, a similarity tree was first calculated for the *T. cruzi* trans-sialidase gene family using the Tamura-Nei genetic distance model and the neighbor-joining tree constructing method (Saitou and Nei 1987). DAWG v1.2 (Cartwright 2005) was then used to generate simulated *T. cruzi* trans-sialidase gene families. The 4 recombination-negative datasets were generated containing random point mutation rates of 0%, 1%, 5%, 10%, respectively. Homologous recombination events or gene conversion events were separately introduced into the 0% point mutation negative data to generate the 2 recombination-positive datasets. To introduce artificial gene conversion recombination events, a subsequence from a randomly chosen donor gene is extracted, and used to replace the homologous subsequence in a randomly chosen recipient gene. To introduce artificial homologous recombination (reciprocal exchange)

events, homologous subsequences from 2 randomly chosen genes are extracted and exchanged. Normal distribution is used to model the length of the recombination region. Exponential distribution is used to model the number of recombination events per gene. Exponential distribution with probability density function $y = 0.52622\, e^{-0.52622x}$ is a good approximation of the distribution of frequency of number of recombination events per gene observed in our preliminary analysis. A custom PERL program script was used to automate the process of introducing multiple recombination events in a single dataset. The PERL script takes 4 parameters, percentage of gene conversion recombination "g%", total recombination gene number "n", recombination region mean length "l" and standard deviation of recombination region length "std". The program randomly chose g% of the genes to introduce recombination into. For each recombinant gene, the program sampled the exponential distribution described above to determine the number of recombination events to introduce into the gene. For each recombination event, g% probability to introduce gene conversion event, and 1-g% probability to introduce homologous recombination events (reciprocal exchange), the length of recombination event was sampled from the normal distribution (l, std2).

Simulated datasets were analyzed by our recombination detection pipeline, the results were then compared to the documented recombination events during the generation of the simulated datasets. Sensitivity is defined as the ratio of detected recombination events over all recombination events in the simulated dataset, and specificity was defined as the ratio of the correctly identified recombination events over all detected positives events. In negative datasets, all detected recombination genes were considered false positives. In recombination positive datasets, only the detected events

that had a recombination breakpoint within 200bp distance from the real recombination break point were considered to match (or correctly identified) the documented recombination events. All matched (correctly identified) recombination events were considered true positives and those that did not match were considered false positives.

Evaluation of the pipeline with simulated data indicated dependable performance (Table 1) with a sensitivity of 0.93 and 0.73 for gene-conversion-recombinant genes and gene-conversion-recombinant events, respectively and a specificity of 0.54 and 0.44 for gene-conversion-recombinant genes and gene-conversion-recombinant events, respectively. Note that the event-specificity is impacted by the difficulties of pinpointing the recombination breakpoint between sequences with less informative variations (e.g. random mutations that obscure sequence origins). The false positive rates are below 0.003 for all negative datasets tested. Notably, our recombination pipeline is not intended to detect reciprocal exchange of DNA sequences between 2 TcTS genes, as is shown by simulation results (Table 1).

Frequent intra gene family recombination events in the TcTS gene family

Using our recombination detection pipeline to analyze the 3209 TcTS genes, we document clear evidence of recombination in 787 trans-sialidase genes with a total of 2087 recombination events; 783 and 749 genes have participated at least once as major donor and minor donors, respectively. Of all the 2087 recombination events, 1742 events have a p-value less than 0.0001 with one or more of the algorithms in RDP and 680 events have p-values <0.0001 for all algorithms. All 2087 recombination events also have multiple-test corrected p-values under 0.01. Thus, these recombination events are

detected with very high confidence. The 787 mosaic genes identified have between 1 and 12 recombination events each, with an average of 2.65 recombination events per gene (Figure 4). The signals of mosaic patches in TcTS genes can be confirmed by manual inspection at the nucleotide level and Figure 5 shows an example. Note that the break-points of recombinant regions frequently lack a distinct boundary but instead have short regions that are not encoded by either the major donor or minor donor (Figure 6). Participation in recombination (either as the mosaic product or as a major or minor donor) was higher for previously annotated TcTS genes relative to the newly annotated genes, although a significant proportion of this difference was due to the smaller average size of the newly annotated TcTS genes. Recombination events were more likely to be detected between full length and other TcTS genes of 2500 bases or more (Figure 7). However, 25% of the donors contributing to mosaic full length TcTS genes were of <2500 bases, demonstrating that the incomplete TcTS gene fragments in the *T. cruzi* genome play an active role in the diversity of the expressed, full length genes. Within a select set of TcTS genes grouped into 8 subsets based upon multiple component analysis (Freitas, dos Santos et al. 2011), recombination was exactly 4 times more likely to occur between members of the same group (444 events) than between groups (111 events). It is also notable that recombination events both contributed to the production of full length TcTS proteins from partial gene donors (Figure 8a) as well as partial genes from full length genes due to shifts in reading frame as a result of recombination (Figure 8b). Thus there is an active exchange between the full-length TcTS genes encoding functional proteins and partial TcTS genes and pseudogenes dispersed throughout the *T. cruzi* genome.

Materials and Methods

Data source

Sequences of 1430 annotated TcTS gene family members were obtained from version 5 of the *T. cruzi* CL-Brener strain genome. Sequences of 1779 newly annotated TcTS gene family members were obtained from Weatherly et al. (unpublished).

Codon substitution analysis

TcTS genes were aligned using the codon alignment option in the MUSCLE algorithm (Edgar 2004) implemented by Geneious v5.6.5 (free version). Resulting alignments were used for codon substitution analysis by KaKs_Calculator v1.2 (Zhang, Li et al. 2006) using the MS method (Model Selection according to the AICc). Heat map was plotted by R v3.06 using custom scripts.

Phylogenetic tree of related TcTS sequences

To obtain clusters of highly-similar TcTS suitable for phylogenetic tree construction, nucleotide sequence of 930 TcTS genes between 1000bp and 3000 bp were aligned to each other using MEGABLAST (Zhang, Schwartz et al. 2000) with non-default parameter -v 2000 -b 2000, and the 149 sequences with high scoring segment pairs (HSPs) covering more than 70% of the query sequence were selected. The setting of 70% HSP was chosen for subsequent tree inference because it strikes a balance between alignment quality (higher consensus identity, fewer gaps) and homolog numbers (149 homologs yielding high quality trees). Protein sequences of 149 highly-similar TcTS

were aligned using MUSCLE algorithm (Edgar 2004) implemented in Geneious v5.6.5. Alignment of the TcTS sequences was manually edited with alignment viewer in Geneious v5.6.5, Bayesian tree inference was performed using MrBayes v3.2.1 (Ronquist, Teslenko et al. 2012) parallel version on XSEDE (Extreme Science and Engineering Discovery Environment) via CIPRES Science Gateway (Miller, Pfeiffer et al. 2010), non-default parameters are generalized time-reversible (GTR) substitution model (nst=6), gamma rate variation with 8 discrete categories, nchains=5, temp=0.3 (more MCMC chains and higher heating temp are required to reach convergence), 10,000,000 MCMC generations, burn-in fraction 0.5. Phylogenetic trees were visualized by Figtree v1.40 (http://tree.bio.ed.ac.uk/software/figtree/). Treeness was calculated by TreeStat v1.2 (http://tree.bio.ed.ac.uk/software/treestat/)

A computational pipeline to assess gene conversion events in TcTS gene family

The PERL program script files that implement our recombination detection pipeline have been deposited to Github (https://github.com/duopeng/Recombination-detection-pipeline-for-large-gene-families).

Evaluate specificity and sensitivity of recombination detection pipeline

DAWG v1.2 (Cartwright 2005) was used to generate simulated *T. cruzi* trans-sialidase gene families using the following parameters: GTR evolution model (Tavaré 1986), alpha=0.5, root sequence length 2500.

CHAPTER 3

DISCUSSION

In the evolutionary arms race between the host immune system and pathogens, mammals have incorporated immense complexity into the immune system. The multi-gene segmental organization of immunoglobulin and T cell receptor genes, combined with recombination, can produce many millions of distinct antigen receptors capable of detecting and neutralizing a diverse array of pathogens. Pathogens, including eukaryotic parasites, have also evolved mechanisms to incorporate diversity into their structures in order to avoid immune destruction. The archetypal example of antigenic variation is the variant surface glycoproteins (VSG) gene family of *T. brucei*, which has more than 2000 members. Although its expression is mono-allelic, the relatively frequent switch of the expressed VSG gene enables *T. brucei* to maintain its persistence despite the constant surveillance by host immune responses. The creation of mosaic VSGs by recombination among VSG family members (Hall, Wang et al. 2013, Mugnier, Cross et al. 2015) is also crucial for parasite persistence, as host immunity to previously expressed VSGs (Magez, Schwegmann et al. 2008) accumulates over the time of infection. Likewise, *Plasmodium sp.* have evolved smaller multi-gene families encoding cell surface proteins that depend on a high frequency of recombination to produce new immune-evading variants (Freitas-Junior, Bottius et al. 2000, Barry, Leliwa-Sytek et al. 2007).

T. cruzi encodes several large, highly variant gene families, including the largest known family of variant surface protein-coding genes, the TcTS gene family. The mostly GPI-anchored TcTS surface proteins have several essential functions in the parasite. The presumed founding members of this family provide *T. cruzi* the ability to acquire sialic acid from host donor molecules and move it to terminal glycans in a sialyl-transferase reaction. This process is apparently necessitated by the fact that *T. cruzi* lacks the ability to produce its own sialic acid. In the absence of sialic acid, *T. cruzi* trypomastigotes are highly susceptible to host complement activation and lysis and are also poorly invasive. However, only a small proportion of the immense TcTS gene family, perhaps as few as 10-15 genes, actually encode enzymatically active TcTS. The remainder have been ascribed a number of other activities; among the more important, as lectins potentially involved in host cell attachment (Oppezzo, Obal et al. 2011). Also, it is important to note that *T. cruzi* expresses the TcTS genes from multiple alleles concurrently rather than serially as in the case for the surface variant antigens in *Plasmodium* and *T. brucei*.

Despite the importance of these TcTS activities for the survival of *T. cruzi*, it is difficult to reconcile the huge commitment in terms of genome space that *T. cruzi* makes to the maintenance of TcTS-family genes with these noted TcTS functions alone. The original assembly of the reference CL Brener genome tallied 1430 TcTS genes, approximately half of which were annotated as "pseudogenes". A recent study by Weatherly et al. (unpublished) remapped the original CL Brener sequence reads and re-analysis for TcTS-like sequences more than doubled the number of full and partial TcTS genes identifiable in the CL Brener genome to the current 3209 unique genes. Further, proteome analysis also confirmed the expression of at least a portion of these newly

identified full length and truncated TcTS family genes. Certainly *T. cruzi* should not require >3000 full and partial length TcTS family genes to carry out the functions of sialic acid transfer and cell adhesion, and even if all the proposed functions of TcTS proteins are considered, performance of these activities perhaps requires a handful of genes but not >3000. This degree of gene diversity within individuals of a single species is nearly unprecedented, rivaled only by the ~2700 member VSG gene family in the related *T. brucei* (Cross, Kim et al. 2014).

The critical requirement for the cell surface expression of TcTS activity presents a dilemma for a pathogen like *T. cruzi*. On the one hand, lack of TcTS activity significantly reduces the ability of *T. cruzi* to survive in mammalian hosts. On the other, expression of a single species of TcTS protein on the parasite surface would provide a tantalizing target for the host immune system and an effective response to such a target could prove deadly for *T. cruzi*. Indeed TcTS proteins are significant targets of anti-*T. cruzi* immune responses (Martin, Weatherly et al. 2006) and vaccination against TcTS proteins is somewhat effective, at least in preventing potentially lethal infections (reviewed in (Vazquez-Chagoyan, Gupta et al. 2011)). Thus there is a strong pressure from host immune responses on TcTS proteins, and as with other cases of antigenic variation, this is the likely driving force for the retention and diversification that ultimately created this immense family of genes.

The mechanism of the expansion and continued diversification of the TcTS family in *T. cruzi* is gene duplication, mutation and recombination among family members. We present strong evidence for 787 mosaic trans-sialidase genes resulting from 2087 separate recombination events. Mosaicism is also a prominent feature of *T. brucei*

VSGs (Marcello and Barry 2007). Due to the conservative calling of recombination by gene conversion, ours is a minimal estimate; other (gene conversion) recombination events are likely obscured by sequence divergence resulting from accumulating mutations in both donor and recipient genes. Despite the frequency and wide scope of recombinations, all the TcTS gene family members, including enzymatically active TcTS, inactive, truncated molecules, and gene fragments incapable of producing a protein product, retain a similar domain structure. The retention of the common features of TcTS genes and proteins across the entire family likely reflects constraints imposed by the non-enzymatic functions of TcTS, such as binding to (host) sialyl and beta-galactopyranosyl residues (Todeschini, Girard et al. 2002, Todeschini, Dias et al. 2004), as well as the contribution of pseudogenes/partial genes to the formation of new mosaic TcTS genes. Substantial homology is generally required for gene conversion-based recombination (Liskay, Letsou et al. 1987, Chen, Cooper et al. 2007) and this requirement would work to sustain the common structure of the TcTS genes. The fact that intra-gene-family recombination events among TcTS genes are 4 times more likely to occur between genes in the same similarity group identified by Freitas et al. (Freitas, dos Santos et al. 2011) than between members of different groups supports this point. Importantly our analysis additionally supplies evidence that the structurally conserved but sequence variant non-coding gene fragments are derived from and are recycled back into functional (i.e., full length, expressed) TcTS genes. Indeed, nearly half of the 437 full-length protein-coding TcTS identified as mosaics have incorporated TcTS pseudogenes and/or partial genes. Thus, the non-coding TcTS fragments undoubtedly acts as a repository of diversity for the generation of new and antigenically diverse TcTS genes. Similar integration of

pseudogene sequences into mosaic VSGs has been documented for *T. brucei* (Hall, Wang et al. 2013).

In this thesis work, I developed a computational pipeline to inspect TcTS gene family for gene conversion style recombination events. This recombination detection pipeline first uses BLAST to find local sequence matches between individual TcTS genes. Then, for each local sequence match, we group involved-TcTS genes into a subset, between which recombination is likely to produce the mosaic patch (local sequence match). Then for each TcTS subset, we aligned the sequences and subject the alignment to a recombination detection program (RDP) that implements many classical recombination detection algorithms. If a recombination event is cross-validated by multiple recombination detection algorithms, we perform an additional round of validation by realigning just the 2 recombination donors and the mosaic gene and then subjecting the alignment to recombination analysis by RDP. This final validation step further removes false-positives caused by alignment artifacts. By starting from every possible mosaic-forming event in the gene family and validating through 2 rounds of rigorous recombination analysis, we combined both statistical support and search-comprehensiveness in our recombination detection pipeline for very-large gene families. Though search-comprehensiveness is our goal, our pipeline would still underestimate the recombination events in a large gene family since the rigorous statistical threshold would miss many weaker recombination signals; and recombination events could be overlapping, masking each other and blurring the recombination breakpoint. Thus, we are likely able to see only a small part of the complex TcTS recombination landscape. The current analysis is also restricted to a single *T. cruzi* isolate, the genome reference CL

Brener clone. Each *T. cruzi* clone will generate a distinct pattern of recombination and thus its own set of TcTS genes, resulting in incredibly diverse populations of parasites within and between hosts. Similar detailed analysis to that conducted here in other isolates is expected to confirm this species-wide diversity.

This thesis work represents the first attempt to thoroughly analyze the TcTS gene family for recombination events. Using a novel computational pipeline, we documented frequent intra-family mosaicism in the TcTS gene family with signatures of gene conversion recombination, which is a potent mechanism for generation variations in this surface protein family. Examining the contribution of recombination to spreading protein sequence variation would require analysis of recombination profile of TcTS gene family in different time points of a chronic infection. If recombination could spread protein sequence variation to expressed TcTS, the selective pressure to present variant antigens exerted by host immune system would favor the survival of parasites with altered TcTS gene family recombination profile. Our pipeline can greatly facilitate the analysis of recombination profile of TcTS gene family, provided that high assembly-quality genome sequence is supplied.

Solid proof of immune evasion endowed by simultaneously expressing multiple TcTS variant antigens awaits procedures which can drastically reduce the number of simultaneously expressed TcTS genes. The CRISPR-Cas9 gene editing system may provide that opportunity. The thesis author has recently shown the ability in *T. cruzi* to substantially decrease the functional activity of proteins encoded by over 50 genes using the CRISPR-Cas9 system and just 3 guide RNAs (Peng, Kurup et al. 2015). The most highly conserved motifs in the TcTS proteins allow a relatively small number of gRNAs

to target > 1,000 TcTS genes. Thus, we believe it is well within reach to ask if mutation of large numbers of TcTS genes compromises the immune evasion capabilities of *T. cruzi*.

GLOSSARY

Major donor: In a gene conversion event, a major donor is a gene that receives sequence from a non-self origin. A significant proportion of the resulting mosaic gene sequence is "donated" by the major donor.

Minor donor: In a gene conversion event, a minor donor is a gene that donates a copy of a relative small proportion of its sequence to a recipient gene.

Mosaic gene: Mosaic gene is a gene with sequence from two or more origins.

REFERENCES

Agusti, R., et al. (2004). "Lactose derivatives are inhibitors of Trypanosoma cruzi trans-sialidase activity toward conventional substrates in vitro and in vivo." Glycobiology **14**(7): 659-670.

Albertti, L. A., et al. (2010). "Role of host lysosomal associated membrane protein (LAMP) in Trypanosoma cruzi invasion and intracellular development." Microbes Infect **12**(10): 784-789.

Atwood, J. A., 3rd, et al. (2005). "The Trypanosoma cruzi proteome." Science **309**(5733): 473-476.

Azuaje, F. J., et al. (2007). "In silico, biologically-inspired modelling of genomic variation generation in surface proteins of Trypanosoma cruzi." Kinetoplastid Biol Dis **6**: 6.

Barry, A. E., et al. (2007). "Population genomics of the immune evasion (var) genes of Plasmodium falciparum." PLoS Pathog **3**(3): e34.

Boni, M. F., et al. (2007). "An exact nonparametric method for inferring mosaic structure in sequence triplets." Genetics **176**(2): 1035-1047.

Buscaglia, C. A., et al. (1999). "Tandem amino acid repeats from Trypanosoma cruzi shed antigens increase the half-life of proteins in blood." Blood **93**(6): 2025-2032.

Cartwright, R. A. (2005). "DNA assembly with gaps (Dawg): simulating sequence evolution." Bioinformatics **21 Suppl 3**: iii31-38.

Carvalho, S. T., et al. (2010). "A new class of mechanism-based inhibitors for Trypanosoma cruzi trans-sialidase and their influence on parasite virulence." Glycobiology **20**(8): 1034-1045.

Chen, J. M., et al. (2007). "Gene conversion: mechanisms, evolution and human disease." Nat Rev Genet **8**(10): 762-775.

Chuenkova, M. V., et al. (2001). "Trypanosoma cruzi trans-sialidase: a potent and specific survival factor for human Schwann cells by means of phosphatidylinositol 3-kinase/Akt signaling." Proc Natl Acad Sci U S A **98**(17): 9936-9941.

Chuenkova, M. V. and M. PereiraPerrin (2009). "Trypanosoma cruzi targets Akt in host cells as an intracellular antiapoptotic strategy." Sci Signal **2**(97): ra74.

Confalonieri, A. N., et al. (1983). "Sialoglycolipids in Trypanosoma cruzi." Biochem Int **7**(2): 215-222.

Cross, G. A., et al. (2014). "Capturing the variant surface glycoprotein repertoire (the VSGnome) of Trypanosoma brucei Lister 427." Mol Biochem Parasitol **195**(1): 59-73.

Cross, G. A. and G. B. Takle (1993). "The surface trans-sialidase family of Trypanosoma cruzi." Annu Rev Microbiol **47**: 385-411.

Dc-Rubin, S. S. and S. Schenkman (2012). "T rypanosoma cruzi trans-sialidase as a multifunctional enzyme in Chagas' disease." Cell Microbiol **14**(10): 1522-1530.

de Titto, E. H. and F. G. Araujo (1988). "Serum neuraminidase activity and hematological alterations in acute human Chagas' disease." Clin Immunol Immunopathol **46**(1): 157-161.

Edgar, R. C. (2004). "MUSCLE: multiple sequence alignment with high accuracy and high throughput." Nucleic Acids Res **32**(5): 1792-1797.

El-Sayed, N. M., et al. (2005). "The genome sequence of Trypanosoma cruzi, etiologic agent of Chagas disease." Science **309**(5733): 409-415.

Erdmann, H., et al. (2009). "Sialylated ligands on pathogenic Trypanosoma cruzi interact with Siglec-E (sialic acid-binding Ig-like lectin-E)." Cell Microbiol **11**(11): 1600-1611.

Freitas-Junior, L. H., et al. (2000). "Frequent ectopic recombination of virulence factor genes in telomeric chromosome clusters of P. falciparum." Nature **407**(6807): 1018-1022.

Freitas, L. M., et al. (2011). "Genomic analyses, gene expression and antigenic profile of the trans-sialidase superfamily of Trypanosoma cruzi reveal an undetected level of complexity." Plos One **6**(10): e25914.

Gjini, E., et al. (2012). "The impact of mutation and gene conversion on the local diversification of antigen genes in African trypanosomes." Mol Biol Evol **29**(11): 3321-3331.

Hall, J. P., et al. (2013). "Mosaic VSGs and the scale of Trypanosoma brucei antigenic variation." PLoS Pathog **9**(7): e1003502.

Jackson, A. P., et al. (2012). "Antigenic diversity is generated by distinct evolutionary mechanisms in African trypanosome species." Proc Natl Acad Sci U S A **109**(9): 3416-3421.

Kahn, S., et al. (1991). "The major 85-kDa surface antigen of the mammalian-stage forms of Trypanosoma cruzi is a family of sialidases." Proc Natl Acad Sci U S A **88**(10): 4481-4485.

Kipnis, T. L., et al. (1981). "Enzymatic treatment transforms trypomastigotes of Trypanosoma cruzi into activators of alternative complement pathway and potentiates their uptake by macrophages." Proc Natl Acad Sci U S A **78**(1): 602-605.

Larkin, M. A., et al. (2007). "Clustal W and Clustal X version 2.0." Bioinformatics **23**(21): 2947-2948.

Leguizamon, M. S., et al. (1999). "Trans-sialidase from Trypanosoma cruzi induces apoptosis in cells from the immune system in vivo." J Infect Dis **180**(4): 1398-1402.

Lieke, T., et al. (2011). "Invasion of Trypanosoma cruzi into host cells is impaired by N-propionylmannosamine and other N-acylmannosamines." Glycoconj J **28**(1): 31-37.

Liskay, R. M., et al. (1987). "Homology requirement for efficient gene conversion between duplicated chromosomal sequences in mammalian cells." Genetics **115**(1): 161-167.

Magez, S., et al. (2008). "The role of B-cells and IgM antibodies in parasitemia, anemia, and VSG switching in Trypanosoma brucei-infected mice." PLoS Pathog **4**(8): e1000122.

Marcello, L. and J. D. Barry (2007). "Analysis of the VSG gene silent archive in Trypanosoma brucei reveals that mosaic gene expression is prominent in antigenic variation and is favored by archive substructure." Genome Res **17**(9): 1344-1352.

Martin, D. and E. Rybicki (2000). "RDP: detection of recombination amongst aligned sequences." Bioinformatics **16**(6): 562-563.

Martin, D. and R. Tarleton (2004). "Generation, specificity, and function of CD8+ T cells in Trypanosoma cruzi infection." Immunol Rev **201**: 304-317.

Martin, D. L., et al. (2006). "CD8+ T-Cell responses to Trypanosoma cruzi are highly focused on strain-variant trans-sialidase epitopes." PLoS Pathog **2**(8): e77.

Martin, D. P., et al. (2010). "RDP3: a flexible and fast computer program for analyzing recombination." Bioinformatics **26**(19): 2462-2463.

Martin, D. P., et al. (2005). "A modified bootscan algorithm for automated identification of recombinant sequences and recombination breakpoints." AIDS Res Hum Retroviruses **21**(1): 98-102.

Miller, M. A., et al. (2010). Creating the CIPRES Science Gateway for inference of large phylogenetic trees. Gateway Computing Environments Workshop (GCE), 2010. New Orleans, LA**:** 1- 8.

Ming, M., et al. (1993). "Mediation of Trypanosoma cruzi invasion by sialic acid on the host cell and trans-sialidase on the trypanosome." Mol Biochem Parasitol **59**(2): 243-252.

Mucci, J., et al. (2002). "Thymocyte depletion in Trypanosoma cruzi infection is mediated by trans-sialidase-induced apoptosis on nurse cells complex." Proc Natl Acad Sci U S A **99**(6): 3896-3901.

Mucci, J., et al. (2006). "The trans-sialidase from Trypanosoma cruzi triggers apoptosis by target cell sialylation." Cell Microbiol **8**(7): 1086-1095.

Mugnier, M. R., et al. (2015). "The in vivo dynamics of antigenic variation in Trypanosoma brucei." Science **347**(6229): 1470-1473.

Ohta, T. (1992). "A statistical examination of hypervariability in complementarity-determining regions of immunoglobulins." Mol Phylogenet Evol **1**(4): 305-311.

Oppezzo, P., et al. (2011). "Crystal structure of an enzymatically inactive trans-sialidase-like lectin from Trypanosoma cruzi: the carbohydrate binding mechanism involves residual sialidase activity." Biochim Biophys Acta **1814**(9): 1154-1161.

Padidam, M., et al. (1999). "Possible emergence of new geminiviruses by frequent recombination." Virology **265**(2): 218-225.

Padilla, A. M., et al. (2009). "Insufficient TLR activation contributes to the slow development of CD8+ T cell responses in Trypanosoma cruzi infection." J Immunol **183**(2): 1245-1252.

Parham, P. and T. Ohta (1996). "Population biology of antigen presentation by MHC class I molecules." Science **272**(5258): 67-74.

Peng, D., et al. (2015). "CRISPR-Cas9-mediated single-gene and gene family disruption in Trypanosoma cruzi." MBio **6**(1): e02097-02014.

Pereira, M. E. (1983). "A developmentally regulated neuraminidase activity in Trypanosoma cruzi." Science **219**(4591): 1444-1446.

Posada, D. and K. A. Crandall (2001). "Evaluation of methods for detecting recombination from DNA sequences: computer simulations." Proc Natl Acad Sci U S A **98**(24): 13757-13762.

Previato, J. O., et al. (1985). "Incorporation of sialic acid into Trypanosoma cruzi macromolecules. A proposal for a new metabolic route." Mol Biochem Parasitol **16**(1): 85-96.

Ronquist, F., et al. (2012). "MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space." Syst Biol **61**(3): 539-542.

Rosenberg, I. A., et al. (1991). "Differential expression of Trypanosoma cruzi neuraminidase in intra- and extracellular trypomastigotes." Infect Immun **59**(1): 464-466.

Roth, C., et al. (1991). "Antigenic variation in Trypanosoma equiperdum." Res Microbiol **142**(6): 725-730.

Rubin-de-Celis, S. S., et al. (2006). "Expression of trypomastigote trans-sialidase in metacyclic forms of Trypanosoma cruzi increases parasite escape from its parasitophorous vacuole." Cell Microbiol **8**(12): 1888-1898.

Ruef, B. J., et al. (1994). "Expression and evolution of members of the Trypanosoma cruzi trypomastigote surface antigen multigene family." Mol Biochem Parasitol **63**(1): 109-120.

Ruiz Diaz, P., et al. (2015). "Trypanosoma cruzi trans-sialidase prevents elicitation of Th1 cell response via interleukin 10 and downregulates Th1 effector cells." Infect Immun **83**(5): 2099-2108.

Saitou, N. and M. Nei (1987). "The neighbor-joining method: a new method for reconstructing phylogenetic trees." Mol Biol Evol **4**(4): 406-425.

Schauer, R., et al. (1983). "The occurrence of N-acetyl- and N-glycoloylneuraminic acid in Trypanosoma cruzi." Hoppe Seylers Z Physiol Chem **364**(8): 1053-1057.

Schenkman, R. P., et al. (1993). "Mammalian cell sialic acid enhances invasion by Trypanosoma cruzi." Infect Immun **61**(3): 898-902.

Schenkman, S., et al. (1991). "A novel cell surface trans-sialidase of Trypanosoma cruzi generates a stage-specific epitope required for invasion of mammalian cells." Cell **65**(7): 1117-1125.

Schenkman, S., et al. (1992). "Trypanosoma cruzi trans-sialidase and neuraminidase activities can be mediated by the same enzymes." J Exp Med **175**(2): 567-575.

Scudder, P., et al. (1993). "Enzymatic characterization of beta-D-galactoside alpha 2,3-trans-sialidase from Trypanosoma cruzi." J Biol Chem **268**(13): 9886-9891.

Smith, J. M. (1992). "Analyzing the mosaic structure of genes." J Mol Evol **34**(2): 126-129.

Takle, G. B. and G. A. Cross (1991). "An 85-kilodalton surface antigen gene family of Trypanosoma cruzi encodes polypeptides homologous to bacterial neuraminidases." Mol Biochem Parasitol **48**(2): 185-198.

Tavaré, S. (1986). "Some Probabilistic and Statistical Problems in the Analysis of DNA Sequences." Lectures on Mathematics in the Life Sciences (American Mathematical Society) **17**: 57-86.

Taylor, J. E. and G. Rudenko (2006). "Switching trypanosome coats: what's in the wardrobe?" Trends Genet **22**(11): 614-620.

Thompson, C. B. (1992). "Creation of immunoglobulin diversity by intrachromosomal gene conversion." Trends Genet **8**(12): 416-422.

Thon, G., et al. (1989). "Antigenic diversity by the recombination of pseudogenes." Genes Dev **3**(8): 1247-1254.

Todeschini, A. R., et al. (2004). "Enzymatically inactive trans-sialidase from Trypanosoma cruzi binds sialyl and beta-galactopyranosyl residues in a sequential ordered mechanism." J Biol Chem **279**(7): 5323-5328.

Todeschini, A. R., et al. (2002). "trans-Sialidase from Trypanosoma cruzi binds host T-lymphocytes in a lectin manner." J Biol Chem **277**(48): 45962-45968.

Tribulatti, M. V., et al. (2005). "The trans-sialidase from Trypanosoma cruzi induces thrombocytopenia during acute Chagas' disease by reducing the platelet sialic acid contents." Infect Immun **73**(1): 201-207.

Vazquez-Chagoyan, J. C., et al. (2011). "Vaccine development against Trypanosoma cruzi and Chagas disease." Adv Parasitol **75**: 121-146.

Vercelli, C. A., et al. (2005). "Trypanosoma cruzi trans-sialidase inhibits human lymphocyte proliferation by nonapoptotic mechanisms: implications in pathogenesis and transplant immunology." Transplant Proc **37**(10): 4594-4597.

Weirather, J. L., et al. (2012). "Mapping of VSG similarities in Trypanosoma brucei." Mol Biochem Parasitol **181**(2): 141-152.

Zhang, Z., et al. (2006). "KaKs_Calculator: calculating Ka and Ks through model selection and model averaging." Genomics Proteomics Bioinformatics **4**(4): 259-263.

Zhang, Z., et al. (2000). "A greedy algorithm for aligning DNA sequences." J Comput Biol **7**(1-2): 203-214.

a

Datasets/window size	Total gene number	Number of falsely detected recombinant gene	False positive rate
0% additional point mutation	1446	2	0.0014
1%additional point mutation	1446	3	0.0021
5% additional point mutation	1446	4	0.0028
10%additional point mutation	1446	0	0
Binning scan 50bp window size 20bp slide window	1446	4	0.0028
Binning scan 100bp window size 40bp slide window	1446	1	0.0007
Gene conversion	1446	3	0.002
Reciprocal exchange	1446	92	0.06

b

Types of recombination	Sensitivity (gene)	Specificity (gene)	Sensitivity (event)	Specificity (event)
Reciprocal exchange	0.01	0.003	0.003	0.0007
Gene Conversion recombination	0.93	0.54	0.73	0.44
Gene Conversion recombination (binning scan 50bp window size, slide by 20bp)	0.93	0.57	0.74	0.35
Gene Conversion recombination (binning scan 100bp window size, slide by 40bp)	0.85	0.45	0.62	0.55

Table 1: Summary of evaluation of the recombination detection pipeline using simulated data. a) false positive rates, b) sensitivity, and **s**pecificity of the recombination detection pipeline using simulated datasets of trans-sialidase gene family members, recombination positive data contains 200 recombinant genes with 1000 recombinant events in a total of 3209 synthetic TcTS sequences.

Algorithm name & reference	Parameters (that deviate from the default)
RDP (Martin and Rybicki 2000)	window size=125
GENECONV (Padidam, Sawyer et al. 1999)	none
Chimaera (Posada and Crandall 2001)	variable sites per window=42
MaxChi (Smith 1992)	variable sites per window=42
Bootscan (Martin, Posada et al. 2005)	window size=125, step size=20
3SEQ (Boni, Posada et al. 2007)	none

Table 2: Algorithms and respective parameters for the RDP package.

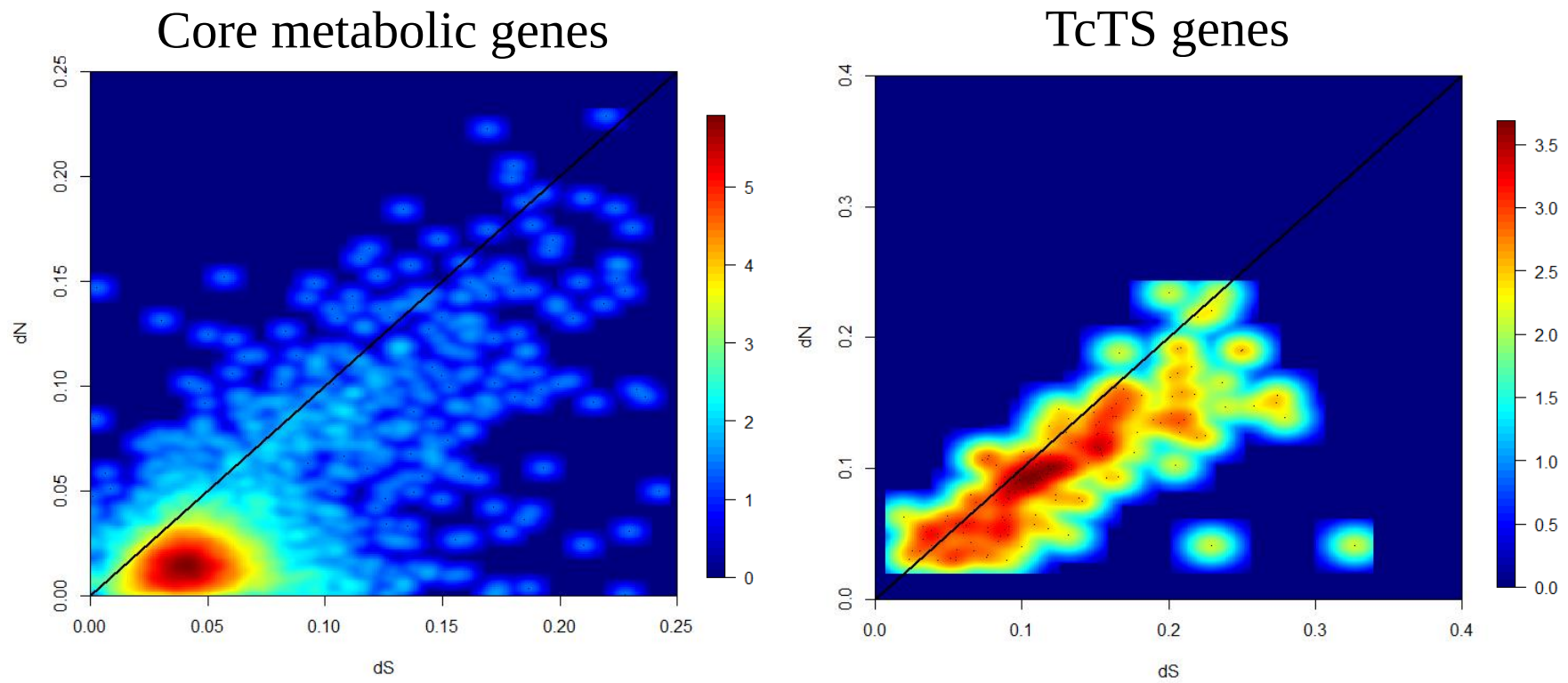

Figure 1: Heat map of synonymous and non-synonymous substitution rates (dS and dN respectively) of core metabolic genes and TcTS genes.

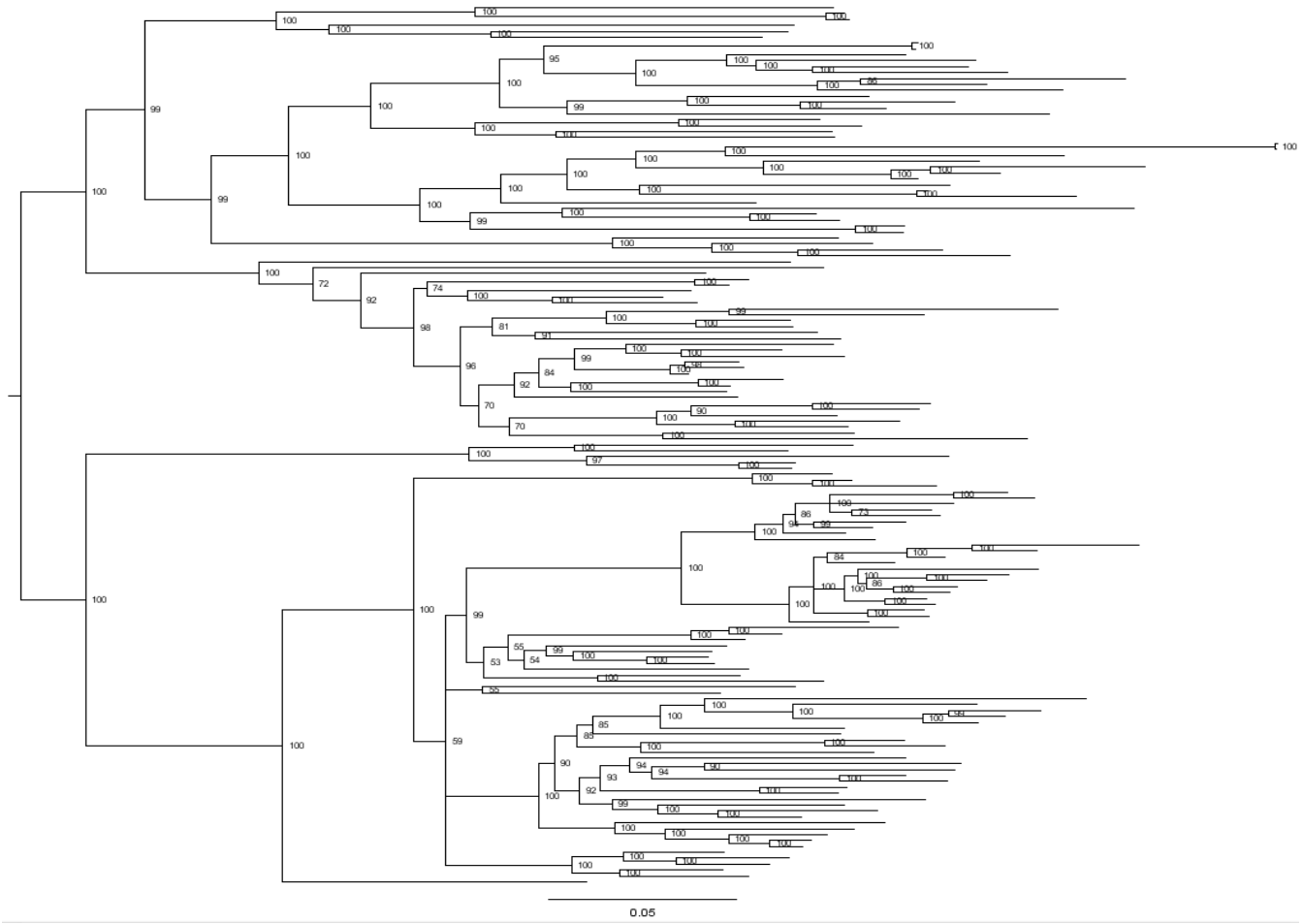

Figure 2: Bayesian phylogenetic tree of 149 closely related TcTS protein sequences showing long terminal branches. Treeness (= 0.363), is calculated by dividing the cumulative internal branch length by the total length of the tree.

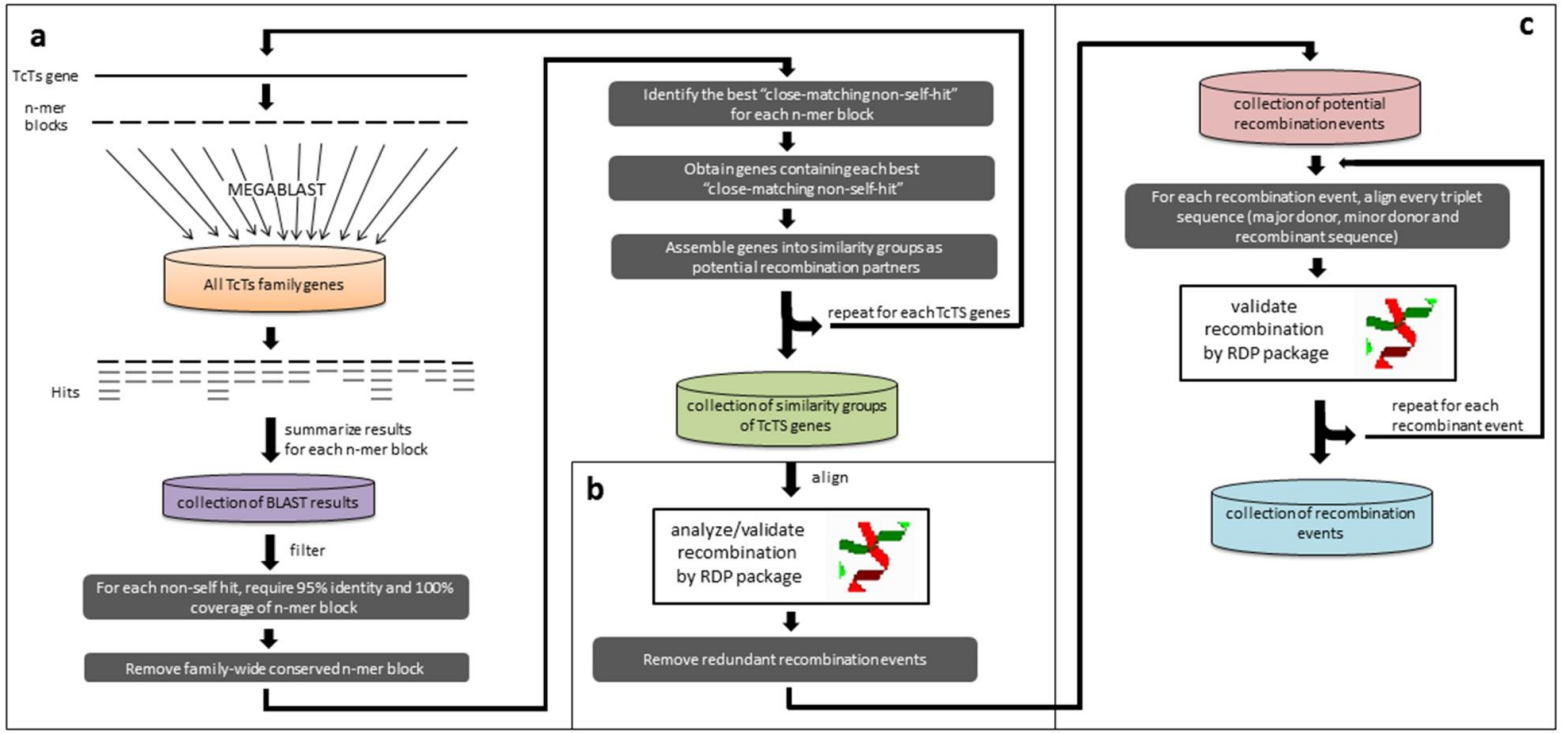

Figure 3: Workflow of recombination detection pipeline for the TcTS gene family. a) TcTS sequences are binned into similarity sets by scanning each TcTS gene for closest-matching non-self sequences within the TcTS gene family; b) RDP is used to identify possible recombinant regions, recombination breakpoints, and the major and minor sequence donor(s) among the TcTS similarity sets; c) Mosaic gene, major donor and minor donor from each detected recombination event in b) are realigned to obtain higher quality alignment, which will be analyzed by RDP to confirm corresponding recombination event.

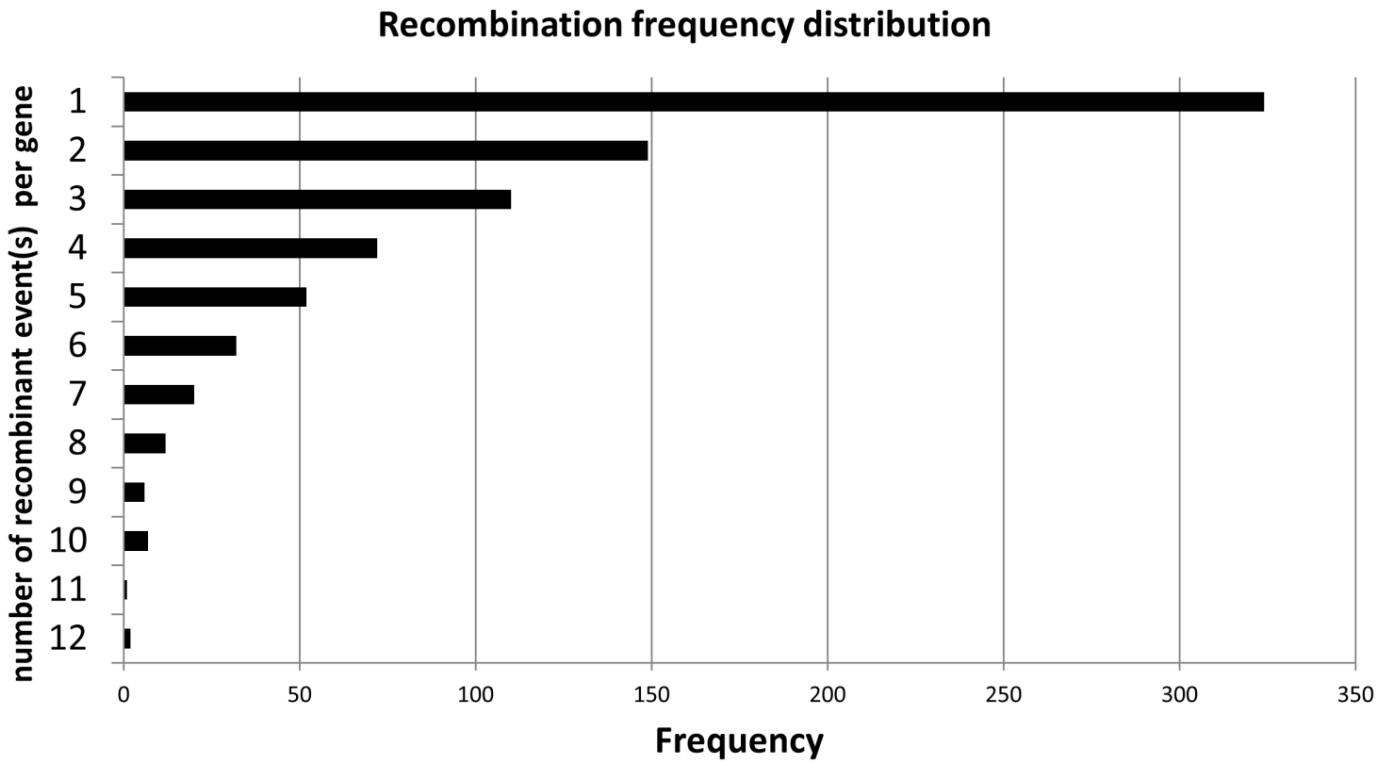

Figure 4: Recombination frequency distribution. The distribution of recombination frequency summarized for each of all TcTS gene analyzed in this study.

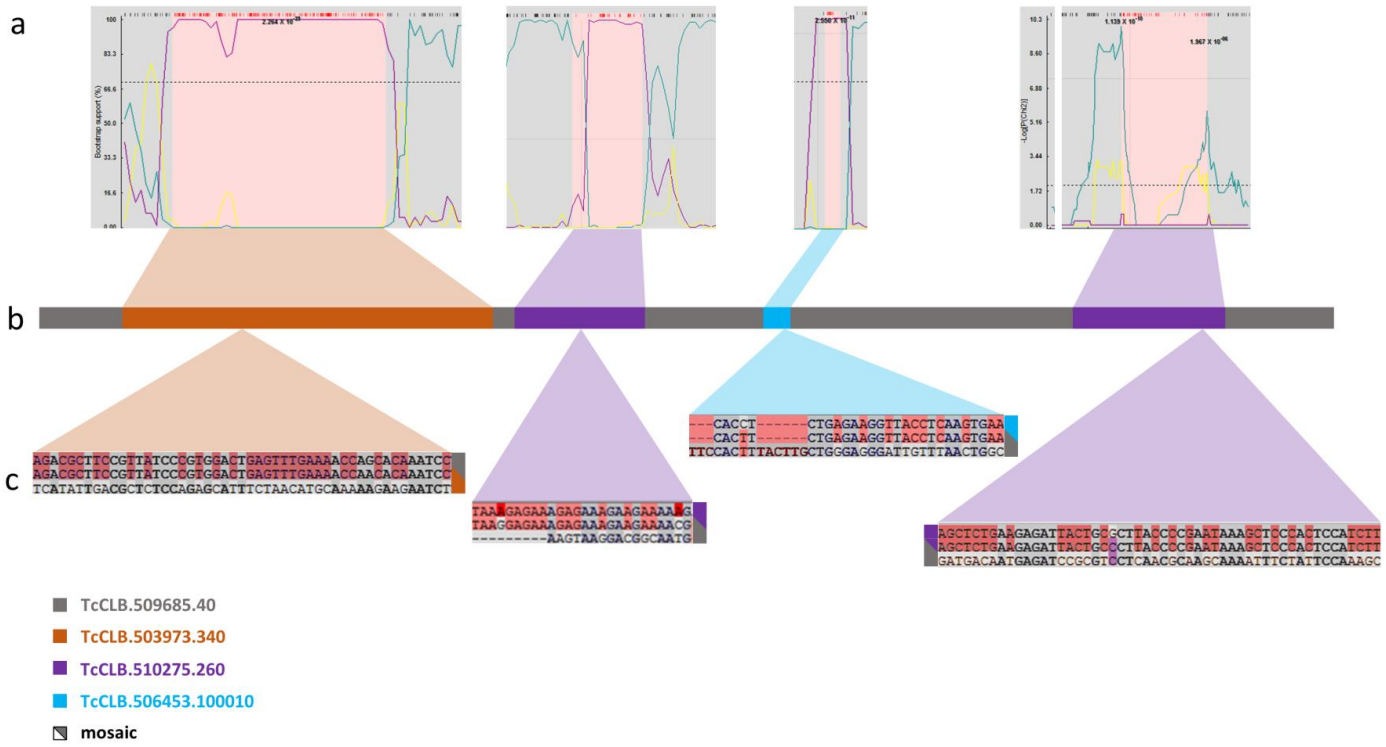

Figure 5: A mosaic TcTS gene showing evidence for recombination events and example sequence comparisons within each minor donor sequence. a) BOOTSCAN and MAXICHI output showing alterations in bootstrap support and peak log(p(χ^2)) values, respectively, at mosaic region boundaries. b) The TcCLB.509685.40 mosaic gene with donor contributions indicated by colors. c) Partial alignment of mosaic regions with respective donors. In each sequence triplet alignment, the middle sequence is the mosaic gene, shading highlights identical nucleotides of mosaic region with major (top) and minor (bottom) donor.

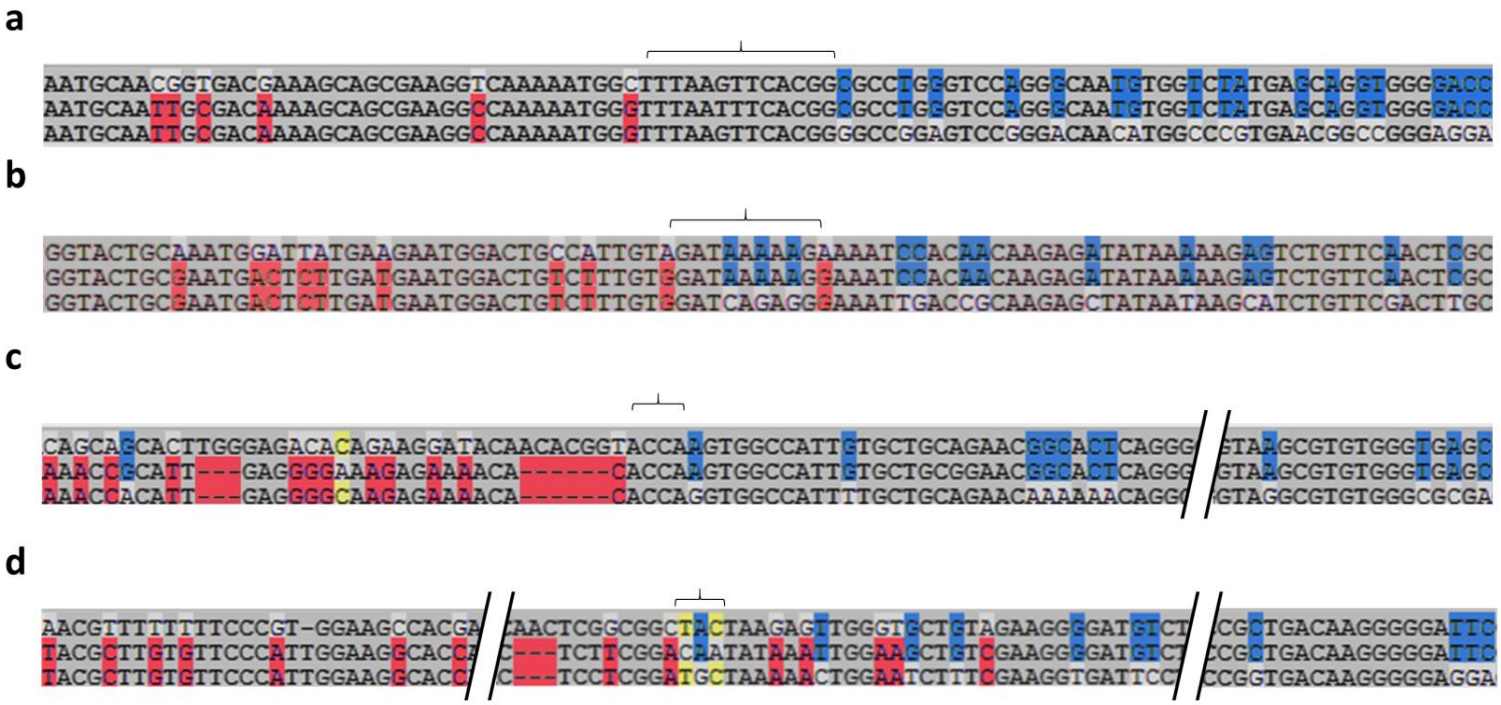

Figure 6: Examples of boundary regions in mosaic TcTS. From the mosaic TcCLB.509685.40 gene in Figure 4, boundary regions on one end of each of the 4 donor sequences are shown. In each sequence triplet alignment, the middle sequence is the mosaic gene, red and blue shading highlight identical nucleotides of mosaic region with major donor and minor donors. The brackets indicate regions most likely containing the recombination breakpoint. a) Donor sequence = TcCLB.503973.340; b) Donor sequence = TcCLB.510275.260; c) Donor sequence = TcCLB.506453.100010; d) Donor sequence = TcCLB.510275.260.

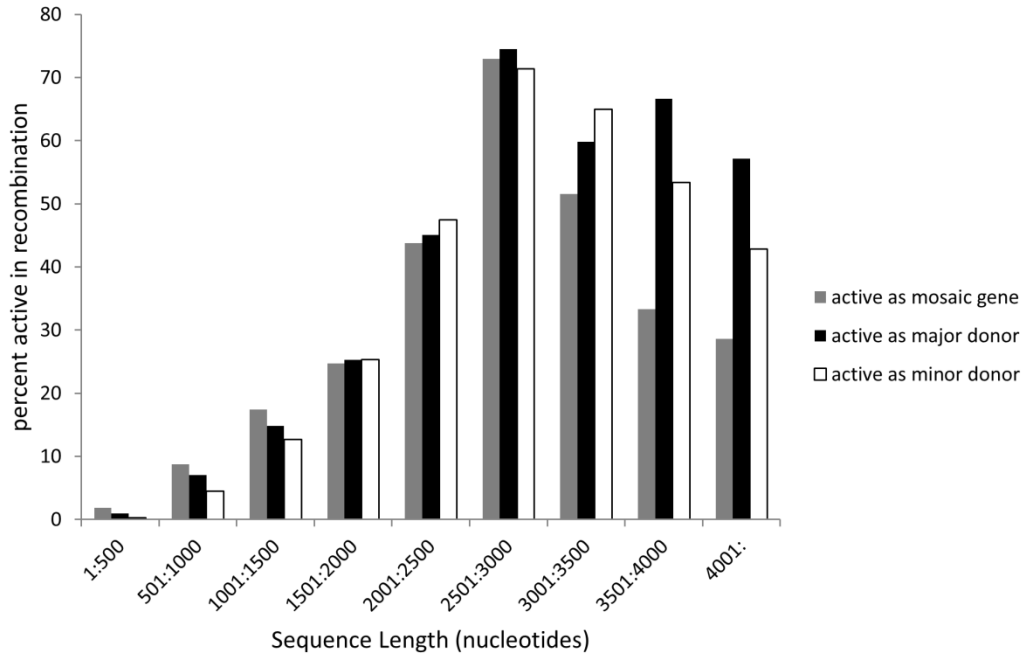

Figure 7: TcTS genes participating (being active) in recombination in relation to gene size. The fraction of TcTS genes in each size class that served as a minor donor, major donor or are the mosaic product of recombination. Percent active as mosaic gene was calculated by dividing the number of mosaic gene by the total number of TcTS genes for each category; Percent active as major donor was calculated by dividing the number of genes that took role (at least once) as major donor by the total number of TcTS genes for each category; Percent active as minor donor was calculated by dividing the number of genes that took role (at least once) as minor donor by the total number of TcTS genes for each category.

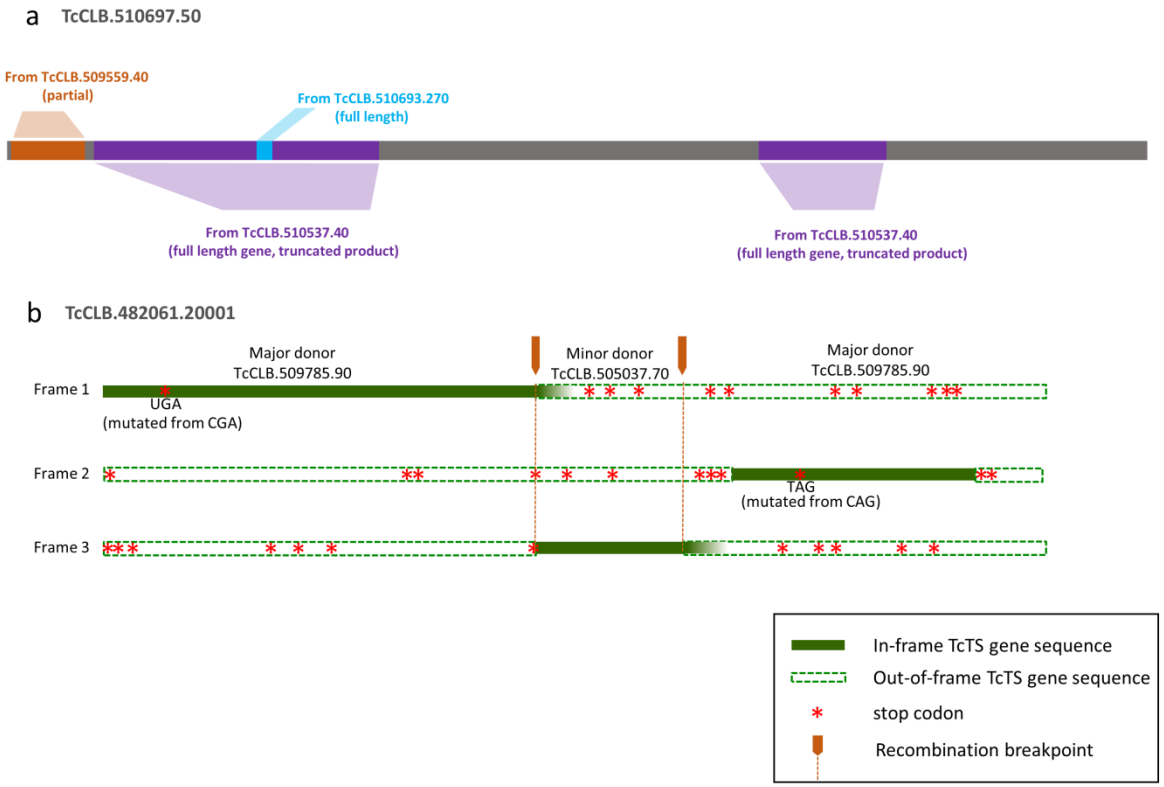

Figure 8: Productive and nonproductive recombination events. a) The full length, protein coding TcTS TcCLB.510697.50 has incorporated mosaic parts from partial TcTS gene and full length TcTS donors. b). Introduction of sequence from the full-length TcCLB.505037.70 gene into the wrong reading frame of the major donor TcCLB.509785.90 results in a partial TcCLB.482061.20001 product. Also note that 2 in-frame stop codons have resulted from single point mutations in the major donor supplied sequence.