genes

MDPI

Article

Local Genomic Instability of the *SpTransformer* Gene Family in the Purple Sea Urchin Inferred from BAC Insert Deletions

Megan A. Barela Hudgell [1,†], Farhana Momtaz [1,‡,§], Abiha Jafri [1,§,‖], Max A. Alekseyev [2] and L. Courtney Smith [1,*]

[1] Department of Biological Sciences, George Washington University, Washington, DC 20052, USA; mhudgell@unm.edu (M.A.B.H.); momtazf@gwmail.gwu.edu (F.M.)
[2] Department of Mathematics and the Computational Biology Institute, George Washington University, Washington, DC 20052, USA; maxal@gwu.edu
[*] Correspondence: csmith@gwu.edu
[†] Current addresses: Department of Biology, University of New Mexico, Albuquerque, NM 87131, USA.
[‡] Current addresses: Columbia Heights Education Campus, Washington, DC 20010, USA.
[§] These authors contributed equally to this work.
[‖] Current addresses: School of Osteopathic Medicine, Campbell University, Buies Creek, NC 27506, USA.

Abstract: The *SpTransformer* (*SpTrf*) gene family in the purple sea urchin, *Strongylocentrotus purpuratus*, encodes immune response proteins. The genes are clustered, surrounded by short tandem repeats, and some are present in genomic segmental duplications. The genes share regions of sequence and include repeats in the coding exon. This complex structure is consistent with putative local genomic instability. Instability of the *SpTrf* gene cluster was tested by 10 days of growth of *Escherichia coli* harboring bacterial artificial chromosome (BAC) clones of sea urchin genomic DNA with inserts containing *SpTrf* genes. After the growth period, the BAC DNA inserts were analyzed for size and *SpTrf* gene content. Clones with multiple *SpTrf* genes showed a variety of deletions, including loss of one, most, or all genes from the cluster. Alternatively, a BAC insert with a single *SpTrf* gene was stable. BAC insert instability is consistent with variations in the gene family composition among sea urchins, the types of *SpTrf* genes in the family, and a reduction in the gene copy number in single coelomocytes. Based on the sequence variability among *SpTrf* genes within and among sea urchins, local genomic instability of the family may be important for driving sequence diversity in this gene family that would be of benefit to sea urchins in their arms race with marine microbes.

Keywords: *Strongylocentrotus purpuratus*; short tandem repeats; tandem duplications; long-read assembly; large DNA deletions

check for updates

Citation: Barela Hudgell, M.A.; Momtaz, F.; Jafri, A.; Alekseyev, M.A.; Smith, L.C. Local Genomic Instability of the *SpTransformer* Gene Family in the Purple Sea Urchin Inferred from BAC Insert Deletions. *Genes* **2024**, *15*, 222. https://doi.org/10.3390/genes15020222

Academic Editor: Sónia C. Da Silva Andrade

Received: 9 January 2024
Revised: 2 February 2024
Accepted: 5 February 2024
Published: 9 February 2024

1. Introduction

There are many examples of gene copy number expansion into immune response families that function in innate immune systems among organisms. Many examples include cell surface, cytoplasmic, and secreted pathogen recognition receptors that bind microbes and/or their molecules and activate the innate immune response in the presence of possible pathogens or opportunists (e.g., [1–3]). The resulting gene families are under selection pressure for genes to be retained or to be deleted from the genome and perhaps to diversify based on pathogens and/or environmental variations. The outcome for the increased size and structure of the gene families results in benefits to the host for a broad recognition of the associated microbial populations and survival in its environment. However, the underlying mechanisms for the initial expansion of a single gene copy into a family is not understood. We have proposed previously that genomic instability is part of the complexity that results in gene expansion [4]. Genomic instability (also referred to as genomic fragility) has been the subject of many studies, a significant number debating its origin. Instability is associated with a high density of genomic sequence repeats [5], which are considered to be rapidly evolving compared to the rest of the genome [6]. These repeats are thought to be generated through processes such as conversion, recombination, slipping misalignments, and single-strand annealing [6,7]. Long regions of repeats can result in secondary structures, such as Z-DNA, bulge loops, hairpin loops, four loop junctions, and G-quadruplexes, that may impair cellular processes (reviewed in [6]). The impairment of cellular processes, including DNA replication and repair, can lead to segmental duplications [6,8–11], elevated recombination rates [12,13], duplications of tRNA genes [14,15], non-homogeneity of gene distribution [16], and expanded regulatory regions [17,18].

The *Transformer* (*Trf*) gene family (previously termed the *185/333* genes) encodes immune response proteins and has been identified in several species of euechinoids [19–22]. The *SpTrf* genes are present as a family in the purple sea urchin and are upregulated swiftly by phagocytes in adults responding to several different pathogen-associated molecular patterns [23–25] and marine bacteria [26–30], and by filipodial cells in the blastocoel of larvae [31]. Several aspects of the structure of the genes and the *SpTrf* family in the purple sea urchin, *Strongylocentrotus purpuratu*s, are consistent with genomic instability. The genes are small, with two exons, of which the second shows significant sequence diversity based on a variety of attributes [19]. It is composed of blocks of similar but non-identical sequences termed elements that are present in mosaic structures that are different among different genes (Figure 1A). Both tandem and interspersed repeats are also present in the second exon. The hypothetical evolutionary history of the tandem repeats suggests local duplications, ectopic duplications and insertions, deletions, and repeat recombination that resulted in two to four imperfect tandem repeats in the extant family [32,33]. In addition to the theoretical recombination among repeats, evidence of possible recombination among genes has also been reported [33]. Although there is significant sequence variability among the genes, they are 88% similar [19]. Consequently, the genes themselves may also be viewed as repeats. The repeats associated with the gene family, along with the sequence similarities among the genes themselves, are consistent with a region of local genomic instability in which evolutionary processes such as recombination may have key functions in sequence diversification and the evolution of both the genes and the family.

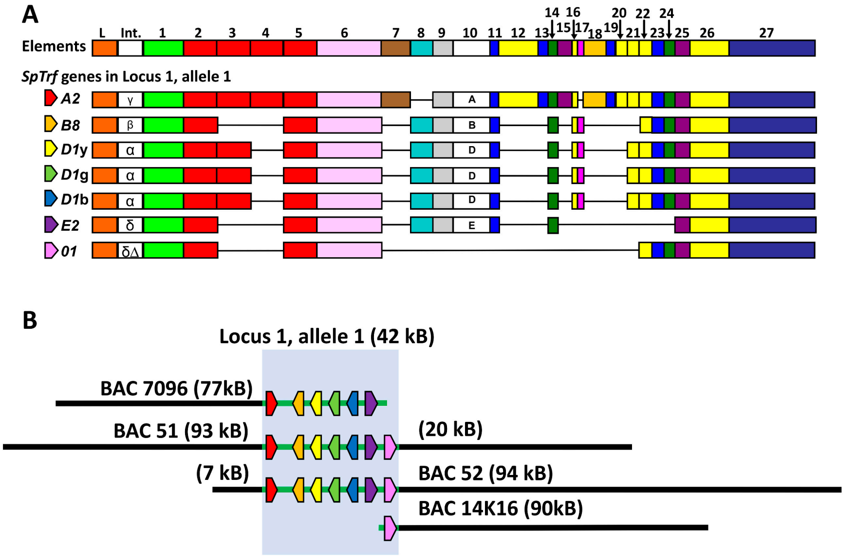

Figure 1. Graphical representation of BAC inserts that contain *SpTrf* gene locus 1, allele 1. (**A**) A graphical alignment of selected *SpTrf* genes showing the element pattern of each gene according to the repeat-based alignment [19]. Genes are composed of two exons; the first encodes the leader (L), and the second encodes the mature protein. All genes have a single, short intron (int) with several identifiable sequences that are indicated by the letters [19]. Exon 2 is a mosaic of elements (all of which are shown with numbers at the top) and is variable among the genes. Element 10 is labeled with different letters that indicate different sequences and define the element pattern. The horizontal

black lines indicate missing elements. This figure is modified from Figure 6B in [34]. (**B**) The overlaps among the BAC inserts containing allele 1 from locus 1 are highlighted with the blue box. The black lines to the left and right of the blue box indicate the size of the flanking regions of the inserts that are not part of the gene cluster. This figure is modified from Figure 4A in [35].

The characteristics of the *SpTrf* gene family structure are also consistent with genomic instability. The sequenced family, consisting of 17 alleles, is arranged in the sequenced genome (version 5.0) in two loci [36,37]. The genes are clustered and tightly linked, with intergenic distances ranging from 3.0 kb to 12 kb [32,35,38]. Locus 1 is composed of seven alleles on one chromosome and a mismatch of six alleles on the other (Figure 1B), whereas locus 2 has two genes, each with associated alleles [32]. Short tandem repeats (STRs) composed of two nucleotides, GA, surround each gene, and three regions of GA repeats of 3 kb to 4 kb are present in locus 2 in locations that appear to correspond with the locations of genes in locus 1 [32,35]. A different STR with the repeated sequence of GAT is present in multiple locations within both loci. In addition to the variety of repeats within and surrounding the genes, segmental duplications of 4.3 kb are also present that include duplicated genes, some so recent that their sequences are nearly identical [38], whereas other duplications have quite different sequences [32]. The genome assembly and the variety of BAC inserts that have been sequenced illustrate that the *SpTrf* family loci are riddled with a wide variety of repeats [32,35,37,38].

Genomic instability has been suggested previously, based on the descriptions of the genes and the *SpTrf* gene family structure described above [4]. Quantitative PCR was used to estimate the gene copy numbers in genomes from individual sea urchins, which indicated about 40 to 60 genes per genome [25]. However, the current genome assembly (version 5.0) shows only nine genes in two loci, many of which are annotated incorrectly (Echino-base.org, accessed on 26 October 2023). This underrepresentation has been attributed in part to a computational assembly problem that recognizes the multiple *SpTrf* genes as repeats or alleles and collapses them into single consensus genes, which appear as mixed sequences of two or more alleles or genes [35,38]. Furthermore, the *SpTrf-01* gene identified in a BAC insert sequence and a member of allele 1 in locus 1 [35] is not present in the genome. In an effort to overcome these assembly problems and present a corrected gene family structure, the BAC library that was sequenced for genome assembly (13X coverage of the genome; [39]) was screened for clones harboring *SpTrf* sequences [35]. It was notable that although 75 clones screened positive for *SpTrf* sequences, only 27 BAC clones supported PCR amplification of *SpTrf* sequences after *E. coli* was grown for BAC DNA isolation and analysis. Although this difference might have been the result of a failure of the PCR to amplify many genes for a variety of possible reasons, we propose that an alternative explanation is the instability of the BAC inserts based on the repeats and tight gene clustering that resulted in DNA deletions. This notion is in agreement with the underrepresentation of the *SpTrf* gene copy number in the sequenced genome that was assembled from overlapping BAC insert end sequences [37,40,41]. This hypothesis was tested with repeated inoculation and growth of *E. coli* harboring BAC clones for 10 days. BAC DNA with multiple *SpTrf* genes that was isolated from single colonies after the 10-day growth period had a variety of deletions. Alternatively, the BAC insert with only a single SpTrf gene was stable. Based on these results, we propose that local genomic instability is active in the *SpTrf* gene family in *S. purpuratus* and may be under cellular control to block deletion of the entire gene family. We propose that it is an underlying mechanism for the variability among the genes in the family, which includes sequence diversification that is a benefit for sea urchins in the arms race with marine microbial pathogens.

2. Methods

2.1. BAC Plasmids, DNA Isolation, and Sequencing

BAC clones were chosen based on their *SpTrf* gene copy numbers (Table 1) [35], which ranged in gene content from a single gene to the full cluster of seven genes in locus 1, allele 1 (Figure 1A). *E. coli* harboring pBACe3.6 plasmids with sea urchin genomic (g)DNA inserts were spread on LB plates with 12.5 µg/mL chloramphenicol (LB/chlor) and grown overnight at 37 °C. Single colonies were inoculated into 5 mL of LB/chlor media and grown overnight at 37 °C with rotation. Bacteria in 1 µL of media were re-inoculated in 5 mL fresh media and grown overnight, which was repeated for 10 days. Bacteria in the final broth culture were spread on LB/chlor plates, single colonies (34–40) were selected per BAC and expanded in LB/chlor media, and BAC plasmid DNA was isolated by the alkaline lysis method as described in [19]. All control BACs (BAC-con) were isolated from *E. coli* harboring the pBACe3.6 plasmids with sea urchin gDNA inserts after a single day of growth as described above. BAC DNA was digested with *Not*I to release the insert from the plasmid and in combination with *Xho*I or *Sac*II, and the inserts were evaluated by pulsed field gel electrophoresis (PFGE; CHEF-DR II Chiller System, Bio-Rad item #1703727) with 1% pulsed field certified agarose (Bio-Rad Laboratories, Hercules, CA, USA) in 0.3X concentration Tris borate EDTA (TBE) gel running buffer (27 mM Tris (pH 7.4), 27 mM boric acid, 0.6 mM EDTA). The PFGE parameters were 6 V/cm, with a switch time of 1–15 s over 16 h [35]. After separation, the gels were soaked in ethidium bromide solution and imaged with a UV imaging system (Gel Logic 1500 Imaging System, Kodak, Rochester, NY, USA). The locations of possible deletions were predicted by virtual restriction enzyme digests (https://nc3.neb.com/NEBcutter; accessed on 18 May 2018; New England Biolabs, Ipswich, MA, USA) based on the full-length insert sequence of each BAC, based on a previous report [42] or as reported here.

Table 1. BACs with multiple *SpTrf* genes show changes in insert sizes and gene copy numbers after a 10-day growth period.

		BAC Insert Screens		**SpTrf Gene Copies in BAC Inserts**		
BAC Clone Name [1]	**GenBank Accession Number**	**Colonies Screened for BAC Insert Size**	**BACs with Deletions**	**Colonies Screened**	**Gene Copies Expected**	**Gene Copies in 8 BACs with Short Inserts**
BAC-7096	BK007096	40	4 (10%)	8	3	3, 2, 2, 3, 3, 3, 3, 3
BAC-51	KU668451	34	1 (3%)	8	4	4, 4, 4, 3, 4, 4, 4, 4
BAC-52	KU668452	40	3 (7.5%)	8	4	4, 0, 4, 0, 4, 1, 4, 2
BAC-67	PP082967	40	0 (0%)	8	1	1, 1, 1, 1, 1, 1, 1, 1

[1] Abbreviated BAC accession numbers are used as clone names.

The *SpTrf* genes in the BAC inserts were identified by PCR amplicon sizes using primers that amplified all *SpTrf* genes based on annealing sites in the untranslated regions (5′UTR-forward: TAGCATCGGAGAGACCT; 3′UTR-reverse: AAATTCTACACCTCGGC-GAC), as described previously [25]. Amplicons were separated by gel electrophoresis and visualized with the Kodak UV imaging system.

2.2. BAC Insert Assembly from Sequencing Reads, Alignments, and Dot Plots

BAC DNA (n = 3) from single colonies was isolated after 10 days of re-inoculations for clones that showed smaller inserts compared to the full-length original BACs. In addition, BAC-con DNA (n = 2) from single colonies was isolated after a single day of growth and showed no insert size differences compared to the full-length original BAC inserts. BACs were grown in LB/chlor media overnight at 37 °C, and the plasmid DNA was isolated by a NucleoBond BAC100 high molecular weight DNA kit (Machery-Nagel Inc., Allentown, PA, USA) and evaluated for *E. coli* genomic DNA contamination by PCR (primers: Ecoli-1F, CGAAGCGACTGGAGCATGTG; Ecoli-1R, ACGCCACATTCGC-CAATTC) compared to amplicons of the plasmid (pBACe3.6F, AGCCGTGTAACCGAG-CATAGC; pBACe3.6R, GGAACATGACGGTATCTGCGAG). The reaction used PrimeSTAR GXL DNA polymerase (Takara Bio USA, San Jose, CA, USA), and the PCR program was an initial DNA melt at 98 °C for 30 s, 30 cycles of 98 °C for 10 s, 60 °C for 15 s, and 68 °C for 1 min, with a 68 °C extension and 4 °C hold. The amplicons for both pairs of primers were the same size, and inserts from BAC DNA with low levels of contaminating genomic DNA from *E. coli* were sequenced at the University Maryland Institute for Genome Sciences (https://marylandgenomics.org/) using long-read technology (Pacific Biosciences, Menlo Park, CA, USA). BAC insert sequences were assembled using Canu version 2.2 (HiCanu; https://github.com/marbl/canu/releases, accessed on 27 September 2022) [43–45] using the following parameters: genome size = 0.03 M, corErrorRate = 0.045, batOptions = "−eg 0.0 −sb 0.001 −dg 3 −dr 0 −ca 2000 −cp 200", and mhapPipe = false.

Contigs covering each of the BAC inserts that were returned from the HiCanu assembly were aligned by hand in Molecular Evolutionary Genetics Analysis X (MEGAX) against the relevant original BAC insert sequence [35], and a consensus sequence was generated using EMBOSS Cons (https://www.ebi.ac.uk/Tools/msa/emboss_cons/ accessed on 27 September to 15 November 2023) from contigs returned from HiCanu assembly. The consensus sequence was used for further analysis. Sequence assemblies of BAC inserts were verified by BLAST searches of the sequencing reads against the original BAC sequences (GenBank accession numbers KU668451 and KU668452) that have been reported previously [35,38]. Insert sequences for BAC-51-15 and BAC-52-2b are available from GenBank (accession numbers PP082968 and PP082969, respectively). Sequence reads for BAC-42 and BAC-44 are available as raw sequence reads from GenBank (BioSample accession numbers SAMN39322606 and SAMN39322605, respectively).

Dot plots were generated using the YASS genomic similarity search tool (https://bioinfo.univ-lille.fr/yass/index.php, accessed on 3 October 2023) [46] to visualize the deletions in each BAC with a short insert against the original BAC insert sequence. The e-value threshold was set to e^{-30}, with the rest of the parameters left at standard settings (scoring matrix: +5, −4, −3, −4; gap costs: −16, −4; X-drop threshold: 30).

Raw PacBio reads for the BAC inserts were mapped against the relevant original BAC insert sequence [35] using minimap2 2.1 with preset parameters [47,48]. The output file was converted from a .sam file to a .bam file, sorted, and indexed using samtools 1.6 [49] before visualization in The Integrative Genomics Viewer (version 2.15.2) [50–53].

2.3. Southern Blots and Riboprobes

BAC clones were digested with *Sal*I and *Not*I, and fragments were separated by gel electrophoresis and transferred to a GeneScreen Plus hybridization membrane (PerkinElmer, Waltham, MA, USA) by capillary blotting [35,54]. The filter was evaluated with ^{32}P-riboprobes generated with RNA polymerases from linearized gene clones that served as templates to incorporate ^{32}P labeled ribonucleotides as described in [23,35,55]. The gene clones chosen for templates had an *A6* element pattern (GenBank accession number EF607716.1), a *B3* element pattern (EF607770.1), and a *D1* element pattern (EF607784.1) [19]. After hybridization with the probes, the filter was exposed to X-ray film at −80 °C, which was processed with developer and fixer, scanned with epi-white light, and imaged with the Kodak Gel Logic 1500 Imaging System.

3. Results

3.1. Sea Urchin Genomic DNA Harboring the SpTrf Gene Family Is Unstable

Genomic instability is predicted for regions of DNA that contain many types of repeats including tightly linked genes with similar sequences [6,56], such as the *SpTrf* genes. An initial characterization of BAC DNA was carried out by *Not*I restriction digests, which released the insert from the pBACe3.6 vector to evaluate size. This approach identified 3% to 10% of the colonies from which BAC DNA was isolated after the 10-day growth period and had inserts that were smaller than expected when they included more than one *SpTrf* gene (Table 1). BAC-51-15 (see Table 1 for the BAC naming conventions) had a small insert compared to BAC-51-con (Figure 2A, blue vs. white arrows), and the corresponding gene amplicons indicated that the *SpTrf-A2* gene was missing (Figure 2B, blue vs. white arrows). The other BACs that originated from BAC-51 all had full-length inserts and a full complement of genes (Figure 2A,B). BAC DNA isolated from *E. coli* colonies that contained

BAC-52 showed a variety of changes to the insert sizes (Figure 2C). BAC-52-19, BAC-52-2b, and BAC-52-4 all had small inserts compared to BAC-52-con (Figure 2C, colored arrows) and showed varying numbers of missing *SpTrf* genes (Figure 2D, colored arrows). BAC-52-19 did not support amplification of any *SpTrf* gene, indicating that all were missing (Figure 2D, red arrow). There was a single gene amplicon from BAC-52-2b, which indicated that *SpTrf-A2* was present, but the other genes were missing (Figure 2D, green arrow). BAC-52-4 (Figure 2C,D, yellow arrows), however, showed all gene amplicons, indicating that the full complement of *SpTrf* genes were present (Figure 2D, yellow arrow). Of additional note was BAC-52-4c, which had a full-length insert (Figure 2C, purple arrow), but none of the *SpTrf* genes were amplified (Figure 2D, purple arrow). Al-though the BACs with small inserts showed a variable loss of *SpTrf* genes, the majority of the deletions involved the *SpTrf* gene cluster. In contrast, BAC DNA isolated from multiple colonies that contained BAC-67 with only a single *SpTrf* gene showed no deletions (Table 1, Figure 2E), and the single *SpTrf* gene was maintained (Figure 2F). These results suggested that the repeats within the *SpTrf* cluster were associated with most of the DNA deletions, whereas a single gene separated from the rest of the cluster and with fewer repeats in the insert did not result in DNA deletions.

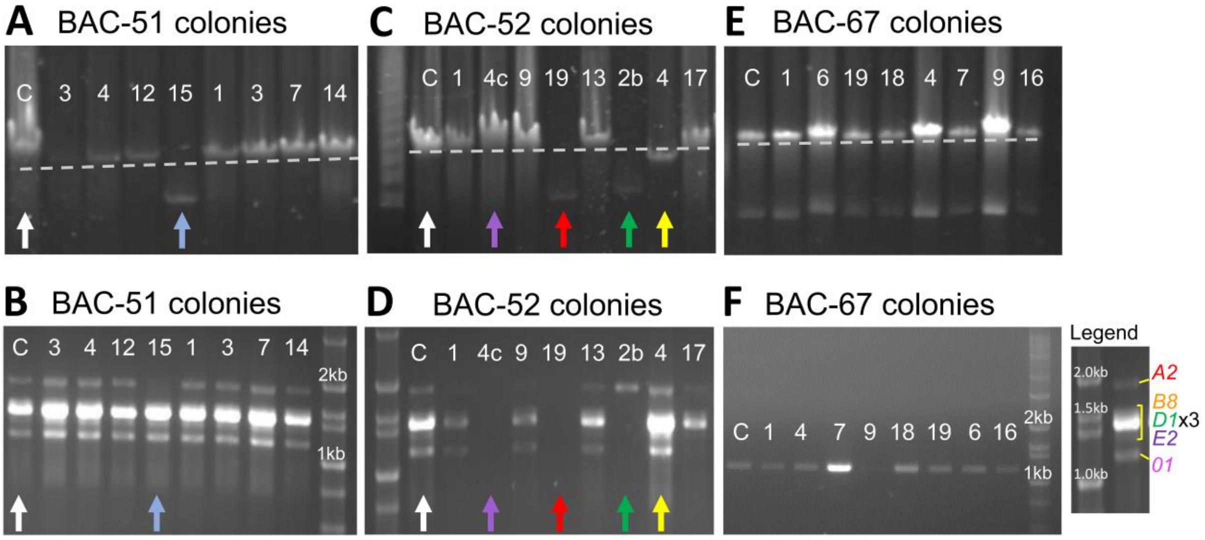

Figure 2. BACs isolated from some colonies after 10 days of growth have small inserts. (**A**,**C**,**E**) Representative *Not*I digests of four BACs containing *SpTrf* genes in locus 1, allele 1 identify BACs with decreased insert sizes. BAC-con clones (C lanes, white arrows) that were grown overnight once show full-length inserts. The C lanes are followed by eight representative samples of BAC DNA isolated from single colonies after 10 days of growth. The colored arrows indicate BACs with small inserts. The first lane in (**C**) shows a PFGE lambda ladder (Bio-Rad Laboratories, Hercules, CA, USA). (**B**,**D**,**F**) PCR amplicons of genes in BAC inserts illustrate the genes in the cluster and identify the genes that are missing after 10 days of growth. Genes were amplified using primers specific for *SpTrf* genes (see Materials and Methods). The colored arrows in (**A**,**C**) correspond to the arrows in (**B**,**D**), respectively. Colored arrows indicate the BACs in which the gene copy number is different from the BAC-con clones (white arrows). The first or last lanes of the gels show the DNA ladder (Hi Lo DNA Standard, Fisher Scientific, Hampton, NH, USA). The legend for *SpTrf* gene amplicons shows the bands that correlate with individual genes based on sizes. *SpTrf-A2* is the largest gene; the cluster of bands of intermediate size includes *SpTrf-B8*, *-D1*, and *-E2* (three *D1* genes result in the strongest bands in the center); and *SpTrf-01* is the smallest. The marker is the Hi Lo DNA standard.

3.2. Deletions to the BAC Inserts Are Positioned in a Variety of Locations

Although the analysis of BAC insert size and gene copy number indicated a variety of DNA deletions, these results did not identify the locations of the deletions or whether there could be more than one deletion in an individual BAC insert. To address this question, the

BAC-con clones and three BAC clones with expected deletions were digested with either *Xho*I/*Not*I or *Sac*II/*Not*I. To predict which fragments corresponded to regions of the BAC inserts to aid in the evaluation of the actual digests, virtual digests were used to evaluate full-length BAC insert sequences for BAC-51 and BAC-52 (Table 1). The virtual double digests generated a wide size distribution of bands and located the seven *SpTrf* genes in specific bands (Figure 3A). While most of the *SpTrf* genes were positioned on the same fragment because of their tight clustering (gene colors in Figure 3A,B are indicated in Figure 1), *SpTrf-A2* and *SpTrf-01* were more likely to be located on different fragments, which correlated with their distant locations at the edges of the cluster. These two genes have larger intergenic regions of 12 kb and 7.3 kb, respectively [35]. However, the *Xho*I/*Not*I digest for BAC-52 showed all genes on the same fragment except for *SpTrf-A2* (Figure 3A). The virtual digest results provided a framework for interpreting the fragments resulting from the actual digests.

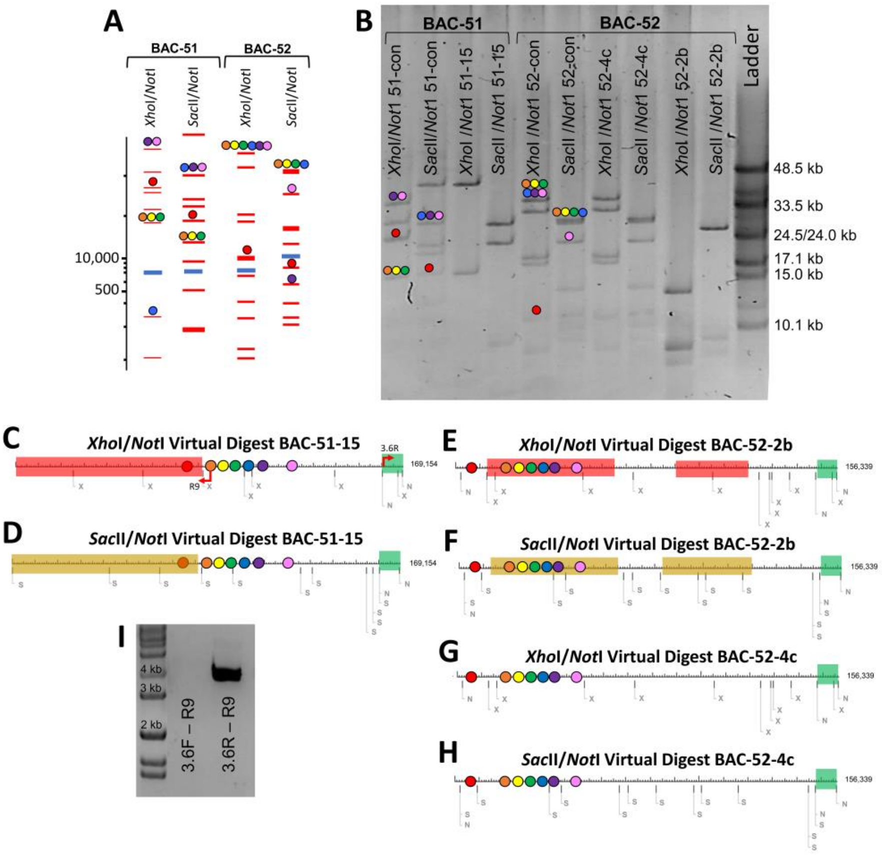

Figure 3. Locations of insert deletions are predicted by restriction digests. (**A**) *Xho*I/*Not*I and *Sac*II/*Not*I virtual digests (https://nc3.neb.com/NEBcutter/ accessed on 18 May 2018) of full-length BAC insert sequences show the fragment sizes and on which fragments the *SpTrf* genes are located. (**B**) *Not*I double digests with *Xho*I or *Sac*II show altered band sizes from BAC clones with small inserts and multiple *SpTrf* genes. BAC-51-con and BAC-52-con with full-length inserts show bands of expected size based on the virtual digests. The ladder is lambda DNA (monocot λ mix; New England Bio-Labs). (**C**–**H**) Maps of virtual digests predict the locations of the deleted regions. The maps are shown in linear format, even though the BAC DNA is circular. The areas in red and yellow indicate the predicted deletions based on results in (**B**). The *SpTrf* genes are indicated by colored dots based on the gene colors in Figure 1. The green segment indicates the pBACe3.6 vector. Restriction endonuclease sites in the BAC DNA are indicated as N, *Not*I; X, *Xho*I; S, *Sac*II. (**I**) PCR is used to orient and verify the size change in BAC-51-15 (see **C**). Primers (pBACe3.6F or pBACe3.6R) that surround the vector and the R9 primer that is located within each gene (see red arrows in (**C**)) confirm a ~90 kb deletion in BAC-51-15 that results in a 4 kb amplicon. The gel image is edited to delete irrelevant lanes and bring the DNA standard (Hi Lo; Fisher Scientific) next to the lanes of interest.

The actual digests with *Not*I and either *Xho*I or *Sac*II of the DNA from BAC-51-con, BAC-52-con, and BACs with short inserts (Figure 3B) were compared to the virtual digests (Figure 3A). Differences were used to identify which bands were absent or had changed in size to deduce which regions of the BAC inserts had undergone deletions. Two of the three BACs showed evidence of deletions. The digests of BAC-51-15 indicated a large deletion in which most of the expected bands were missing (Figure 3B). The *Xho*I/*Not*I digest showed a loss of all but one band and a large change in size of another, which the *Sac*II/*Not*I digest showed as a loss of all but three bands. When these deletions and size changes were mapped onto the virtual digest of BAC-51, results predicted a deletion of ~90 kb including the region expected to contain *SpTrf-A2* (Figure 3B–D). This verified the *SpTrf* gene amplicon results for BAC-51-15 in which the *SpTrf-A2* gene was missing (Figure 2D), thereby locating the position of the deleted region within the insert (Figure 3C,D). This deletion was also verified by PCR using the R9 primer (the annealing site is within the coding region of all *SpTrf* genes) and either the pBACe3.6F primer or the pBACe3.6R primer (annealing sites are at the ends of the vector; see red arrows in Figure 3C). An amplicon of 4 kb was generated from BAC-51-15 with pBACe3.6R, which was only feasible if a large deletion had brought the R9 annealing site in *SpTrf-B8* (orange dot) into close proximity to the vector (Figure 3C,I). Results for BAC-52-2b suggested two deletions, of which one removed all genes except for *SpTrf-A2* (Figure 3B,E,F). The second deletion in BAC-52-2b did not alter the *SpTrf* gene cluster; however, it indicated that multiple deletion events could occur in BAC inserts. BAC-52-4c was digested to determine whether any deletions were present given that it showed no change in insert size based on the *NotI* digest compared to the BAC-52-con (Figure 2C, purple arrow) and that it failed to amplify any of the *SpTrf* genes (Figure 2D, purple arrow). Results indicated that there were no differences in the digests for BAC-52-4c and the digests for the BAC-52-con (Figure 3G,H), suggesting that it had no deletions and that the absence of gene amplicons was likely a technical PCR failure (Figure 2D, purple arrow). These findings predicted that inserts that were smaller than expected for some BAC clones were the outcome of deletions. This also suggested that BAC inserts with multiple *SpTrf* genes that are associated with many types of repeats in the gene cluster [35,38] were the basis of the deletions. This was supported by results from BAC-67 with a single *SpTrf* gene and far fewer associated repeats, which did not undergo deletions (Figure 2E,F).

3.3. BAC Insert Sequencing and Assembly Verifies Gene Loss and Identifies the Edges of Deletions

To verify the results indicating BAC insert deletions based on changes in insert sizes and gene copy numbers, five BACs were sequenced (PacBio) and assembled into full-length insert sequences. BAC-51-15 and BAC-52-2b were selected based on predicted deletions, BAC-52-4c was selected because it maintained the full-length insert after the 10-day growth period but did not appear to have maintained the *SpTrf* genes (Figure 2C,D), and BAC-51-con and BAC-52-con were selected after growth for a single day. Notably, the preparation of large quantities of BAC DNA for sequencing required at least one and often more than one round of growth to acquire enough DNA of sequencing quality. The assembled insert sequences were aligned by hand to either BAC-51 or BAC-52 (Table 1) to verify the locations of the deletions (Supplementary File). Unexpectedly, all sequenced BACs had deletions, including BAC-51-con, BAC-52-con, and BAC-52-4c, which had been identified as full-length without deletions (Figure 2A,C; Table 2). Dot plots were used to illustrate the *SpTrf* gene cluster in BAC-51 and BAC-52 (Figure 4A,B) and to show the locations of deletions in the BACs that underwent the 10-day growth period (Figure 4C–G). The single deletion in BAC-51-15 was consistent with the size and location predicted by the virtual digests and included *SpTrf-A2* (Figure 4C). BAC-52-2b showed one large deletion that removed all of the *SpTrf* genes except for *SpTrf-A2* (Figure 4D) that was not consistent with the virtual digest predictions that suggested two smaller deletions (Figure 3E,F). It appeared that the two predicted deletions may have progressed to a single larger deletion that deleted the region between the two smaller deletions during the preparation for sequencing. Surprisingly, during the preparation of DNA for BAC-51-con, BAC-52-con, and BAC-52-4c for sequencing, these BACs appeared to have also acquired deletions. The insert assembly and dot plot for BAC-51-con showed that, rather than a full-length insert with all *SpTrf* genes in the cluster, three deletions had occurred that truncated *SpTrf-A2* and *SpTrf-E2* and deleted *SpTrf-01* (Figure 4E). Similarly, the dot plot for BAC-52-con compared to BAC-52 also showed a deletion, which truncated *SpTrf-E2* and deleted five genes, although *SpTrf-01* remained (Figure 4F). Furthermore, dot-plot results for BAC-52-4c compared to BAC-52 showed that there were three deletions that removed about 100 kb, including *SpTrf-01*, and the other two deletions were located outside of the gene cluster (Figure 4G). This size change was not evident in the gel image of the *Not*I released insert, which appeared as the same size as BAC-52 (Figure 2C, purple vs. white arrows). The inconsistency within the results required further analysis to identify the source.

Table 2. The *SpTrf* gene complement is variable among BAC inserts.

BAC Clone Name [1]	Accession Number	Growth Period (Days)	BAC Insert Size; PFGE [2], Sequence	*SpTrf* Genes in Locus 1 Allele 1	Analyses [3] to Identify *SpTrf* Gene Complement
BAC-51	KU668451	na [4]	Full-length 157,542 bp	All genes present	*Not*I digest Gene amplicons
BAC-51-con [5]	na [6]	1	Full-length 115,850 bp [7]	All genes present	*Not*I digest Gene amplicons Virtual digests Insert sequence
BAC-51-15	PP082968	10	Short 67,180 bp [8]	Deletion of *A2* only	*Not*I digest Gene amplicons Virtual digests Insert sequence
BAC-52	KU668452	na [4]	Full-length 144,728 bp	All genes present	Gene amplicons
BAC-52-con [5]	na [6]	1	Full-length 124,749 bp [7]	All genes present	*Not*I digest Gene amplicons Virtual digests Insert sequence
BAC-52-4c	na [6]	10	Full-length 76,739 bp [7]	All genes deleted [9]	*Not*I digest Gene amplicons Insert sequence
BAC-52-2b	PP082969	10	Short 36,374bp [8]	All genes deleted except *A2*	*Not*I digest Gene amplicons Virtual digests Insert sequence
BAC-52-19	na	10	Short Not sequenced	All genes deleted	*Not*I digest Gene amplicons
BAC-52-4	na	10	Short Not sequenced	All genes present	*Not*I digest Gene amplicons Virtual digests

[1] Abbreviated BAC accession numbers are used as clone names. [2] PDFE, pulsed field gel electrophoresis. [3] Virtual digests and gene amplicons are shown in Figures 2 and 3. [4] na, not applicable. The BAC insert was not sequenced for this study; it was acquired from GenBank. [5] con, control. BAC DNA isolated from a single colony grown for a single day served as the control. [6] The assembly artifacts precluded submission to GenBank. [7] The BAC insert length reported here includes deletion artifacts resulting from the assembly process. [8] Insert sequence and deletions are shown in the Supplementary File. [9] This is likely a technical PCR failure and a false negative (see Figure 2).

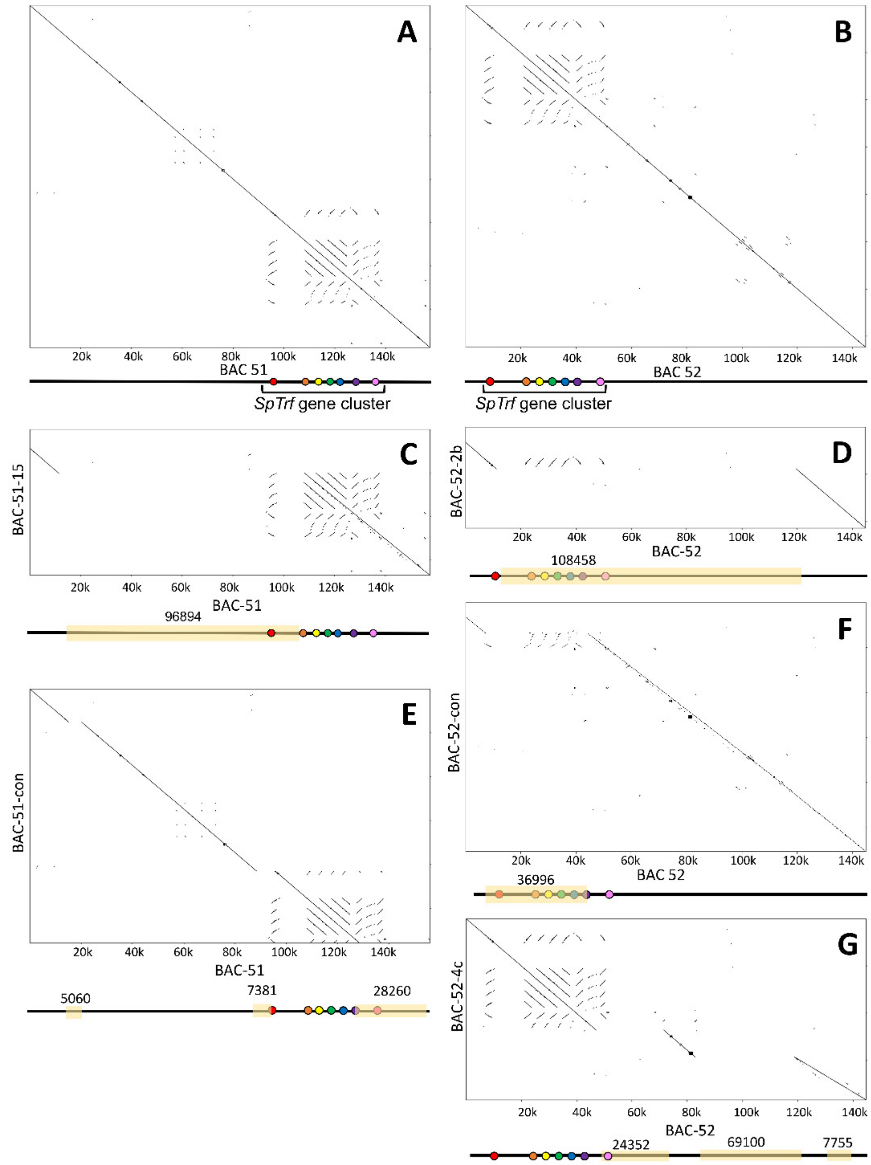

Figure 4. Dot plots of sequenced BAC inserts identify deletions by comparisons to the full-length BAC insert sequences. Dot plots compare each full-length BAC insert to itself and to other BAC inserts that show deletions. The gene cluster maps of BAC-51 or BAC-52 are shown at the bottom of each dot plot with the location of each gene indicated by colored dots (based on the gene colors in Figure 1). Deletions identified by the dot plots are indicated by yellow highlights in the cluster maps and the number of nucleotides in each deletion is indicated. (**A**) BAC-51 vs. self. (**B**) BAC-52 vs. self. (**C**) BAC-51 vs. BAC-51-15. (**D**) BAC-52 vs. BAC-52-2b. (**E**) BAC-52 vs. BAC-52-con. (**F**) BAC-52 vs. BAC-52-con. (**G**) BAC-52 vs. BAC-52-4c.

Because deletions in the full-length BAC inserts did not fit previous analyses of these inserts and because of reported difficulties in assembling sequence reads that include repeats [45,57], as we describe above, verification of the assemblies was required. Raw sequence reads from all five BACs were mapped against the reference sequences submitted to GenBank for BAC-51 and BAC-52 (see Table 1 for accession numbers). Results showed that BAC-51-15 and BAC-51-2b had deletions (Figure 5) in agreement with predictions from gene amplicons and both actual and virtual digests (Figures 2 and 3). BAC-51-con, BAC-52-con, and BAC-52-4c, which were predicted to be full-length with the full complement of genes (Figures 2 and 3), did not show deletions based on the mapping results (Figure 5), which was not in agreement with dot-plot comparisons of the assemblies (Figure 4). The insert assemblies for BAC-51-con, BAC-52-con, and BAC-52-4c were likely the outcome of poor-quality sequence reads that the assembler program omitted when assembling the BAC insert sequences. Consequently, the deletions in these BACs were deemed to be assembly artifacts and illustrated the necessity to employ multiple means to verify sequence assemblies that include repeats. Indeed, the most common reason for low- or poor-quality sequence reads is the presence of highly repetitive stretches of sequences [57], which is the case for these BAC inserts that cover the *SpTrf* gene locus. Furthermore, poor-quality sequences in certain regions of the BAC sequences may be due to variability in deletion presence and location among the individual BAC clones isolated from individual bacterial cells that are present collectively in a single culture. The outcome for sequencing the BAC DNA with inserts that are unstable may be that some of the inserts are full-length and others have random deletions, leading to poor-quality sequence reads at specific locations. Overall, the BAC inserts with predicted and verified deletions showed that one or more *SpTrf* genes were removed.

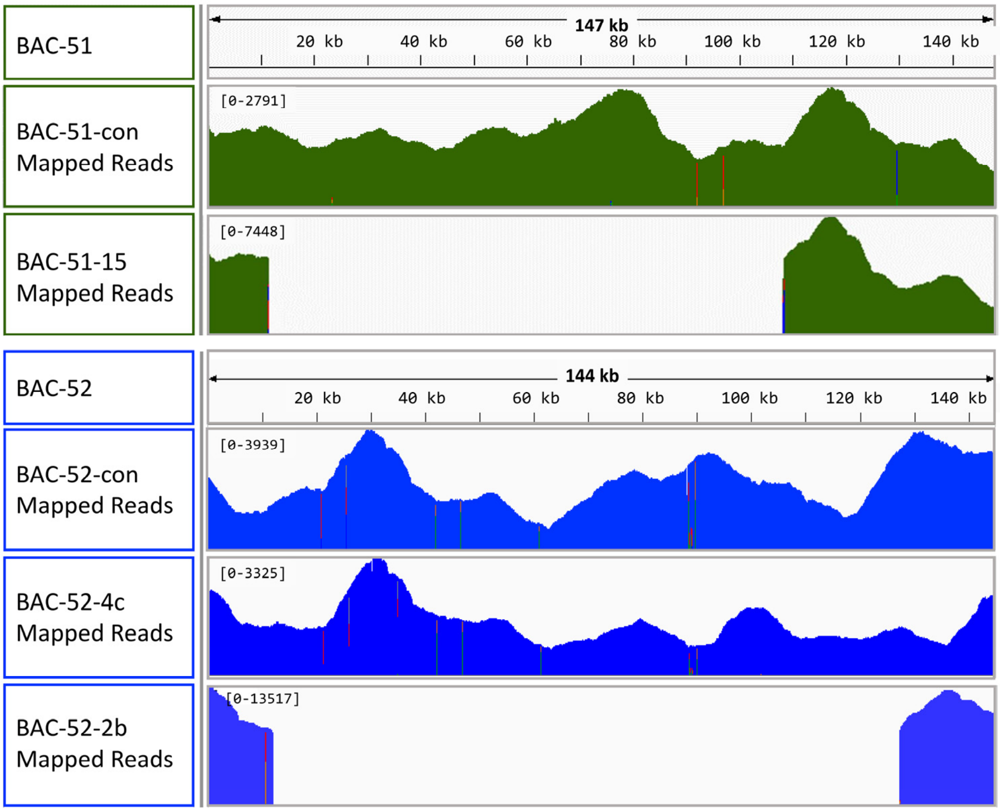

Figure 5. BAC insert assemblies generate false deletions from low-quality sequencing reads. Sequence reads for each BAC are mapped onto reference sequences for BAC-51 (green maps) or BAC-52 (blue maps) that have been reported previously [35]. The lengths of the reference sequences for BAC-51 and BAC-52 are indicated and a ruler is included for every 20 kb. Mapping histograms for each BAC compared to the reference sequence indicate the depth of sequencing coverage and reads per position from the raw PacBio sequencing data. Higher peaks indicate greater coverage, while the absence of peaks indicates no sequence coverage. The colored bars within the histograms indicate nucleotide positions that have increased nucleotide variability across sequence reads. The height of each colored bar indicates the number of sequences for a specific nucleotide. Bar colors indicate nucleotides: red (T), orange (G), light green (A), and dark blue (C).

3.4. BAC Deletions Are Flanked by STRs

Many reports of genomic instability and DNA deletions have focused on defects in DNA repair and replication or the locations of deletions and their associations with specific sequences, including STRs, other types of repeats, and poly G sequences that can form G-quadruplexes [58,59]. Consequently, the verified assemblies for BAC-51-15 and BAC-52-2b enabled an investigation of the sequences at the edges of the deletions. The two verified deletions in the BAC inserts identified from sequence alignments were located at or near regions containing STRs or polynucleotides. The first and larger deletion in BAC-51-15 was associated with CT STRs, which spanned 120 base pairs (bp) at the 5′ end of the deletion and a region of CT and TA STRs of over 425 bp at the 3′ end (Table 3; Supplementary File). The second and smaller deletion that was not verified by mapping results and was likely an assembly artifact based on poor-quality sequence reads (Figure 5) was not associated with repetitive sequences (Table 3; Supplementary File). BAC-52-2b had a single deletion, which was bracketed by poly G and poly C (poly G on the complementary strand; see Supplementary File) (Table 3). Poly G stretches presented the possibility of G-quadruplex formation that may impede DNA replication, leading to genomic instability [60]. In general, the locations of the verified deletions in the BAC inserts were associated with STRs and polynucleotides, in agreement with reports of DNA deletion associated with instability hotspots [6,58].

Table 3. Regions in BAC insert sequences that are the origins of deletions [1].

BAC	Deletion	Local Sequence at Locations of Deletions [2]
51 vs. 51-15	First, 5′ end	**CTCTCTCTCT**CCCTTT**CTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCT**//C**CTCT**AC**CTCTCT**ACTGTAT
	First, 3′ end	**CTCT**CACTTTCTCCTTT**CTCTCTCTCTCTCTCTCTCTCT**//ATCTAT**CTCTATCTCT**ACC
	Second, 5′ end	CT**CACA**GTGTG**AAAA**TGATTTCACA**GTGT**G//ACCAAACTGTGAT**AAA**TAGATTTT**CACAGTGT**G
	Second, 3′ end	TTTACAGTATACCAGAAGTCATTCTTCCCCGATTC//CATCATGCTTTGGAACTTGCTACCTGCTGGA
52 vs. 52-2b	First, 5′ end	**TTTCTTTTTT**GCTCAAC**GGGGGGGGGGG**//TAGTGCATGCGCGGATCCAGGGGAGG**CCCCCCGCCCCCC**AAAA
	First, 3′ end	TACGCAGGA**TTTTTT**CAAGT**GGGGGGGGGGGG**//GGGTTTAACA**TTTTTAAA**TCGGCCG**AAAATTT**CGCAT

[1] See the Supplementary File for the full-length sequence alignments. Bold font shows the locations of STRs or polynucleotide sequences. [2] //, indicates the location of the deletion. Sequences to the 3′ side of the deletion are present in the full-length BAC but are missing in the BACs that contain deletions.

3.5. Some BAC Inserts Are Deleted Prior to Analysis

The initial screen of the BAC library of sea urchin genomic DNA identified 75 BACs with *SpTrf* gene sequences; however, only 27 BAC clones supported PCR amplification of the *SpTrf* genes [35], of which we report results for BAC-51 and BAC-52 above. The clones that did not support amplification of the *SpTrf* genes were evaluated by Southern blots, and BAC-42 and BAC-44 were identified as containing *SpTrf* gene sequences (Figure 6). Based on these conflicting results, these BAC clones were submitted for long-read sequencing. The sequence reads for these BAC inserts could not be assembled into a single sequence and were instead assembled into eight contigs for BAC-44 and three contigs for BAC-42. No *SpTrf* gene sequences were identified within these assemblies, which contradicted the results in the Southern blot. The inconsistencies with BAC-42 and BAC-44 among the other BACs that failed to amplify *SpTrf* sequences suggested that the initial library screens had identified *SpTrf* gene sequences in 75 BAC inserts, but that during the growth and isolation of the BAC DNA for analysis by PFGE and PCR, the inserts underwent deletions. The inference was that these and the other 47 BAC clones were particularly unstable.

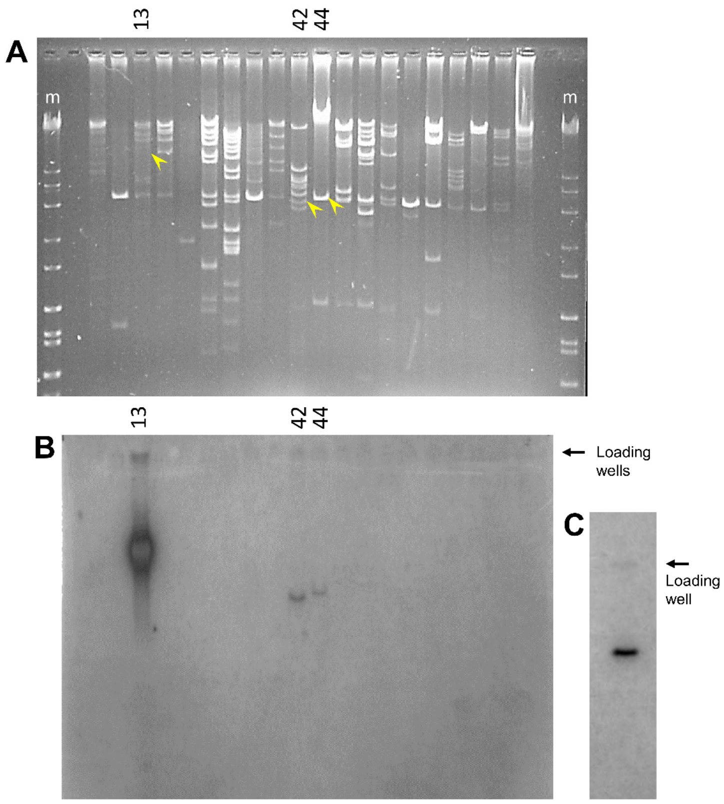

Figure 6. Many BAC inserts do not show *SpTrf* sequences by Southern blot. (**A**) BAC clones that did not amplify *SpTrf* sequences by PCR were digested with *Sal*I and *Not*I, and the fragments were separated by gel electrophoresis, transferred to nylon filters, and evaluated with ^{32}P-riboprobes. A subset of those BAC clones are shown. The yellow arrowheads indicate bands with *SpTrf* sequences that correspond to the bands in (**B**). The marker lanes (m) are Hi Lo DNA standards (Fisher Scientific). (**B**) The probes hybridize to three BAC inserts. The raw reads for BAC-42 and BAC-44 are available as BioSamples in GenBank, accession numbers SAMN39322606 and SAMN39322605, respectively. Preliminary sequence analysis of BAC-13 indicates that it includes a region of allele 2 for the *SpTrf* gene cluster in locus 1. Consequently, because this study focused on allele 1 of locus 1, long-read sequencing of the BAC-13 insert was not pursued. (**C**) BAC-7096 with six *SpTrf* genes (see Table 1, Figure 1B) [35,38] is the positive control for the Southern blot.

4. Discussion

4.1. Instability of BAC Inserts When Hosted by E. coli

Genomic instability can be lethal when it results in severe genomic fragmentation; however, genomic instability is also a source of potentially beneficial adaptions ([32,61]; reviewed in [62]). For the examples presented here, the BAC inserts are of no benefit to the bacteria; rather, it is the vector which contains the antibiotic resistance that is of benefit to the bacteria under the growth conditions in the presence of chloramphenicol. BAC inserts with repeats are unstable in prokaryotes ([63]; this study), and BACs with smaller inserts benefit *E. coli* because cells with smaller inserts can replicate the BAC DNA more quickly and with lower metabolic cost and therefore may proliferate faster than other cells with larger BAC inserts. The findings presented here suggest that BAC clones with multiple *SpTrf* genes are inherently unstable. Furthermore, deletions within the inserts may progress from multiple small deletions to larger deletions that incorporate the smaller deletions. This was likely the case for BAC-52-2b, with two small deletions identified in early analyses and a single deletion identified in the assembled insert sequence as one large deletion that incorporated the small deletions. Results suggest that the involvement of the STRs and polynucleotides, such as multiple Gs, may promote additional, larger deletions after the initial smaller deletions are initiated. We propose that insert deletions in different BAC clones are initiated at different time points during the 10-day growth period. Perhaps early small deletions are more difficult to detect using our initial methods of restriction digests, whereas later, larger deletions in the same BAC are easier to detect. However, this pattern is not necessarily discernable from our dataset. We propose that the terminal condition of the BAC inserts is the deletion of all *SpTrf* genes, including the associated STRs and other types of repeats, as suggested from the insert sequences of BAC-42 and BAC-44 that do not include *SpTrf* sequences. Once the repeat sequences are deleted, subsequent deletions may cease to occur. Overall, the *SpTrf* gene cluster in BAC inserts appears to be very unstable and tends to undergo deletions. This instability has likely impacted the sequencing phase of the sea urchin genome assembly that employed the BAC library and was one aspect of the poor assembly of *SpTrf* gene loci of only 9 *SpTrf* genes, when 50 to 60 genes have been predicted in the family [33].

4.2. Genomic Instability of the SpTrf Gene Family May Underlie Variation in Gene Family Structure and Expression in Sea Urchin Cells

Although our findings suggest that the BAC inserts that include *SpTrf* gene clusters are unstable in *E. coli*, the key question is not about the growth benefits of smaller BACs for *E. coli* but whether local genomic instability applies to the *SpTrf* gene loci in the sea urchin genome. If the *SpTrf* gene family is unstable, and given the results presented here suggesting that the genes tend to be deleted from the BAC inserts, how might instability benefit the innate immune response of sea urchins? Genomic instability has been suggested as a source of beneficial genomic variation [62], and the STRs that surround genes and segmental duplications, the repeats within coding regions, and the sequence similarities among genes may all underlie gene duplications and/or deletions and changes in copy number plus subsequent sequence variations among genes [32,38]. STRs and other repeats have been the basis for proposed local genomic instability [4] and a theoretical evolutionary history of the *SpTrf* gene family [32]. Furthermore, there are three regions of GA STRs of several kilobases each that flank the genes in the *SpTrf* locus 2. These STR islands are located in positions that correspond to genes in locus 1, and this has led us to postulate that they may be the remnants of gene deletions that occurred in locus 2 [32,35].

Deletions and local genomic instability for the *SpTrf* gene clusters in the genome is in accord with previous reports on *SpTrf* gene expression and *SpTrf* gene copy numbers in single coelomocytes [29,42]. Immune challenge of sea urchins significantly increases *SpTrf* gene expression [23,24,64], SpTrf protein production [26], and the numbers of SpTrf$^+$ coelomocytes in the CF and in other tissues [27,42,65]. Similar results for the *Trf* families have been reported for the sea urchins *Heliocidaris erythrogramma* [20] and *Paracentrotus lividus* [22]. It was assumed that individual phagocytes would express multiple *SpTrf* genes to drive swift responses to invading pathogens. However, when single phagocytes were evaluated for *SpTrf* gene expression, only identical *SpTrf* transcript sequences were identified for individual cells, implying expression of a single *SpTrf* gene per cell [29]. Al-though this result is consistent with gene regulation by promoters and enhancers, there may be an alternative explanation. In an approach to address the question of *SpTrf* gene regulation vs. gene deletion, single small phagocytes were sorted based on the SpTrf proteins on the surface, and red spherule cells that do not express *SpTrf* genes [29] but can be sorted based on the red color of echinochrome, which is characteristic of these cells [42]. Single sperm cells were employed as the control. Because the genes are small enough to be amplified by PCR, variations in amplicon sizes (see the legend in Figure 2) show that the arrays of genes are different among many of the coelomocytes, whereas they are all identical for each sperm for a given animal [42]. Furthermore, the *SpTrf* gene copy number is reduced in most coelomocytes compared to sperm. When the results for *SpTrf* gene expression and gene copy number are integrated, the interpretation suggests that the coelomocytes alter the *SpTrf* gene family, which may restrict and enhance the expression of a single gene per cell. These results are consistent with the notion of local genomic instability of the *SpTrf* gene family and the hypothesis that putative DNA repair mechanisms are employed by the coelomocytes to balance control of locus instability with enhancing *SpTrf* gene sequence diversification. The outcome would be to the advantage of sea urchins in the arms race between their innate immune system and potential pathogens. Parallel results of local genomic instability have been reported for the *agglutinin-like sequence* multi-gene family in *Candida albicans* that contain tandem repeats and encode variations in adhesion proteins for binding to and colonizing host endothelial and epithelial cells, which is an example of a pathogen–host arms race ([66]; reviewed in [62]). Similarly, genes in the fungal *hnwd* family, which functions in allorecognition specificity, encode a series of WD40 repeats that form β propeller structures and show local instability, driving variations in the numbers and sequences of repeats and altering interactions with conspecifics [67]. The killer immunoglobulin-like receptor (*KIR*) locus functions in natural killer cells (reviewed by [68]) and displays gene sequence diversity, tight gene clustering of less than 3 kb of intergenic space, and extraordinarily fast evolutionary change to allelic sequences and the locus structure, as suggested by extensive crossing over [69,70]. The range of repeats in the *KIR* locus is consistent with local genomic instability that drives the varia-bility [71]. The first intron of the *KIR* genes is a minirepeat composed entirely of 30–60 repeats of 19–20 bp [71], and the locus has a number of transposable elements [13], which are repeats that may also result in *KIR* locus instability. Hence, the vertebrate *KIR* and the sea urchin *SpTrf* families both have extensive repeats that may be involved in their respective sequence diversification that may be beneficial in the host-pathogen arms race [72].

Hotspots of genomic instability that are often associated with STRs result in slow or poor DNA replication due to replication fork stalling or reversal, leading to double-strand DNA breaks [73,74]. Non-B DNA structures, such as Z-DNA, hairpins from inverted repeats, and poly G stretches that may form G-quadruplexes can also be the basis for genomic instability at hotspots (reviewed in [75]). G-quadruplexes can block DNA replication fork progression, leading to local genomic instability when helicases fail to unwind G quadruplexes in eukaryotes ([76,77]; reviewed in [78]) and in *E. coli* [79,80]. We report both possibilities for verified DNA deletions in the BAC inserts; CT STRs are associated with the deletion site in BAC-51-15, and poly G regions are associated with the deletion site in BAC-51-2b. In general, the wide variety of DNA repeats that are associated with the *SpTrf* gene family has suggested local genomic instability [32,35,38]. The structural appearance and organization of the family is consistent with instability of the genes located in segmental duplications, STRs that flank both genes and segmental duplications, tandem and interspersed repeats in the coding sequences, sequence variations among the genes, and proposed gene deletions in locus 2.

DNA damage resulting in local genomic instability includes (i) DNA replication stress at fragile sites composed of STRs and other types of repeats [6], (ii) R loops of DNA–RNA hybrids that can lead to double-strand breaks [81,82] and double-strand breaks that occur during general DNA replication (reviewed in [83,84], (iii) highly expressed genomic regions resulting in DNA damage from transcription–replication conflicts [85,86], and (vi) DNA tangles resolved by cleavage followed by incorrect re-ligation [87]. DNA damage and genomic instability are counteracted by DNA repair mechanisms and expression in response to DNA damage in eukaryotes [88]. Genomic instability of the *SpTrf* gene clusters that could result in the deletion of entire loci must be regulated in some way, given that the gene family is maintained in genomes of euechinoid species. Molecular control by coelomocytes to promote, block, or repair chromosomal changes in genomic DNA regions of the *SpTrf* gene clusters may correlate with variations in the level of expression for the DNA repair mechanisms. Hence, there may be a correlation between the expression level of the *SpTrf* genes, which are upregulated significantly in response to immune challenge (reviewed in [89]), and the sea urchin DNA repair mechanisms that may control or regulate local instability of the *SpTrf* gene clusters. Because the *SpTrf* gene family shows an identical composition among single sperm cells from individual sea urchins [42] and given that DNA damage is associated with DNA replication, DNA repair genes may be expected to show elevated expression during mitosis and meiosis in gonads to maintain the structure and membership of the *SpTrf* gene family. Conversely, because individual coelomocytes do not appear to maintain the *SpTrf* gene family equally among cells, DNA repair gene expression may be reduced and/or variable among cells in the axial organ and pharynx where the coelomocytes proliferate [27]. DNA repair genes were identified in initial annotations of the *S. purpuratus* genome sequence (version 2.1) [90], although their expression has not been investigated for the tissues and organs of adult *S. purpuratus*. However, gene expression related to DNA repair has been documented for sea urchin larvae and coelomocytes responding to exposure to genotoxic chemicals [91]. Searches of the *S. purpuratus* genome sequence (version 5.0; www.Echinobase.org; [36], as of 15 December, 2023) result in a number of genes that encode proteins with putative DNA repair functions, including general DNA repair, DNA repair and recombination, excision repair, mismatch repair, double-strand break repair, and DNA cross-link repair, in addition to exonuclease and helicase functions (see also Table S2 in [90]).

Local genomic instability in the *S. purpuratus* genome may not be random, based on the findings from BAC clone insert instability in *E. coli*, yet sea urchin cells must have a means to control or regulate instability to take advantage of the characteristics of the *SpTrf* gene clusters, which are riddled with a wide range of repeats with predicted locations of duplications, deletions, and insertions [32,38,42]. In the case of the *SpTrf* genes in euechinoids, we propose that local genomic instability is an initial and required parameter to drive gene sequence diversification in the population that results in an immune response gene family that keeps pace in the arms race with the marine microbes with which sea urchins share their habitat.

5. Conclusions

Genomic instability is driven by the presence of repeats in local regions of the genome. The *SpTrf* gene family, which functions in immune response in the purple sea urchin, is surrounded by multiple types, sizes, and numbers of repetitive sequences, including flanking STRs. We show that the inserts of BAC clones with multiple *SpTrf* genes are unstable in *E. coli* and that with continued growth over time, one or more of the *SpTrf* genes are deleted. The deleted regions are commonly bracketed by STRs or polyG stretches. These results suggest that the repeat-riddled region in the sea urchin genome that includes the *SpTrf* gene family is locally unstable. This may result in expanding, contracting, or maintaining members of the *SpTrf* gene family, which is likely to be beneficial in the arms race against pathogens.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/genes15020222/s1. Supplementary File: BAC insert sequence alignments.

Author Contributions: Conceptualization, M.A.B.H. and L.C.S.; Methodology, M.A.B.H.; Validation, M.A.B.H., F.M. and A.J.; Formal Analysis, M.A.B.H.; Investigation, M.A.B.H., F.M. and A.J.; Data Curation, M.A.B.H.; Writing—Original Draft Preparation, M.A.B.H. and L.C.S.; Writing—Review and Editing, M.A.B.H. and L.C.S.; Visualization, M.A.B.H., F.M., A.J. and L.C.S.; Supervision, L.C.S.; Project Administration, L.C.S.; Funding Acquisition, M.A.B.H., M.A.A. and L.C.S. All authors have read and agreed to the published version of the manuscript.

Funding: The research was supported by awards from the US National Science Foundation (IOS 1550747; IOS 1855747) to L.C.S., the Cross Disciplinary Research Fund from the Columbian College of Arts and Sciences at George Washington University to M.A.A. and L.C.S., a Wilber V. Harlan Graduate Scholarship, a Mortensen Award from the Department of Biology, a Dissertation Fellowship from the Columbian College of Arts and Sciences at GWU to M.A.B.H., and a National Science Foundation postdoctoral award (NSF Postdoctoral Fellowship in Biology FY 2022; 2208923) to M.A.B.H.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The sequence data generated in this research are available from GenBank. See Tables 1 and 2 for accession numbers. The raw sequence reads for BAC-42 and BAC-44 are available from GenBank (BioSample accession numbers SAMN39322606 and SAMN39322605, respectively). The GenBank accession numbers for BAC-51-15 is PP082968, and BAC-52-2b is PP082969. The BAC clones used in this analysis and their genomic DNA library locations are as follows: BAC-51 location is 10B1; BAC-52 location is 4074J14; BAC-67 location is 14K16; BAC-42 location is 4069G2; BAC-44 location is 4069C2, and BAC-13 location is 3020I13. See also: https://www.echinobase.org/echinobase/.

Acknowledgments: Nouf Alhenaky and Meghana Valay at GWU participated in early work on this project.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Buckley, K.M.; Rast, J.P. Diversity of animal immune receptors and the origins of recognition complexity in the deuterostomes. *Dev. Comp. Immunol.* **2015**, *49*, 179–189. [CrossRef]
2. Buckley, K.M.; Dooley, H. Immunological diversity is a cornerstone of organismal defense and allorecognition across metazoa. *J. Immunol.* **2022**, *208*, 203–211. [CrossRef]
3. Zhang, L.; Li, L.; Guo, X.; Litman, G.W.; Dishaw, L.J.; Zhang, G. Massive expansion and functional divergence of innate immune genes in a protostome. *Sci. Rep.* **2015**, *5*, 8693. [CrossRef]
4. Oren, M.; Barela Hudgell, M.A.; Golconda, P.; Lun, C.M.; Smith, L.C. Genomic instability and shared mechanisms for gene diversification in two distant immune gene families: The echinoid *185/333* and the plant NBS-LRR. In *The Evolution of the Immune System: Conservation and Diversification*; Malagoli, D., Ed.; Elsevier-Academic Press: London, UK, 2016; pp. 295–310.
5. Myers, S.; Spencer, C.; Auton, A.; Bottolo, L.; Freeman, C.; Donnelly, P.; McVean, G. The distribution and causes of meiotic recombination in the human genome. *Biochem. Soc. Trans.* **2006**, *34*, 526–530. [CrossRef]
6. Balzano, E.; Pelliccia, F.; Giunta, S. Genome (in)stability at tandem repeats. *Semin. Cell Dev. Biol.* **2021**, *113*, 97–112. [CrossRef]
7. Bzymek, M.; Lovett, S.T. Instability of repetitive DNA sequences: The role of replication in multiple mechanisms. *Proc. Natl. Acad. Sci. USA* **2001**, *98*, 8319–8325. [CrossRef]
8. Armengol, L.; Pujana, M.A.; Cheung, J.; Scherer, S.W.; Estivill, X. Enrichment of segmental duplications in regions of breaks of synteny between the human and mouse genomes suggest their involvement in evolutionary rearrangements. *Hum. Mol. Genet.* **2003**, *12*, 2201–2208. [CrossRef]
9. Koszul, R.; Dujon, B.; Fischer, G. Stability of large segmental duplications in the yeast genome. *Genetics* **2006**, *172*, 2211–2222. [CrossRef]
10. San Mauro, D.; Gower, D.J.; Zardoya, R.; Wilkinson, M. A hotspot of gene order rearrangement by tandem duplication and random loss in the vertebrate mitochondrial genome. *Mol. Biol. Evol.* **2006**, *23*, 227–234. [CrossRef]
11. Zhao, H.; Bourque, G. Recovering genome rearrangements in the mammalian phylogeny. *Genome Res.* **2009**, *19*, 934–942. [CrossRef]
12. Myers, S.; Bottolo, L.; Freeman, C.; McVean, G.; Donnelly, P. A fine-scale map of recombination rates and hotspots across the human genome. *Science* **2005**, *310*, 321–324. [CrossRef]
13. Traherne, J.A.; Martin, M.; Ward, R.; Ohashi, M.; Pellett, F.; Gladman, D.; Middleton, D.; Carrington, M.; Trowsdale, J. Mechanisms of copy number variation and hybrid gene formation in the *KIR* immune gene complex. *Hum. Mol. Genet.* **2010**, *19*, 737–751. [CrossRef]
14. Lecompte, O.; Ripp, R.; Puzos-Barbe, V.; Duprat, S.; Heilig, R.; Dietrich, J.; Thierry, J.-C.; Poch, O. Genome evolution at the genus level: Comparison of three complete genomes of hyperthermophilic archaea. *Genome Res.* **2001**, *11*, 981–993. [CrossRef]
15. Eichler, E.E.; Sankoff, D. Structural dynamics of eukaryotic chromosome evolution. *Science* **2003**, *301*, 793–797. [CrossRef]
16. Peng, Q.; A Pevzner, P.; Tesler, G. The fragile breakage versus random breakage models of chromosome evolution. *PLOS Comput. Biol.* **2006**, *2*, e14. [CrossRef]
17. Kikuta, H.; Laplante, M.; Navratilova, P.; Komisarczuk, A.Z.; Engström, P.G.; Fredman, D.; Akalin, A.; Caccamo, M.; Sealy, I.; Howe, K.; et al. Genomic regulatory blocks encompass multiple neighboring genes and maintain conserved synteny in vertebrates. *Genome Res.* **2007**, *17*, 545–555. [CrossRef]
18. Mongin, E.; Dewar, K.; Blanchette, M. Long-range regulation is a major driving force in maintaining genome integrity. *BMC Evol. Biol.* **2009**, *9*, 203. [CrossRef] [PubMed]
19. Buckley, K.M.; Smith, L.C. Extraordinary diversity among members of the large gene family, *185/333*, from the purple sea urchin, *Strongylocentrotus purpuratus*. *BMC Mol. Biol.* **2007**, *8*, 68. [CrossRef] [PubMed]
20. Roth, M.O.; Wilkins, A.G.; Cooke, G.M.; Raftos, D.A.; Nair, S.V. Characterization of the highly variable immune response gene family, *He185/333*, in the sea urchin, *Heliocidaris erythrogramma*. *PLoS ONE* **2014**, *9*, e62079. [CrossRef] [PubMed]
21. Wang, U.; Ding, J.; Liu, Y.; Liu, X.; Chang, Y. Isolation of immune-relating 185/333-1 gene from sea urchin (*Strongylocentrotus intermedius*) and its expression analysis. *J. Ocean Univ. China* **2016**, *15*, 163–170. [CrossRef]

22. Yakovenko, I.; Donnyo, A.; Ioscovich, O.; Rosental, B.; Oren, M. The diverse transformer (Trf) protein family in the sea urchin *Paracentrotus lividus* acts through a collaboration between cellular and humoral immune effector arms. *Int. J. Mol. Sci.* **2021**, *22*, 6639. [CrossRef]
23. Nair, S.V.; Del Valle, H.; Gross, P.S.; Terwilliger, D.P.; Smith, L.C. Macroarray analysis of coelomocyte gene expression in response to LPS in the sea urchin. Identification of unexpected immune diversity in an invertebrate. *Physiol. Genom.* **2005**, *22*, 33–47. [CrossRef]
24. Terwilliger, D.P.; Buckley, K.M.; Brockton, V.; Ritter, N.J.; Smith, L.C. Distinctive expression patterns of *185/333* genes in the purple sea urchin, *Strongylocentrotus purpuratus*: An unexpectedly diverse family of transcripts in response to LPS, β-1,3-glucan, and dsRNA. *BMC Mol. Biol.* **2007**, *8*, 16. [CrossRef]
25. Terwilliger, D.P.; Buckley, K.M.; Mehta, D.; Moorjani, P.G.; Smith, L. Unexpected diversity displayed in cDNAs expressed by the immune cells of the purple sea urchin, *Strongylocentrotus purpuratus*. *Physiol. Genom.* **2006**, *26*, 134–144. [CrossRef]
26. Brockton, V.; Henson, J.H.; Raftos, D.A.; Majeske, A.J.; Kim, Y.-O.; Smith, L.C. Localization and diversity of 185/333 proteins from the purple sea urchin–unexpected protein-size range and protein expression in a new coelomocyte type. *J. Cell Sci.* **2008**, *121*, 339–348. [CrossRef]
27. Golconda, P.; Buckley, K.M.; Reynolds, C.R.; Romanello, J.P.; Smith, L.C. The axial organ and the pharynx are sites of hematopoiesis in the sea urchin. *Front. Immunol.* **2019**, *10*, 870. [CrossRef]
28. Dheilly, N.M.; Haynes, P.A.; Bove, U.; Nair, S.V.; Raftos, D.A. Comparative proteomic analysis of a sea urchin (*Heliocidaris erythrogramma*) antibacterial response revealed the involvement of apextrin and calreticulin. *J. Invertebr. Pathol.* **2011**, *106*, 223–229. [CrossRef]
29. Majeske, A.J.; Oren, M.; Sacchi, S.; Smith, L.C. Single sea urchin phagocytes express messages of a single sequence from the diverse *Sp185/333* gene family in response to bacterial challenge. *J. Immunol.* **2014**, *193*, 5678–5688. [CrossRef] [PubMed]
30. Sherman, L.S.; Schrankel, C.S.; Brown, K.J.; Smith, L.C. Extraordinary diversity of immune response proteins among sea urchins: Nickel-isolated Sp185/333 proteins show broad variations in size and charge. *PLoS ONE* **2015**, *10*, e0138892. [CrossRef] [PubMed]
31. Ho, E.C.H.; Buckley, K.M.; Schrankel, C.S.; Schuh, N.W.; Hibino, T.; Solek, C.M.; Bae, K.; Wang, G.; Rast, J.P. Perturbation of gut bacteria induces a coordinated cellular immune response in the purple sea urchin larva. *Immunol. Cell Biol.* **2016**, *94*, 861–874. [CrossRef] [PubMed]
32. Barela Hudgell, M.A.; Smith, L.C. Sequence diversity, locus structure, and evolutionary history of the *SpTransformer* genes in the sea urchin genome. *Front. Immunol.* **2021**, *12*, 744783. [CrossRef]
33. Buckley, K.; Munshaw, S.; Kepler, T.; Smith, L. The *185/333* gene family is a rapidly diversifying host-defense gene cluster in the purple sea urchin *Strongylocentrotus purpuratus*. *J. Mol. Biol.* **2008**, *379*, 912–928. [CrossRef]
34. Smith, L.C. Innate immune complexity in the purple sea urchin: Diversity of the *Sp185/333* system. *Front. Immunol.* **2012**, *3*, 70. [CrossRef] [PubMed]
35. Oren, M.; Barela Hudgell, M.A.; D'allura, B.; Agronin, J.; Gross, A.; Podini, D.; Smith, L.C. Short tandem repeats, segmental duplications, gene deletion, and genomic instability in a rapidly diversified immune gene family. *BMC Genom.* **2016**, *17*, 900. [CrossRef]
36. I Arshinoff, B.; A Cary, G.; Karimi, K.; Foley, S.; Agalakov, S.; Delgado, F.; Lotay, V.S.; Ku, C.J.; Pells, T.J.; Beatman, T.R.; et al. Echinobase: Leveraging an extant model organism database to build a knowledgebase supporting research on the genomics and biology of echinoderms. *Nucleic Acids Res.* **2022**, *50*, D970–D979. [CrossRef]
37. Sodergren, E.; Weinstock, G.M.; Davidson, E.H.; Cameron, R.A.; Gibbs, R.A.; Angerer, R.C.; Angerer, L.M.; Arnone, M.I.; Burgess, D.R.; Coffman, J.A.; et al. The genome of the sea urchin *Strongylocentrotus purpuratus*. *Science* **2006**, *314*, 941–952. [CrossRef]
38. A Miller, C.; Buckley, K.M.; Easley, R.L.; Smith, L.C. An *Sp185/333* gene cluster from the purple sea urchin and putative microsatellite-mediated gene diversification. *BMC Genom.* **2010**, *11*, 575. [CrossRef] [PubMed]
39. Buckley, K.M.; Dong, P.; Cameron, R.A.; Rast, J.P. Bacterial artificial chromosomes as recombinant reporter constructs to investigate gene expression and regulation in echinoderms. *Brief. Funct. Genom.* **2018**, *17*, 362–371. [CrossRef]
40. Cameron, R.A.; Mahairas, G.; Rast, J.P.; Martinez, P.; Biondi, T.R.; Swartzell, S.; Wallace, J.C.; Poustka, A.J.; Livingston, B.T.; Wray, G.A.; et al. A sea urchin genome project: Sequence scan, virtual map, and additional resources. *Proc. Natl. Acad. Sci. USA* **2000**, *97*, 9514–9518. [CrossRef] [PubMed]
41. Cameron, R.A.; Rast, J.P.; Brown, C.T. Genomic resources for the study of sea urchin development. *Methods Cell Biol.* **2004**, *74*, 733–757. [CrossRef]
42. Oren, M.; Rosental, B.; Hawley, T.S.; Kim, G.-Y.; Agronin, J.; Reynolds, C.R.; Grayfer, L.; Smith, L.C. Individual sea urchin coelomocytes undergo somatic immune gene diversification. *Front. Immunol.* **2019**, *10*, 1298. [CrossRef] [PubMed]
43. Koren, S.; Rhie, A.; Walenz, B.P.; Dilthey, A.T.; Bickhart, D.M.; Kingan, S.B.; Hiendleder, S.; Williams, J.L.; Smith, T.P.L.; Phillippy, A.M. De novo assembly of haplotype-resolved genomes with trio binning. *Nat. Biotechnol.* **2018**, *36*, 1174–1182. [CrossRef]
44. Nurk, S.; Walenz, B.P.; Rhie, A.; Vollger, M.R.; Logsdon, G.A.; Grothe, R.; Miga, K.H.; Eichler, E.E.; Phillippy, A.M.; Koren, S. HiCanu: Accurate assembly of segmental duplications, satellites, and allelic variants from high-fidelity long reads. *Genome Res.* **2020**, *30*, 1291–1305. [CrossRef] [PubMed]
45. Koren, S.; Walenz, B.P.; Berlin, K.; Miller, J.R.; Bergman, N.H.; Phillippy, A.M. Canu: Scalable and accurate long-read assembly via adaptive *k*-mer weighting and repeat separation. *Genome Res.* **2017**, *27*, 722–736. [CrossRef] [PubMed]

46. Noé, L.; Kucherov, G. YASS: Enhancing the sensitivity of DNA similarity search. *Nucleic Acids Res.* **2005**, *33*, W540–W543. [CrossRef] [PubMed]
47. Li, H. Minimap2: Pairwise alignment for nucleotide sequences. *Bioinformatics* **2018**, *34*, 3094–3100. [CrossRef]
48. Li, H. New strategies to improve minimap2 alignment accuracy. *Bioinformatics* **2021**, *37*, 4572–4574. [CrossRef]
49. Danecek, P.; Bonfield, J.K.; Liddle, J.; Marshall, J.; Ohan, V.; Pollard, M.O.; Whitwham, A.; Keane, T.; McCarthy, S.A.; Davies, R.M.; et al. Twelve years of SAMtools and BCFtools. *GigaScience* **2021**, *10*, giab008. [CrossRef]
50. Robinson, J.T.; Thorvaldsdottir, H.; Turner, D.; Mesirov, J.P. igv.js: An embeddable JavaScript implementation of the Integrative Genomics Viewer (IGV). *Bioinformatics* **2022**, *39*, btac830. [CrossRef]
51. Robinson, J.T.; Thorvaldsdóttir, H.; Wenger, A.M.; Zehir, A.; Mesirov, J.P. Variant Review with the Integrative Genomics Viewer (IGV). *Cancer Res.* **2017**, *77*, 31–34. [CrossRef]
52. Robinson, J.T.; Thorvaldsdóttir, H.; Winckler, W.; Guttman, M.; Lander, E.S.; Getz, G.; Mesirov, J.P. Integrative genomics viewer. *Nat. Biotechnol.* **2011**, *29*, 24–26. [CrossRef] [PubMed]
53. Thorvaldsdóttir, H.; Robinson, J.T.; Mesirov, J.P. Integrative Genomics Viewer (IGV): High-performance genomics data visualization and exploration. *Brief. Bioinform.* **2013**, *14*, 178–192. [CrossRef]
54. Sambrook, J.; Fritsch, E.F.; Maniatas, T. *Molecular Cloning: A Laboratory Manual*; Cold Spring Harbor Laboratory Press: New York, NY, USA, 1989.
55. Smith, L.C.; Shih, C.-S.; Dachenhausen, S.G. Coelomocytes express SpBf, a homologue of factor B, the second component in the sea urchin complement system. *J. Immunol.* **1998**, *161*, 6784–6793. [CrossRef] [PubMed]
56. Ahmad, A.; Kabir, M.A.; Kravets, A.; Andaluz, E.; Larriba, G.; Rustchenko, E. Chromosome instability and unusual features of some widely used strains of *Candida albicans*. *Yeast* **2008**, *25*, 433–448. [CrossRef] [PubMed]
57. Korlach, J.; Biosciences, P. Understanding accuracy in SMRT sequencing. *Pac. Biosci.* **2013**, *2013*, 1–9.
58. Aguilera, A.; García-Muse, T. Causes of genome instability. *Annu. Rev. Genet.* **2013**, *47*, 1–32. [CrossRef] [PubMed]
59. Lopes, J.; Piazza, A.; Bermejo, R.; Kriegsman, B.; Colosio, A.; Teulade-Fichou, M.-P.; Foiani, M.; Nicolas, A. G-quadruplex-induced instability during leading-strand replication. *EMBO J.* **2011**, *30*, 4033–4046. [CrossRef]
60. Kruisselbrink, E.; Guryev, V.; Brouwer, K.; Pontier, D.B.; Cuppen, E.; Tijsterman, M. Mutagenic capacity of endogenous G4 DNA underlies genome instability in FANCJ-defective, *C. elegans*. *Curr. Biol.* **2008**, *18*, 900–905. [CrossRef]
61. Smith, A.C.; Morran, L.T.; Hickman, M.A. Host Defense mechanisms induce genome instability leading to rapid evolution in an opportunistic fungal pathogen. *Infect. Immun.* **2022**, *90*, e0032821. [CrossRef]
62. Dunn, M.J.; Anderson, M.Z. To repeat or not to repeat: Repetitive sequences regulate genome stability in *Candida albicans*. *Genes* **2019**, *10*, 866. [CrossRef]
63. Bichara, M.; Wagner, J.; Lambert, I. Mechanisms of tandem repeat instability in bacteria. *Mutat. Res. Mol. Mech. Mutagen.* **2006**, *598*, 144–163. [CrossRef]
64. Rast, J.P.; Pancer, Z.; Davidson, E.H. New approaches towards an understanding of deuterostome immunity. *Curr. Top. Microbiol. Immunol.* **2000**, *248*, 3–16. [CrossRef]
65. Majeske, A.J.; Oleksyk, T.K.; Smith, L.C. The *Sp185/333* immune response genes and proteins are expressed in cells dispersed within all major organs of the adult purple sea urchin. *Innate Immun.* **2013**, *19*, 569–587. [CrossRef]
66. Zhang, N.; Harrex, A.L.; Holland, B.R.; Fenton, L.E.; Cannon, R.D.; Schmid, J. Sixty alleles of the *ALS7* open reading frame in *Candida albicans*: *ALS7* is a hypermutable contingency locus. *Genome Res.* **2003**, *13*, 2005–2017. [CrossRef] [PubMed]
67. Chevanne, D.; Saupe, S.J.; Clavé, C.; Paoletti, M. WD-repeat instability and diversification of the *Podospora anserina hnwd* non-self recognition gene family. *BMC Evol. Biol.* **2010**, *10*, 134. [CrossRef] [PubMed]
68. Bruijnesteijn, J.; de Groot, N.G.; Bontrop, R.E. The genetic mechanisms driving diversification of the *KIR* gene cluster in primates. *Front. Immunol.* **2020**, *11*, 582804. [CrossRef] [PubMed]
69. Uhrberg, M. The *KIR* gene family: Life in the fast lane of evolution. *Eur. J. Immunol.* **2005**, *35*, 10–15. [CrossRef]
70. Martin, A.M.; Kulski, J.K.; Gaudieri, S.; Witt, C.S.; Freitas, E.M.; Trowsdale, J.; Christiansen, F.T. Comparative genomic analysis, diversity and evolution of two *KIR* haplotypes A and B. *Gene* **2004**, *335*, 121–131. [CrossRef]
71. Wilson, M.J.; Torkar, M.; Haude, A.; Milne, S.; Jones, T.; Sheer, D.; Beck, S.; Trowsdale, J. Plasticity in the organization and sequences of human *KIR/ILT* gene families. *Proc. Natl. Acad. Sci. USA* **2000**, *97*, 4778–4783. [CrossRef]
72. Smith, L.C. Diversification of innate immune genes: Lessons from the purple sea urchin. *Dis. Model. Mech.* **2010**, *3*, 274–279. [CrossRef]
73. Gadgil, R.; Barthelemy, J.; Lewis, T.; Leffak, M. Replication stalling and DNA microsatellite instability. *Biophys. Chem.* **2017**, *225*, 38–48. [CrossRef]
74. Schwartz, M.; Zlotorynski, E.; Goldberg, M.; Ozeri, E.; Rahat, A.; le Sage, C.; Chen, B.P.; Chen, D.J.; Agami, R.; Kerem, B. Homologous recombination and nonhomologous end-joining repair pathways regulate fragile site stability. *Genes Dev.* **2005**, *19*, 2715–2726. [CrossRef] [PubMed]
75. Spiegel, J.; Adhikari, S.; Balasubramanian, S. The structure and function of DNA G-quadruplexes. *Trends Chem.* **2020**, *2*, 123–136. [CrossRef] [PubMed]
76. Lopez, C.R.; Singh, S.; Hambarde, S.; Griffin, W.C.; Gao, J.; Chib, S.; Yu, Y.; Ira, G.; Raney, K.D.; Kim, N. Yeast Sub1 and human PC4 are G-quadruplex binding proteins that suppress genome instability at co-transcriptionally formed G4 DNA. *Nucleic Acids Res.* **2017**, *45*, 5850–5862. [CrossRef] [PubMed]

77. van Wietmarschen, N.; Merzouk, S.; Halsema, N.; Spierings, D.C.J.; Guryev, V.; Lansdorp, P.M. BLM helicase suppresses recombination at G-quadruplex motifs in transcribed genes. *Nat. Commun.* **2018**, *9*, 271. [CrossRef] [PubMed]
78. Masai, H.; Tanaka, T. G-quadruplex DNA and RNA: Their roles in regulation of DNA replication and other biological functions. *Biochem. Biophys. Res. Commun.* **2020**, *531*, 25–38. [CrossRef] [PubMed]
79. Rawal, P.; Kummarasetti, V.B.R.; Ravindran, J.; Kumar, N.; Halder, K.; Sharma, R.; Mukerji, M.; Das, S.K.; Chowdhury, S. Genome-wide prediction of G4 DNA as regulatory motifs: Role in *Escherichia coli* global regulation. *Genome Res.* **2006**, *16*, 644–655. [CrossRef] [PubMed]
80. Parekh, V.J.; Węgrzyn, G.; Arluison, V.; Sinden, R.R. Genomic instability of G-quadruplex sequences in *Escherichia coli*: Roles of DinG, RecG, and RecQ helicases. *Genes* **2023**, *14*, 1720. [CrossRef]
81. Paull, T.T. RNA–DNA hybrids and the convergence with DNA repair. *Crit. Rev. Biochem. Mol. Biol.* **2019**, *54*, 371–384. [CrossRef]
82. García-Muse, T.; Aguilera, A. R loops: From physiological to pathological roles. *Cell* **2019**, *179*, 604–618. [CrossRef]
83. Ortega, P.; Gómez-González, B.; Aguilera, A. Rpd3L and Hda1 histone deacetylases facilitate repair of broken forks by promoting sister chromatid cohesion. *Nat. Commun.* **2019**, *10*, 5178. [CrossRef]
84. Gómez-González, B.; Aguilera, A. Break-induced RNA–DNA hybrids (BIRDHs) in homologous recombination: Friend or foe? *Embo Rep.* **2023**, *24*, e57801. [CrossRef] [PubMed]
85. Kim, N.; Jinks-Robertson, S. Transcription as a source of genome instability. *Nat. Rev. Genet.* **2012**, *13*, 204–214. [CrossRef] [PubMed]
86. Gómez-González, B.; Aguilera, A. Transcription-mediated replication hindrance: A major driver of genome instability. *Genes Dev.* **2019**, *33*, 1008–1026. [CrossRef] [PubMed]
87. Canela, A.; Maman, Y.; Jung, S.; Wong, N.; Callen, E.; Day, A.; Kieffer-Kwon, K.-R.; Pekowska, A.; Zhang, H.; Rao, S.S.P.; et al. Genome organization drives chromosome fragility. *Cell* **2017**, *170*, 507–521. [CrossRef] [PubMed]
88. Ortega, P.; Gómez-González, B.; Aguilera, A. Heterogeneity of DNA damage incidence and repair in different chromatin contexts. *DNA Repair.* **2021**, *107*, 103210. [CrossRef] [PubMed]
89. Smith, L.C.; Lun, C.M. The *SpTransformer* gene family (formerly *Sp185/333*) in the purple sea urchin and the functional diversity of the anti-pathogen rSpTransformer-E1 protein. *Front. Immunol.* **2017**, *8*, 725. [CrossRef] [PubMed]
90. Fernandez-Guerra, A.; Aze, A.; Morales, J.; Mulner-Lorillon, O.; Cosson, B.; Cormier, P.; Bradham, C.; Adams, N.; Robertson, A.J.; Marzluff, W.F.; et al. The genomic repertoire for cell cycle control and DNA metabolism in *S. purpuratus*. *Dev. Biol.* **2006**, *300*, 238–251. [CrossRef]
91. Reinardy, H.C.; Bodnar, A.G. Profiling DNA damage and repair capacity in sea urchin larvae and coelomocytes exposed to genotoxicants. *Mutagenesis* **2015**, *30*, 829–839. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

Supplementary File; BAC insert sequence alignments

Local genomic instability of the SpTransformer gene family in the purple sea urchin inferred from BAC insert deletions

Megan A. Barela Hudgell, Farhana Momtaz, Abiha Jafri, Max A. Alekseyev, L. Courtney Smith

Methods

The sequence reads for each BAC insert were assembled into contigs as described in the main paper. The consensus sequences were aligned by hand in Molecular Evolutionary Genetics Analysis X (MEGAX) against the original BAC insert sequences (GenBank Accession numbers KU668451 [BAC-51], KU668452 [BAC-52]) (1). The alignments shown below were generated in BioEdit (version 5.0.9), formatted and exported as rtf files, and imported into this document.

BAC insert sequence alignments

Alignment 1. BAC-51 and BAC-51-15

```
                     18910     18920     18930     18940     18950     18960     18970     18980     18990     19000
                ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|
BAC-51          TCATTCAACATACCGTAAAATAGCTTCATCATGTTCTTTCAGAAGATTTTAATTTAAAATATTTATTGTGGTTGGGGTATGTTTTCTCCAACTGCAAATG
BAC-51-15       TCATTCAACATACCGTAAAATAGCTTCATCATGTTCTTTCAGAAGATTTTAATTTAAAATATTTATTGTGGTTGGGGTATGTTTTCTCCAACTGCAAATG

                     19010     19020     19030     19040     19050     19060     19070     19080     19090     19100
                ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|
BAC-51          ATGTGCATTATATTTTATAAGGTTCATTTTAAATAAGAAACTGTAATAACCTAGTATTTTTGATCAAATATGTTGCTCAGAAGTCTTGAACAGCATTAAT
BAC-51-15       ATGTGCATTATATTTTATAAGGTTCATTTTAAATAAGAAACTGTAATAACCTAGTATTTTTGATCAAATATGTTGCTCAGAAGTCTTGAACAGCATTAAT
```

```
                     19110     19120     19130     19140     19150     19160     19170     19180     19190     19200
                ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|
BAC-51          ACACGACTTAACTACAGACCCTGATTGAATTCCATTTTCAGTCAGCATTGTTGCATAAATTCATTTATTTATGACTTATTTGACTTCTGCATTAAAACAG
BAC-51-15       ACACGACTTAACTACAGACCCTGATTGAATTCCATTTTCAGTCAGCATTGTTGCATAAATTCATTTATTTATGACTTATTTGACTTCTGCATTAAAACAG

                     19210     19220     19230     19240     19250     19260     19270     19280     19290     19300
                ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|
BAC-51          CTTGACAACTTGTTCAGAGGCACCCCCATCGGAAGTGACTTTGTGGGACTAGAGATCACCAGCAGTCGCTTTGATACTAATAGTAACTACGTGATTGGCT
BAC-51-15       CTTGACAACTTGTTCAGAGGCACCCCCATCGGAAGTGACTTTGTGGGACTAGAGATCACCAGCAGTCGCTTTGATACTAATAGTAACTACGTGATTGGCT

                     19310     19320     19330     19340     19350     19360     19370     19380     19390     19400
                ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|
BAC-51          ACTCCATCTACCTGTGTGCTACCTCCCACTACAATGAGGAGAGCATCAAGGATACCTTCAATAGCGCCCTCACAGGGGCTAACATGGACACCTTTGCCGC
BAC-51-15       ACTCCATCTACCTGTGTGCTACCTCCCACTACAATGAGGAGAGCATCAAGGATACCTTCAATAGCGCCCTCACAGGGGCTAACATGGACACCTTTGCCGC

                     19410     19420     19430     19440     19450     19460     19470     19480     19490     19500
                ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|
BAC-51          TTCAAATGTCATTGTGACTGATTCACTTGACTTCAGTAAGTCACTCTACCGAGTTCCAGTTTTTTTAAGGGCAGCAAACATAGACTATCTGTCTAATATT
BAC-51-15       TTCAAATGTCATTGTGACTGATTCACTTGACTTCAGTAAGTCACTCTACCGAGTTCCAGTTTTTTTAAGGGCAGCAAACATAGACTATCTGTCTAATATT

                     19510     19520     19530     19540     19550     19560     19570     19580     19590     19600
                ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|
BAC-51          TTATTGTAATCACTCCATGGCAGTGTTTCGCAAAGACTAAGATCGACTTTATGTTGGACTTAACTATGGCAGGCCATTCATGCCACTGTTTGCTCTCACC
BAC-51-15       TTATTGTAATCACTCCATGGCAGTGTTTCGCAAAGACTAAGATCGACTTTATGTTGGACTTAACTATGGCAGGCCATTCATGCCACTGTTTGCTCTCACC

                     19610     19620     19630     19640     19650     19660     19670     19680     19690     19700
                ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|
BAC-51          ATTTACTGGCATTGGTCTCTCAAGGGACGTACAAATCATGGTCACAAATCTACAGATTTTACATTTTTGGATAATTCATTATTGAAGGAAAAATATTATC
BAC-51-15       ATTTACTGGCATTGGTCTCTCAAGGGACGTACAAATCATGGTCACAAATCTACAGATTTTACATTTTTGGATAATTCATTATTGAAGGAAAAATATTATC

                     19710     19720     19730     19740     19750     19760     19770     19780     19790     19800
                ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|
BAC-51          TAAAGATCATAATGGTGAGAGCAAATAGTGGCATGAACGGCCTGCCATAGTTAAGTCCAACTTAAAGTTGGACTTTCCTCTCTCTCACATGGTGCACTCG
BAC-51-15       TAAAGATCATAATGGTGAGAGCAAATAGTGGCATGAACGGCCTGCCATAGTTAAGTCCAACTTAAAGTTGGACTTTCCTCTCTCTCACATGGTGCACTCG

                     19810     19820     19830     19840     19850     19860     19870     19880     19890     19900
                ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|
BAC-51          CAAGCCTACTGTATAAACTTAAAGTTTATCTTACTCTATGTGAAACACTGCCCATGTCATATTATAATTTGCCTTGTTTGTTGGTGTTCAAATAAACACA
BAC-51-15       CAAGCCTACTGTATAAACTTAAAGTTTATCTTACTCTATGTGAAACACTGCCCATGTCATATTATAATTTGCCTTGTTTGTTGGTGTTCAAATAAACACA

                     19910     19920     19930     19940     19950     19960     19970     19980     19990     20000
                ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|
BAC-51          ATCAGGATGACACAAGTACATGAATTATATATTTCATGGAAACATCATCATTCTTCTTAGCTCAAAGGTAGATAAGGATGCATTCAATAGATCCAGCCAT
BAC-51-15       ATCAGGATGACACAAGTACATGAATTATATATTTCATGGAAACATCATCATTCTTCTTAGCTCAAAGGTAGATAAGGATGCATTCAATAGATCCAGCCAT

                     20010     20020     20030     20040     20050     20060     20070     20080     20090     20100
                ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|
BAC-51          ATAATGAGATCCTCCCAGCACGGTGTTGGTTCATGTATTTGACAGCTATATAATTAACTCATGGTCTCTACTCTCTCTCTCTTTCTCTCTCTTTCTCTCT
BAC-51-15       ATAATGAGATCCTCCCAGCACGGTGTTGGTTCATGTATTTGACAGCTATATAATTAACTCATGGTCTCTACTCTCTCTCTCTTTCTCTCTCTTTCTCTCT
```

```
                     20110     20120     20130     20140     20150     20160     20170     20180     20190     20200
                ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|
BAC-51          CTCTCTTTCTCTCTCTCTCTCTCTCTCTGTCTCTTTCTCTCTCTCTCCCTTTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCCTCTACCTCTCTA
BAC-51-15       CTCTCTTTCTCTCTCTCTCTCTCTCTCTGTCTCTTTCTCTCTCTCTCCCTTTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCT--------------
```

Deletion 1, 5' end

```
                     20210     20220     20230     20240     20250     20260     20270     20280     20290     20300
                ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|
BAC-51          CTGTATGTGTCTGTTTATCTCTCTACAGATGTGGTAGAGTTTAGGGGTCAGTTTACACTGAAGAGCATTGAGAATCCTGGAACCGTTCCTGTCGAATACC
BAC-51-15       ----------------------------------------------------------------------------------------------------

                     20310     20320     20330     20340     20350     20360     20370     20380     20390     20400
                ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|
BAC-51          AATCATGCTGCTCAGATCCATCCAGCAGTACCTACCAAGCACTTCAGAACAGAATCTATGATCATGTAAGTATACCAAGATAGTTATATATAGTCAAATT
BAC-51-15       ----------------------------------------------------------------------------------------------------
```

Sequences (96,854 bp) in BAC-51 that are deleted in BAC-51-15 are omitted

```
                     116610    116620    116630    116640    116650    116660    116670    116680    116690    116700
                ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|
BAC-51          GAGAGAGAGAGAGAGATCCTGCTACTACTACTGTCGCTAGATTACTGCTATAAAAACTACTATTTCTTAAAAGGCAAGTACACTCCAAAAATACCTTACT
BAC-51-15       ----------------------------------------------------------------------------------------------------

                     116710    116720    116730    116740    116750    116760    116770    116780    116790    116800
                ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|
BAC-51          TTCAATAGAAATCAGACAATATGTAATCAGACAATTATTTCCCTCACTTTCACTATTATATAAAATTTCTTTTGTCTATCTTCTTCTTCTCTCTCTCCCC
BAC-51-15       ----------------------------------------------------------------------------------------------------

                     116810    116820    116830    116840    116850    116860    116870    116880    116890    116900
                ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|
BAC-51          CCCTCTCTCTCTCTCTCTCTCTCTCTCACACACACACACACAAACGTACACACTCACACACACACTCTCTCTCTCTCCCTCTCTCCATCTCACTCTCATCTC
BAC-51-15       ----------------------------------------------------------------------------------------------------

                     116910    116920    116930    116940    116950    116960    116970    116980    116990    117000
                ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|
BAC-51          TGTCTTTGTCTCTTTGTCTCTCCCTCCCCCCCCCTCTCTCTCTCTCTCTATATATAATCAGACAATATATAATACCTATATCCCTCACTCTCACTTTCTC
BAC-51-15       ----------------------------------------------------------------------------------------------------

                     117010    117020    117030    117040    117050    117060    117070    117080    117090    117100
                ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|
BAC-51          ACTCCCTCTCGTTTTTTTTCTTTCTCTCTCTCTCTCTCTCTCTATCTATCTCTATCTCTACCTATCTCTCTCTCTCTCTTTCTTTTTCTCTCTCTCTC
BAC-51-15       ---------------------------------------------ATCTATCTCTATCTCTACCTATCTCTCTCTCTCTCTTTCTTTTTCTCTCTCTCTC
```

Deletion 1, 3' end

```
                     117110    117120    117130    117140    117150    117160    117170    117180    117190    117200
                ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|
BAC-51          TCTCTATCAATCTCTATCTCTACCTATCTCTCTCTCTCTCTCTCTCTTTCTCTCTCTCTCTCTCTCTCTCTTATCTCAATCGATCAATATCGTGAT
BAC-51-15       TCTCTATCAATCTCTATCTCTACCTATCTCTCTCTCTCTCTCTCTCTTTCTCTCTCTCTCTCTCTCTCTCTTATCTCAATCGATCAATATCGTGAT
```

```
                       117210    117220    117230    117240    117250    117260    117270    117280    117290    117300
                ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|
BAC-51          CAATCCAGCTGCATAAGGAAAACAATTATTAGGGGCTATAATATTGAACGCTTCACAAGTATTAGTATGAGTTCAGTTATATATTTTTATTTGACAAAGT
BAC-51-15       CAATCCAGCTGCATAAGGAAAACAATTATTAGGGGCTATAATATTGAACGCTTCACAAGTATTAGTATGAGTTCAGTTATATATTTTTATTTGACAAAGT

                       117310    117320    117330    117340    117350    117360    117370    117380    117390    117400
                ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|
BAC-51          CTTGACAAATTCAGTATCAATAAGAGTTTAACAATTCGTCAACATTTCATGCACCTTTCACTGGCTAAGAGTCTCCTGTATCTGAATTTTAGACACAATA
BAC-51-15       CTTGACAAATTCAGTATCAATAAGAGTTTAACAATTCGTCAACATTTCATGCACCTTTCACTGGCTAAGAGTCTCCTGTATCTGAATTTTAGACACAATA

                       117410    117420    117430    117440    117450    117460    117470    117480    117490    117500
                ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|
BAC-51          ATGTTTTAGTCTCTGGTAAGATATGCATTCTGTCCAAGGAAAGACATTAGTGACTATTAGTATATTTACACAAGCAATCTTTTTTTATACAAATTGTACA
BAC-51-15       ATGTTTTAGTCTCTGGTAAGATATGCATTCTGTCCAAGGAAAGACATTAGTGACTATTAGTATATTTACACAAGCAATCTTTTTTTATACAAATTGTACA

                       117510    117520    117530    117540    117550    117560    117570    117580    117590    117600
                ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|
BAC-51          CCTCGGCGACCACATTGACGTCTTCTTCATTGACTGCGATCTCAACCACTTCAGGTGGTGAAGATGTCGTTACGCTTTCCGTCGTTGGATGCTCCTTCTC
BAC-51-15       CCTCGGCGACCACATTGACGTCTTCTTCATTGACTGCGATCTCAACCACTTCAGGTGGTGAAGATGTCGTTACGCTTTCCGTCGTTGGATGCTCCTTCTC

                       117610    117620    117630    117640    117650    117660    117670    117680    117690    117700
                ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|
BAC-51          CTCACTTTCATTCTCACCCCAACGTCCTCGGTTACCATAGGGATGACGGTGGCCATCACGCCTGGGAGATCCTTCTTCGGTTCTGTTGCGTCTGCCGAAG
BAC-51-15       CTCACTTTCATTCTCACCCCAACGTCCTCGGTTACCATAGGGATGACGGTGGCCATCACGCCTGGGAGATCCTTCTTCGGTTCTGTTGCGTCTGCCGAAG

                       117710    117720    117730    117740    117750    117760    117770    117780    117790    117800
                ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|
BAC-51          GGACGGTCTCCGAAAGGCTTTCTACCGAAGGGGTTGACCCAGAAGGGCCTCATCTCAAACATTGGTCTGTCCTGATCTCCATCTCCGGTCTTGTTGTGAT
BAC-51-15       GGACGGTCTCCGAAAGGCTTTCTACCGAAGGGGTTGACCCAGAAGGGCCTCATCTCAAACATTGGTCTGTCCTGATCTCCATCTCCGGTCTTGTTGTGAT

                       117810    117820    117830    117840    117850    117860    117870    117880    117890    117900
                ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|
BAC-51          GACGATGGGGGTGATCTCCCGTCTCATTATGACCCTGGTGACCTTCTGTGTGGTTGCGTCTGCCGAAGTGGTTGAACCGGAAGGGCCTCGTATCATTTGG
BAC-51-15       GACGATGGGG-TGATCTCCCGTCTCATTATGACCCTGGTGACCTTCTGTGTGGTTGCGTCTGCCGAAGTGGTTGAACCGGAAGGGCCTCGTATCATTTGG

                       117910    117920    117930    117940    117950    117960    117970    117980    117990    118000
                ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|
BAC-51          TTTGTCCTGATCTTGATCTCCCGTCTCATTATGTCCTTGGTGACCCTCTGTCTGGTTATGATGACGGTGGTGGTGGCCTTGGTGGCCATGGTGGCCATGA
BAC-51-15       TTTGTCCTGATCTTGATCTCCCGTCTCATTATGTCCTTGGTGACCCTCTGTCTGGTTATGATGACGGTGGTGGTGGCCTTGGTGGCCATGGTGGCCATGA

                       118010    118020    118030    118040    118050    118060    118070    118080    118090    118100
                ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|
BAC-51          TGGTGAGGGTGAGGACGGCCATCCTCCTCGTTGCGTTCATTGCGCTGACCGAACGGTTGTTCCTCTGGTCGGTCCTGAGGTGGACCTTGACGGTGACCAT
BAC-51-15       TGGTGAGGGTGAGGACGGCCATCCTCCTCGTTGCGTTCATTGCGCTGACCGAACGGTTGTTCCTCTGGTCGGTCCTGAGGTGGACCTTGACGGTGACCAT

                       118110    118120    118130    118140    118150    118160    118170    118180    118190    118200
                ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|
BAC-51          GATGACGACGTCCATGACCATCAAACTGACCGCGCCCTCCTAGACCATCACCAATTTGTTGGGAAGCATCAGTTTCTTCTTCTCCATCACCACGTCTTCC
BAC-51-15       GATGACGACGTCCATGACCATCAAACTGACCGCGCCCTCCTAGACCATCACCAATTTGTTGGGAAGCATCAGTTTCTTCTTCTCCATCACCACGTCTTCC
```

```
                       118210    118220    118230    118240    118250    118260    118270    118280    118290    118300
                ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|
BAC-51          TCCTTGGCCGAAGAAAGGTCTTCCTCCAGCACCATCTGGTCTGGAGCCACCAAATCCAGGTCCGTCGAATCTCCTACCACCCATCGGACCGCCATTTTGT
BAC-51-15       TCCTTGGCCGAAGAAAGGTCTTCCTCCAGCACCATCTGGTCTGGAGCCACCAAATCCAGGTCCGTCGAATCTCCTACCACCCATCGGACCGCCATTTTGT

                       118310    118320    118330    118340    118350    118360    118370    118380    118390    118400
                ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|
BAC-51          CTGCGTCCATCCATATGTGGGGCACCAAATCCAGGTCCATCGAACCTCCTTCCACCCATTGGTCCACCATCTTGCCTAGATCCACCCATCTGCATTCCAC
BAC-51-15       CTGCGTCCATCCATATGTGGGGCACCAAATCCAGGTCCATCGAACCTCCTTCCACCCATTGGTCCACCATCTTGCCTAGATCCACCCATCTGCATTCCAC

                       118410    118420    118430    118440    118450    118460    118470    118480    118490    118500
                ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|
BAC-51          CAGGCCTTCCTCCAAAGCGACCTTGTCCTCTCTCTCTGCCATTCTCATTTCCTCGCCGTTCATTGAAATCTCTTCGTGCGTGAGCTGTAAGGGTTAATGA
BAC-51-15       CAGGCCTTCCTCCAAAGCGACCTTGTCCTCTCTCTCTGCCATTCTCATTTCCTCGCCGTTCATTGAAATCTCTTCGTGCGTGAGCTGTAAGGGTTAATGA

                       118510    118520    118530    118540    118550    118560    118570    118580    118590    118600
                ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|
BAC-51          AATAATCGAATTAATTAAGTAATCACATTCTGCATTATTATATTTGGCTCTACATTGGTTTATAGCAGGTGACAAATAAGAATGGTAACTCGCCTTCTTA
BAC-51-15       AATAATCGAATTAATTAAGTAATCACATTCTGCATTATTATATTTGGCTCTACATTGGTTTATAGCAGGTGACAAATAAGAATGGTAACTCGCCTTCTTA

                       118610    118620    118630    118640    118650    118660    118670    118680    118690    118700
                ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|
BAC-51          TATTTTT-TCCCTTTATGTTTAATCCTAATCCGTAAGTGCGGTTCGATATTCAAAACGCCTAATTCTGACTGAATTAAATTTGATTACCAGATTGAACGT
BAC-51-15       TATTTTTTTCCCTTTATGTTTAATCCTAATCCGTAAGTGCGGTTCGATATTCAAAACGCCTAATTCTGACTGAATTAAATTTGATTACCAGATTGAACGT

                       118710    118720    118730    118740    118750    118760    118770    118780    118790    118800
                ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|
BAC-51          AAACATAATACACATCCTTTGTCATTCAGAAAACGGATTGGTGTAGGCTTAATAAATACAATTAATTATATTGTGTAACGAACAATTTAAAAATGCATAT
BAC-51-15       AAACATAATACACATCCTTTGTCATTCAGAAAACGGATTGGTGTAGGCTTAATAAATACAATTAATTATATTGTGTAACGAACAATTTAAAAATGCATAT

                       118810    118820    118830    118840    118850    118860    118870    118880    118890    118900
                ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|
BAC-51          TACAAACTATGAAATAATCATATTTACTTCTGTGAGCCCTTCGTTGGCTTTATATTTAGCACTTATCAAGTAATACCGAGTAATAATTTGATTTCTTACC
BAC-51-15       TACAAACTATGAAATAATCATATTTACTTCTGTGAGCCCTTCGTTGGCTTTATATTTAGCACTTATCAAGTAATACCGAGTAATAATTTGATTTCTTACC

                       118910    118920    118930    118940    118950    118960    118970    118980    118990    119000
                ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|
BAC-51          CGAGATAGCAAGAGCAGCCACAATGGCAACGATCAATGTTGCTTTCACCTCCATGTTTGTAAGGTCTCTCCGATGCTACAAGCTTTCTCTAGATTCGTTG
BAC-51-15       CGAGATAGCAAGAGCAGCCACAATGGCAACGATCAATGTTGCTTTCACCTCCATGTTTGTAAGGTCTCTCCGATGCTACAAGCTTTCTCTAGATTCGTTG

                       119010    119020    119030    119040    119050    119060    119070    119080    119090    119100
                ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|
BAC-51          CCTTCCAAGAGAGAACTAGCTCCAAACTTAACAACTACCTGCTGGGCCACTGAATTTATAGGTTTTCTACCCTAGATTGATATCTCACACTAGTACCAGA
BAC-51-15       CCTTCCAAGAGAGAACTAGCTCCAAACTTAACAACTACCTGCTGGGCCACTGAATTTATAGGTTTTCTACCCTAGATTGATATCTCACACTAGTACCAGA
```

Matched sequences (37,500 bp) between the BACs are omitted

```
                       156610    156620    156630    156640    156650    156660    156670    156680    156690    156700
                ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|
BAC-51          GGGCACTTATCAGAGAAGAAAAGGAGTAGTTATAAAAAGGAG---AAGGGGCACTCTGTTTTAATGAAAAGCGCATTTATTAGAAAGGTAAAGGGGCACT
BAC-51-15       GGGCACTTATCAGAGAAGAAAAGGAGTAGTTATAAAAAG-AGAAGAAGGGGCACTCTGTTTTAATGAAAAGCGCATTTATTAGAAAGGTAAAGGGGCACT
```

```
                       156710    156720    156730    156740    156750    156760    156770    156780    156790    156800
                ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|
BAC-51          TGTCAAAGTAAAACGAGCATCTCTCATATTTGAATAGGGGCAATTATCGGACGTGAAAGAAGCACTAGTCTTATCTGAAATGAAAAGGAGCAACTTCTTT
BAC-51-15       TGTCAAAGTAAAACGAGCATCTCTCATATTTGAATAGGGGCAATTATCGGACGTGAAAGAAGCACTAGTCTTATCTGAAATGAAAAGGAGCAACTTCTTT

                       156810    156820    156830    156840    156850    156860    156870    156880    156890    156900
                ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|
BAC-51          TTACCTGAAAGGGGTCATTCACCAGAGCTGAAAAGAGGCATTTATGAAAAGGAGAAGGGGATCTTTGTTTTAATGGGG-AACTAACCATCAGGAAAAGGG
BAC-51-15       TTACCTGAAAGGGGTCATTCACCAGAGCTGAAAAGAGGC-TTTATGAAAAGGAGAAGGGGATCTTTGTTTTAATGGGGGAACTAACCATCAGGAAAAGGG

                       156910    156920    156930    156940    156950    156960    156970    156980    156990    157000
                ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|
BAC-51          GCACCTCTCAAATGTTAATAAGGCAATTATCAGAGGTGAAGGGGGCACTTATTACAAGTGAAAAGGCGCACGTATCGTACCTGAAAAGAGGCACTTACTA
BAC-51-15       GCACCTCTCAAATGTTAATAAGGCAATTATCAGAGGTGAAGGGGGCACTTATTACAAGTGAAAAGGCGCACGTATCGTACCTGAAAAGAGGCACTTACTA

                       157010    157020    157030    157040    157050    157060    157070    157080    157090    157100
                ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|
BAC-51          GACATTAAAAGAGGCACACAGCATGAGTAGTAAACACAGCCGTGGAAAAATGTAAGCAACACAAGTTCGTATATAGTCGTACGATGAGCGATTGAAACAA
BAC-51-15       GACATTAAAAGAGGCACACAGCATGAGTAGTAAACACAGCCGTG-AAAAATGTAAGCAACACAAGTTCGTATATAGTCGTACGATGAGCGATTGAAACAA

                       157110    157120    157130    157140    157150    157160    157170    157180    157190    157200
                ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|
BAC-51          CGAAGTCTGACGGATCTTTGGTCGAAGAGCACACCTGTCAGCACTGGCGTACTGGTAGGGGGCTAGGGG-AGGGGGGGGGG-CGCCATCCCCCAAATAAA
BAC-51-15       CGAAGTCTGACGGATCTTTGGTCGAAGAGCACACCTGTCAGCACTGGCGTACTGGTAGGGGGCTAGGGGGAGGGGGGGGGGGCGCCATCCCCCAAATAAA

                       157210    157220    157230    157240    157250    157260    157270    157280    157290    157300
                ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|
BAC-51          ATTGAACCATGACATATTTTCAATTCCATATACAAGTGATGTATAAATGTTCGTGATATCACATGATTTTCAACAAAATGCGTTTTTCTCCATTTCTGTT
BAC-51-15       ATTGAACCATGACATATTTTCAATTCCATATACAAGTGATGTATAAATGTTCGTGATATCACATGATTTTCAACAAAATGCGTTTTTCTCCATTTCTGTT

                       157310    157320    157330    157340    157350    157360    157370    157380    157390    157400
                ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|
BAC-51          GCAAACTGACACATTCCCGAAGCCAGGAAGAGCATATGTTTTTTTTATAATCACTAAGGGGCCTTATAGTGGCCACTAACAGTAACACTTGTATAAAAAA
BAC-51-15       GCAAACTGACACATTCCCGAAGCCAGGAAGAGCATATGTTTTTTTTATAATCACTAAGGGGCCTTATAGTGGCCACTAACAGTAACACTTGTATAAAAAA

                       157410    157420    157430    157440    157450    157460    157470    157480    157490    157500
                ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|
BAC-51          ACACACTCATTCACAGTGTGAACACACAGTTCACACTGTGAACGGAGCCTTTTCACAGTGTGGACACTGTCACAATGTGGTAAA--CCACACAGTATTCA
BAC-51-15       ACACACTCATTCACAGTGTGAACACACAGTTCACACTGTGAACGGAGCCTTTTCACAGTGTGGACACTGTCACAATGTGGTAAAAACCACACAGTATTCA

                       157510    157520    157530    157540    157550    157560    157570    157580    157590    157600
                ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|
BAC-51          CATTGTAAAATTCGATTTCACAGTGTGAATGTACCTGCTCACAGTGTGAAAATGATTTCACAGTGTGACCAAACTGTGATAAATAGATTTTCACAGTGTG
BAC-51-15       CATTGTAAAATTCGATTTCACAGTGTGAATGTACCTGCTCACAGTGTGAAAATGATTTCACAGTGTG---------------------------------
```

Deletion 2, 5’ end, this is likely an assembly artifact

```
                       157610    157620    157630    157640    157650    157660    157670    157680    157690    157700
                ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|
BAC-51          AATACCTTACGACATACATTTCACGGTGTGAACAGTAAATTTCTCAGTGTGGCCACTCTGTGAAAAATACTTCACATTGTGAAAAAATAGAATTCACAGT
BAC-51-15       ----------------------------------------------------------------------------------------------------
```

```
                    157710    157720    157730    157740    157750    157760    157770    157780    157790    157800
                ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|
BAC-51          ATGGCCGCTGGAATACTAGTGTTTTGGATCGAGGCCAATCAAGTTTCCGTGATAACAGACGAATATGTACAACTACATGTTGACGGATGAGAACATAATT
BAC-51-15       ----------------------------------------------------------------------------------------------------

                    157810    157820    157830    157840    157850    157860    157870    157880    157890    157900
                ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|
BAC-51          GGAATGGCGGACATGGCCTGGTGAAGTTTCCCAACGGCTAGGATAGCTATCAGGCGAGGGTGTTGCCGGGCTGCTGCCGGGCTGTCATTGAATGGTGAGT
BAC-51-15       ----------------------------------------------------------------------------------------------------
```

Sequence (2,270 bp) in BAC-51 that deleted from BAC-51-15 is omitted, this is likely an assembly artifact

```
                    159510    159520    159530    159540    159550    159560    159570    159580    159590    159600
                ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|
BAC-51          TCCTGGAGTATGCGTGCGTGCACAGTGCGGGATCCTCACACCCAGGACAACATACACAGGCTGGAAATGGTACAGCTCCGCAACGCCCAGTTTATCACTG
BAC-51-15       ----------------------------------------------------------------------------------------------------

                    159610    159620    159630    159640    159650    159660    159670    159680    159690    159700
                ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|
BAC-51          GAGATCACCATCATACCAGCAGCATCGCACCAATGCTGCAACATCTCAAATGGCCATCTCTCCGAGAACGCAGGGCACGGTTTAAGATGGTGATAGTGTA
BAC-51-15       ----------------------------------------------------------------------------------------------------

                    159710    159720    159730    159740    159750    159760    159770    159780    159790    159800
                ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|
BAC-51          CGGAACTGTCAACCTGCACTCAGAGTTTCTCAAAATGTACCTTCTATCGGTGTTAACATCCTACCATGGTCATACCCAGTAATTCCAAATACAATTTACA
BAC-51-15       ----------------------------------------------------------------------------------------------------

                    159810    159820    159830    159840    159850    159860    159870    159880    159890    159900
                ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|
BAC-51          ATGACTCTCGTATACCAGAAGTCATTCTTCCCCGATTCCATCATGCTTTGGAACTTGCTACCTGCTGGATTTGTCAACTGTACTTTCGCTAAAGCTTTTA
BAC-51-15       --------------------------------------CATCATGCTTTGGAACTTGCTACCTGCTGGATTTGTCAACTGTACTTTCGCTAAAGCTTTTA
```

Deletion 2, 3’ end, this is likely an assembly artifact

```
                    159910    159920    159930    159940    159950    159960    159970    159980    159990    160000
                ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|
BAC-51          AGCAAGAGGTACAAAATTTCCATCTCCGTTAGAAGGGGAAGAAGGATTTTTAACTGCTCATAGAATATTATAGAAAGTTTTAAATTCGCACTGTATATTC
BAC-51-15       AGCAAGAGGTACAAAATTTCCATCTCCGTTAGAAGGGGAAGAAGGATTTTTAACTGCTCATAGAATATTATAGAAAGTTTTAAATTCGCACTGTATATTC

                    160010    160020    160030    160040    160050    160060    160070    160080    160090    160100
                ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|
BAC-51          GCGACACCAGACCTGACGTACGTGATACACCATACGGTGGACTGGAGTGATATCTCAAGATCTGGGGGATGATAATTCGGGCTTTAACAAATTGTGTGAT
BAC-51-15       GCGACACCAGACCTGACGTACGTGATACACCATACGGTGGACTGGAGTGATATCTCAAGATCTGGGGGATGATAATTCGGGCTTTAACAAATTGTGTGAT

                    160110    160120    160130    160140    160150    160160    160170    160180    160190    160200
                ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|
BAC-51          AGTGGAGGTATTATTTTTGACTGCACAGTCGGAAGTTCCATCTAGTTATAATATCCTAACATCAGTATATTATCGCTAGAACAAAAAATCATAAGTTACT
BAC-51-15       AGTGGAGGTATTATTTTTGACTGCACAGTCGGAAGTTCCATCTAGTTATAATATCCTAACATCAGTATATTATCGCTAGAACAAAAAATCATAAGTTACT
```

```
                      160210    160220    160230    160240    160250    160260    160270    160280    160290    160300
                  ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|
BAC-51            TATGTAGTTTAGGAAGTACATATATTTTTGTATGCCATTCTGCTTTTTTCTGTGCCACCGTATATAGAGTGTATGGTTATTTATTTGTAAAAAGATGAGC
BAC-51-15         TATGTAGTTTAGGAAGTACATATATTTTTGTATGCCATTCTGCTTTTTTCTGTGCCACCGTATATAGAGTGTATGGTTATTTATTTGTAAAAAGATGAGC

                      160310    160320    160330    160340    160350    160360    160370    160380    160390    160400
                  ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|
BAC-51            AAAATCTGTTTTTTTTCTCATCTTCAGGTATTTATTAGTATTAGTTGTAGAAATTTATGTAGCCATAACAGTAAAACGTGTCTGCCTAAAATAGATTTGT
BAC-51-15         AAAATCTGTTTTTTTTCTCATCTTCAGGTATTTATTAGTATTAGTTGTAGAAATTTATGTAGCCATAACAGTAAAACGTGTCTGCCTAAAATAGATTTGT

                      160410    160420    160430    160440    160450    160460    160470    160480    160490    160500
                  ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|
BAC-51            TTTCAACTTTTTTTCAAGATGTATTGACAGAGATATTTTGAAACTTTTTATCATTTTAATAGGTTTATGAAGAACCTGTATGTGCCAAAGATTTTCTGAT
BAC-51-15         TTTCAACTTTTTTTCAAGATGTATTGACAGAGATATTTTGAAACTTTTTATCATTTTAATAGGTTTATGAAGAACCTGTATGTGCCAAAGATTTTCTGAT

                      160510    160520    160530    160540    160550    160560    160570    160580    160590    160600
                  ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|
BAC-51            ACATTGCAATTGCCAAAGTTACATGTCATGTGTGGATATTCTATACTTGGAAAATGTGACTAGCCATACTAATGTATCTTCATAAATTTAATATACCTAC
BAC-51-15         ACATTGCAATTGCCAAAGTTACATGTCATGTGTGGATATTCTATACTTGGAAAATGTGACTAGCCATACTAATGTATCTTCATAAATTTAATATACCTAC

                      160610    160620    160630    160640    160650    160660    160670    160680    160690    160700
                  ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|
BAC-51            AGAAATGTTTAAGAGATATTGCCAGAGATGCTCTAAAACTATTTT-ATTACTGACGTAACATGCTAATGTGCTTTTATTATGTTTTACGAGTGTCAGTAC
BAC-51-15         AGAAATGTTTAAGAGATATTGCCAGAGATGCTCTAAAACTATTTTTATTACTGACGTAACATGCTAATGTGCTTTTATTATGTTTTACGAGTGTCAGTAC

                      160710    160720    160730    160740    160750    160760    160770    160780    160790    160800
                  ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|
BAC-51            GAGTTCAGGGAAATATTGATGATATGTGCAATTTTTGCCGTCACTGTAATTTATTCCTCGATGTTATACAGTGGAACCTCTATAACTCTATTTCACAAGT
BAC-51-15         GAGTTCAGGGAAATATTGATGATATGTGCAATTTTTGCCGTCACTGTAATTTATTCCTCGATGTTATACAGTGGAACCTCTATAACTCTATTTCACAAGT

                      160810    160820    160830    160840    160850    160860    160870    160880    160890    160900
                  ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|
BAC-51            CCACGTTTTTCTGTAATTCATTGTTTTAAACCCTTGTACAATAAGATATCTGATATAAAGAGGTCAATTTAGTGGACACCAATGTGCTCATTATAATGAG
BAC-51-15         CCACGTTTTTCTGTAATTCATTGTTTTAAACCCTTGTACAATAAGATATCTGATATAAAGAGGTCAATTTAGTGGACACCAATGTGCTCATTATAATGAG

                      160910    160920    160930    160940    160950    160960    160970    160980    160990    161000
                  ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|
BAC-51            GTTTCACTGTAGATAGCAAGTTCTGTTTTCAGAAGTATCTCTACTATCTAATTTAAACTGTTTGGATTCATTTGTCTTTGTTTTCATTCTGGTTCACATC
BAC-51-15         GTTTCACTGTAGATAGCAAGTTCTGTTTTCAGAAGTATCTCTACTATCTAATTTAAACTGTTTGGATTCATTTGTCTTTGTTTTCATTCTGGTTCACATC

                      161010    161020    161030    161040    161050    161060    161070    161080    161090    161100
                  ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|
BAC-51            ATGTTGGCATTCACAGAGTAAAATTTGTTTATCTGTAAGTATAGTATTACTGCAATCTTTGTATACCTAAGAGGTATATGTTCATAAGTCCTTTTAGCGA
BAC-51-15         ATGTTGGCATTCACAGAGTAAAATTTGTTTATCTGTAAGTATAGTATTACTGCAATCTTTGTATACCTAAGAGGTATATGTTCATAAGTCCTTTTAGCGA

                      161110    161120    161130    161140    161150    161160    161170    161180    161190    161200
                  ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|
BAC-51            AAAGTACATCAATATAAGCTTTTGAAATACATTTTAAGTTACCTCATAAAGGCTATAACATGTATATGTTTAAGATTACTGTATTATTTTCAGATGTTTC
BAC-51-15         AAAGTACATCAATATAAGCTTTTGAAATACATTTTAAGTTACCTCATAAAGGCTATAACATGTATATGTTTAAGATTACTGTATTATTTTCAGATGTTTC
```

```
                 161210    161220    161230    161240    161250    161260    161270    161280    161290    161300
                 ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|
BAC-51           GGGGGTATATTTCTATTTAAGTTGAGTTTATAAGTGAGACATTGATTATAAACCTTTTCTCCTTCGTATTTATACGTACTAAGCTCTACATGACTACATT
BAC-51-15        GGGGGTATATTTCTATTTAAGTTGAGTTTATAAGTGAGACATTGATTATAAACCTTTTCTCCTTCGTATTTATACGTACTAAGCTCTACATGACTACATT

                 161310    161320    161330    161340    161350    161360    161370    161380    161390    161400
                 ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|
BAC-51           TTCATAGTTTGAAATATTTCACACAGAATTCAAAAACTAAACTTGTATCACAGTGTGGTGGGCGTTTCACAGTGTGGTAATCTCATACAATATAATGTGT
BAC-51-15        TTCATAGTTTGAAATATTTCACACAGAATTCAAAAACTAAACTTGTATCACAGTGTGGTGGGCGTTTCACAGTGTGGTAATCTCATACAATATAATGTGT

                 161410    161420    161430    161440    161450    161460    161470    161480    161490    161500
                 ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|
BAC-51           CACAATGTAACGTAGTGGGCGTTTCACAGTGTGTTTGGTGTTTCACAGTGTGGTGGGCATTTCACTGTGTGAGTGTTCACAATATTAAATTGGGGCGTTT
BAC-51-15        CACAATGTAACGTAGTGGGCGTTTCACAGTGTGTTTGGTGTTTCACAGTGTGGTGGGCATTTCACTGTGTGAGTGTTCACAATATTAAATTGGGGCGTTT

                 161510    161520    161530    161540    161550    161560    161570    161580    161590    161600
                 ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|
BAC-51           TCACATTGTGGTTGCTTCGTTTTACAGTGTGGTTCACAGTGTGGATATCTACACTGAAACGACGCGACCACAATGTGAAAATACCCCAATGTAACATTGT
BAC-51-15        TCACATTGTGGTTGCTTCGTTTTACAGTGTGGTTCACAGTGTGGATATCTACACTGAAACGACGCGACCACAATGTGAAAATACCCCAATGTAACATTGT

                 161610    161620    161630    161640    161650    161660    161670    161680    161690    161700
                 ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|
BAC-51           GGGCGTTTCACAGTGTGTTTGCTGTTTCAAAGTGTGCTGGGCATTTCACTGTGTGAGTGTTTACAATGTTAAATTGGGGCGTTTCACATTGTGGTCGCTT
BAC-51-15        GGGCGTTTCACAGTGTGTTTGCTGTTTCAAAGTGTGCTGGGCATTTCACTGTGTGAGTGTTTACAATGTTAAATTGGGGCGTTTCACATTGTGGTCGCTT

                 161710    161720    161730    161740    161750    161760    161770    161780    161790    161800
                 ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|
BAC-51           CGTTTCACAGTTGGGTTCACAGTATGAAACGACGCGACCACAATGTGAAAACGCCCCAATTTAACATTGTGGGCGTTACACAGTGTGGTGGGCATTTCAC
BAC-51-15        CGTTTCACAGTTGGGTTCACAGTATGAAACGACGCGACCACAATGTGAAAACGCCCCAATTTAACATTGTGGGCGTTACACAGTGTGGTGGGCATTTCAC

                 161810    161820    161830    161840    161850    161860    161870    161880    161890    161900
                 ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|
BAC-51           TGTGTGAGTGTTCACAATGTTAAATTGGGGCGTTTTCACAGTGTGATTCACACTGTGTTTCACACTGTGAATGAGTGTTGTGGTCTGGTCTGGTCGCGAC
BAC-51-15        TGTGTGAGTGTTCACAATGTTAAATTGGGGCGTTTTCACAGTGTGATTCACACTGTGTTTCACACTGTGAATGAGTGTTGTGGTCTGGTCTGGTCGCGAC
```

Alignment 2. BAC-52 and BAC-52-2b

```
                9910       9920       9930       9940       9950       9960       9970       9980       9990      10000
          ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
BAC52     CATACTGAAT ACTAAAAATT GTAAAATAAA -GATGTAGCT GAACTCATAC GAAGGTAATG TGCAATAATA AAGCCCCGAA TATTTTCCTG ACGCAGATGA
BAC52-2b  CATACTGAAT ACTAAAAAAT TGTAAAATAA AGATGTAGCT GAACTCATAC GAAGGTAATG TGCAATAATA AAGCCCCGAA TATTTTCCTG ACGCAGATGA

               10010      10020      10030      10040      10050      10060      10070      10080      10090      10100
          ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
BAC52     ATTGATCATT GTTATTGAAT CAATGAGGAA GACGAAATGA AGGGAAAAGG AATACAATGT ATATAGAGAT AGAGAGAGAG AGAGAGAGGG --ATAGAGAG
BAC52-2b  ATTGATCATT GTTATTGAAT CAATGAGGAA GACGAAATGA AGGGAAAAGG AATACAATGT ATATAGAGAT AGAGAGAGAG AGAGAGAGAG GGATAGAGAG
```

```
                 10110      10120      10130      10140      10150      10160      10170      10180      10190      10200
          ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
BAC52     GGATATATAT ATATATATAT ATATATATAG AGAGAGAGAG AGAGAGAGAG AGA--GAATG AGATAGAGAG AAATAAAAGA GAGAGAGAGA GAA-GAAAGA
BAC52-2b  GGATATATAT ATATATATAT ATATATATAT AGAGAGAGAG AGAGAGAGAG AGA--GAATG AGATAGAGAG AAATAAAAGA GAGAGAGAGA GAAAGAAAGA

                 10210      10220      10230      10240      10250      10260      10270      10280      10290      10300
          ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
BAC52     GAGAGG-AGA GGTGAGATAG ATAGAAAGGG TAGAGAGAGA TGAAGAGACA AAAGAAATAG AGATATGTAG AGATAGAGAG AGCGAGAGGG AGAGACAAAG
BAC52-2b  GAGAGGGAGA GGTGAGATAG ATAGAAAGGG TAGAGAGAGA TGAAGAGACA AAAGAAATAG AGATATGTAG AGATAGAGAG AG--AGAGGG AGAGACAA-G

                 10310      10320      10330      10340      10350      10360      10370      10380      10390      10400
          ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
BAC52     AGACAGAGAG ACAAAGAGAA TATGAGAGTG AGATGGAGAG AGATATGAGT GTGAGAGAAA GAGG-AGAGA GAGAGAGAGA GAGAGAGAGA GAGAAGATAG
BAC52-2b  AGACAGAGAG ACAAAGAGAA TATGAGAGTG AGATGGAGAG AGATATGAGT GTGAGAGAAA GAGGGAGAGA GAGAGAGAGA GAGAGAGAGA GAGAAGATAG

                 10410      10420      10430      10440      10450      10460      10470      10480      10490      10500
          ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
BAC52     ATAAGATTAT TTATATATAC AATGCGTATA AAAAAAAGTG TGCCCAACTT TAGTAACCTC TACTTAAAAA TTTATAACAT ATAAACTGAT ATCCTGTCTA
BAC52-2b  ATAAGATTAT TTATATATAC AATGCGTATA AAAAAAAGTG TGCCCA-CTT TAGTAACCTC TACTTAAAAA TTTATAACAT ATAAACTGAT ATCCTGTCTA

                 10510      10520      10530      10540      10550      10560      10570      10580      10590      10600
          ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
BAC52     ATGATTTTAT ATTCATAATT GTACTGCCAT GGATGATTGA GCAGCAGGCT TTTTATAAAG TTTTTTCAAA ATCCTTTTTG AACCAAGTTT GGGCCGAAGC
BAC52-2b  ATGATTTTAT ATTCATAATT GTACTGCCAT GGATGATTGA GCAGCAGGCT TTTTATAAAG TTTTTTCAAA ATCCTTTTTG AACCAAGTTT GGGCCGAAGC

                 10610      10620      10630      10640      10650      10660      10670      10680      10690      10700
          ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
BAC52     AATGGATGCG GGTCAAAAGT CATTATGTGT GGGTTATCTC ATTCCATAAA CAATTGTTCT CTTATGAATA CTAACTATTA GGCTGTTGAA GATAAAAGGT
BAC52-2b  AATGGATGCG GGTCAAAAGT CATTATGTGT GGGTTATCTC ATTCCATAAA CAATTGTTCT CTTATGAATA CTAACTATTA GGCTGTTGAA GATAAAAGGT

                 10710      10720      10730      10740      10750      10760      10770      10780      10790      10800
          ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
BAC52     GTGAATATCA ACTCAAGCTA TAAAGGAAAT TTTATCACAA ACAAAATGTA TGATATTCTC TTTGTTATGA TGAATAAAAT CTTAAATTCA TTCGTATCCC
BAC52-2b  GTGAATATCA ACTCAAGCTA TAAAGGAAAT TTTATCACAA ACAAAATGTA TGATATTCTC TTTGTTATGA TGAATAAAAT CTTAAATTCA TTCGTATCCC

                 10810      10820      10830      10840      10850      10860      10870      10880      10890      10900
          ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
BAC52     TTGTTTAAGG AAACCTTTCT CAGTGATAAA AAACCGATCG GGACGTTATC TTCTTTATTA AAGTCTCATC GTAATAATAA TATGAAAAAT TTATATAGCG
BAC52-2b  TTGTTTAAGG AAACCTTTCT CAGTGATAAA AA--CGATCG GGACGTTATC TTCTTTATTA AAGTCTCATC GTAATAATAA TATGAAAAAT TTATATAGCG

                 10910      10920      10930      10940      10950      10960      10970      10980      10990      11000
          ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
BAC52     CTTGTGACAA AAGTTTCAAA GCACTCGTGT GTTCGTTCCT GCATTTGGAT GTAGTAACCT TTGAGTTTTC ATTCGCTATT CAATATAAAC ACCATAATGT
BAC52-2b  CTTGTGACAA AAGTTTCAAA GCACTCGTGT GTTCGTTCCT GCATTTGGAT GTAGTAACCT TTGAGTTTTC ATTCGCTATT CAATATAAAC ACCATAATGT

                 11010      11020      11030      11040      11050      11060      11070      11080      11090      11100
          ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
BAC52     GCTCATTACG CGTGTGAAAT TGTGTAGAGG ACAAGTGAGC AGAGAGAAAT TAGATATAAT GAAACAAGAA GTTAGTTGGC TGTTCAAAAT AACTAAAAGT
BAC52-2b  GCTCATTACG CGTGTGAAAT TGTGTAGAGG ACAAGTGAGC AGAGAGAAAT TAGATATAAT GAAACAAGAA GTTAGTTGGC TGTTCAAAAT AACTAAAAGT
```

```
                11110      11120      11130      11140      11150      11160      11170      11180      11190      11200
           ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
BAC52      TTTCCCCGGG GTCGTAACAA GGTTCTGCGA AACAGGGGCG GATCTAGCCG GCGGCGAGGG TGGGGGGGGG GGGG----CA ATTTAAGAAA ATAGTTAGCG
BAC52-2b   TTTCCCCGGG GTCGTAACAA GGTTCTGCGA AACAGGGGCG GATCTAGCCG GCGGCGAGGG TGGGG-AGGG GGGGGGGGCA ATTTAAGAAA ATAGTTAGCG

                11210      11220      11230      11240      11250      11260      11270      11280      11290      11300
           ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
BAC52      CCGAATTAGC CGGCGAAAAA GCAATAGGGG GGG--TTAAT GAGCAATAAA TTGTTATTTC TTTTTTGCTC AACGGGGGGG GGGGTAGTGC ATGCGCGGAT
BAC52-2b   CCGAATTAGC CGGCGAAAAA GCAATAGGGG GGGGGTTAAT GAGCAATAAA TTGTTATTTC TTTTTTGCTC AACGGGGGGG GGGG------ ----------
                                                                                                   Deletion 1, 5’ end

                11310      11320      11330      11340      11350      11360      11370      11380      11390      11400
           ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
BAC52      CCAGGGGAGG CCCCCCGCCC CCCAAAAAAG TTTTTAATTT TTTGTTTTTT TAAATGGAGA CAGATAAAAA TTTAGTGCTC ACTACCACCC CCCCCCCCCC
BAC52-2b   ---------- ---------- ---------- ---------- ---------- ---------- ---------- ---------- ---------- ----------

                11410      11420      11430      11440      11450      11460      11470      11480      11490      11500
           ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
BAC52      CCCCCTACTG AGCAAAATTT ATCGGCCGGC ACGATTTTCG AATTTCACCA CGCTAAATTA AAAATTGATT TCAAATTTGG GTCTCCCCTA AGGAATCCTG
BAC52-2b   ---------- ---------- ---------- ---------- ---------- ---------- ---------- ---------- ---------- ----------

                11510      11520      11530      11540      11550      11560      11570      11580      11590      11600
           ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
BAC52      GACCCGCGCC AGTAGTAAGC ACTATTTTTA TTTACTGAAC CCCCCCCCTT GAGCAAAAAC AAAAAGAAGG TAAGATTGTG GATACTATCT TTGTTTAAAT
BAC52-2b   ---------- ---------- ---------- ---------- ---------- ---------- ---------- ---------- ---------- ----------

               119410     119420     119430     119440     119450     119460     119470     119480     119490     119500
           ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
BAC52      TGCCTGAAAA AAAAATCACA GTAGAAATAT CCCCTGAAAA GGATTGATCA TCGTTTTCGA AATATTACCG GAGGTAGCTG GAAATTTGTT GATAAATATT
BAC52-2b   ---------- ---------- ---------- ---------- ---------- ---------- ---------- ---------- ---------- ----------

               119510     119520     119530     119540     119550     119560     119570     119580     119590     119600
           ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
BAC52      CCATGCCACA TGTGTAGTTT AATATTAGCC TACCCAATTT CGTTTATGTG TGCACTAAAA AGTCATTTGT CAGCATTTTC AGACATTCTG ATTAAATTTC
BAC52-2b   ---------- ---------- ---------- ---------- ---------- ---------- ---------- ---------- ---------- ----------

               119610     119620     119630     119640     119650     119660     119670     119680     119690     119700
           ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
BAC52      TTGTAATTTT GATGAATTCA AACCTTACAC TTCTGTACTA AACAGGCGCG TACGCAGGAT TTTTTCAAGT GGGGGGGGGG GGGGGTTTAA CATTTTTAAA
BAC52-2b   ---------- ---------- ---------- ---------- ---------- ---------- ---------- ---------- --GGGTTTAA CATTTTTAAA
                                                                                    Deletion 1, 3’ end

               119710     119720     119730     119740     119750     119760     119770     119780     119790     119800
           ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
BAC52      TCGGGCCGAA AATTTCGCAT CGACTCAGCC AGCCGCTGAA TAAGCGGGGG GGGGGGGGGG GG--AAGAGG ACACTTTTTC TTTTCTTTTT TTGGTCTCGA
BAC52-2b   TCGGGCCGAA AATTTCGCAT CGACTCAGCC AGCCGCTGAA TAAGCGGGGG GGGGGGGGGG GGGGAAGAGG ACACTTTTTC TTTTCTTTTT TTGGTCTCGA
```

```
                 119810     119820     119830     119840     119850     119860     119870     119880     119890     119900
           ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
BAC52      AATTTGAAAA TTTGACATTT TGCTCTGTTG GGGGGGGGGG GG-TAAGGTC GCCTTTTTTA GGTCAGCCAT GGGAATAGTT TTTTT-ATTA TTATTTTTTT
BAC52-2b   AATTTGAAAA TTTGACATTT TGCTCTGTTG GGGGGGGGGG GGGTAAGGTC GCCTTTTTTA GGTCAGCCAT GGGAATAGTT TTTTT-ATTA TTATTTTTTT

                 119910     119920     119930     119940     119950     119960     119970     119980     119990     120000
           ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
BAC52      TTT-ATTATT CAAAAAAACA AAAAACAAAA CAAAACAAAG GGGGGGGGGG TTGTTTTTTT AGGGGGGTTT ATACACATAA AAATACCAAA GGGGGGGGG-
BAC52-2b   TT--ATTATT CAAAAAAACA AAAAACAAAA CAAAACAAAG GGGGGGGGGG TTGTTTTTTT AGGGGGGTTT ATACACATAA AAATACCAAA GGGGGGGGGG

                 120010     120020     120030     120040     120050     120060     120070     120080     120090     120100
           ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
BAC52      TTAACCCCTA AACACCCCCC CCCCCCC--T GCGTACGCGC CTGGTACTAC ATACATATTA AACCTTTCTA TGTAAATTTT AGCCCATTTC TTTCAAAACG
BAC52-2b   TTAACCCCCT AAACACCCCC CCCCCCCCCT GCGTACGCGC CTGGTACTAC ATACATATTA AACCTTTCTA TGTAAATTTT AGCCCATTTC TTTCAAAACG

                 120110     120120     120130     120140     120150     120160     120170     120180     120190     120200
           ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
BAC52      TAATCACATG ACCAGTACTA GATTTATTTT GAAACATCCG TCTTGATGTT TATAATGATA GATGACGTTA TACGACCCTG TCTTATGGGG TACCTTGTCA
BAC52-2b   TAATCACATG ACCAGTACTA GATTTATTTT GAAACATCCG TCTTGATGTT TATAATGATA GATGACGTTA TACGACCCTG TCTTATGGGG TACCTTGTCA

                 120210     120220     120230     120240     120250     120260     120270     120280     120290     120300
           ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
BAC52      TTCGTGCCAT TTTGGGGGGA GGGTGGAGAC CCGCAGACTT CCCCCTCTAG AACCACTACT GAATACCACT GAACGCAACT GAAGAATATG TTATCGTCAT
BAC52-2b   TTCGTGCCAT TTTGGGGGGA GGGTGGAGAC CCGCAGACTT CCCCCTCTAG AACCACTACT GAATACCACT GAACGCAACT GAAGAATATG TTATCGTCAT

                 120310     120320     120330     120340     120350     120360     120370     120380     120390     120400
           ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
BAC52      AGAGGTCGGT TATACATTTT TTTCCATGCA GGTTTTCATT CACTGTGAGG TGCTCATTTG TAATGATAAT GACCCATCGT CTCGGTGTTC CCAGGGATGT
BAC52-2b   AGAGGTCGGT TATACATTTT TTTCCATGCA GGTTTTCATT CACTGTGAGG TGCTCATTTG TAATGATAAT GACCCATCGT CTCGGTGTTC CCAGGGATGT

                 120410     120420     120430     120440     120450     120460     120470     120480     120490     120500
           ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
BAC52      AAGTCTAGAT TCAGACGTGG TAGCCGTCAC ACCCGTGGAG CAAGCTCCAC TCCTCATCTA ATCTCTAATG GACCTCTTTC AACTGCCCAC ACTATGCATG
BAC52-2b   AAGTCTAGAT TCAGACGTGG TAGCCGTCAC ACCCGTGGAG CAAGCTCCAC TCCTCATCTA ATCTCTAATG GACCTCTTTC AACTGCCCAC ACTATGCATG

                 120510     120520     120530     120540     120550     120560     120570     120580     120590     120600
           ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
BAC52      CCGCAGAAGC ACACAACTCT GGTAATTACC TTTTTAAAGA ACACAACACT TAACTTTAAA GGTGTCGTGT TTGTCTAGTG AAAACGTCGT TGGACTGTGG
BAC52-2b   CCGCAGAAGC ACACAACTCT GGTAATTACC TTTTTAAAGA ACACAACACT TAACTTTAAA GGTGTCGTGT TTGTCTAGTG AAAACGTCGT TGGACTGTGG
```

Reference

1. Oren M, Barela Hudgell MA, D' Allura B, Agronin J, Gross A, Podini D, et al. Short tandem repeats, segmental duplications, gene deletion, and genomic instability in a rapidly diversified immune gene family. BMC Genomics. 2016;17:900.