Note

Under-representation of repetitive sequences in whole-genome shotgun sequence databases: an illustration using a recently acquired transposable element

Akihiko Koga

Division of Genome Diversity, Primate Research Institute, Kyoto University,
Inuyama City 484-8506, Japan

Running title: Under-representation of repetitive sequences

Correspondence to Akihiko Koga at:
E-mail <koga@pri.kyoto-u.ac.jp>
Phone <+81 568 63 0526>
Fax <+81 568 62 9554>

Abstract: It is widely accepted in a conceptual framework that repetitive sequences, especially those with high sequence homogeneity among copies, tend to be under-represented in whole-genome shotgun sequence databases, because of the difficulty of assembling sequence reads into contigs. Although this is easily inferred, there is no quantitative illsutration of this phenomenon. An example using a currently used database is expected to contribute to the intuitive understanding of how serious the under-representation is. The present study provides the first quantitative example (in the case of 16 copies of virtually identical, 4.7-kb sequences in a genome of 7 x 10^8 bp) by comparing the results of BLAST searches of a sequence database (contig N50 9.8 kb) with those of Southern blot analysis of genomic DNA. This has revealed that the internal regions of the repetitive sequences are under-represented to a striking extent.

Keywords: contig, database, repetitive sequences, under-representation, transposon.

The number of nucleotides whose sequence can be determined by a single reading pass is limited, and certainly far smaller than the genome size. It is therefore necessary to assemble single reads into contigs, relying upon overlapping segments. One key factor in the production of reliable contigs is the number of sequence reads: any increase in this number is expected to contribute to longer contigs and more thorough elimination of sequencing errors. There is one problem; however, that is not solved simply by increasing the number of reads, and it involves dispersed repetitive sequences whose lengths are greater than those of single reads. Such sequences, esepcially when homogeneous among copies, are an obstacle to contig formation, raising the possibility of incorrect assembly. This is a major disadvantage of whole-genome shotgun sequencing compared to hierarchical shotgun sequencing. Such a situation could occur even in the latter strategy if multiple copies of repetitive sequences are contained in a DNA clone to be cut into short fragments. An expected outcome of this difficulty in contig formation is the under-representation of repetitive sequences in the resultant databases. Though clearly inferred, this phonomenon has not previously been demonstrated in a currently used sequence database.

The *Tol2* transposable element of the medaka fish, *Oryzias latipes*, (Koga et al. 1996) is an ideal means for demonstrating this phenomenon. This element belongs to the *hAT* (*hobo*/*Activator*/*Tam3*) family of DNA-based transposable elements (Arensburger et al. 2011); being 4.7 kb in length, it is larger than the usual range of single reads (500-1000 bp in chain-termination sequencing and 20-400 bp in next-generation sequencing). Ten to 30 *Tol2* copies are present in a diploid medaka genome, and their nucleotide sequences are highly homogeneous among copies, exhibiting virtually no sequence variation (Koga and Hori 1999). As we have previously proposed, the reason for this sequence homogeneity is the recent occurrence of the invasion of this element into the medaka genome (Koga et al. 2000). This explanation is consistent with our observation that the element is active and that it causes mutations on host genes when inserted (Koga and Hori 1997; Iida et al. 2004) or excised (Iida et al. 2005; Koga et al. 2006).

The medaka genome sequencing project was begun in 2002, applying the whole-genome shotgun sequencing strategy and using the Hd-rR strain as the source of genomic DNA (Kasahara et al 2007; Kobayashi and Takeda 2008). The genome database of version 1.0, which is contained in Ensembl (http://www.ensembl.org/Multi/blastview) release 63 (June 2011), was created in 2006. Its N50 sizes are 5.1 Mb, 1.41 Mb, and 9.8 kb for ultracontigs, scaffolds, and contigs, respectively. A BLAST search against this database, using the entire *Tol2* sequence

(GenBank accession number D84375; 4682 bp) as query, resulted in more than 10 hits for the left and right terminal regions, but few or no hits for internal regions (data not shown). A more detailed analysis, sending 200-bp blocks separately to the database, clearly showed a U-shaped distribution of hits over the element sequence (Figure 1). It is noteworthy that there was no hit for the region between nucleotides 2210 and 4082.

The homogeneity of the *Tol2* element's nucleotide sequence has already been demonstrated, as mentioned above. The present report provides further experimental evidence for the homogeneity of the nucleotide sequence. Genomic Southern blot analysis was performed, using the same Hd-rR strain that served as the DNA source of the medaka genome sequence database. Probes used were the three portions of the *Tol2* element shown in Figure 1 (probes L, C and R). The first trial of Southern blot analysis was as follows: digestion of genomic DNA of the Hd-rR strain with *Bam*HI (no cutting site in the D84375 sequence), pulsed-field gel electrophoresis in three lanes on an agarose gel, transfer to a nylon membrane, and then hybridization separately with each of the three probes. The band patterns on the resulting autodiagrams appeared identical among the three probes, but it was difficult to identify individual bands because more than 10 bands whose size was 10 kb or greater were clustered in an upper region where the resolution was relatively poor (data not shown). The use of *Kpn*I (no cutting site in the D84375 sequence) in addition to *Bam*HI led to little improvement (data not shown). To overcome this problem, we used *Dra*II and *Bgl*II, each of which has a single cutting site (nucleotides 2226/2227 and 3006/3007, respectively) in the D84375 sequence. Digestion with these enzymes, in addition to *Bam*HI, was expected to produce bands of smaller sizes that could be distinguished from one another through regular, constant-voltage gel electrophoresis. As shown in Figure 2, DNA digested with *Bam*HI and *Bgl*II yielded identical band patterns between probe L and probe C. The number of bands observed was 16. This result indicates that the Hd-rR strain carries 16 *Tol*2 copies in a haploid genome, and that all of them have both the probe-L region and the probe-C region. Similarly, hybridization after digestion with *Bam*HI and *Dra*II produced identical band patterns, consisting of 16 bands, between probe C and probe R. Southern blot analysis thus showed that the Hd-rR strain harbors 16 *Tol*2 copies and that every copy carries the three probed regions.

These results constitute the first quantitative illustration of the inferred under-representation of repetitive sequences in a whole-genome shotgun database. In addition to the "treated" medaka genome database contained in Ensembl, we conducted a BLAST search against the Trace Archives medaka database (http://www.ncbi.nlm.nih.gov/Traces/home/), a collection of raw data from single reads, using the entire *Tol2* sequence as query. This search resulted in hits on several partially or totally overlapping sequence reads, and every portion of the *Tol2* sequence was involved in multiple reads (data not shown). Thus, the under-representation in the treated database occurs not because of a difference in the generation efficiency among the DNA fragments to be used as templates for the sequencing reaction, but because of the assembly treatment performed by computer. A likely explanation is that many of the reads corresponding to the *Tol2* internal regions remain suspended because of insufficient credibility about linkage to other reads or contigs. In the case of the medaka genome project, compilation of sequence data was conducted by the RAMEN genome assembler (Kasahara et al. 2007). Absence of a contig containing the entire *Tol2* sequence in the resultant draft genome does not imply a limit in the assembly performance of this assembler, but rather shows its high level of ability to avoid misassembly.

The distribution of hits shown in Figure 1 is U-shaped at a first glance, but further examination reveals the following features: (*i*) the number of hits decreases inwards within the left terminal region (800 bp from the left end), (*ii*) the same situation is true in the right terminal region (800 bp from the right end), (*iii*) there is a low-frequency but continuous appearance of hits in the left half of the internal region (nucleotides 801-2200), and (*iv*) there is a complete absence of hits in the right half of the internal region (nucleotides 2201-3882). The first two of these features are consistent with the explanation that *Tol2*-flanking chromosomal sequences, which are expected to differ from one *Tol2* copy to another, are key factors in determining linkage to other reads or contigs. The difference in hit frequency between the left and right halves of the internal region may be due to specific characteristics of the sequence of the *Tol2* element. As shown in Figure 1, *Tol2* contains a pair of internal inverted repeats (IIRs). The repeats are 302 bp (nucleotides 1434–1735) and 303 bp (1786–2088), respectively, and the two units are separated by 50 bp (1736–1785), possibly forming a hairpin structure (Izsvak et al. 1999). Inverted repeat structures are known to be fragile during plasmid amplification in host bacteria (Doherty et al. 1993). Actually, we often encountered rearrangements of the IIR region in the early period of our study of this element, which we eventually began to avoid by using a host bacterial strain that carries the *uvrC* and *umuC* mutations (the SURE strain). Host bacterial strains commonly used for library preparation, such as the DH10B strain, do not have these mutations. One possible explanation of the low-frequency occurrence of IIRs and their surrounding regions is that "variation" is generated while clones or subclones to be used for sequencing are prepared, and that sequence reads from "variants" that were identical by chance were regarded as descendants of a single original sequence.

The second purpose of the present report was to asses the possibility of misassembly caused by artifact variation; this remains no more than a speculation, but the under-representation of repetitive sequences is now clear. This illustration was successful because medaka contains copies of a highly homogeneous transposable element which, as we have proposed, is a recent invader of the genome.

It is unlikely that repetitive sequences that are short enough to be included in a single read tend to be under-represented in sequence databases. An example is the SINEs (short interspersed elements), which are mostly a few hundred base pairs in length (Goodier and Kazazian 2008). LINEs (long interspersed elements), LTR (long terminal repeat) retrotransposons, retroviruses, and DNA-based transposable elements often consist of copies of several kilobase pairs. The results shown in the present report reveal the necessity, in quantitative analysis of these kinds of elements, of considering the possibility of under-representation. Of these repetitive sequences, DNA-based transposable elements would be the sequences most likely to be affected by a tendency toward under-representation. This is because internal deletion is common in this class of transposable elements. The great majority of elements of this class comprise complete and defective copies, with a good correspondence to autonomous and nonautonomous copies, respectively. Defective copies are generated from complete copies or already-defective copies, and deletion of internal regions is the most common cause of this transition (Fedoroff et al. 1983; Streck et al. 1986; Warren et al. 1994). The mechanism underlying such deletions is considered to be the premature interruption of gap repair after excision of the element (Rubin and Levy 1997). When a relatively infrequent occurrence of internal regions is found in database mining, careful validation must be applied to determine whether it is due to the accumulation of internally deleted copies in the genome, or to the tendency toward under-representation of sequence databases, or to both.

Acknowledgements

The author is grateful to A. Shimada, H. Takeda, and H. Hori for helpful discussions. This work was supported by grant No. 23657165 from the MEXT of Japan to A. K., and by the Global COE program A06 of Kyoto University.

References

Arensburger, P., Hice, R. H., Zhou, L., Smith, R. C., Tom, A. C., Wright, J. A., Knapp, J., O'Brochta, D. A., Craig, N. L., and Atkinson, P. W. 2011. Phylogenetic and functional characterization of the hAT transposon superfamily. Genetics, **188**: 45-57.

Doherty, J. P., Lindeman, R., Trent, R. J., Graham, M. W., and Woodcock, D. M. 1993. *Escherichia coli* host strains SURE and SRB fail to preserve a palindrome cloned in lambda phage: improved alternate host strains. Gene, **124**: 29-35.

Fedoroff, N., Wessler, S., and Shure, M. 1983. Isolation of the transposable maize controlling elements Ac and Ds. Cell, **35**: 235–242.

Goodier, J. L., and Kazazian, H. H. Jr. 2008. Retrotransposons revisited: the restraint and rehabilitation of parasites. Cell, **135**: 23-35.

Iida, A., Inagaki, H., Suzuki, M., Wakamatsu, Y., Hori, H., and Koga, A 2004. The tyrosinase gene of the i^{b} albino mutant of the medaka fish carries a transposable element insertion in the promoter region. Pigment Cell Res. **17**: 158-164.

Iida, A., Takamatsu, N., Hori, H., Wakamatsu, Y., Shimada, A., Shima, A., and Koga, A. 2005. Reversion mutation of i^{b} oculocutaneous albinism to wild-type pigmentation in medaka fish. Pigment Cell Res. **18**: 382-384.

Izsvak, Z., Ivics, Z., Shimoda, N., Mohn, D., Okamoto, H., and Hackett, P. B. 1999. Short inverted-repeat transposable elements in teleost fish and implications for a mechanism of their amplification. J. Mol. Evol. **48**: 13-21.

Kasahara, M., Naruse, K., Sasaki, S., Nakatani, Y., Qu, W. et al. 2007. The medaka draft genome and insights into vertebrate genome evolution. Nature, **447**: 714-719.

Kobayashi, D., and Takeda, H. 2008. Medaka genome project. Brief. Funct. Genomics **7**: 415-426.

Koga, A., and Hori, H. 1999. Homogeneity in the structure of the medaka fish transposable element *Tol2*. Genet. Res. Camb. **73**: 7-14.

Koga, A, Iida, A., Hori, H., Shimada, A., and Shima, A. 2006. Vertebrate DNA transposon as a natural mutator: the medaka fish *Tol2* element contributes to genetic variation without recognizable traces. Mol. Biol. Evol. **23**: 1414-1419.

Koga, A., Shimada, A., Shima, A., Sakaizumi, M., Tachida, H., and Hori H. 2000. Evidence for recent invasion of the medaka fish genome by the *Tol2* transposable element. Genetics, **155**: 273-281.

Koga, A., Suzuki M., Inagakin H., Besshon Y.n and Hori, H. 1996. Transposable element in fish. Nature, **383**: 30.

Koga, A., Wakamatsu, Y., Sakaizumi, M., Hamaguchi, S., and Shimada A. 2009. Distribution of complete and defective copies of the *Tol1* transposable element in natural populations of the medaka fish *Oryzias latipes*. Genes Genet. Sys. **84**: 345-352.

Streck, R. D., MacGaffey, J. E., and Beckendorf, S. K. 1986. The structure of hobo transposable elements and their insertion sites. EMBO J. **5**: 3615–3623.

Warren, W. D., Atkinson, O. W., and O'Brochta, D. A. 1994. The *Hermes* transposable element from the house fly, *Musca domestica*, is a short inverted repeat-type element of the *hobo*, *Ac*, and *Tam3* (*hAT*) element family. Genet. Res. Camb. **64**: 87–97.

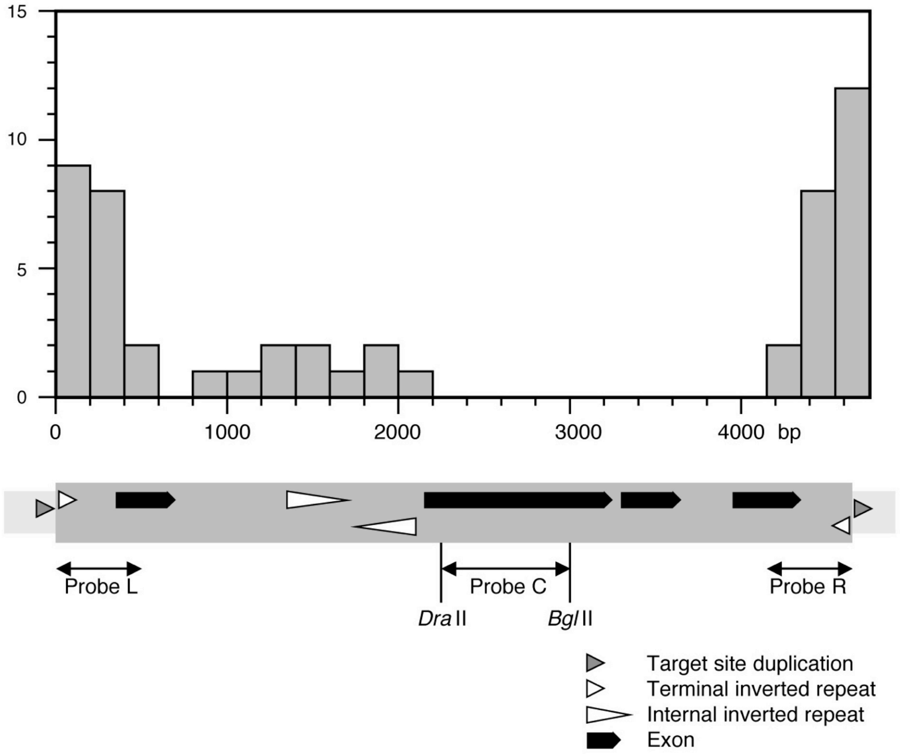

Figure 1. Distribution of the numbers of hits over the entire *Tol2* sequence. The structure of the *Tol2* element is shown under the horizontal axis of the graph. Segments of 200 bp from the left end to the center and from the right end to the center of the 4682-bp *Tol2* sequence were sent to the medaka database of Ensembl, with default settings, for BLAST hits. Successful alignments were defined as those consisting of more than 190 nucleotides and exhibiting sequence similarity of more than 95%. Such alignments were counted in the list of search results obtained, and the numbers are indicated by the height of the bars in the graph. The segmentation of the entire sequence into 200 bp from respective ends yielded a partial overlap at the center because the length of the element is not a multiple of 200. There were no hits for these two overlapping segments. Regions used for probes L (nucleotides 1-458), C (2227-3006), and R (4267-4682) are indicated by double-headed arrows under the element. Positions of the cutting sites of the restriction enzymes *Dra*II (2226/2227) and *Bgl*II (3006/3007) are also shown.

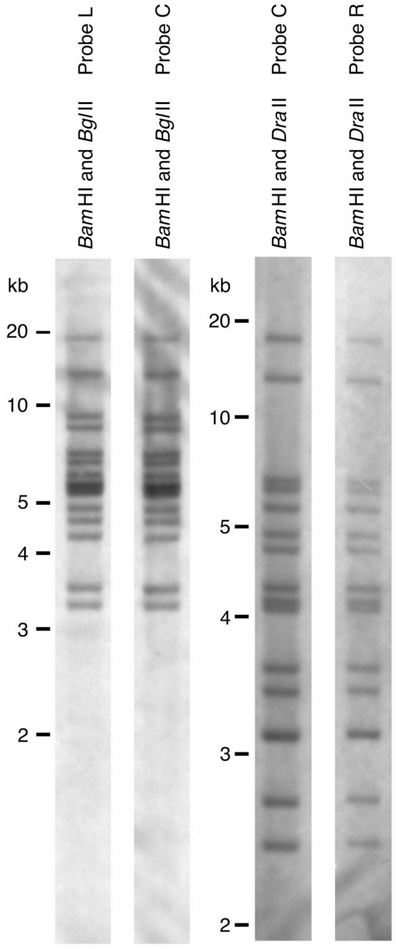
*Bam*HI and *Bgl*II Probe L
*Bam*HI and *Bgl*II Probe C
*Bam*HI and *Dra*II Probe C
*Bam*HI and *Dra*II Probe R
kb
20
10
5
4
3
2
kb
20
10
5
4
3
2

Figure 2. Southern blot hybridization for distribution of sizes of restriction fragments containing parts of the *Tol2* element. Genomic DNA (5 µg for each lane) was digested with the restriction enzymes and hybridized with the probes shown above the panels. Methods were as described in Koga et al. (2009). Each hybridization band is expected to represent a restriction fragment consisting of part of *Tol2* and its flanking chromosomal region. Size differences among bands are thought to be caused by differences in the latter. The relatively thick bands, two in the lanes for *Bam*HI and *Bgl*II, and one in the lanes for *Bam*HI and *Dra*II, were assumed to represent two restriction fragments.