Jonathan King
MCB 149, Wikipedia Article
May 4, 2016

Tandem Duplications

Summary

A tandem duplication (TD) is a rare but significant type of genetic duplication event that occurs when a region of DNA is duplicated such that the copies are in close proximity to each other.

Relative to other kinds of major structural variation and gene duplication events, tandem duplications are unique because they are not deleterious, can range in size from whole genes to short sequences (such as internal tandem duplications), and often introduce novel genetic sequences into the genome. Because tandem duplications have the ability to dramatically alter the function of proteins, they are of great significance in the studies of evolutionary genetics and, more recently, cancer biology.

Examples

Evolutionary Genetics

Gene duplication is thought to play an important role in the evolution of an organism because once duplicated, a region of DNA may accumulate beneficial mutations while the original copy is left intact. It can be inferred that the genes in most gene families arose from sequential duplication since they are often arranged in "head-to-tail" fashion (Hu et al., 1992).

In his book, *Human Gene Evolution*, David Cooper gives a few examples (see **Figure A**) of how tandem duplications in ancestral genomes may have given rise to many of the multigene clusters seen today.

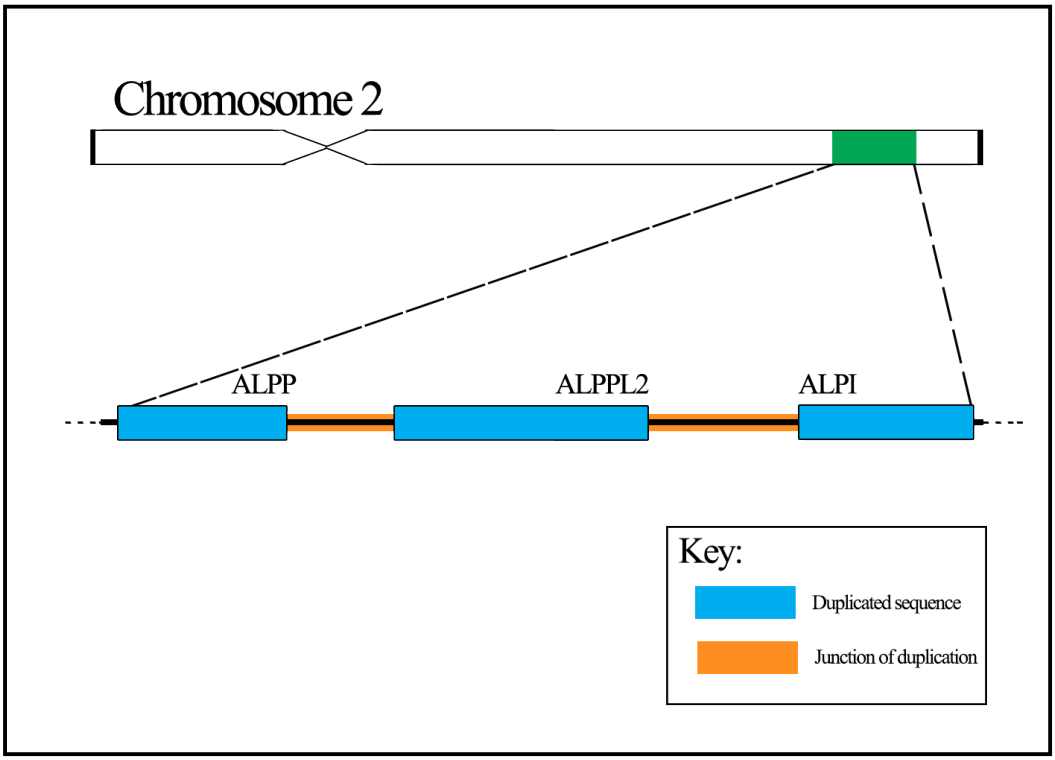

Figure A: *An example of an ancestral non-internal Tandem Duplication that may have given rise to the alkaline phosphatase gene cluster (genes ALPP, ALPI, ALPPL2) at locus 2q37 (Cooper, 1999).*

Gene families likely caused by ancestral tandem duplication include:

- the immunoglobulin genes (IGHA, IGHD, IGHG),
- T-cell receptor genes (TCRA, TCRD, TCRB, TCRG),
- and the carcinoembryonic antigen gene family (Cooper, 1999).

Tandem Inverted Duplications

Tandem Inverted Duplications (TIDs) are defined by a region that has been copied multiple times but in alternating orientations (normal->inverted->normal). Although rare in populations without strong selective pressure, TIDs are common in cases where increasing gene copy number confers accelerated growth in an organism (Kugelberg et al., 2010).

A well-documented example is that of the bacteria *Salmonella enterica*, which creates TIDs in the *lac* operon to improve the organism's fitness under selection (Kugelberg et al., 2010). Research suggests that TIDs might be the result of multiple rounds of duplication and deletion, which ultimately causes the inversion and rearrangement of the copies into the most stable structural form (Reams and Roth, 2015).

Internal Tandem Duplications

Though duplication of DNA can sometimes be beneficial for an organism, the copied region of DNA may also have negative consequences if, for example, the duplication happens within an exon, a portion of a gene that encodes for amino acids. Such is the case for Internal Tandem Duplications (ITDs). With ITDs, the duplication may cause extra amino acids to be inserted into the peptide sequence, potentially altering the protein's conformation or, more drastically, resulting in a frameshift mutation which produces a truncated protein (Hu et al., 1992).

Generally speaking, duplications and amplifications of sequences in cancer cells often allow tumor cells to confer resistance to many drugs that would otherwise inhibit their growth (Reams and Roth, 2015). Internal Tandem Duplications, in particular, can be considered oncogenic gain-of-function mutations (Roy et al., 2015) because they often confer new functionality that may facilitate unregulated growth of cells in cancer patients. As such, research suggests that ITDs may serve as an effective "molecular diagnostic test" for certain cancers, some of which are discussed below (Roy et al., 2015).

ITD Example 1: FMS-like tyrosine kinase 3 *(FLT3)*

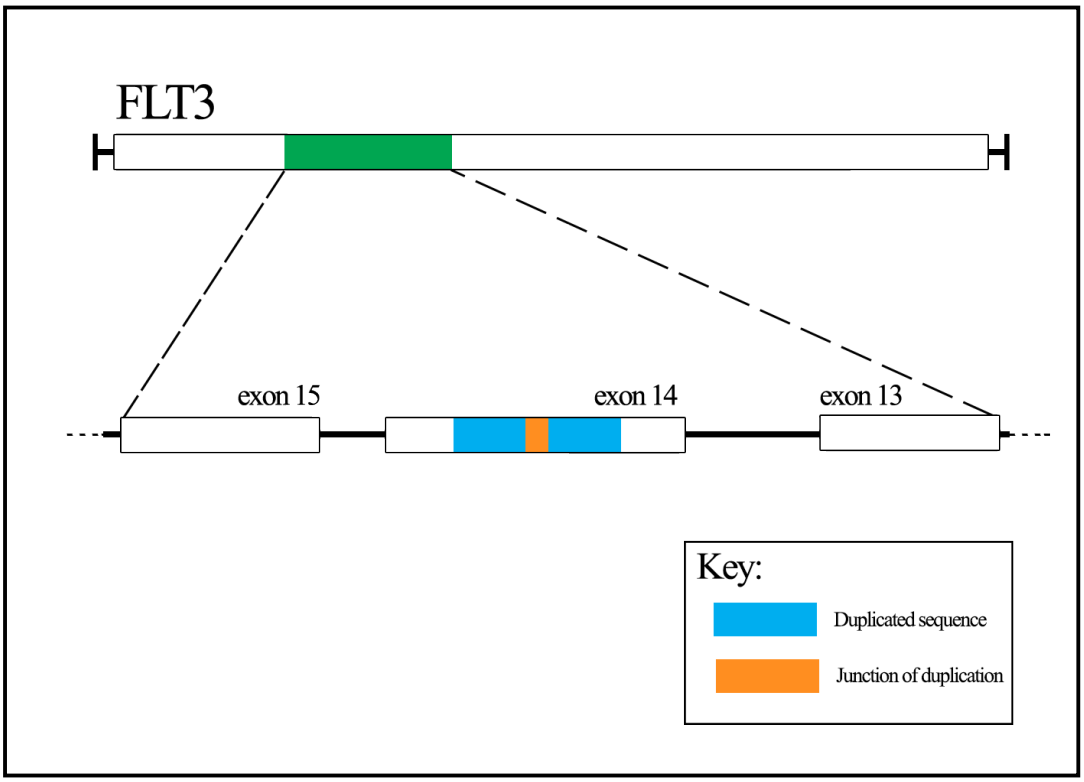

Figure B: *An Internal Tandem Duplication in the highly studied FLT3 gene associated with Acute Myeloid Leukemia. The duplication is considered internal because the duplicated sequence is from a code determining region, i.e., an exon. The orange region between duplications is not always present and may contain DNA that does not match the reference sequence.*

ITDs in the FLT3 gene have repeatedly been demonstrated to be associated with poor outcomes for patients that have Acute Myeloid Leukemia (AML) (Whitman et al., 2001). Many of the ITDs occur in exons 13, 14, and 15 of the gene (Hirade et al., 2015). One example of an ITD in FLT3 is shown in **Figure B**. Tandem duplications within this region cause an upregulation of the gene RUNX1, a transcription factor responsible for blood formation (Hirade et al., 2015). The high expression of RUNX1 results in a deregulation of cell proliferation, which the FLT3 ITD mutant cells take advantage of (Hirade et al., 2015).

Researchers have speculated that this ITD event may be caused by DNA replication error, though such a mechanism has not been proven. One such theory of its genesis is as follows:

> [The region consisting of amino acids] D593 to K602 potentially forms a palindromic intermediate. If a lagging strand makes a hairpin during DNA replication and the following mismatch repair system is impaired, the tandemly duplicated fragment will be fixed in DNA. If FLT3/ITD occurs in an out-of-frame manner, the leukemia cell carrying it does not acquire a growth advantage and will not be selected (Kiyoi et al., 1998).

ITD Example 2: BCL-6 co-repressor (BCOR)

Clear Cell Sarcoma of the Kidney (CCSK) is a rare renal tumor found in children. Although correct diagnosis of CCSK is critical to patient survival (Astolfi et al., 2015), it is often difficult to differentiate from other renal tumors. Consequently, researchers have turned

to the use of Whole Transcriptome Sequencing (WTS) to identify ITDs in the BCOR gene, genetic anomalies present in nearly all patients with CCSK (Astolfi et al., 2015, Roy et al., 2015, Ueno-Yokohata et al., 2014). The ITDs are usually around 100 base-pairs (bp) in length–similar in size to those found in FLT3–and may cause CCSK by altering or disrupting the behavior of the PRC1.1/BCOR complex (Roy et al., 2015). All reported ITDs in BCOR have been found to be duplicated in frame (Astolfi et al., 2015).

Detecting Tandem Duplications

Biological Assays

Despite the improvement of sequencing technologies to detect structural variants in DNA, independent verification of the results must often accompany a diagnosis in a clinical setting. When duplication events are larger than 3 to 5 megabases, it is possible to visually determine the genotype microscopically (Gu et al., 2008). However, considering most Tandem Duplications are significantly smaller than this, more sensitive verification techniques must be used.

Most of the time, verification is carried out using Polymerase Chain Reaction (PCR) experiments. Traditional or "Real Time" PCR analysis can detect Tandem Duplications by comparing the size of the PCR products. Using primers that lie outside the region of duplication, a TD positive PCR product will be significantly larger than a TD negative sample. Unfortunately, traditional PCR is relatively insensitive for this purpose and is only

able to detect 1 in 100 cells containing the TD (Grunwald et al., 2014). Because somatic mutations, as opposed to those in the germline, are only present in a handful of cells, it becomes important to use an even more sensitive experiment.

A new method was developed in 2014 by Grunwald et al. called TD-PCR. This method relies on the use of inverted PCR primers to amplify DNA *only* in the presence of a duplication, thus raising the sensitivity of the test to a single molecule. In one study, TD-PCR demonstrated its effectiveness by detecting TDs in 25% of the patients for whom traditional PCR failed (Grunwald et al., 2014).

Analysis of Sequencing Data

High throughput sequencing of DNA is quickly appearing as a popular alternative for detecting the presence of TDs in a sample. Sequencing is already considered a robust method for detecting Single Nucleotide Polymorphisms (SNPs), but many existing algorithms struggle to detect tandem duplications since they often vary in size, location, and frequency.

The program "Genomon ITDetector", developed by researchers at the University of Tokyo, first uses BLAST Local Alignment Tool (BLAT) to align reads to the reference genome and, second, analyzes the reads that have misaligned or "soft-clipped" sequences (Chiba et al., 2015). **Figure C** demonstrates how the duplicated region is flanked on the left and right by soft-clipped reads. This information can then be extracted algorithmically and reported as potential ITDs. Because Genomon ITDetector relies on the reads to be aligned prior to analysis, choosing different alignment tools can result in dramatically different results,

especially when many alignment algorithms do not have well defined behavior for addressing soft-clipped reads.

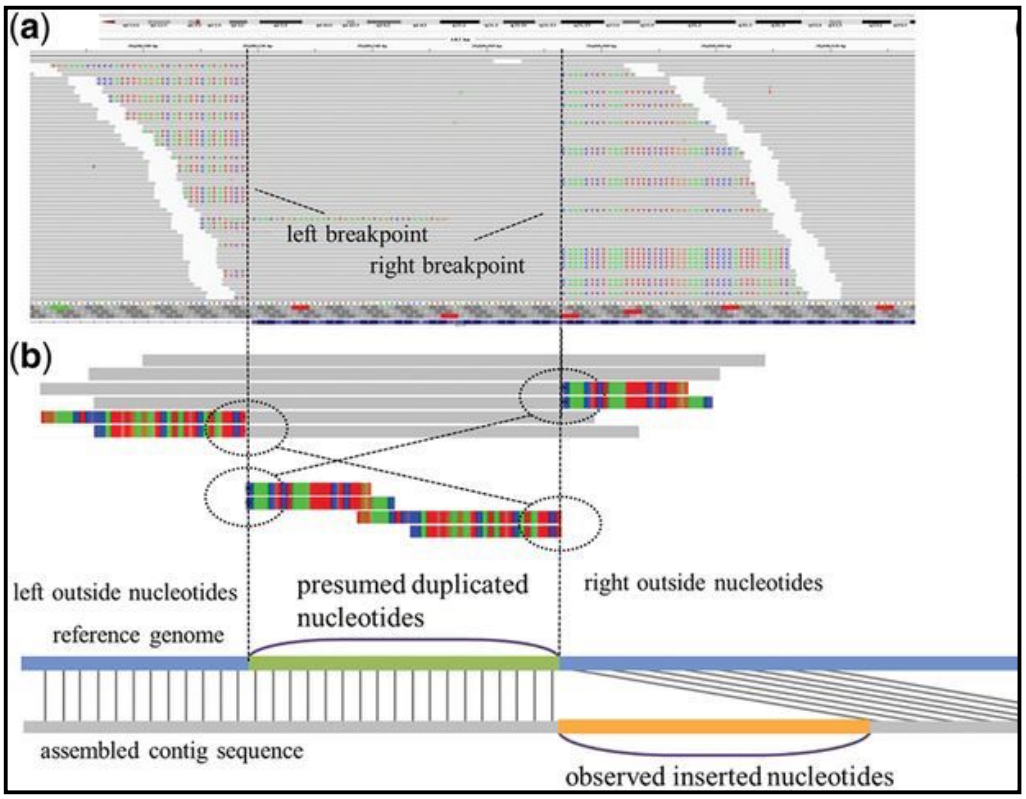

Figure C: *Using soft-clipped reads after alignment to detect the presence of ITDs. Soft-clipped reads are rainbow colored in regions that do not match the reference sequence. In this case, the mismatches are complete duplications of a region to the immediate left or right. This figure is borrowed from Chiba et al.'s* Genomon ITDetector *paper.*

Researchers from The National Institute of Agronomic Research in Paris, France developed a related program called "ReD Tandem" which also finds tandem duplications (Audemard et al., 2012). This program was not intended to detect somatic mutations like Genomon ITDetector, but, instead, tries to predict evolutionarily recent duplication events.

ReD Tandem uses a minimum cost flow based algorithm to identify sequences that are most likely TDs.

Next, researchers at EMBL Outstation European Bioinformatics Institute in Cambridge, UK have developed a similar program entitled "Pindel" (Ye et al., 2009). Pindel presumably gets its name from its pattern ("P") based methodology for finding potential insertions or deletions ("indels") (Ye et al., 2009). Pindel's approach is known to work especially well with large sized insertions but struggles to find insertions and deletions 1 to 5bp in length (Ghoneim et al., 2014).

In the cases where detecting tandem duplications is of clinical importance (i.e. finding events associated with disease), algorithms for detecting tandem duplications are often tested on sets of simulated reads where the event of interest is first "introduced" into the sample *in silico* and then obscured by randomly generated sequencing errors (Ghoneim et al., 2014). After generating synthetic reads using a program such as ART (Huang et al., 2011), algorithm designers can then compare the reported duplication events with the events introduced at the start of the experiment.

Mechanisms of Duplication

In order to determine the effects and possible mechanisms of tandem duplications, scientists employ several techniques, the most common of which is transcriptome sequencing. Transcriptome sequencing can elucidate the mechanistic details of genetic events by

measuring the levels of gene expression in samples that are known to contain the event of interest (Astolfi et al., 2015). In addition, analyzing the sequences at the junction of duplication (denoted by the orange portions in **Figures A** and **B**) can also help reveal key details about the formation of tandem duplications (Reams and Roth, 2015). However, junction sequences may be deceptive in that they may have been transformed since the duplication event (Reams and Roth, 2015).

When tandem duplications were first discovered by Calvin Bridges in the Bar gene, they were thought to only be the product of unequal homologous recombination (Bridges, 1936). Though this is true for many duplication events, some tandem duplications are caused by other mechanisms (Reams and Roth, 2015), for example, transposition or failures in replication.

Transposable Elements

Transposable elements may cause tandem duplications either by providing regions of homology for homologous recombination or by transposition in which a region is directly copied from one area of the chromosome to another. The duplication detected by Bridges was later determined to arise from an interaction between two transposable elements that flanked the Bar gene (Tsubota et al., 1989). Interestingly, the phenotype of Bar mutants is characterized not by the duplicated region of DNA, but by the junction *between* the duplications, once again underscoring the importance of junction sequences in understanding tandem duplications (Tsubota et al., 1989).

Replication Errors

Another possible mechanism of duplication comes from the functional failure of enzymes responsible for DNA replication. The most prominent model based on this idea is called Fork Stalling and Template Switching (FoSTeS) in which the DNA replication fork "stalls" at a given position, which causes the lagging strand to disengage from the replication complex. Then, the lagging strand anneals to another nearby replication fork where DNA synthesis resumes and introduces a duplicate sequence into the DNA (Gu et al., 2008). The rate at which FoSTeS occurs *in vivo* remains to be determined.

References

Astolfi, A., Melchionda, F., Perotti, D., Fois, M., Indio, V., Urbini, M., Genovese, C., Collini, P., Salfi, N., and Nantron, M. et al. (2015). Whole transcriptome sequencing identifies BCOR internal tandem duplication as a common feature of clear cell sarcoma of the kidney. Oncotarget.

Audemard, E., Schiex, T., and Faraut, T. (2012). Detecting long tandem duplications in genomic sequences. BMC Bioinformatics *13*, 83.

Bridges, C. (1936). THE BAR "GENE" A DUPLICATION. Science *83*, 210-211.

Chiba, K., Shiraishi, Y., Nagata, Y., Yoshida, K., Imoto, S., Ogawa, S., and Miyano, S. (2014). Genomon ITDetector: a tool for somatic internal tandem duplication detection from cancer genome sequencing data. Bioinformatics *31*, 116-118.

Cooper, D. (1999). Human gene evolution (Oxford: Bios Scientific Publishers).

Ghoneim, D., Myers, J., Tuttle, E., and Paciorkowski, A. (2014). Comparison of insertion/deletion calling algorithms on human next-generation sequencing data. BMC Research Notes *7*, 864.

Grunwald, M., Tseng, L., Lin, M., Pratz, K., Eshleman, J., Levis, M., and Gocke, C. (2014). Improved FLT3 Internal Tandem Duplication PCR Assay Predicts Outcome after Allogeneic Transplant for Acute Myeloid Leukemia. Biology Of Blood And Marrow Transplantation *20*, 1989-1995.

Gu, W., Zhang, F., and Lupski, J. (2008). Mechanisms for human genomic rearrangements. Pathogenetics *1*, 4.

Hirade, T., Abe, M., Onishi, C., Taketani, T., Yamaguchi, S., and Fukuda, S. (2015). Internal tandem duplication of FLT3 deregulates proliferation and differentiation and confers resistance to the FLT3 inhibitor AC220 by Up-regulating RUNX1 expression in hematopoietic cells. Int J Hematol *103*, 95-106.

Hu, X. and Worton, R. (1992). Partial gene duplication as a cause of human disease. Human Mutation *1*, 3-12.

Huang, W., Li, L., Myers, J., and Marth, G. (2011). ART: a next-generation sequencing read simulator. Bioinformatics *28*, 593-594.

Kiyoi, H., Towatari, M., Yokota, S., Hamaguchi, M., Ohno, R., Saito, H., and Naoe, T. (1998). Internal tandem duplication of the FLT3 gene is a novel modality of elongation mutation which causes constitutive activation of the product. Leukemia *12*, 1333-1337.

Kugelberg, E., Kofoid, E., Andersson, D., Lu, Y., Mellor, J., Roth, F., and Roth, J. (2010). The Tandem Inversion Duplication in Salmonella enterica: Selection Drives Unstable Precursors to Final Mutation Types. Genetics *185*, 65-80.

Reams, A. and Roth, J. (2015). Mechanisms of Gene Duplication and Amplification. Cold Spring Harbor Perspectives In Biology *7*, a016592.

Roy, A., Kumar, V., Zorman, B., Fang, E., Haines, K., Doddapaneni, H., Hampton, O., White, S., Bavle, A., and Patel, N. et al. (2015). Recurrent internal tandem duplications of BCOR in clear cell sarcoma of the kidney. Nature Communications *6*, 8891.

Tsubota, S., Rosenberg, D., Szostak, H., Rubin, D., and Schedl, P. (1989). The cloning of the Bar region and the B breakpoint in Drosophila melanogaster: evidence for a transposon-induced rearrangement. Genetics *122*, 881-890.

Ueno-Yokohata, H., Okita, H., Nakasato, K., Akimoto, S., Hata, J., Koshinaga, T., Fukuzawa, M., and Kiyokawa, N. (2015). Consistent in-frame internal tandem duplications of BCOR characterize clear cell sarcoma of the kidney. Nature Genetics *47*, 861-863.

Whitman, S., Archer, K., and Feng, L. (2001). Absence of the Wild-Type Allele Predicts Poor Prognosis in Adult de Novo Acute Myeloid Leukemia with Normal Cytogenetics and the Internal Tandem Duplication of FLT3. Cancer Research *61*.

Ye, K., Schulz, M., Long, Q., Apweiler, R., and Ning, Z. (2009). Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. Bioinformatics *25*, 2865-2871.