Identification of segmental duplications in the human genome

Dr. Ananda Fonseka BDS, MSc
Dental Officer (Medical Informatics), Ministry of Health, Colombo, Sri Lanka
E-Mail address: fdananda1234@gmail.com

Sri Lanka Journal of Bio-Medical Informatics 2012;**3**(1):12-19
DOI: http://dx.doi.org/10.4038/sljbmi.v3i1.2467

Abstract

Sequencing the human genome early in this century marked one of the greatest achievements in the advancement of science. Soon it opened up a great opportunity to the scientific community worldwide with a massive amount of biological data to be explored.

Segmental duplications are one of the structural variations found in the genome sequence. It accounts for about 5% of genome and with relatively long segments of sequence of more than 1KB in length and more than 90% of sequence identity. Segmental duplications are distributed within the chromosome and throughout the genome and are mostly found in the sub-telometric and pericentrometric regions. Segmental duplications play a major role in the evolutionary history and its association with genetic disorders.

This review intends to examine the methodology followed in previous studies to achieve their objective of identifying segmental duplications. Finding similar sequences has been done through local sequence alignment algorithms, most frequently with the blast programme followed by global alignment algorithms. Further, some other tools such as Blat, DupMasker and MUMmer have also been used.

Keywords **-** Segmental Duplications; Human Genome; Genetic Diseases; Software Tools; Algorithms

Introduction

In 2003, the International Human Genome Sequencing Consortium announced that they had completed the standard reference human genome, according to the guidelines of the original Human Genome Project, with 99.99% accuracy. The completed sequence revealed a code over three billion bases long. Only about 2% of the genome encodes for proteins. The final sequence was found to contain only 25,000 genes(1).

Repeat sequences that do not code for proteins; accounts for more than 50% of the human genome. Though repeats are believed to have no direct function they may influence the structure and the dynamics of chromosomes.

These repeats may reshape the genome by re-arranging the genome by creating new genes or modifying and reshuffling the existing genes.

Repeats are divided in to five groups:

- Transposon-derived repeats
- Pseudogenes
- Simple repeats
- Segmental duplications

- Tandem repeats[2]

Segmental duplications (SDs), also known as low copy repeats (LCRs), are common architectural features of many genomes. Segmental duplications are long DNA sequence that map to, two or more genomic locations. They can contain any part of genomic DNA, including genic sequence and common repeats. They can be organised in tandem or be mapped to interspersed locations[3].

Segmental duplications are defined as large regions of ≥1Kb in length and ≥90% of sequence identity. They are concentrated in the pericentrometric regions and sub-telometric regions in the chromosomes. They are distributed intrachromosomally as well as interchromosomally. Human segmental duplications architecture differs from other mammalian genomes in its complexity and frequency of occurring large blocks of duplications[5] . They consist of 4-5% of the genome. As more than 50% euchromatic gaps contain undefined duplications, estimates of the content of segmental duplications may continue to rise incrementally to have an upper limit of 6%[3].

Identification and characterisation of segmental duplications are important as they play a role in:

- creating problems in mis-assignment and mis-assembly in the reconstructing human genome,
- evolutionary relationships (recent genomic changes might have contributed significantly to species divergence between humans and apes), and
- phenotypes of gene expression and genetic diseases.

Once they have been formed, segmental duplications are subject to the laws of evolution that affect all genomic sequences, like base-pair substitutions, insertions, deletions and retrotransposition. Highly identical segmental duplications also mutate by homology driven processes. Such processes contribute in several ways to structural differences in the architecture of segmental duplication regions, both within and between primate species.

Homology between segmental duplications can initiate non allelic homogenous recombination, which occurs through the alignment of highly similar segmental duplications followed by paralogous recombination. The type of rearrangement that occurs as a result of this (which can be duplication, deletion, inversion or translocation) depends on the location and orientation of the segmental duplications. Such recombination in meiosis is well recognised as the major cause of genomic disorders, particularly recurrent ones. Duplication mediated rearrangement typically occurs between segmental duplications that have alignments of >95% identity and >10 kb in length, with the largest and most identical segmental duplications showing the highest rates of non-allelic homogenous recombination. Given the tendency to recurrent rearrangement of such regions, Segmental duplications can be considered as a source of structural variation among humans[3]. Such structural variations may end up in dosage imbalance of genetic materials or generation of new gene products. These changes in the genome result in genetic diseases[4].

Genomic disorder	**Chromosomal rearrangement**	**Chromosomal location**	**Rearrangement size (MB)**
Charcot-Marie-Tooth disease type 1A (CMT1A)	Intersticial duplication	17p12	1.5
Hereditary neuropathy with pressure palsies(HNPP)	Deletion	17p12	1.5
Smith-Magenis syndrome (SMS)	Deletion	17p11.2	5
Duplication 17p11.2	Intersticial duplication	17p11.2	5
Neurofibromatosis type (NF1)	Deletion	17q11.2	1.5
Prader-Willi syndrome (PWS)	Deletion	15q11-15q13	4
Angleman Syndrome (AS)	Deletion	15q11-15q13	4
Inverted duplication 15	Supermumerary marker	15q11-15q14	4
(inv duplications (15))	chromosome	7q11.23	1.6
Williams-Beuren syndrome (WBS)	Deletion		
DiGeorge and velocardiofacial syndromes (DGS/VCFS)	Deletion	22q11.2	3
Cat eye syndrome(CES)	• Supermumerary marker • chromosome	22q11.2	3
X-linked ichthyosis	Deletion	Xp22	1.9
Haemophilia	Inversion	Xq28	0.5

Table 1. Genomic disorders mediated by segmental duplications[(4)]

Previous studies

Many studies have been done for identification and characterisation of segmental duplications since the time of the assembled human genome. Determination of segmental duplications is not trivial, as it is very difficult to differentiate paralogous segments from allelic overlaps, especially when the similarity fall within 99.5% involving potential sequence mis-assignment errors in the human genome[(6)]. Specialised methods will be required to integrate these regions into the reference human genome sequence(7). It is expected that when assembly coverage reaches its maximum, the content of segmental duplication would rises up to 6%.

For the past 10 years, studies have been done to identify and characterise segmental duplications in relation to individual chromosomes (7, 16, 17 etc. including X and Y) and genome wide basis(5, 5, 7, 8) and to find segmental duplication's associated with genomic diseases. In recent years many researchers have attempted to study the complex patterns of organisation of segmental duplications and to correlate it with evolutionary relationships(5). Furthermore, segmental duplications of apes have been studied for comparative analysis(5). Segmental duplications of other mammalians and eukaryotes also have been studied.

Assemblies used

Studies have made use of various human genome assemblies NCBI build 23(Sep. 2000)(7) 28(Dec. 2001), 29(Apr. 2002), 30(Jun. 2002)(6), 33(Apr. 2003)(8), 34(Jul. 2003)(8), 35(May. 2004)(5) and data sets of known human segmental duplication(9).

Results and analysis

It was estimated that segmental duplications of ≥1Kb in length and ≥ 90% in identity in the human genome accounted for 5 to 6% (7,8) of the genome. Segmental duplications of ≥5 Kb in length and ≥ 90% of identity accounted for 3.53%. 1530 of intra-chromosomal duplications accounting for 2.64% and 1637 inter chromosomal duplications (total 3167) accounting for 1.44% were also found (7). Another study using the same criteria has reported a total of 3119 duplications accounting for 3.9% of the genome. Intra-chromosomal duplications were 1.7 times more numerous than inter-chromosomal duplications(5). The differences in content may be attributed to the duplications length and method used in the analysis.

The largest duplication in the human genome spans 1.5MB on chromosome Y and the average duplication length is 33.7 KB. Chromosomal distribution of duplications is not uniform, with chromosome Y and 9 having the greatest duplication content (53.0 and 9.0%, respectively), and chromosome 3 having the least duplication content (0.7%)

Peri-centromeric and sub-telometric regions have increased repeats and these repeat dense regions contributed to evolution of the human genome(6). Intra-chromosomal duplications are enriched in both peri-centromeric and sub-telometric regions, whereas inter-chromosomal duplications are found more commonly (6:1)(3) in the peri-centromeric regions(3).

Regions containing recent segmental duplication can harbour rapidly evolving hominoid specific genes as well as novel gene families that are unique to primates(6). Gene duplication followed by functional speciation has been considered a major evolutionary force for gene innovation. These genes embedded within recent genomic duplications may end up in adaptations specific to primate evolution(13).

Using NCBI refseq annotation, 1152 human genes that were mapped to duplication regions have been identified. Out of these 1152 genes, 475 genes were fully located within duplication regions and were the best candidate for recent whole genome duplications. It has been found that there was a significant increase in gene duplication for genes involved in immune defense and reproduction in functional analysis using the ‘gene ontology consortium database[6].

For the determination of the organisation and structure of human segmental duplications accurately, a high quality genome assembly is required[8].

Tools and software

Blast[9,10,11,12] RepeatMasker[18,19,20] DupMasker[20] MUMmer[15,16], Blat[17] and Perl suits.

Methodology

Most methods have used pair-wise comparison combined with algorithms that find large blocks with high identity.

RepeatMasker, Blast and GS Aligner/Align

Chromosome sequences masked repeats using RepeatMasker have been compared against itself with chromosome-wide Blast to detect intra-chromosomal segmental duplications and pair-wise sequence alignment to 24 chromosomes to detect inter-chromosomal segmental duplications[6].

Once downloaded, each chromosome is divided into segments of 400Kb[7], 500Kb[8] or is use as a whole chromosome to start sequence comparison[6].

Then RepeatMasker is applied to mask high copy repeats for sequences with repeats. Resulting unique genome sequences have undergone global BLAST similarity alignments with reduced affine gap extension parameters which allow large gap 1Kb to be traversed[7] with default parameters[6,8]. BLAST was done for all segments or whole chromosome to itself and against all other segments and chromosomes.

The BLAST results have been parsed for alignment with >1 kb of aligned bases and >88% identity. Each alignment has been re-inserted with the high-copy repeats and then alignment end trimming had been done with the programme blast end trim. End trimming more precisely defines the alignment end positions, which may have been incorrect as a result of the relaxed gap parameters used or because the true end positions resided in a high copy repeat. Blast end trim is a heuristic programme that attempted to extend the alignment (up to 2 Kb) beyond the defined end position using global alignments generated by **ALIGN** (Myers and Miller 1988). When extension failed, the length of the attempted extension is recursively decreased until it converges on a given end position. After trimming, ALIGN is used to generate global alignments from which statistics were calculated using the programme align scorer (J.A. Bailey, unpubl.). Global alignments that equal or exceed the threshold of 1000 bases is aligned and those with ≥90% identity (ie. gaps excluded) are retained for further analysis. Generation of global alignments also act as a safeguard against false positives from BLAST analysis[7].

In other research with segments of 500Kb size, from the BLAST results, self-hits of each DNA segment and hits with less than 90% similarity have been discarded. The remaining BLAST hits that are less than 50 Kb apart on the same chromosome are combined into one tentative duplication block. After this step, the sequences of each block pair plus the 10 Kb sequences from each side of the block is taken out. The GS-aligner programme (Shih and Li 2003) is used to align the two sequences of each block pair. The GS-aligner produces HSP (High-Scoring segment Pairs) and non-HSP regions. HSP regions are highly similar regions without gaps, whereas non-HSP regions have a lower similarity and may contain gaps. Two HSP regions with > 90% sequence similarity are combined if the non HSP region between them also has a sequence similarity ≥90%. However, non HSP regions may have a similarity lower than 90% because of random fluctuations. To be more vigorous, a binomial test is applied to each non-HSP region and if the sequence similarity is not significantly lower (P, 0.05) than 90%, the two flanking HSP regions and the non-HSP region are combined into one segment. The alignment end is extended by a dynamic programming for up to 5 Kb outward from both ends of the alignment while maintaining the requirement ≥90% sequence similarity during the extension process(8).

All BLAST results are subsequently parsed to eliminate low quality and fragmented alignments under the following criteria: BLAST results having ≥ 90% sequence identity, ≥80 bp in length, and with expected value $\leq 10^{-30}$. Each BLAST report is sorted by chromosomal coordinates. All identical hits (same coordinate alignments), including suboptimal BLAST alignments recognised by multiple, overlapping alignments, as well as mirror hits (reverse coordinate alignments) from the BLAST results of the intra chromosomal set are removed. Contiguous alignments separated by a distance of less than 3 KB, then 5 KB and subsequently 9 KB are joined (stepwise) into modules in order to traverse masked repetitive sequences and to overcome breaks in the BLAST alignments caused by insertions/deletions and sequence gaps. Such contiguous sequence alignment modules represent sequence similarity between the subject and query chromosome sequence in question at their respective positional coordinates(6).

MUMmer

MUMmer is a system of algorithms for the comparison of large scale genome sequences developed by The Institute of Genetic Research (TIGR)(18). MUMmer version 3.0 is open source with publicly available codes for redistribution and with modular extensibility. It is based on generation of suffix tree data structure and is 4 - 110 times faster than BLAST. Its memory usage is less than 4 GB of real memory(18).

Using minimum human genome sequence it can be queried against itself. In a benchmark test of MUMmer, the time taken to build the suffix tree was 4.7 hours, while query time was 101.5 hours and memory used had not exceeded 3.9 GB on a single server type computer, so that total time taken was 4.5 days to compare the entire human genome(18).

BLAT

BLAT is a very effective tool for nucleotide alignment of genomic DNA taken from the same species. It is more accurate and orders of magnitude faster than other published tools. BLAT working in translated mode is capable of rapidly aligning data across vertebrate species without significant compromise. While TBLASTX can be configured to be more sensitive than BLAT, at settings commonly used for mammal comparisons, BLAT runs approximately 50 times faster. BLAT implements a very quick algorithm for finding multiple perfect matches, which allows the search stage to be specific enough that the genome itself can be kept on disk and only the index kept in RAM in memory in the client/server mode(19). The BLAT software in source and executable form is available without charge (http://www.soe.ucsc.edu /kent) for nonprofit, academic and personal uses. BLAT has been used in an annotated package of programmes and scripts to detect human segmental duplications.

Conclusion

Many research projects have been carried out to explore segmental duplications or low copy repeats as they show an association with genomic diseases. Identification and characterisation of these is the foundation on which future work rely.

Researchers have used BLAST frequently as their tool to analyse sequence. Other tools include BLAT, MUMmer, and ALIGN.

The latest version of MUMmer has proved to be a fast and versatile tool to identify human segmental duplications even on a single computer. Furthermore, it includes a module to display the matches.

References

1. The Human Genome Project. Available from: http://www.sanger.ac.uk/about/history/hgp/

2. International Human Genome Sequencing Consortium: Initial sequencing and analysis of the human genome. Nature 2001; **409**: 861-921.

3. Bailey JA, Eichler EE. Primate segmental duplications: crucibles of evolution, diversity and disease. Nature 2006; **7:** 552-564.

4. Emanuel BS, Shaikh TH. Segmental duplications an 'expanding' role in genomic instability and disease. Genetics 2001; **2:** 791-800.

5. Jiang Z, Tang H, Ventura Mutations, Cardone MF, Marques-Bonet T, *et al* Ancestral reconstruction of segmental duplications reveals punctuated cores of human genome evolution. Nature Genetics 2007; **39**: 1361-1368.
   Doi: http://dx.doi.org/10.1038/ng.2007.9

6. Cheung J, Estivill X, Khaja R, MacDonald J.R, Lau K, *et al*. Genome-wide detection of segmental duplications and potential assembly errors in the human genome sequence. Genome Biology 2003; **4**: R25.1-25.10.

7. Bailey JA, Yavor AM, Massa HF, Trask BJ, Eichler EE. Segmental Duplications: Organization and Impact within the Current Human Genome Project Assembly. Genome Research 2001; **11**: 1005-17.
   Doi: http://dx.doi.org/10.1101/gr.GR-1871R

8. Zhang L, Lu H.H.S, Chung W, Yang J and Li W: Patterns of Segmental Duplication in the Human Genome. Molecular Biology and Evolution 2005; **22**(1): 135-41.

Doi: http://dx.doi.org/10.1093/molbev/msh262

9. Madden T. Chapter 16 - The BLAST Sequence Analysis Tool. The NCBI Handbook. Available from: http://www.ncbi.nlm.nih.gov/books/NBK21097/

10. Blast - Basic Local Alignment Search Tool from Advanced Biocomputing, LLC. Available from: http://blast.advbiocomp.com/doc/README.html

11. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Molecular Biology 1990; **215**: 403-10.

12. Bailey JA, Gu Z, Clark RG, Reinert K, Samonte RV, *et al.* Recent Segmental Duplications in the Human genome. Science 2002; **297**: 1003-07.
Doi: http://dx.doi.org/10.1126/science.1072047

13. Jiang Z, Hubley R, Smit A, and Eichler EE. DupMasker: A tool for annotating primate segmental duplications. Genome Research 2008; **18:** 1362-68.
Doi: http://dx.doi.org/10.1101/gr.078477.108

14. Delcher A, Phillippy A, Carlton J and Salzberg L. Fast algorithm for large-scale genome alignment and comparison. Nucleic Acid Research 2002; **30**: 2478-83.
Doi: http://dx.doi.org/10.1093/nar/30.11.2478

15. Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, *et al.* Versatile and open software for comparing large genomes. Genome Biology 2004; **5**: R12.1-12.9.

16. Kent WJ, Blat - the Blast-like Alignment Tool. 2002; 1-9.
Available from: http://www.ncbi.nlm.nih.gov/pubmed/11932250

17. Tarailo-Graovac M, Chen N. Using Repeat Masker to Identify Repetitive Elements in Genomic Sequences. Current Protocols in Bioinformatics 2009; **4:** 10.1-10.14.
Doi: 10.1002/0471250953.bi0410s25.

18. Saha S, Bridges S, Magbanua ZV, and Peterson DG. Empirical comparison of ab initio repeat finding programs. Nucleic Acid Research 2008; **36**: 2284-94.
Doi: http://dx.doi.org/10.1093/nar/gkn064

19. Saha S, Bridges S, Magbanua ZV, and Peterson DG. Computational Approaches and Tools Used in Identification of Dispersed Repetitive DNA Sequences. Tropical Plant Biology 2008; **1:** 85-96.
Doi: http://dx.doi.org/10.1007/s12042-007-9007-5