

**Genome Analysis**

**ABWGAT: Anchor Based Whole Genome Analysis Tool.**

Sarbashis Das $^1$, Anchal Vishnoi $^1$ and Alok Bhattacharyya $^{1,2,*}$

$^1$ Center for Computational Biology and Bioinformatics, School of Information Technology, Jawaharlal Nehru University, New Delhi 110067

$^2$ School of Life Sciences, Jawaharlal Nehru University, New Delhi, 110067, India.

Associate Editor: Prof. Alfonso Valencia

**ABSTRACT**

Summary: Large numbers of genomes are being sequenced regularly and the rate will go up in future due to availability of new genome sequencing techniques. In order to understand genotype to phenotype relationships it is necessary to identify sequence variations at the genomic level. Alignment of a pair of genomes and parsing the alignment data is an accepted approach for identification of variations. Though there are a number of tools available for whole genome alignment none of these allows automatic parsing of the alignment and identification of different kinds of genomic variants with high degree of sensitivity. Here we present a simple web based interface for whole genome comparison named ABWGAT (Anchor Based Whole Genome Analysis Tool) that is simple to use. The output is a list of variations such as SNVs, INDELs, Repeat expansion and Inversion.

Availability: The web server is freely available to non-commercial users at the following address:

http://abwg yaşayan.ac.in/~sarba/

Contact: dsarbashis@gmail.com

Supplementary Information: Supplementary data are available at http://abwg yaşan.ac.in/~sarba/cgi-bin/abwgc_retrival.cgi using job id 524, 526 and 528.

**1 INTRODUCTION**

A number of bacterial genomes have been sequenced so far and with the advent of new genome sequencing techniques many more are being added regularly. Particularly, a number of strains with different phenotypes, such as resistance to drugs are being sequenced in order to understand relation between genotype and phenotype. There are a number of tools that can help to analyse genomic sequences or more specifically can align two or more genome sequences in order to identify the sequence variations (Blanchette M, 2007). In general the alignments generated by these algorithms need to be parsed to list the sequence variations. Moreover, most of the methods are not able to identify all the variations, such as single nucleotide variations (SNV), indels, repeat expansion and divergent region. There are also highly useful visualization tools available for comparative analysis of multiple genomes such as K-BROWSER (Chakrabati K and Pachter L, 2004) and MAUVE (Darling C. et al, 2004). These tools help to identify gross changes and not list all the sequence differences that may exist. We had developed an algorithm ABWGC (Anchor Based Whole Genome comparison) for pair wise whole genome comparison and identification of all sequence variants (Vishnoi A. et al, 2007). It was shown to be useful for identification of genomic variations in related organisms including different species.

**2 METHODS**

In this report we describe ABWGAT (Anchor-Based Whole Genome Analysis Tool) and a pipeline for automated identification and listing of genomic variations in a pair wise manner. ABWGAT, based on a modified version of ABWGC algorithm, can identify insertion, deletion, single nucleotide variations (SNV), repeat expansion and inversion. A flow diagram describing the different modules of ABWGAT is shown in Fig. 1. The basic features of ABWGC, is maintained in this tool. The process of identification of random anchors from reference and query genomes is essentially as described before (Vishnoi A. et al, 2007). The anchor length of 100 nucleotides was found to be about optimum with respect to noise and computation time. It was decided on the basis of experimentation using different genomes and due to low probability of finding by chance a match for this length of sequence in a genome. This can be shown as follows. The match of an anchor in a genome has binomial distribution, due to large size of genomes. This can be approximated to a Poisson distribution. If the size of genome is 4,500,000 (generally the size of a large bacterial genome), the probability of finding a fixed given sequence of length 100 in a genome of this size is less than $2.8 \times 10^{-53}$ by a simple Poisson approximation of a binomial.

The distance between consecutive anchors of both the genomes can be used to determine if there is any insertion or deletion in any of the sequence. If the lengths are exactly same it is possible that there may still be single nucleotide changes. Presence of SNVs in these sequences was determined by comparing dinucleotide frequencies of the two sequences. The exact position of the SNV is estimated by using a suffix-tree based algorithm (Gusfield, D. 1997). The same algorithm is also used to find short inversion or indels (<7 nucleotides). When the inter anchor length difference is larger than 7 nucleotides a global alignment algorithm (EMBOSS package http://emboфесс.sourceforge.net/) is used to identify the positions of SNVs and indels. Translocation, recombination and large inversions are detected by identifying anchor synteny breakpoints in the two genomes.

$^a$To whom correspondence should be addressed.
Many of the sequence variations in the genomes are due to expansion or contraction of tandem repetitive elements. Inter anchor regions suspected to have indels are also analyzed for tandem repeats using “tandem repeat finder” (Benson G, 1999). The output is parsed to find presence of tandem repeats, number of copies in the two inter anchor regions, length and genomic positions.

3 IMPLEMENTATION

The program runs on a Linux workstation and the web server is accessible through internet via Perl and CGI based web interface. The main codes are written in C++ and Perl. Most of the codes run in parallel using MPI (Message Passing Interface). To implement MPI we used MPICH package (http://www.mcs.anl.gov/research/projects/mpi/mpich1/). All complete genome sequences and original annotation files were downloaded from NCBI (ftp://ftp.ncbi.nih.gov/genomes/Bacteria/). The genomes are selected using drop down menus. User defined sequences can also be uploaded using a menu driven buttons. For accurate analysis the user has to make sure that the two sequences are not very different from each other or from selected genomic sequences. The user can select the type of analysis, i.e. the nature of variations (SNV, insertion, inversion and repeat expansion). The results are retrieved later using the job id number provided at the time of submission.

4 ANALYSIS OF GENOMES BY ABWGAT

The server was used to analyse genomic variations among strains of three different bacterial species. These species are *Acinetobacter*, *Francisella* and *Vibrio*. Query and subject genomes were selected from the drop down menu. In the current implementation only analysis of different strains of a species is permitted. The summary of the results of analysis of three pairs of genomes are shown in the table. 1. Depending upon the species the number of variants was found to be quite different. For example, the number of indels in *A. baumannii* was much higher than that of other two organisms and *V. cholerae* displayed a very high number of SNVs. The output of the web server contains position in the genome, type of nucleotide change, gene name (if not intergenic) and predicted functions. The complete results can be viewed in supplementary material.

5 CONCLUSION

In this report a web server for identification of genomic variations using a pair of completely sequenced genomes has been presented. The usefulness of ABWGAT is the ease by which comparative genomics analysis can be carried out without prior knowledge of bioinformatics or computation methods. The output comes out in listed format for easy interpretation unlike other comparative genomics tools that displays output either graphically or table format in the webpage.

<table>
<thead>
<tr>
<th></th>
<th>Total No. of SNVs</th>
<th>Total No. of INDELs</th>
<th>Total No. of Repeat expansion</th>
<th>Total No. of Inversion Long, short</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Acinetobacter baumannii</em> AB0057 and <em>Acinetobacter baumannii</em> B307_0294</td>
<td>335</td>
<td>415</td>
<td>6</td>
<td>5, 660</td>
</tr>
<tr>
<td><em>Francisella tularensis holarctica</em> and <em>Francisella tularensis holarctica</em> FTA</td>
<td>853</td>
<td>67</td>
<td>5</td>
<td>0, 29</td>
</tr>
<tr>
<td><em>Vibrio Cholerae</em> M66_2 and <em>Vibrio Cholerae</em> 0395</td>
<td>4259</td>
<td>92</td>
<td>7</td>
<td>3, 63</td>
</tr>
</tbody>
</table>

Table 1. The results of analysis of three pairs of genomes.

ACKNOWLEDGEMENTS

We thanks to CDAC (Center for Development of Advance Computing) for help in implementing parallel computing.

Funding: The authors thank DBT and COE (DBT) for support in carrying out this work.

Conflict of Interest: none declared.
REFERENCES


