AlignDB: a computational protocol for analyzing the relationship between indels and SNPs in genome sequences

Qiang Wang
Department of Biology, Nanjing University
Jian-Qun Chen
Department of Biology, Nanjing University
Dacheng Tian
Department of Biology, Nanjing University

Method Article

Keywords: mutation hotspot, indel, insertion/deletion, SNP

Posted Date: September 22nd, 2008

DOI: https://doi.org/10.1038/nprot.2008.208

Abstract

Introduction

AlignDB is developed to investigate the distribution of single-nucleotide changes around insertion/deletions \(indels) in genome comparisons. There are two set of analysis, two-way and three-way, in AlignDB. The two-way analysis indicates that nucleotide divergence \(D) is substantially elevated surrounding indels and decreases monotonically to near-background levels over several hundred bases. D is significantly correlated with both size and abundance of nearby indels. In a comparison of closely related species, the three-way analysis is available. We find that derived nucleotide substitutions surrounding indels occur in significantly greater numbers on the lineage containing the indel than on the one containing the ancestral \(non-indel) allele; the same holds within species for single-nucleotide mutations surrounding polymorphic indels. AlignDB is freely available on request from the authors. The parameters are fully modifiable. The following protocol describes how to use AlignDB to do two-way and three way genomic analyses for indels.

Equipment

Hardware: * A machine capable of running a modern OS. \(e.g. Window 2000 or newer, Linux kernel 2.6 or newer, OS X, etc.) * 1 Gb of RAM. * 80 Gb of hard drive space. Software: * Perl 5.8.x or better. * MySQL 5.0.x. * ClustalW alignment program. * Microsoft Excel or OpenOffice.org. * AlignDB itself. Data: * Whole genome alignment. * Genomic annotation. We use Ensembl as the genomic annotation source.

Procedure

1. Install all AlignDB requirement software and dependent Perl modules. Read the README file in the AlignDB package root directory. 2. Install AlignDB itself. Follow the steps in doc/tutorial.pdf and use the example data in data/ directory coming with AlignDB for a trial run. 3. Prepare a set of input data in Blastz .axt file format. For example, you can download the completed genomic alignments from the UCSC Genome Bioinformatics Site. 4. Starting the GUI shell. Enter the base directory where you installed AlignDB. Type the follow command: >perl gui\gui2.pl The GUI interface shows in Fig.1. DO NOT close the command line window, AlignDB need it. 5. Two-way comparison steps. Step 1: test your server. Click on the “DB Server” notebook tab, and check MySQL server setting. Step 2: initiate your database. Then click the “Init. alignDB” button, which will call init\init_alignDB.pl to build the skeleton of the database. Step 3: generate two-way analysis database. Click the “Gene. alignDB” button which will call init\gene_alignDB.pl to analysis alignments and build the database. Step 4: update features using Ensembl annotations. Click the “Feature” tab, change nothing and click the “Upd. feature” button, which will run init\update_feature.pl. Step 5: update misc features. Click the “Misc” tab. Click “Upd. slippage” to run init\update_indel_slippage.pl and then click “Upd. isw” to run init\update_isw_indel_id.pl. Step 6: do statistics and get results. Then click “Common stat” button. An Excel workbook with 25 worksheets will be generated. Step 7: charting. If you run AlignDB on Windows, just like me, you could click the “Common

chart” button, which will run stat\common_chart_factory.pl that call Excel to draw charts in .xls files by OLE. Fig.2 is an example workbook. 6. Three-way comparison steps. Three-way comparison needs three genome sequences to accomplish. So you should add a new genome sequence into the previous two-way comparison. This can be accomplished by joining two two-way comparisons, which share a common target sequence. Step 1: build another two-way comparison. You can just do steps 1-3 and omit step 4-7 if you don’t need statistical analysis for the new two-way database. Step 2: generate three-way analysis database. Load two completed databases into “first db” and “second db”. Generate “goal db” name with the default naming rule. Leave the “T/Q/R:” options unchanged, and click the “Ref outgroup” button. See Fig.3 for an example. Step 3: update two-way features. Three-way comparison database is also a legal two-way database. Do step 4-5 of two-way comparison to update the new three-way database. Step 4: update three-way features. Go to “Misc” tab in the “Three-way” pane and click the “Upd. CpG” button to run extra\update_snp_cpg.pl. Step 5: do statistics and get results. Then click “Three stat” button which will run stat\three_stat_factory.pl. An Excel workbook with 11 worksheets will be generated. Step 6: charting. This step is almost the same as step 7 in Two-way comparison. Just click the “Three chart” button, which will run stat\three_chart_factory.pl.

Figures

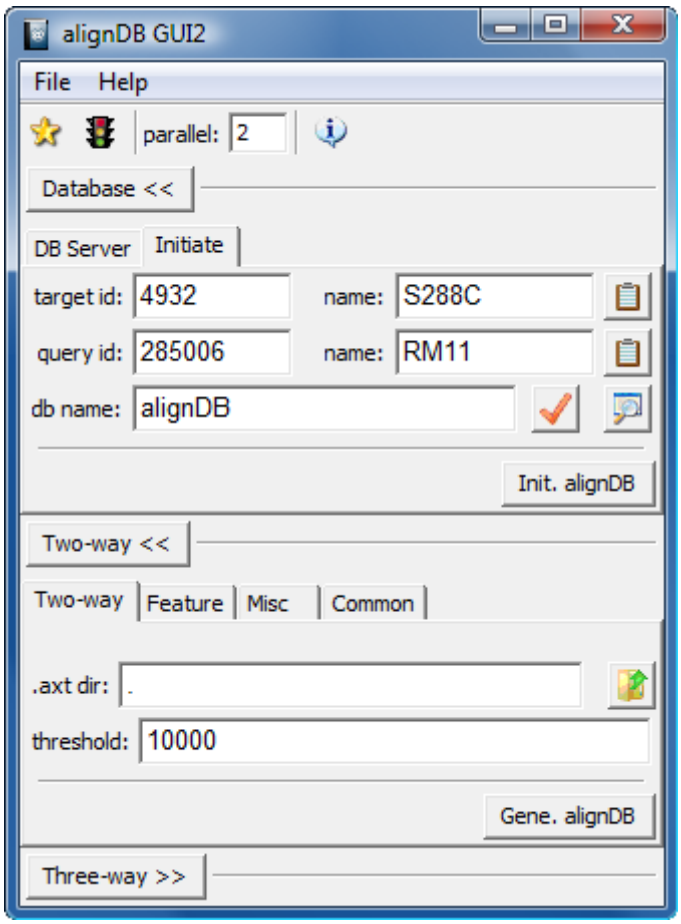
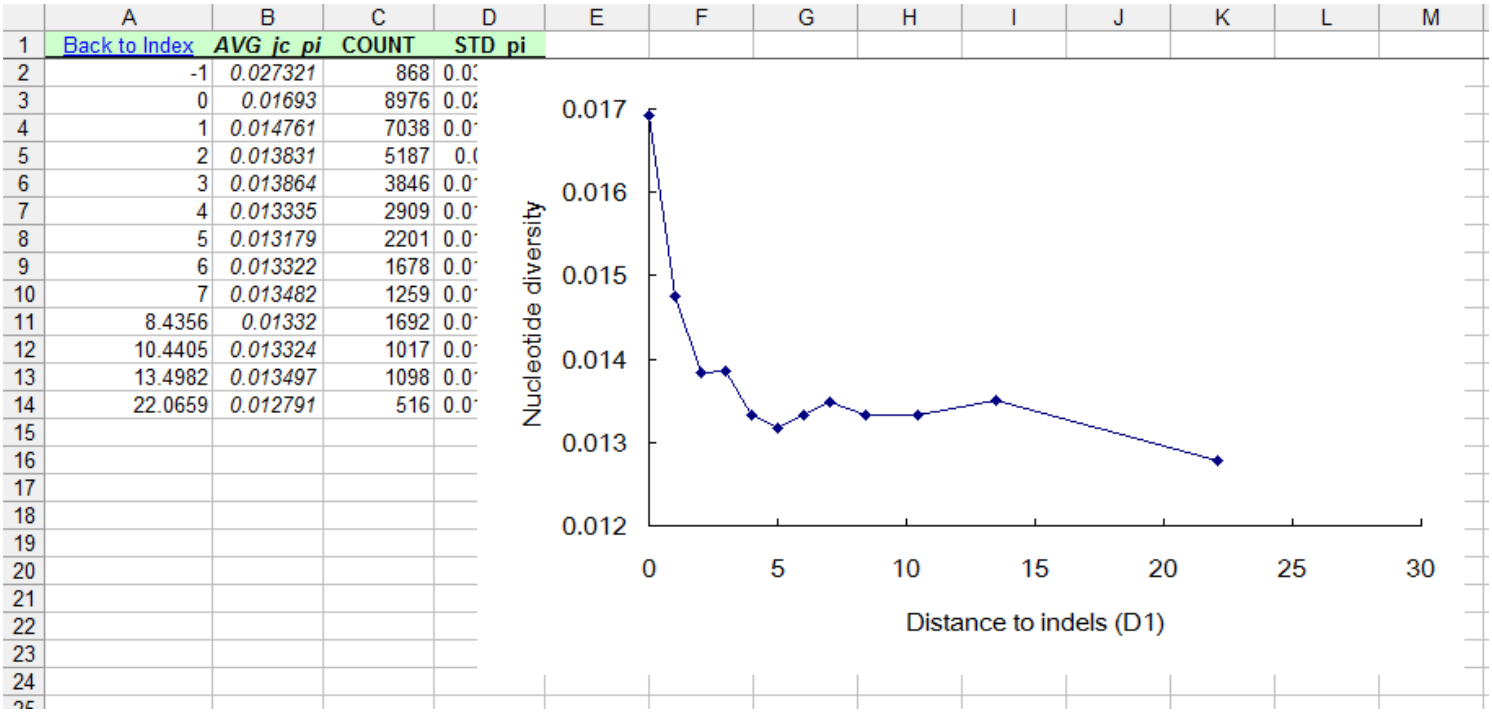


Figure 1

The GUI shell of AlignDB.



	A	B	C	D
1	Back to Index	*AVG_jc_pi*	COUNT	STD_pi
2	-1	*0.027321*	868	0.0
3	0	*0.01693*	8976	0.0
4	1	*0.014761*	7038	0.0
5	2	*0.013831*	5187	0.0
6	3	*0.013864*	3846	0.0
7	4	*0.013335*	2909	0.0
8	5	*0.013179*	2201	0.0
9	6	*0.013322*	1678	0.0
10	7	*0.013482*	1259	0.0
11	8.4356	*0.01332*	1692	0.0
12	10.4405	*0.013324*	1017	0.0
13	13.4982	*0.013497*	1098	0.0
14	22.0659	*0.012791*	516	0.0

Figure 2

Result workbook of two-way analysis.

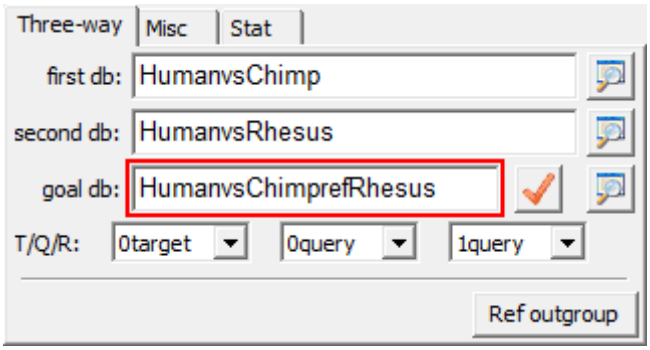


Figure 3

Build a three-way database from two-way databases.