BIOINFORMATICS APPLICATIONS NOTE Vol. 28 no. 7 2012, pages 1021–1023
doi:10.1093/bioinformatics/bts063

Genome analysis

Advance Access publication February 2, 2012

ArchTEx: accurate extraction and visualization of next-generation sequence data

William K. M. Lai[1,2,†], Jonathan E. Bard[2,†] and Michael J. Buck[1,2,3,*]

[1]Department of Biochemistry, [2]Center of Excellence in Bioinformatics and Life Sciences, State University of New York at Buffalo, 701 Ellicott St, Buffalo, NY 14203 and [3]Molecular Epidemiology and Functional Genomics, Roswell Park Cancer Institute, Buffalo, NY 14263, USA

Associate Editor: Michael Brudno

ABSTRACT

Motivation: The extension of mapped sequence tags is a common step in the analysis of single-end next-generation sequencing (NGS) data from protein localization and chromatin studies. The optimal extension can vary depending on experimental and technical conditions. Improper extension of sequence tags can obscure or mislead the interpretation of NGS results. We present an algorithm, ArchTEx (Architectural Tag Extender), which identifies the optimal extension of sequence tags based on the maximum correlation between forward and reverse tags and extracts and visualizes sites of interest using the predicted extension.

Availability and implementation: ArchTEx requires Java 1.6 or newer. Source code and the compiled program are freely available at http://sourceforge.net/projects/archtex/

Contact: mjbuck@buffalo.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on October 25, 2011; revised on January 4, 2012; accepted on January 27, 2012

1 INTRODUCTION

Next-generation sequencing (NGS) is now being used with great success in conjunction with laboratory techniques such as chromatin immunoprecipitation (ChIP) (Robertson *et al*., 2007), micrococcal nuclease (MNase) digestion (Schones *et al*., 2008), and FAIRE (Gaulton *et al*., 2010). A standard high-throughput NGS run can typically produce several million short-read sequence tags (Mardis, 2008). The number and length of the sequence tags varies depending on the particular experiment and the platform on which it was run (Mardis, 2008). A standard run on a single lane of the Illumina HiSeq2000 platform produces upwards of 200 million tags ranging anywhere from 36 to 100 bp in length. These tags are then typically aligned to a reference genome using a short-read tag alignment algorithm such as Eland or Bowtie (Langmead *et al*., 2009). Further analysis is often dependent on the nature of the experiment and the personal preferences of the investigator.

The use of sequence tag extension in NGS analysis is used to adjust for the length of DNA that was sequenced and to smooth data at portions of the genome which may have been poorly sampled (Pepke *et al*., 2009). Tag extensions are justified by the relatively short sequence tag lengths produced by the majority of high-throughput platforms. The short sequence tag represents the 5′ end of a larger DNA strand that was sequenced. The length of the extension should reflect the true length of the DNA population that was sequenced. These extended tags are then typically overlaid on top of each other in a genome-wide tag frequency map. Regions of interest can then be examined independently or as an average across multiple similarly classified regions.

Currently, researchers will extend their mapped sequence tags based on known biology such as 146 bp extensions in the case of a mononucleosome sequencing run (Zhang *et al*., 2009) or an extension based on the estimated length of the DNA during the pre-sequencing library preparation protocol (Pepke *et al*., 2009). This article describes an algorithm, ArchTEx (Architectural Tag Extender), which identifies the average length of the DNA fragments that were sequenced using cross-correlation of single-read sequencing. ArchTEx also provides a method for quick extraction and visualization of individual sites of interest using the optimal tag extension as predicted by cross-correlation. ArchTEx can output the results in a format readable by other NGS analysis packages (Lai and Buck, 2010), clustering software, and can be uploaded to UCSC genome browser.

2 IMPLEMENTATION

2.1 Architecture

ArchTEx was designed and coded in Java 1.6. Graphical visualizations of the data are implemented through the open source java packages JFreechart and JCommon (JFree; JFree). The input file is a compiled BAM file produced by the short-read tag alignment algorithm Bowtie (Langmead *et al*., 2009). BAM file management is controlled by the picard-tools java package (samtools). The number of CPU cores to utilize can also be specified by the user for improved run times. ArchTEx can be run on any platform possessing Java 1.6 or newer.

2.2 Design

The main workflow of DNA length prediction performed by ArchTEx is depicted in Figure 1. ArchTEx begins by randomly sampling uniquely mapped sequence tags from each chromosome from the input BAM alignment file. By default, the sequence tags from 10 to 50 kb regions are randomly sampled from each chromosome. Repeated random sampling helps reduce the dangers

*To whom correspondence should be addressed.

†The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors

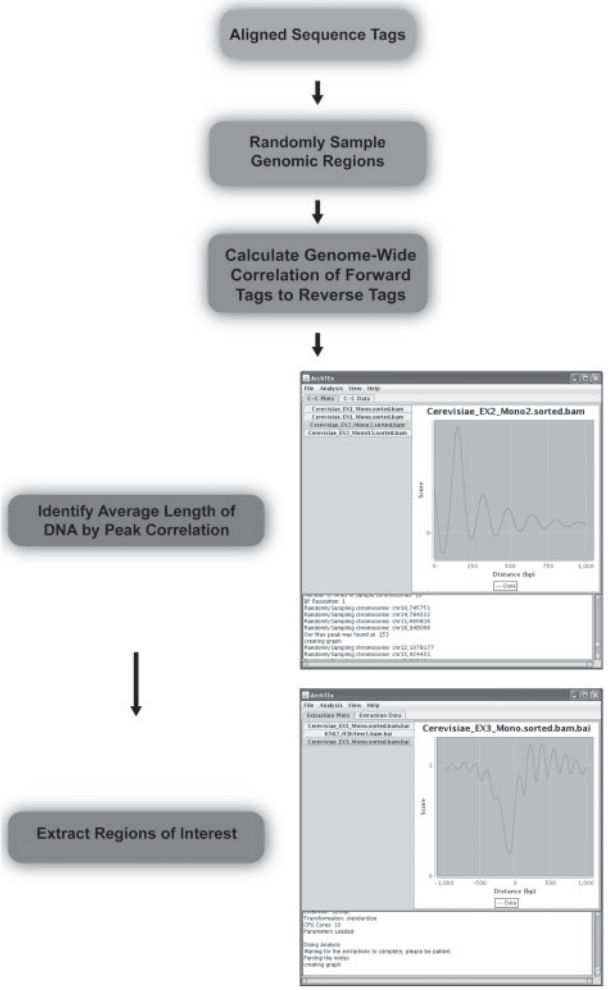

Fig. 1. After initial sorting and parsing of all sequence tags, the genome-wide cross-correlation calculated at each base pair shift from 1 to 1000 bp between forward and reverse sequence tags. Sample output after cross-correlation of an *S.cerevisiae* MNase-Seq experiment.

of outliers skewing the cross-correlation. Each random sample is then parsed into two separate files containing the forward and reverse tags, respectively. Each forward and reverse subfile is then sorted by sequence tag start site in order to allow every tag to be processed in order of genomic position. ArchTEx then counts how many tag start sites are present for every base pair for both forward and reverse strands. The data from each of the random samples are then merged and the reflective Pearson correlation is then calculated between these tag populations by comparing the tag counts at each base [Equation (1)].

$$R = \frac{\sum(\text{Forward Tag Count} \times \text{Reverse Tag Count})}{\sqrt{\sum(\text{Forward Tag Count})^2 \times \sum(\text{Reverse Tag Count})^2}} \quad (1)$$

The strand shift is calculated by adjusting which reverse tag counts are correlated to the forward tags by sliding the reverse strand one base and performing the correlation to the forward tag counts. This process is calculated for all strand shifts from 1 to 1000 bp. ArchTEx then outputs the genome-wide correlations of the forward tags to the reverse tags at all strand shifts and the strand shift which produced the highest genome-wide correlation.

2.3 Runtime

Estimated runtimes for ArchTEx's DNA length estimation depend on the number of random regions sampled as well as the number of CPU cores devoted to the program. The speed of ArchTEx decreases as the number of random regions sampled increases. Correspondingly, increasing the number of cores used by ArchTEx will result in a decrease in computational time. In addition, larger BAM files will reduce the speed of ArchTEx due to the computational requirements of parsing large amounts of reads. For example, ArchTEx using 2 cores on a human BAM file containing 25 million sequence tags will take ~3 min to complete. Extracting 27 000 transcription start sites (TSSs) in human from the same file using 2 CPU's takes ~1 min.

2.4 Output

ArchTEx graphically displays the cross-correlation plot showing the peaks of predicted DNA length as depicted in Figure 1. The cross-correlation data at each base pair are also made immediately available for download for further analysis. ArchTEx allows for the immediate extraction of sites of interest using any input BAM file and a user-specified tag extension. This data are graphically presented by ArchTEx and made available for download for alternative analysis. MNase-Seq data from for TSSs in *Saccharomyces cerevisiae* are shown in Figure 1 (Rizzo *et al*., 2011).

3 RESULTS

The accuracy of ArchTEx was confirmed for ChIP-Seq and MNase-Seq datasets by comparing ArchTEx's extended data from single-end reads to the actual sequenced fragments from paired-end experiments (Ercan *et al*., 2011; Kent *et al*., 2011; Wang *et al*., 2010) (Supplementary Figs 1–3). The problems of improper extensions such as peak shifting and bi-modal peak formation may be avoided using the predicted extension provided by ArchTEx (Supplementary Fig. 4), although these problems may persist if the length of DNA in the experiment is highly variable. ArchTEx provides immediate visualization of the tag distribution at individual regions of interest. This will be helpful in analysis of a variety of different NGS experiments including but not limited to ChIP-Seq, MNase-Seq and FAIRE-Seq.

Funding: National Science Foundation (grant IIS1016929 to M.J.B.)

Conflict of Interest: none declared.

REFERENCES

Ercan,S. *et al.* (2011) High nucleosome occupancy is encoded at X-linked gene promoters in C. elegans. *Genome Res.*, **21**, 237–244.

Gaulton,K.J. *et al.* (2010) A map of open chromatin in human pancreatic islets. *Nat. Genet.*, **42**, 255–259.

Jfree Jcommon-1.0.16. http://www.jfree.org/jcommon/index.html.

Jfree Jfreechart-1.0.13. http://www.jfree.org/jfreechart/.

Kent,N.A. *et al.* (2011) Chromatin particle spectrum analysis: a method for comparative chromatin structure analysis using paired-end mode next-generation DNA sequencing. *Nucleic Acids Res.*, **39**, E26.

Lai,W.K. and Buck,M.J. (2010) Archalign: coordinate-free chromatin alignment reveals novel architectures. *Genome Biol.*, **11**, R126.

Langmead,B. *et al.* (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.*, **10**, R25.

Mardis,E.R. (2008) Next-generation DNA sequencing methods. *Annu. Rev. Genomics Hum. Genet.*, **9**, 387–402.

Pepke,S. *et al.* (2009) Computation for ChIP-Seq and RNA-Seq studies. *Nat. Methods*, **6**, S22–S32.

Rizzo,J.M. *et al.* (2011) Tup1 stabilizes promoter nucleosome positioning and occupancy at transcriptionally plastic genes. *Nucleic Acids Res.*, **39**, 8803–8819.

Robertson,G. *et al.* (2007) Genome-wide profiles of Stat1 DNA association using chromatin immunoprecipitation and massively paralle sequencing. *Nat. Methods*, **4**, 651–657.

Samtools Picard-Tools-1.52. http://picard.sourceforge.net/.

Schones,D.E. *et al.* (2008) Dynamic regulation of nucleosome positioning in the human genome. *Cell*, **132**, 887–898.

Wang,C. *et al.* (2010) An effective approach for identification of in vivo protein-DNA binding sites from paired-end ChIP-Seq data. *BMC Bioinformatics*, **11**, 81.

Zhang,Y. *et al.* (2009) Intrinsic histone-DNA interactions are not the major determinant of nucleosome positions in vivo. *Nat. Struct. Mol. Biol.*, **16**, 847–852.