* + - 1. overview of chose methods

|  |  |  |  |
| --- | --- | --- | --- |
| Method | Descriptions | Command | Source file |
| GReEn | Copy model and Arithmetic coding | GReEnC -v -o  GReEnD -v -o | ftp://ftp.ieeta.pt/~ap/codecs/GReEn1.tar.gz |
| GDC | LZ-like, Delta encoding and Huffman coding | gdc c -ma1000000000 -mp20,4,20 -rn1 -hs8 -rm0 (compression for H. sapiens)  gdc c -ma1000000000 -rn1 (compression for rice and E.coli)  gdc d | http://sun.aei.polsl.pl/gdc |
| DELIMINATE | Delta encoding and 7z | delim a  delim e | http://metagenomics.atc.tcs.com/compression/DELIMINATE/ |
| gzip | [Huffman coding](http://en.wikipedia.org/wiki/Huffman_coding) and [LZ77](http://en.wikipedia.org/wiki/LZ77_and_LZ78)  Version 1.4 | gzip -6 default compression mode | http://www.gzip.org/ |
| bzip2 | BWT, [move-to-front transform](http://en.wikipedia.org/wiki/Move-to-front_transform) and [Huffman](http://en.wikipedia.org/wiki/Huffman_coding) coding.Version 1.0.6 | bzip2 -9 default compression mode | http://www.bzip.org/ |
| xz | LZMA2 algorithm  Version 5.1.0alpha | xz -6 default compression mode | http://tukaani.org/xz/ |
| pbzip2 | Parallel BZIP2 uses libbzip2.Version v1.1.8 | Default mode bzip2 -9 with 4 processors | http://compression.ca/pbzip2/ |
| mgzip | a multi-processor capable .gz file creator.Version 1.2c | Default mode gzip -6 and 4 worker threads | https://github.com/jerodsanto/mgzip |
| pigz | parallel implementation of gzip.Version 2.3.1 | Default mode gzip -8 and 4 worker threads | [http://www.zlib.net/pigz](http://www.zlib.net/pigz/pigz-2.3.1.tar.gz) |

|  |  |  |
| --- | --- | --- |
| **Method** | **Descriptions** | **Source file** |
| **GReEn** | Copy model and Arithmetic coding | ftp://ftp.ieeta.pt/~ap/codecs/GReEn1.tar.gz |
| **GDC** | LZ-like, Delta encoding and Huffman coding | http://sun.aei.polsl.pl/gdc |
| **DELIMINATE** | Delta encoding and 7z | http://metagenomics.atc.tcs.com/compression/DELIMINATE/ |
| **gzip** | [Huffman coding](http://en.wikipedia.org/wiki/Huffman_coding) and [LZ77](http://en.wikipedia.org/wiki/LZ77_and_LZ78)  Version 1.4 | http://www.gzip.org/ |
| **bzip2** | BWT, [move-to-front transform](http://en.wikipedia.org/wiki/Move-to-front_transform) and [Huffman](http://en.wikipedia.org/wiki/Huffman_coding) coding.Version 1.0.6 | http://www.bzip.org/ |
| **xz** | LZMA2 algorithm  Version 5.1.0alpha | http://tukaani.org/xz/ |
| **pbzip2** | Parallel BZIP2 uses libbzip2.Version v1.1.8 | http://compression.ca/pbzip2/ |
| **mgzip** | a multi-processor capable .gz file creator.Version 1.2c | https://github.com/jerodsanto/mgzip |
| **pigz** | parallel implementation of gzip.Version 2.3.1 | [http://www.zlib.net/pigz](http://www.zlib.net/pigz/pigz-2.3.1.tar.gz) |

* + - 1. overview of the testing datasets

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Target Genome | Number | File size (Byte) | Reference | Data source |
| Escherichia\_coli | 59 | 299917564 | [MG1655](http://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?mode=Info&id=511145&lvl=3&lin=f&keep=1&srchmode=1&unlock) | NCBI |
| Oryza sativa | 11 | 4165271984 | Build 4.0 | NCBI and Plant Biology Michigan State university |
| Homo sapiens | 11 | 32993445836 | GRCh37.p13 | NCBI and Korean Bioinformation Center |

[~~1] J. Shendure and H. Ji~~, "Next-generation DNA sequencing," Nature biotechnology, vol. 26, pp. 1135-1145, 2008.

[~~2] D. C.~~ Koboldt, K. M. Steinberg, D. E. Larson, R. K. Wilson, and E. R. Mardis, "The next-generation sequencing revolution and its impact on genomics," Cell, vol. 155, pp. 27-38, 2013.

[~~3] E. R. Mardis, "A decade/'s perspective on DNA sequencing technology," Nature, vol. 470, pp. 198-203, 2011~~.

[~~4] H. L. Rehm, "Disease-targeted sequencing: a cornerstone in the clinic," Nature Reviews Genetics, vol. 14, pp. 295-300, 2013.~~

[~~5] S. Grumbach and F. Tahi, "Compression of DNA sequences," 1994~~.

[~~6] R. Giancarlo, D. Scaturro, and F. Utro, "Textual data compression in computational biology: a synopsis," Bioinformatics, vol. 25, pp. 1575-1586, 2009~~.

[~~7] R. Giancarlo, D. Scaturro, and F. Utro, "Textual data compression in computational biology: Algorithmic techniques," Computer Science Review, vol. 6, pp. 1-25, 2012~~.

[~~8] X. Chen, S. Kwong, and M. Li, "A compression algorithm for DNA sequences," IEEE engineering in medicine and Biology, vol. 20, pp. 61-66, 2001~~.

[~~9] T. Matsumoto, K. Sadakane, and H. Imai, "Biological sequence compression algorithms," GENOME INFORMATICS SERIES, pp. 43-52, 2000~~.

~~[10] X. Chen, M. Li, B. Ma, and J. Tromp, "DNACompress: fast and effective DNA sequence compression," Bioinfor~~matics, vol. 18, pp. 1696-1698, 2002.

[~~11] G. Korodi and I. Tabus, "An efficient normalized maximum likelihood algorithm for DNA sequence compression," ACM Transactions on Information Systems (TOIS), vol. 23, pp. 3-34, 2005~~.

[~~12] M. D. Cao, T. I. Dix, L. Allison, and C. Mears, "A simple statistical algorithm for biological sequence compression," in Data Compression Conference, 2007. DCC'07, 2007, pp. 43-52~~.

[~~13] S. Christley, Y. Lu, C. Li, and X. Xie, "Human genomes as email attachments," Bioinformatics, vol. 25, pp. 274-275, 2008.~~

[~~14] M. C. Brandon, D. C. Wallace, and P. Baldi, "Data structures and compression algorithms for genomic sequence data," Bioinformatics, vol. 25, pp. 1731-8, Jul 15 2009~~.

[~~15] D. S. Pavlichin, T. Weissman, and G. Yona, "The human genome contracts again," Bioinformatics, vol. 29, pp. 2199-202, Sep 1 2013.~~

[~~16] S. Deorowicz, A. Danek, and S. Grabowski, "Genome compression: a novel approach for large collections," Bioinformatics, vol. 29, pp. 2572-8, Oct 15 2013~~.

[~~17] S. Kuruppu, S. J. Puglisi, and J. Zobel, "Relative Lempel-Ziv compression of genomes for large-scale storage and retrieval," in String Processing and Information Retrieval, 2010, pp. 201-206~~.

[~~18] S. Kuruppu, S. J. Puglisi, and J. Zobel, "Optimized relative Lempel-Ziv compression of genomes," in Proceedings of the Thirty-Fourth Australasian Computer Science Conference-Volume 113, 2011, pp. 91-98.~~

[~~19] G. Navarro and V. Mäkinen, "Compressed full-text indexes," ACM Computing Surveys (CSUR), vol. 39, p. 2, 2007~~.

[~~20] C~~. Wang and D. Zhang, "A novel compression tool for efficient storage of genome resequencing data," Nucleic Acids Research, vol. 39, pp. e45-e45, 2011.

[~~21] S. Deorowicz and S. Grabowski, "Robust relative compression of genomes with random access," Bioinformatics (Oxford, England), vol. 27, p. 2979, 2011~~.

~~[22] A. J.~~ Pinho, D. Pratas, and S. P. Garcia, "GReEn: a tool for efficient compression of genome resequencing data," Nucleic Acids Research, vol. 40, pp. e27-e27, 2011.

~~[23] B~~. G. Chern, I. Ochoa, A. Manolakos, A. No, K. Venkat, and T. Weissman, "Reference Based Genome Compression," IEEE Information Theory Workshop, 2012.

[~~24] S. Wandelt and U. Leser, "Adaptive efficient compression of genomes," Algorithms Mol Biol, vol. 7, p. 30, 2012~~.

[~~25] S. Kuruppu, B. Beresford-Smith, T. Conway, and J. Zobel, "Iterative Dictionary Construction for Compression of Large DNA Data Sets," IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 9, pp. 137-149, 2012~~.

[~~26] M. H. Mohammed, A. Dutta, T. Bose, S. Chadaram, and S. S. Mande, "DELIMINATE--a fast and efficient method for loss-less compression of genomic sequences: sequence analysis," Bioinformatics, vol. 28, pp. 2527-9, Oct 1 2012.~~

[~~27] J. J. Hunt, K.-P. Vo, and W. F. Tichy, "Delta algorithms: An empirical analysis," ACM Transactions on Software Engineering and Methodology (TOSEM), vol. 7, pp. 192-214, 1998~~.

[~~28] T. Bose, M. H. Mohammed, A. Dutta, and S. S. Mande, "BIND - an algorithm for loss-less compression of nucleotide sequence data," J Biosci, vol. 37, pp. 785-9, Sep 2012~~.

[~~29] M. Burrows and D. J. Wheeler, "A block-sorting lossless data compression algorithm," 1994~~.

[~~30] S. Mantaci, A. Restivo, G. Rosone, and M. Sciortino, "An extension of the Burrows Wheeler transform and applications to sequence comparison and data compression," in Combinatorial Pattern Matching, 2005, pp. 178-189~~.

[~~31] M. J. Bauer, A. J. Cox, and G. Rosone, "Lightweight algorithms for constructing and inverting the BWT of string collections," Theoretical Computer Science, vol. 483, pp. 134-148, 2013~~.

[~~32] A.~~ J. Cox, M. J. Bauer, T. Jakobi, and G. Rosone, "Large-scale compression of genomic sequence databases with the Burrows-Wheeler transform," Bioinformatics, vol. 28, pp. 1415-9, Jun 1 2012.

[~~33] K. Daily, P. Rigor, S. Christley, X. Xie, and P. Baldi, "Data structures and compression algorithms for high-throughput sequencing technologies," BMC Bioinformatics, vol. 11, p. 514, 2010~~.

[~~34] M. Hsi-Yang Fritz, R. Leinonen, G. Cochrane, and E. Birney, "Efficient storage of high throughput DNA sequencing data using reference-based compression," Genome Res, vol. 21, pp. 734-40, May 2011~~.

[~~35] D. C. Jones, W. L. Ruzzo, X. Peng, and M. G. Katze, "Compression of next-generation sequencing reads aided by highly efficient de novo assembly," Nucleic Acids Research, vol. 40, pp. e171-e171, 2012~~.

[~~36] R. Giancarlo, S. E. Rombo, and F. Utro, "Compressive biological sequence analysis and archival in the era of high-throughput sequencing technologies," Brief Bioinform, Dec 17 2013~~.

[~~37] Z. Zhu, Y. Zhang, Z. Ji, S. He, and X. Yang, "High-throughput DNA sequence data compression," Brief Bioinform, Dec 3 2013~~.