Table 3 Evaluation of *de novo* assembly result by EST/cDNA data

Dataset	Number	Covered by assembly	with>90% sequence in one scaffold		with>50% sequence in one scaffold	
			Number	Percent	Number	Percent
All	913,423	99.11%	814,237	89.14%	901,893	98.74%
>1 Kb	16,372	99.96%	14,336	87.56%	16,195	98.92%

A genome-scale analysis of the glycosylation and viral susceptibility in the CHO-K1 genome identifies homologs to 99% of the human genes, with 53% and 59% of them expressed respectively. We demonstrated that the expression and activities of these gene products were more important than their presence in the genome for determining the diversity of glycan structures on protein products and viral entry receptors in CHO.

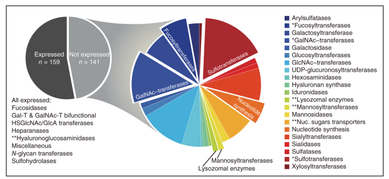


Fig. 1 A global view of the expression of CHO-K1 glycosylation genes. Glycosylation gene classes enriched in expressed genes were denoted with ** and significantly depleted classes in expressed genes were denoted with *.

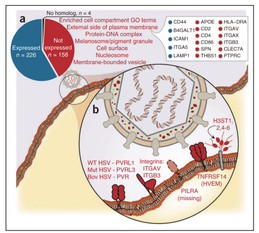


Fig. 2 An assessment of the expression state of viral susceptibility genes in CHO-K1. Blue for expressed and red for not expressed.

References

[1] Walsh G. Biopharmaceutical benchmarks 2010. ***Nat Biotechnol***. 2010, 28(9): 917-924.

[2] Derouazi Mea. Cell Technology for Cell Products. ***Springer Netherlands***. 2007: 443–446.

[3] Hormozdiari F, Alkan C, Eichler EE, *et al*. Combinatorial algorithms for structural variation detection in high-throughput sequenced genomes. ***Genome Res***. 2009, 19(7): 1270-1278.

[4] Li R, Zhu H, Ruan J, *et al*. *De novo* assembly of human genomes with massively parallel short read sequencing. ***Genome Res***. 2010, 20(2): 265-272.

[5] Li Y, Zheng H, Luo R, *et al*. Structural variation in two human genomes mapped at single-nucleotide resolution by whole genome *de novo* assembly. ***Nat Biotechnol***. 2011.

[6] Chen K, Wallis JW, McLellan MD, *et al*. BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. ***Nat Methods***. 2009, 6(9): 677-681.

[7] Ye K, Schulz MH, Long Q, *et al*. Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. ***Bioinformatics***. 2009, 25(21): 2865-2871.

[8] Xu X, Nagarajan H, Lewis NE, *et al*. The genomic sequence of the Chinese hamster ovary (CHO)-K1 cell line. ***Nat Biotechnol***. 2011, 29(8): 735-741.

Contact Us

China (Mainland)
BGI-Shenzhen, Beishan Industrial Zone, Yantian District, Shenzhen, 518083, China
Tel: 400-706-6615
+86-755-25273620
Email: tech@genomics.cn
www.genomics.cn

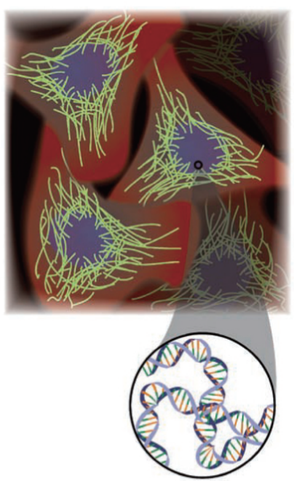
Asia Pacific（Hong Kong）
BGI Hong Kong Co. Limited, 16th Dai Fu Street, Tai Po Industrial Estate, Tai Po, Hong Kong
Tel: +852-3610-3510
Email: bgihk.enquiry@genomics.cn
www.bgisequence.com

North America (Boston)
BGI Americas Corporation, One Broadway, 14th Floor, Cambridge, MA 02142, USA
Tel: +1-617-500-2741
Email: info@bgiamericas.com
www.bgiamericas.com
www.bgisequence.com

Europe (Copenhagen)
Ole Maaløes Vej 3, DK-2200 Copenhagen N, Denmark
Tel: +45-7026 0806
Email: bgieurope@genomics.cn
www.bgisequence.com

华大基因 BGI | Premier Scientific Partner

De novo Sequencing of Cell Line



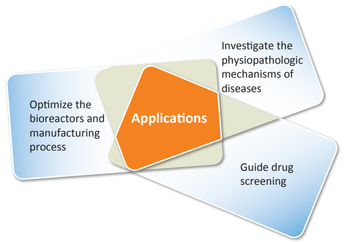
Overview

Cell lines have been widely used in scientific researches or productions relating to molecular mechanism of disease development, selection of biomarkers in drug R&D, bioengineering for producing certain proteins and vaccines, *etc*[1].

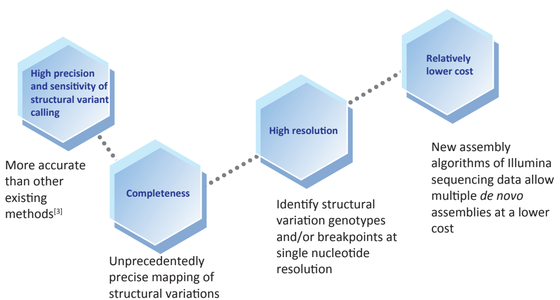
The genomes of cell lines derived from their ancestors may contain large-scale rearrangements that even clonal populations are known to diverge into heterogeneous subpopulations[2]. Therefore, the knowledge of genomic background of cell lines and corresponding clonal populations are quite important for cell line selections and perfections. However, existing re-sequencing-based methods for calling structural variations (SVs) from short sequencing reads have some limitations[3]: favor discovery of particular limited length or types of SVs, unable to identify SVs at single nucleotide resolution, and unable to improve the low accuracy and validation rate of SV identification.

Thus, BGI has successfully launched *de novo* sequencing of cell line service which can conquer the existing limitations of re-sequencing, and accelerate innovation and development in biological and disease research.

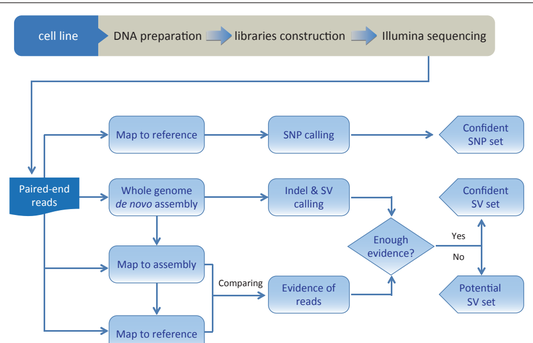
Applications



Advantages



Workflow



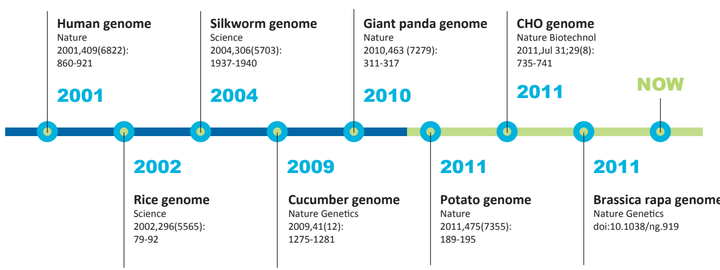
Guarantee

BGI self-developed and widely-applied assembly software SOAPdenovo[4] guarantees the quality of genome assembly by *de novo* whole genome sequencing (the sequencing depth is more than 100×):

Contig N50 length ≥ 20 Kb	Scaffold N50 length ≥ 300 Kb	Coverage of euchromatin area ≥ 95%	Coverage of gene region ≥ 98%	Single base assembly error rate ≤ 10^{-5}

BGI Experiences

BGI has abundant experiences and acknowledged accomplishments in *de novo* sequencing.



BGI Demo cases

1. Structural variation in two human genomes mapped at single-nucleotide resolution by whole genome *de novo* assembly[5]

BGI utilized whole-genome *de novo* assembly of short sequencing reads to map structural variation (SV) in an Asian genome and an African genome, which was published in Nature Biotechnology. Our approach identified small- and intermediate-size structural variants (1–50 kb).Table 1 shows our method had similar sensitivity but an improved precision compared with other existing structural variation detection tools.

Table 1 Comparison of structural variation detection tools between *de novo* assembly-based method, BreakDancer[6] and PIndel[7]

	SV detection tools		
	De novo assembly	BreakDancer	PIndel
Main detectable length rangea	1 bp-50 kbp	>10 bp	1-10 kbp (deletions) 1-16 bp (insertions)
Detectable SV types			
Insertions	Yes	Yes	Yes
Deletions	Yes	Yes	No
Inversions	Yes	Yes	No
Complex	Yes	Yes	No
Precision of breakpoints	Single base	A short ambiguous range	Single base
Genotypes of SV events	Yes	No	Yes
Simulated data			
False-positive rate	1.20%	9.1-10.3%	<2%
False-negative rate	9.60%	26-32%	~20%
Inaccessible missing rateb	24.10%		
Experimental validation			
False-positive rate	2.60%	11-22%	Not evaluated

2.The genomic sequence of the Chinese hamster ovary (CHO) K1 cell line-the first cell line genome obtained from *de novo* sequencing[8]

BGI has successfully presented a draft genomic sequence and comprehensive annotation of the CHO-K1 ancestral cell line. The assembly comprises 2.45 Gb of genomic sequence, with 24,383 predicted genes (Table 2 and 3). Furthermore, we integrated the RNA-Seq data to investigate relevant genes and explained some mechanisms not only involved in glycosylation, which affect therapeutic protein quality, but also viral susceptibility, which is relevant to cell engineering and regulatory concerns.

Table 2 *De novo* assembly result

	Contig Size (bp)	Scaffold Size (bp)	Number of Scaffods
N90	5,118	75,346	3,663
N80	12,695	254,361	1,921
N70	20,335	482,028	1,224
N60	28,784	782,420	831
N50	38,289	1,115,615	567
Total Size	2,367,185,801	2,447,154,408	
Total Number (>2Kb)			14,122

*N50: The contig or scaffold such that 50% of the *de novo* assembled genome lies in blocks of this size or larger. N60 to N90 are also used.