Biotechnology Journal
www.biotechnology-journal.com

Biotechnol. J. 2016, 11

DOI 10.1002/biot.201500455

Biotech Method

Augmenting Chinese hamster genome assembly by identifying regions of high confidence

Nandita Vishwanathan[1], Arpan A. Bandyopadhyay[1], Hsu-Yuan Fu[1], Mohit Sharma[2], Kathryn C. Johnson[1], Joann Mudge[3], Thiruvarangan Ramaraj[3], Getiria Onsongo[4], Kevin A. T. Silverstein[4], Nitya M. Jacob[1], Huong Le[1], George Karypis[2] and Wei-Shou Hu[1]

[1] Department of Chemical Engineering and Materials Science, University of Minnesota, Minneapolis, MN, USA
[2] Department of Computer Science & Engineering, University of Minnesota, Minneapolis, MN, USA
[3] National Center for Genome Resources (NCGR), Santa Fe, New Mexico, USA
[4] Minnesota Supercomputing Institute (MSI), University of Minnesota, Minneapolis, MN, USA

Chinese hamster Ovary (CHO) cell lines are the dominant industrial workhorses for therapeutic recombinant protein production. The availability of genome sequence of Chinese hamster and CHO cells will spur further genome and RNA sequencing of producing cell lines. However, the mammalian genomes assembled using shot-gun sequencing data still contain regions of uncertain quality due to assembly errors. Identifying high confidence regions in the assembled genome will facilitate its use for cell engineering and genome engineering. We assembled two independent drafts of Chinese hamster genome by de novo assembly from shotgun sequencing reads and by re-scaffolding and gap-filling the draft genome from NCBI for improved scaffold lengths and gap fractions. We then used the two independent assemblies to identify high confidence regions using two different approaches. First, the two independent assemblies were compared at the sequence level to identify their consensus regions as "high confidence regions" which accounts for at least 78% of the assembled genome. Further, a genome wide comparison of the Chinese hamster scaffolds with mouse chromosomes revealed scaffolds with large blocks of collinearity, which were also compiled as high-quality scaffolds. Genome scale collinearity was complemented with expressed sequence tags (EST) based synteny which also revealed conserved gene order compared to mouse. As cell line sequencing becomes more commonly practiced, the approaches reported here are useful for assessing the quality of assemblies and potentially facilitate the engineering of cell lines.

Received	20 JUL 2015
Revised	08 JUN 2016
Accepted	14 JUN 2016
Accepted article online	04 JUL 2016

Supporting information available online

Keywords: Annotation · CHO cells · NextGen sequencing · Scaffolds · Synteny

Correspondence: Prof. Wei-Shou Hu, Department of Chemical Engineering and Materials Science, University of Minnesota, 421 Washington Avenue SE, Minneapolis, MN 55455-0132, USA
E-mail: acre@umn.edu

Current adresses: Nandita Vishwanathan, Takeda Pharmaceutical Co. Ltd., 300 Massachusetts Avenue, Cambridge, MA 02139, USA
Huong Le, Amgen Inc., 1 Amgen Center Drive, Thousand Oaks, CA 91320, USA
Nitya M. Jacob, Amgen, One Kendall Square, 360 Binney St, Cambridge, MA 02141, USA
Kathryn C. Johnson, KBI Biopharma, Inc., 1101 Hamlin Rd., Durham, NC 27704, USA

Abbreviations: CHO, Chinese hamster Ovary; **EST**, expressed sequence tags; **LIR**, long insert reads; **SIR**, short insert read

1 Introduction

Chinese hamster ovary (CHO) cell lines have become the most prominent host cell system for recombinant protein production, producing over 60% of all recombinant protein therapeutics produced today [1, 2]. CHO cells are also among the most important cell lines used in biomedical research for decades (reviewed in [3]). Their small number of chromosomes and relative ease of deriving mutant lines made these cell lines a valuable cytogenetic tool [4]. Given the importance of CHO cells in biomedical research and in the biopharmaceutical industry, the genome will serve as a valuable resource to study the extent of genomic rearrangements and consequent transcriptome patterns,

which will further help to understand the underlying genomic foundation that allowed CHO cells to gain such prominence.

A draft assembly of the CHO-K1 genome was available in 2011 [5]. In view of the inherent nature of aneuploidy and continuing chromosomal rearrangement in CHO cells (reviewed in [6]) the Chinese hamster genome was sequenced later [7, 8]. These draft genome sequences can be used as reference genomes for sequencing efforts on CHO cell lines or other cells derived from Chinese hamster. Most of such cell line sequencing efforts will employ shot-gun sequencing using NextGen technologies. Some sequencing efforts may cover only a relatively shallow depth and align the sequencing reads to the reference genome for the purpose of discerning local variability among a number of cell lines. Others may cover a greater sequencing depth and conduct de novo assembly of the genome. The output of a de novo assembly is a set of contiguous sequence fragments referred to as "contigs". Contigs are then ordered and oriented into "scaffolds". These scaffolds have gaps between contigs, representing the regions where the sequence is uncertain but its length can be estimated. Despite the continuing progress on sequencing and assembly, detecting assembly errors, filling in gaps within the assembly and evaluating the contiguity of a large genome like Chinese hamster, still remains a challenge and the quality of an assembly is often difficult to assess [9, 10].

Herein, we report a strategy of identifying high confidence regions in a genome assembly of Chinese hamster. We first de-novo assembled the sequencing reads to yield a draft genome. Another independent draft assembly was generated by re-scaffolding and gap-filling the reference genome assembly AMDS00000000.1 [8] using the same set of sequencing reads to generate another draft genome. The two drafts were then compared to each other to generate the consensus regions. These consensus regions, having been derived from two independent assemblies, are considered as 'high confidence' regions. Comparisons of the gap-filled genome to the mouse genome yielded regions of strong orthology at both sequence and gene order levels. These syntenic regions also account for high confidence regions due to consensus in the large-scale order of these genomic segments across organisms. These two methods are complementary as the former focuses on consensus in smaller scales (sequence level), whereas the later confirms large-scale order (genome/gene level). Both methods enable the confirmation of the high confidence regions of the assembled genome. The identification of such 'high confidence' sequences are especially useful while using large genomes assembled from high throughput sequencing data, since they are prone to sequencing and assembly errors.

2 Materials and methods

2.1 DNA library preparation and genome sequencing of Chinese hamster

DNA was extracted from the liver of a single, highly inbred, female Chinese hamster of the 17A/GY strain (Cytogen Hamsters, West Roxbury, MA). DNA was isolated using the DNeasy Blood and Tissue Kit (Qiagen, Valencia, CA) using standard manufacturer recommended protocols.

The extracted DNA was fragmented by nebulization and ligated with Illumina sequencing adaptors to facilitate sequencing (Illumina, CA). For the short insert libraries, the DNA was size-selected using agarose gel electrophoresis. Fragments with insert sizes of 300, 400 and 500 bp were selected and gel-purified. The size distribution of the DNA fragments was verified in a 2100 Bioanalyzer using a DNA 7500 chip (Agilent, Santa Clara, CA).

For the long insert libraries, the DNA was fragmented, end labelled with biotin, and then size selected for 3 kbp. The DNA fragments were circularized by ligation of either ends of the linear fragment. The DNA was then fragmented and then the portion of the DNA near the site of ligation was selected using streptavidin bound magnetic beads. Illumina sequencing adaptors were ligated to these DNA fragments and then size selected prior to sequencing on an Illumina GAIIx flow cell. DNA-PET was used to generate 10 kbp and 20 kbp long insert mate-pair libraries for use in scaffolding [11]. Summary of the read library repertoire is tabulated in Supporting information, Table S1. Illumina v3 chemistry was used and the base calling was done by the GA 1.4.2 pipeline using recommended standard parameters.

2.2 Data pre-processing and assembly

Prior to assembly, the raw sequencing reads were preprocessed to remove low quality sequences, homopolymers or short tandem repeats, resulting in the loss of 10% of the raw data. The remaining 247 Gbp of sequence were used for assembly. AbySS assembler [14] was used for assembling the Chinese hamster genome. A k-mer sweep was performed between 40 bp and 95 bp with increments of 5 bp (Supporting information, Fig. S1). The best individual assembly was found for a k-mer value of 75 bp. The resulting assembly contained 2.24 Gbp of sequence in ≈495 019 contigs, with an N50 length of 7.363 kbp.

Paired-end sequence data with varying long insert sizes of approximately 3, 10, and 20 kbp were used stepwise from the shortest (3 kbp) to the longest (20 kbp) insert size to link contigs from the assemblies generated using AbySS. Sspace scaffolder [12] was used to generate the final scaffolds which are a series of ordered contigs separated by stretches of N's, where the length of the stretch is an estimate of the distance separating the contigs on

Table 1. Draft genome statistics at different stages of genome assembly enhancement shown in figure 1

Parameter	UMN1.0	Enhancement of the AMDS00000000.1 Chinese hamster genome			
		AMDS00000000.1	(I)	(II)	UMN2.0
Contigs	495 019	176 068	46 025	46 025	34 077
Max. contig (bp)	94 031	219 443	804 483	804 483	1 172 502
Mean contig (bp)	4 535	12 980	50 189	50 189	67 925
Contig N50 (bp)	7 363	27 317	127 489	127 489	185 327
Contig N90 (bp)	2 153	7 453	34 453	34 453	50 199
Total contig length (Gbp)	2.244745772	2.285446913	2.309963092	2.309963092	2.314664150
Assembly GC (%)	41.54	41.39	41.42	41.42	41.43
Scaffolds	31 821	10 868	10 867	7 876	7 865
Max scaffold (bp)	13 559 257	8 324 132	8 320 503	60 518 007	60 517 459
Mean scaffold (bp)	79 139	215 691	215 652	298 330	298 735
Scaffold N50 (bp)	2 202 990	1 579 055	1 578 388	17 211 978	17 210 817
Scaffold N90 (bp)	315 650	415 923	415 119	3 499 887	3 499 883
Total scaffold length (Gbp)	2.518295354	2.344125543	2.343495360	2.349643860	2.349546887
Captured gaps	463 198	165 200	35 158	38 149	26 212
Max. gap (bp)	29 555	20 486	20 068	98 050	97 673
Mean gap (bp)	591	355	954	1 040	1 331
Gap N50 (bp)	2 054	2 422	4 031	4 698	5 047
Total gap length (bp)	273 549 582	58 678 630	33 532 268	39 680 768	34 882 737

the genome. AbySS assembled scaffold set (UMN1.0) produced 31 821 scaffolds with a N50 of 2.2 Mbp (Table 1).

2.3 Enhancing genome assembly by gap-closing and rescaffolding

Short insert reads and long insert reads were used to close the gaps in AMDS00000000.1. The gap fraction was reduced from 2.5 to 1.4% (Stage I) (Table 1) using GapCloser [13]. This gap-closed draft genome sequence was then re-scaffolded using long insert reads using the AbySS-scaffolder [14] (Stage II). AbySS standalone scaffolder was used and the default mapping algorithm was replaced with AbySS-BWA to improve scaffolding. The scaffolding step increased the number of gaps in the assembly. Hence, another round of gap closing was done using another set of independent short insert reads, to yield UMN2.0.

2.4 Identification of syntenic regions in comparison with mouse genome

NUCmer was used identify the regions of homology between the Chinese hamster genome and mouse (GRCm38/mm10) with the following parameters: minimum cluster length of 30, minimum length of a maximal exact match of 10 and maximum gap between two adjacent matches in a cluster of 1000. The output from NUCmer was converted to SyntenyMiner format as outlined on SyntenyMiner's user guide, and further processed by SyntenyMiner (http://syntenyminer.sourceforge.net/) to visualize and compare the alignments between the UMN2.0 scaffolds and mouse chromosomes. Default parameters were used for visualization where the longest common subsequence between two sequences (UMN2.0 scaffold and mouse chromosome) is displayed. Each colored line represents an alignment between the two sequences.

3 Results

3.1 De novo assembly of draft Chinese hamster genome UMN1.0

A total of 273 Gbp of paired-end sequencing reads (sequenced from both ends of each fragment) from libraries of short insert sizes (the distance between the two reads as determined from the size distribution) of 300, 400 and 500 bp as well as long insert sizes of approximately 3, 10, and 20 kbp were used in the assembly (listed in Supporting information, Table S1). Based on the genome size of 2.66 Gbp as estimated from the haploid DNA content [15], the coverage was about 100-fold. AbySS, a parallel short read assembler was used to obtain the de novo assembled contigs using the short insert read (SIR) sequence data. The long insert reads (LIR) were then used to scaffold the assembled contigs using the Sspace scaffolder [12]. The assembly pipeline is provided in Fig. 1A. The resulting UMN1.0 draft genome comprised of 2.52 Gbp of DNA sequence assembled into 31 821 scaffolds after applying a minimum cut-off of 1 kbp on the scaffold length. The scaffold N50 of the resulting draft assembly was 2.2 Mbp. The assembled scaffolds cover approximately 95% of the estimated genome size of 2.66 Gbp. Without any other anchoring information such

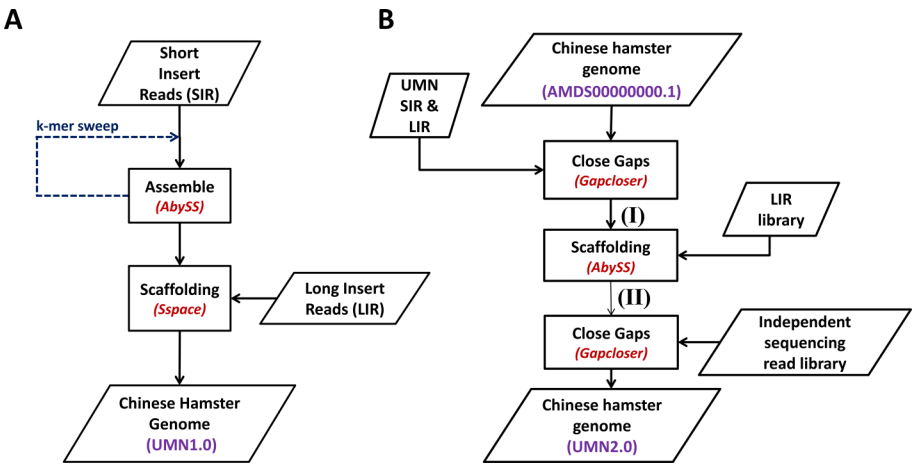

Figure 1. Workflow for genome assembly. (**A**) Workflow for de novo assembly of the Chinese hamster genome using short insert reads (SIR) and long insert reads (LIR) yeilding UMN1.0. (**B**) Workflow for enhancingAMDS00000000.1 to yield UMN2.0. The same set of SIR and LIR used to close gaps in the AMDS00000000.1 Chinese hamster genome draft assembly (I) followed by a scaffolding using LIR (II) and further gap filling using independent short insert reads.

as genetic maps, BAC maps or radiation hybrid maps, the number of scaffolds in the Chinese hamster genome at 32 590 is higher than those in the initial drafts of the mouse and rat genomes [16, 17].

3.2 Improved contiguity in draft Chinese hamster genome by re-scaffolding and gap-filling (UMN2.0)

The same set of short insert reads and long insert reads used in assembly of UMN1.0 were used to fill in the gaps and re-scaffold the AMDS00000000.1 Chinese hamster genome to generate another assembly called UMN2.0 [8] (Fig. 1B). This draft assembly comprised of 7865 scaffolds after applying a minimum cut-off of 1 kbp on the scaffold length. The scaffold N50 of the resulting draft assembly was 17.2 Mbp and the assembled scaffolds cover approximately 87% of the estimated genome size of 2.66 Gbp. In terms of contiguity and total gap length, UMN2.0, is a significant improvement over UMN1.0 and AMDS00000000.1 (Table 1).

3.3 Quality assessment of the Chinese hamster draft genome

3.3.1 High confidence sequence identification from consensus between draft assemblies

In the de novo assembly of a large genome like that of Chinese hamster's, assembly errors are unavoidable. We compared the two independently assembled drafts UMN1.0 and UMN2.0 and identified the regions of consensus between the two drafts (with a stringent criterion of contiguous segments larger than 1 kbp with sequence identity greater than 97% matching to unique reference and query sequence). These regions are likely to be of 'high confidence' and of higher quality. It must also be noted that the 'high confidence region' call is dependent on the stringency of the criterion used for identifying consensus regions and that the regions that do not fall into the 'high confidence regions' are not necessarily of low quality. Using the stringent criteria that we set forth, the consensus regions of one-to-one unique mapping cover about 1.83 Gbp accounting for 77.8% of the UMN2.0 genome (data for 20 largest scaffolds is presented in the Supporting information; the complete data set available at http://dx.doi.org/10.13020/D6Z304). Relaxing the criterion to also include non-unique hits with the same sequence identity criterion, greater than 97%, the consensus regions covered 2.14 Gbp accounting for 91.1% of the UMN2.0 genome.

A similar comparison was done between largest 20 scaffolds of the AMDS00000000.1 draft assembly with UMN1.0 to mark consensus regions on the former draft assembly using the aforementioned criteria. 82% of the total scaffold length consists of consensus regions with the UMN1.0 assembly (Supporting information, Table S2). The bed file of the consensus region is included in the Supporting information. RefSeq transcripts (GCF_000419365.1) were mapped to AMDS00000000.1 and the percentage of transcript length covered in the 'high confidence regions' were computed. Out of 29 764 transcripts with length greater than 500 bp, 25 412 transcripts had at least 70% of their length covered in the high consensus regions (Supporting information, Fig. S2A). This analysis gives an indication that most gene containing regions are assembled with good accuracy.

3.3.2 Quality assessment using sequence homology with mouse

Another assurance of the quality of the assembled genome is the high degree of conservation with the closely related mouse genome. UMN2.0 and mouse (GRCm38/mm10) were compared by whole genome alignments using NUCmer [18]. Given the evolutionary distance between mouse and Chinese hamster, a relatively relaxed seeding parameter set was used instead of the NUCmer default parameter set. Alignment results showed 82% gross genome sequence identity between Chinese hamster and mouse.

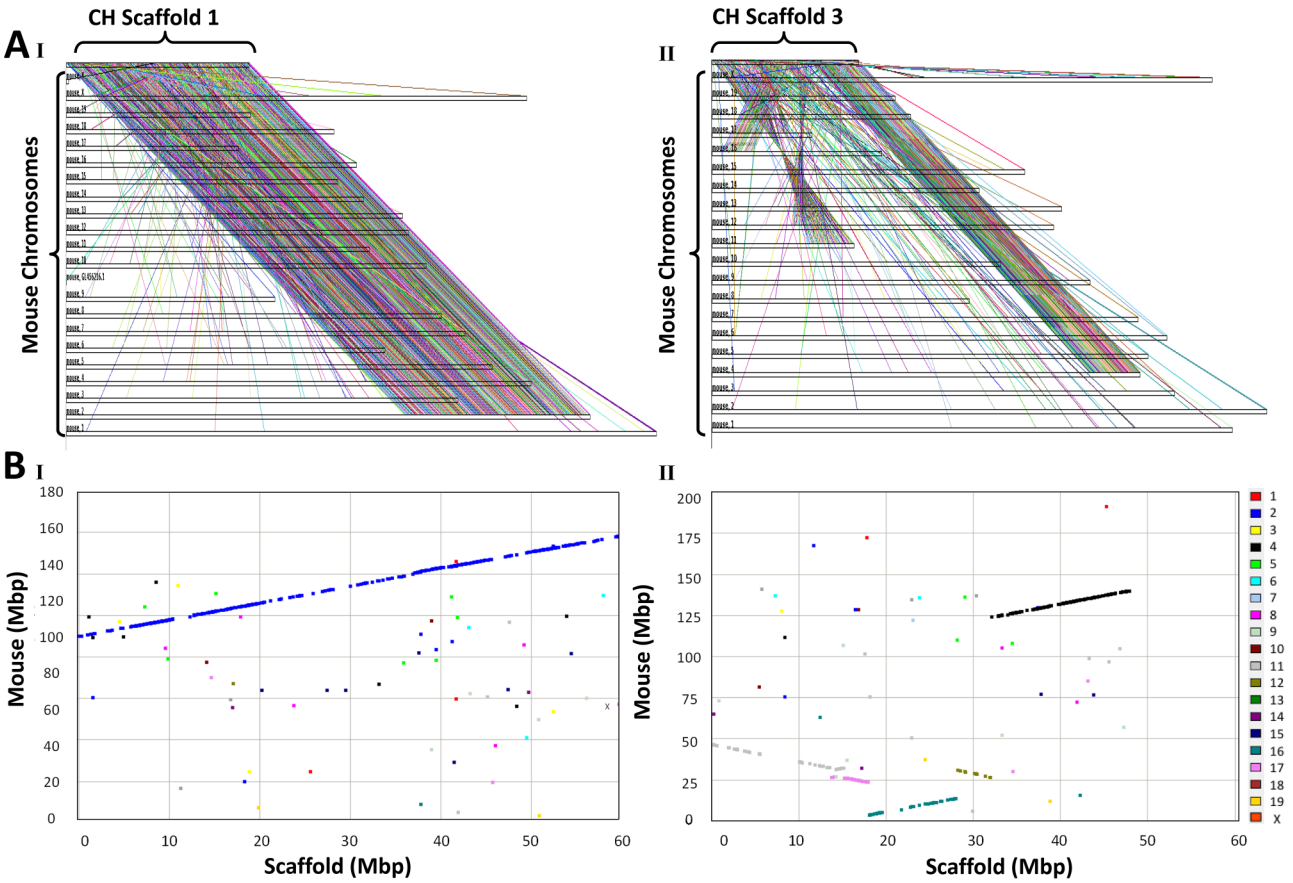

Figure 2. Conservation of collinearity/synteny between hamster and mouse genomes as a method for quality assessment. (**A**) Collinearity based on complete genome alignment with mouse genome. The figure on the left shows the largest scaffold in UMN2.0, CH Scaffold 1 shows a large syntenic block mapping to mouse chromosome 2 (I); and on the right, the third largest scaffold CH Scaffold 3 shows five syntenic blocks mapping to mouse chromosomes 4, 11, 12, 16 and 17 (II). (**B**) Synteny based on gene position (EST alignment) with mouse genome. Similar to (**A**), on the left, the largest scaffold in UMN2.0, CH Scaffold 1 shows a large syntenic block mapping to mouse chromosome 2, consistent with the genome scale synteny (I). The figure on the right shows the third largest scaffold CH Scaffold 3 with five syntenic blocks mapping to mouse chromosomes 4, 11, 12, 16 and 17 (II).

In order to verify the quality of the assembly, the twenty largest scaffolds in UMN2.0 covering 32% of the entire genome length were examined for large-scale order in orthologous regions with mouse using the NUCmer output. Orthologous regions showing preservation of order across the two genomes are highly likely to have been assembled correctly. However, not all correctly assembled regions may have preserved order in orthologous regions due to genomic rearrangements during evolution. Fig. 2A shows the collinear blocks in the first and third largest scaffold to mouse chromosome. Similar plots for the other 18 scaffolds are shown in Supporting information, Fig. S3. Large collinear blocks mapping to mouse chromosomes were seen in all twenty scaffolds. The orientation of two adjacent collinear blocks may be reversed in some cases. Preservation of collinearity of orthologous regions within the genome assemblies of Chinese hamster and mouse gives an additional degree of confidence in the quality of UMN2.0, thus providing yet another method for the affirmation of genome quality.

3.3.3 *Quality assessment using gene order between UMN2.0 and mouse*

Synteny, the order in which genetic loci are organized within the genomes of two species, is a measure of their closeness in having a common ancestry. Given that the mouse genome is the best annotated closest genome available to Chinese hamster, the conservation of syntenic blocks can be used to assess the quality of the assembled Chinese hamster genome. A region in UMN2.0 that is highly consistent in the order of its genetic loci in mouse is also highly likely to have been assembled correctly. In order to perform such a comparison, the Chinese hamster genome (UMN2.0) was annotated using the annotated expressed sequence tags (EST) contigs as described below.

A repertoire of annotated EST contigs described previously in [19], were mapped to UMN2.0 using the spliced

alignment software BLAT [17]. Almost 90% of the contigs (315 432) were aligned to the draft genome. Out of a total of 39 179 mouse genes in the Ensembl database, 24 652 were covered. On mapping the EST contigs greater than 500 bp, 89 690 out of 110 600 EST contigs had at least 70% of their length covered within the high consensus regions (Supporting information, Fig. S2B).

To verify the quality of the scaffolds, the loci of CHO EST contigs annotated with mouse orthologs were used to compare the gene order on the UMN2.0 Chinese hamster scaffold to that on the mouse genome (GRCm38/mm10). In the case that an EST has multiple hits (alignments) to the scaffolds, the hit with the highest confidence (based on the strength of the hit) was selected. Those with lower than 500 bp in the alignment length were discarded. A total of 18 055 mouse orthologs were aligned to 920 scaffolds and the scaffolds that had more than 40 genes were assessed for the conservation of gene order. The position of the gene on the mouse chromosome was plotted against the position in the Chinese hamster scaffolds. Two examples are shown in Fig. 2B where good synteny in terms of large syntenic blocks, can be observed between Chinese hamster and mouse.

In a highly syntenous region, adjacent genes in the Chinese hamster scaffold can also be seen as adjacent genes in the mouse genome. The close synteny is suggestive of high quality scaffolding (Fig. 2B, I). In some regions, genes on a Chinese hamster scaffold map to multiple mouse genome loci, sometimes on different mouse chromosomes (Fig. 2B, II). This is not necessarily an indication of incorrect scaffolding, as there could be many genomic rearrangements or duplication events between Chinese hamster and mouse, considering the evolutionary distance between them. Out of the 147 scaffolds examined, a total of 122 scaffolds showed good synteny conservation, similar to the graph on the left of Fig. 2B. The synteny with mouse adds an additional affirmation of high confidence level of the genome assembly.

4 Discussion

The advances in DNA sequencing technology and the availability of the Chinese hamster genomic sequence are posed to usher cell bioprocessing into a post-genomic era [20, 21]. We describe a strategy for the integration of shotgun sequencing data from Chinese hamster into an existing draft genome to enhance scaffold contiguity by successive scaffolding and gap-filling. Since errors are unavoidable in assemblies derived from shotgun sequencing, the consensus regions of two independent assemblies are identified as high confidence regions. Highly syntenous regions with mouse genome provide additional confidence on the quality of the assembly.

Increasingly genome sequencing is performed on industrial CHO cell lines. The pipeline presented in this study will facilitate attaining an enhanced genome draft and identifying regions of high confidence. For example, regions of consensus were obtained using stringent criteria to compare the CHO-K1 (AFTD00000000.1) [5] and Chinese hamster genome (AMDS00000000.1) [8], 1.72 Gbp (71.9%) of CHO-K1 genome can be identified as consensus regions (data for the 20 largest scaffold in Supporting information. Complete data set is available at http://dx.doi.org/10.13020/D6Z304). When the regions of non-unique hits are included, the regions of consensus increased to 2.16 Gbp (90.1%). The approach reported here will enhance the utility of the genome sequence of cell lines and potentially facilitate cell and genome engineering in CHO cell lines.

The computational resources were provided by the Minnesota Supercomputing Institute. This work was supported by the Consortium for CHO Systems Biotechnology. Huong Le was a recipient of Vietnam Educational Foundation Fellowship.

The authors declare no conflict of interest.

5 References

[1] Bandyopadhyay, A., Fu, H.-Y., Vishwanathan, N., Hu, W.-S., Genomics and systems biotechnology in biopharmaceutical processing. *Chem. Eng. Progr.* 2014, *100*, 45–50.

[2] Kantardjieff, A., Zhou, W., Mammalian cell cultures for biologics manufacturing. *Adv. Biochem. Eng./Biotechnol.* 2014, *139*, 1–9.

[3] Feeney, W. P., Chapter 35 – The Chinese or striped-back hamster, in: Suckow, M. A., Stevens, K. A., Wilson, R. P. (Eds.), *The Laboratory Rabbit, Guinea Pig, Hamster, and Other Rodents*, Academic Press, Boston 2012, pp. 907–922.

[4] Galloway, S. M., Armstrong, M. J., Reuben, C., Colman, S. et al., Chromosome aberrations and sister chromatid exchanges in Chinese hamster ovary cells: Evaluations of 108 chemicals. *Environ. Mol. Mutagen.* 1987, *10 Suppl. 10*, 1–175.

[5] Xu, X., Nagarajan, H., Lewis, N. E., Pan, S. et al., The genomic sequence of the Chinese hamster ovary (CHO)-K1 cell line. *Nat. Biotechnol.* 2011, *29*, 735–741.

[6] Wurm, F., CHO Quasispecies—Implications for manufacturing processes. *Processes* 2013, *1*, 296–311.

[7] Brinkrolf, K., Rupp, O., Laux, H., Kollin, F. et al., Chinese hamster genome sequenced from sorted chromosomes. *Nat. Biotechnol.* 2013, *31*, 694–695.

[8] Lewis, N. E., Liu, X., Li, Y., Nagarajan, H. et al., Genomic landscapes of Chinese hamster ovary cell lines as revealed by the *Cricetulus griseus* draft genome. *Nat. Biotechnol.* 2013, *31*, 759–765.

[9] Baker, S., Joecker, A., Church, G., Snyder, M. et al., Genome interpretation and assembly-recent progress and next steps. *Nat. Biotechnol.* 2012, *30*, 1081–1083.

[10] Denton, J. F., Lugo-Martinez, J., Tucker, A. E., Schrider, D. R. et al., Extensive error in the number of genes inferred from draft genome assemblies. *PLoS Comput. Biol.* 2014, *10*, e1003998.

[11] Fullwood, M. J., Wei, C. L., Liu, E. T., Ruan, Y., Next-generation DNA sequencing of paired-end tags (PET) for transcriptome and genome analyses. *Genome Res.* 2009, *19*, 521–532.

[12] Boetzer, M., Henkel, C. V., Jansen, H. J., Butler, D., Pirovano, W., Scaffolding pre-assembled contigs using SSPACE. *Bioinformatics* 2011, *27*, 578–579.
[13] Tsai, I. J., Otto, T. D., Berriman, M., Improving draft assemblies by iterative mapping and assembly of short reads to eliminate gaps. *Genome Biol.* 2010, *11*, R41.
[14] Simpson, J. T., Wong, K., Jackman, S. D., Schein, J. E. et al., ABySS: A parallel assembler for short read sequence data. *Genome Res.* 2009, *19*, 1117–1123.
[15] Greilhuber, J., Volleth, M., Loidl, J., Genome size of man and animals relative to the plant *Allium cepa*. *Can. J. Genet. Cytol.* 1983, *25*, 554–560.
[16] Gibbs, R. A., Weinstock, G. M., Metzker, M. L., Muzny, D. M. et al., Genome sequence of the Brown Norway rat yields insights into mammalian evolution. *Nature* 2004, *428*, 493–521.
[17] Consortium, M. G. S., Waterston, R. H., Lindblad-Toh, K., Birney, E. et al., Initial sequencing and comparative analysis of the mouse genome. *Nature* 2002, *420*, 520–562.
[18] Kurtz S, Phillippy A, Delcher AL, et al. Versatile and open software for comparing large genomes. *Genome Biology*. 2004;5(2):R12. doi:10.1186/gb-2004-5-2-r12.
[19] Vishwanathan, N., Yongky, A., Johnson, K. C., Fu, H. Y. et al., Global insights into the Chinese hamster and CHO cell transcriptomes. *Biotechnol. Bioeng.* 2014, *112*, 965–976.
[20] Kremkow, B. G., Baik, J. Y., MacDonald, M. L., Lee, K. H., CHOgenome.org 2.0: Genome resources and website updates. *Biotechnol. J.* 2015, *10*, 931–938.
[21] Monger, C., Kelly, P. S., Gallagher, C., Clynes, M. et al., Towards next generation CHO cell biology: Bioinformatics methods for RNA-Seq-based expression profiling. *Biotechnol. J.* 2015, *10*, 950–966.