frontiers in
GENETICS

MINI REVIEW ARTICLE
published: 22 July 2014
doi: 10.3389/fgene.2014.00243

Generation of physical map contig-specific sequences

Yanliang Jiang[1], Peng Xu[1] and Zhanjiang Liu[2]*

[1] Centre for Applied Aquatic Genomics, Chinese Academy of Fishery Sciences, Beijing, China

[2] Aquatic Genomics Unit, The Fish Molecular Genetics and Biotechnology Laboratory, School of Fisheries, Aquaculture and Aquatic Sciences, and Program of Cell and Molecular Biosciences, Auburn University, AL, USA

Edited by:
Yih-Horng Shiao, United States Patent Trademark Office, USA

Reviewed by:
Meredith Emily Kennedy Calvert, Temasek Lifesciences Laboratory, Singapore
Robin Ketteler, University College London, UK

****Correspondence:***
Zhanjiang Liu, Aquatic Genomics Unit, The Fish Molecular Genetics and Biotechnology Laboratory, School of Fisheries, Aquaculture and Aquatic Sciences, and Program of Cell and Molecular Biosciences, Auburn University, Auburn, AL 36849, USA
e-mail: liuzhan@auburn.edu

Rapid advances of the next-generation sequencing technologies have allowed whole genome sequencing of many species. However, with the current sequencing technologies, the whole genome sequence assemblies often fall in short in one of the four quality measurements: accuracy, contiguity, connectivity, and completeness. In particular, small-sized contigs and scaffolds limit the applicability of whole genome sequences for genetic analysis. To enhance the quality of whole genome sequence assemblies, particularly the scaffolding capabilities, additional genomic resources are required. Among these, sequences derived from known physical locations offer great powers for scaffolding. In this mini-review, we will describe the principles, procedures and applications of physical-map-derived sequences, with the focus on physical map contig-specific sequences.

Keywords: physical map contig-specific sequences, BAC end sequences, whole genome sequencing, assembly, scaffolding

INTRODUCTION

Whole genome sequencing is the most robust approach to unraveling the genetic information of an organism. During the past several years, sequencing costs have declined drastically through the use of next-generation sequencing technologies, which include a suite of sequencing platforms, such as Illumina sequencing, SOLiD sequencing, and PacBio sequencing, among quite a few others. As a result, more and more species with biological or economic importance are added to the list of organisms whose whole genomes have been sequenced.

However, each of the next-generation sequencing platforms has distinctive shortcomings. For instance, Illumina sequencing and SOLiD sequencing generate accurate but short tags that are difficult to be assembled (Nagarajan et al., 2010; Luo et al., 2012). PacBio sequencing generates long sequences, but its error rate is relatively high. These intrinsic shortcomings affect the assembly qualities of the whole genome sequences, which in turn limit the applicability of these whole genome sequences for genetic analysis.

The assembly qualities of whole genome sequences are measured by a number of parameters including the following: (1) accuracy; (2) completeness; (3) contiguity; and (4) connectivity. Accuracy refers to the correctness of the sequences. It is an important metric as the miscalling of bases can cause substantial trouble for downstream operations, such as the identification of single nucleotide polymorphisms (SNPs). Sequencing accuracy is primarily intrinsic to the sequencing technology but can also be affected by the quality and quantity of the template DNA. In this regard, the Illumina and SOLiD sequencing platforms provide fairly high-quality sequences, while the calling accuracy of the PacBio and Roche 454 sequencing platforms are relatively low. Completeness refers to the percentage of the total bases of the genome that are represented in the assembly of the whole genome sequences. Completeness is important because analysis involving the genes in missing regions will be difficult. Contiguity refers to the lengths of contiguous sequences. Continuous sequences allow full-length gene sequences, including regulatory sequences, to be obtained from the genome sequences for subsequent analysis. Short contigs pose greater challenges for the assembly of the genome into scaffolds, particularly with regards to correct order and orientation. Connectivity refers to the extent to which contigs are properly linked together and reflect their original genomic locations, sequential order, and orientation. For genetic analysis, connectivity is the most important metric. For instance, association analysis has the capability of revealing the significant SNPs associated with a specific trait. If the significant SNPs are located on genome sequences that are well connected at the chromosomal scale, Manhattan plots can be constructed to determine the distribution of significant SNPs along the chromosome(s). The probabilities of the involved significant SNPs can be examined to determine the location of the most significant SNP and how the linkage disequilibrium decays around that specific SNP, thereby determining the number and location of quantitative trait locus (QTLs) involved with the trait. In contrast, if the genome assembly is highly segmented, many significant SNPs remain as isolated contigs or scaffolds, and it will be difficult to determine the number and the location of the QTLs. Therefore, there is limited use for highly segmented genome assemblies (Sierro et al., 2013).

In addition to the intrinsic characteristics of each sequencing technology, the DNA templates used for sequencing can also add additional complexities. Heterozygous diploid organisms with two sets of similar chromosomes pose a challenge for assembly, because it is difficult to distinguish allelic sequences from paralogous loci with high similarities (Hahn et al., 2014). The largest challenge of whole genome sequence assembly most likely comes from the presence of a large number of repetitive elements. Short tandem repeats over 100-bp long often cause a termination of the sequencing reaction, while longer, interspersed repeated sequences prevent short sequence tags from being assembled into long contigs. Repetitive sequences, such as transposons, in the genome shatter *de novo* assembly, because the sequencing reads are usually not long enough to span the entire series of repetitive sequences plus any unique flanking sequences (Jiang et al., 2013). Such challenges are more significant when dealing with species with complex genomes, such as teleost fish, which go through one or two additional rounds of whole genome duplication (Meyer and Van de Peer, 2005; Steinke et al., 2006; Moghadam et al., 2011; Xu et al., 2011b). Assembly is also particularly problematic for species with large genomes. For example, the Norway spruce has a genome size of 20 Gb, and only 25% of its genome is assembled into scaffolds longer than 10 Kb (Nystedt et al., 2013).

Several approaches are available for providing scaffolding capabilities. These include the generation of mate-paired reads from variable lengths of inserts (Boetzer et al., 2011; Gao et al., 2011; Gritsenko et al., 2012; Williams et al., 2012; Hunt et al., 2014; Kajitani et al., 2014; Zimin et al., 2014) or using transcript sequences (Mortazavi et al., 2010). Mate-paired reads can be generated from Illumina sequencing using libraries of various sizes, by using Fosmid libraries (Williams et al., 2012) or bacterial artificial chromosome (BAC) libraries (Xu et al., 2007; Liu et al., 2009). Although extremely efficient, the use of paired reads alone normally cannot reduce the number of scaffolds down to several thousand, as can be done with physical maps. Therefore, we have taken advantage of the available catfish BAC-based physical maps (Xu et al., 2007) and developed a method for generating BAC-based physical map contig-specific sequences (Jiang et al., 2013). Such physical map contig-specific sequences offer the capability to associate all the related genome sequence contigs/scaffolds belonging to a single physical map contig together, effectively reducing the overall number of scaffolds of the genome sequences. Here we will describe the principles, procedures and applications of physical map-derived sequences.

BAC-BASED PHYSICAL MAPS

A BAC-based physical map consists of contigs of overlapping BAC clone DNA fragments. An acceptable BAC-based physical map usually consists of several thousand contigs. Any gaps can be attributed to missing segments of the genome or to highly competitive regions that cannot be properly assigned to specific contigs. Therefore, physical maps organize the entire genome into several thousand contigs.

Early efforts in whole genome sequencing primarily relied on BAC clones selected from physical maps using a minimal tiling path (MTP, Mahairas et al., 1999; Siegel et al., 1999), and as such, the MTP can be selected through a graph-theoretical approach (Bozdag et al., 2013). Such a sequencing strategy has been referred to as the "clone-by-clone" whole genome sequencing strategy. With this approach, BAC clones selected from the physical map using an MTP are sequenced using random shotgun sequencing and assembly (Lander et al., 2001). The clone-by-clone sequencing strategy reduces the complexity of sequencing and assembly from the genome scale to a BAC clone, thus making it easier to assemble the genome. Such a whole genome sequencing strategy, which utilizes a BAC-based physical map, has been widely used in eukaryotes, such as human (Lander et al., 2001), mouse (Waterston et al., 2002), chicken (International Chicken Genome Sequencing Consortium [ICGSC], 2004), zebrafish (Howe et al., 2013), medaka (Kasahara et al., 2007), *Tetraodon* (Jaillon et al., 2004), *Arabidopsis* (Arabidopsis Genome Initiative [AGI], 2000), and rice (International Rice Genome Sequencing Project [IRGSP], 2005), among many others. However, it is very expensive and labor-intensive, especially for non-model species.

The availability of next-generation sequencing technologies has led to greater efforts in the development of software packages for the assembly of whole genome sequences. However, bioinformatic approaches alone cannot resolve the problems of repetitive sequences, especially with large genomes. As a result, large numbers of contigs have been assembled (reflecting a lower quality) for the whole genome sequences of many species. Further enhancement of the whole genome sequence assemblies is needed to make such assemblies useful. Many scientists have considered coupling traditional approaches with contemporary bioinformatic approaches. As such, physical maps are still crucially useful resources to improve genome assembly, especially for large and complex genomes. For instance, to achieve the assembly of the large barley genome (5.1 Gb), a new strategy was developed to include the construction of a sequence-enriched barley physical map (Mayer et al., 2012). Another important role of physical maps in whole genome sequencing is to orient the assembled contigs/scaffolds. In a pilot study of salmon genome sequencing, a 1-Mb genomic region was sequenced using GS FLX shotgun and long paired-end sequencing, resulting in 175 contigs assembled into four scaffolds, which were then verified and oriented by using a BAC-based physical map and BAC end sequences (BES; Quinn et al., 2008). In another genome sequencing pilot study using catfish, a physical map and BES were used to confirm and order the assembled genome contigs (Jiang et al., 2011). Lewin et al. (2009) concluded that physical maps are indispensable for the precision of genome assemblies, after comparing the quality of the genome assemblies with and without the use of physical maps. Finally, physical maps are essential for assessing the quality of whole genome sequence assemblies (Li et al., 2009; Zhang et al., 2012; Xu et al., 2013; Kim et al., 2014).

BAC END SEQUENCES

Bacterial artificial chromosome end sequences are genomic survey sequences using BAC clones as templates with sequencing primers from the BAC vector. They are important genome resources, and the most useful BESs are mate-paired reads. As such, BESs have been generated from a large number of species (Budiman et al., 2000; Yuan et al., 2000; Zhao et al., 2001; Larkin et al., 2003; Ren et al., 2003; Messing et al., 2004; Xu et al., 2006, 2011a; Liu et al., 2009).

The use of BESs in whole genome sequencing projects was first proposed as a tool for the identification of MTPs (Goff et al., 2002;

Yu et al., 2002). With next-generation sequencing, BESs remain helpful in the assembly and scaffolding process, in particular, for complex and repeat-rich genomes (Feuillet et al., 2011). This is because BESs are paired-reads from large inserts that span a distance of over 100–200 Kb. For instance, the average insert size in the catfish BAC library is 161 Kb (Wang et al., 2007). Mate-paired BESs can be used to combine assembled genome scaffolds into superscaffolds (Quinn et al., 2008; Jiang et al., 2011). Moreover, BESs associated with BAC clones allow them to be related to a physical map, thereby integrating genome sequence contigs/scaffolds with physical maps.

However, a study on catfish demonstrated that BESs are not as powerful as expected when functioning as an anchoring point to link genome contigs to physical maps for two reasons (Jiang et al., 2013): first, BESs are relatively short (Xu et al., 2006; Liu et al., 2009); and second, the number of BESs is still limited because of the high cost associated with generating BESs, and even when all of the BAC clones are sequenced, only two end sequences can be generated per BAC clone. Therefore, additional sequences that are specific for the physical map contigs are needed to enhance the anchoring ability of BAC-associated sequences.

PHYSICAL MAP CONTIG-SPECIFIC SEQUENCES

Although BESs from physical maps can be used as sequence tags to anchor assembled genome sequence contigs to the BAC contigs of physical maps, they account for only 0.5–1% of all genome sequences. We have developed a simple strategy for the rapid generation of extensive sequence tags from the distinct BAC contigs of physical maps, to allow the vast majority of assembled genome contigs to be anchored to physical map contigs, at a relatively low cost (Jiang et al., 2013).

The core principle of physical map contig-specific sequences is to generate next-generation sequences with known tags specific for each of the BAC contigs in a physical map. Briefly, the strategy for generating physical map contig-specific sequences includes six major steps (**Figure 1A**): (1) select and cultivate the BAC clones from each physical map contig using MTP; (2) extract the BAC DNA, and pool the DNA representing the MTP of each BAC contig from the physical map; (3) digest the DNA by using two 4-bp restriction endonucleases with different recognition sites but compatible overhangs; (4) individually ligate the specific barcoded adaptors to the fragments generated from each BAC contig from the physical map; (5) amplify the specific barcoded fragments

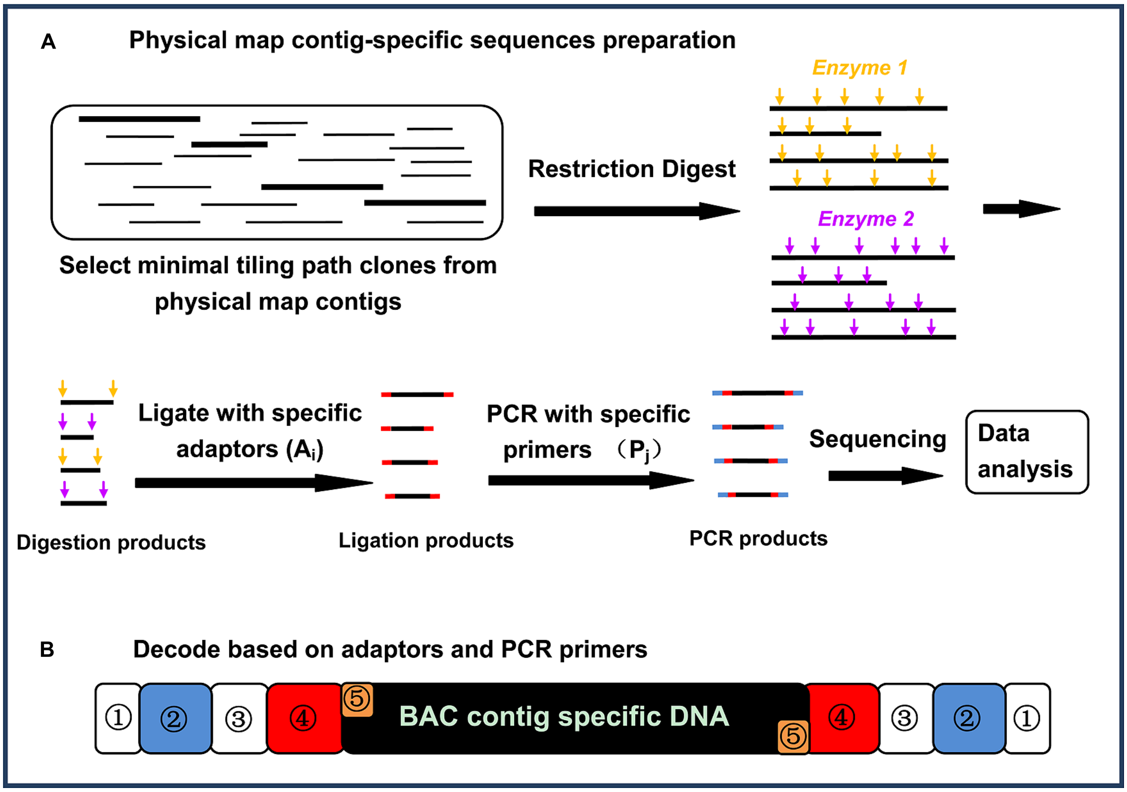
A
Physical map contig-specific sequences preparation
Select minimal tiling path clones from physical map contigs
Restriction Digest
Enzyme 1
Enzyme 2
Ligate with specific adaptors (A_i)
PCR with specific primers （P_j）
Sequencing
Data analysis
Digestion products
Ligation products
PCR products
B
Decode based on adaptors and PCR primers
①
②
③
④
⑤
BAC contig specific DNA
⑤
④
③
②
①

FIGURE 1 | (A) A flow chart illustrating the physical map contig-specific fragment preparation. The BAC clones from each physical map contig were selected, using a minimal tilling path, and then the BAC DNA from each physical map contig were pooled for digestion with two restriction endonucleases. In-house designed adaptors were then ligated to the digestion products, followed by amplification using in-house designed primers. All of the PCR products were pooled together and sequenced (Jiang et al., 2013). **(B)** Decoding based on the adaptors and primers. Each PCR product was attached to a physical map contig-specific tag at both ends. The tag representing the physical map contig ID contains five parts: ① represents the shields DNA, ② represents the specific barcoded sequencing in the primer (Pj), ③ represents the common/complementary sequences between the adaptor and the PCR primer, ④ represents the specific barcoded sequences in the adaptor (Ai), and ⑤ represents the overhangs proximal to the fragments for ligation (Jiang et al., 2013).

via PCR using barcoded PCR primers for the fragments generated from each BAC contig from the physical map; and (6) sequence the PCR-amplified fragments via next-generation sequencing. After sequencing, the sequences can be decoded based on their barcodes to assign them to specific BAC contigs on the physical map.

The highlighted advantages for such a strategy to generate physical map contig-specific sequences are its simplicity and low-cost. A BAC-based physical map normally consists of several thousand BAC contigs, which means that thousands of specific barcodes are required to differentiate each physical map contig-specific sequence. To reduce the total number of barcodes, a two-dimensional tagging strategy was designed, in which there are two separate sets of barcodes; one is attached to the adaptors being ligated to the restriction enzyme digested fragments, and the other is attached to the PCR primer for the amplification of the fragments.

This approach is highly efficient. For instance, we have generated a large number of catfish physical map contig-specific sequences (Jiang et al., 2013) with limited financial resources. The catfish physical map contains 1,824 contigs. If only a single barcode is used, 1,824 specific tags are required. When adopting the two-dimensional tagging strategy, all 1,824 pooled BAC DNA were arrayed into a two-dimensional 38 (row) × 48 (column) setup, using twenty 96-well plates, in which the rows represent one set of tags for adaptors A_i, where $i = 1, 2, 3, \ldots 38$, and the columns represent another set of tags, P_j, where $j = 1, 2, 3, \ldots 48$. As such, each pool of PCR products represents the fragments derived from a single physical map contig with A_i and P_j at the ends. In this way, only 86 (38 + 48) barcodes are needed, but their combination (38 × 48) can generate 1,824 distinct barcodes. As shown in **Figure 1B**, each end of the amplified fragments attached to the specific barcodes consists of five parts: a common sequence that acts as the "shield" to keep the barcodes intact, the specific barcoded sequence in the primer (P_j), the common/complementary sequences between the adaptor and PCR primer, the specific barcoded sequences in the adaptor (A_i), and the overhangs proximal to the fragments to be ligated to the restriction fragments.

One of the most important applications of physical map contig-specific sequences is to associate whole genome sequence contigs into scaffolds. The sequence assemblies obtained from each BAC contig in the physical map can be used to search the contigs in the whole genome sequence using BLAST. Upon receiving hits for two or more of the contigs in the whole genome sequence by one contig in the physical map contig-specific sequences, they are brought together into one contig, thereby reducing the number of contigs in the whole genome sequence. When only one contig from the whole genome sequence is hit, it reveals that the whole genome sequence contig is associated with a specific

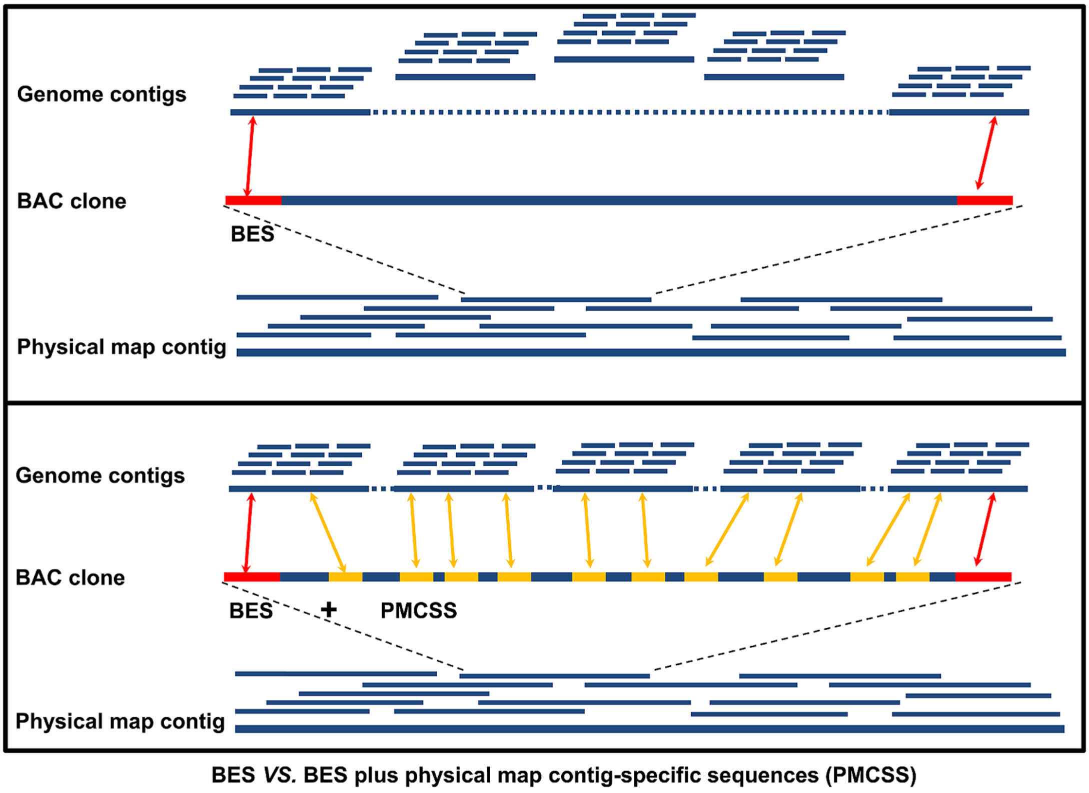

FIGURE 2 | Comparison of the anchoring power of only BESs versus BESs plus physical map contig-specific sequences (PMCSSs). The red bar on the BAC clone represents the BESs, while the yellow bar represents the PMCSSs.

physical map contig. Therefore, the likelihood of "scaffolding" the contigs in the whole genome sequence is increased. For instance, in our study using catfish physical map contig-specific sequences, we compared the power of anchoring the genome contigs by using BESs alone with using both BESs and physical map contig-specific sequences (Jiang et al., 2013; **Figure 2**). With the previously available BES alone, 27,770 whole genome sequence contigs (11% of the whole genome contigs, channel catfish assembly version 1.0, unpublished) had significant hits to the BESs. When the physical map contig-specific sequences were also used, the number of whole genome contigs with significant hits increased to 156,457. In terms of the total length of the genome contigs being scaffolded, over 79% of the assembled whole genome sequences were anchored when using both BESs and the physical map contig-specific sequences, but only 26% of the assembled whole genome sequences were anchored when only BESs were used. To further assess the scaffolding capacity of the physical map contig-specific sequences, we also determined the number of genes that could be anchored to the scaffolds of the whole genome sequences. The number of genes drastically increased from 6,732 when only BESs were used to 16,680 when both BESs and the physical map contig-specific sequences were used (Jiang et al., 2013). All of these results demonstrated the strong anchoring capability of the physical map contig-specific sequences. However, the order and orientation of the whole genome sequence contigs within the physical map contig is still largely unknown, unless the gaps can be filled by physical map contig-specific sequences.

CONCLUSION

Next-generation sequencing technologies have provided unprecedented possibilities for genome sequencing. However, challenges remain in generating well-assembled reference genomes due to the short reads produced via the next-generation sequencing platforms and to the complexities of large eukaryotic genomes with high levels of repetitive elements. For genetic analysis, the anchoring of whole genome sequence contigs and scaffolds to chromosomes is perhaps the most important goal. Among the many different approaches for anchoring whole genome sequences to chromosomes, BES and physical map contig-specific sequences provide great power for linking whole genome shotgun sequence contigs to physical maps, thereby significantly reducing the workload when using genetic linkage mapping to anchor whole genome sequence contigs to chromosomes through the integration of genetic linkage and physical maps. The generation of physical map contig-specific sequences is both technologically simple and cost effective.

ACKNOWLEDGMENTS

This project was partially supported by the Agriculture and Food Research Initiative Competitive Grant no. 2010-65205-20356 and 2012-67015-19410 from the USDA National Institute of Food and Agriculture (NIFA), and by the National High-Technology Research and Development Program of China (863 program; 2011AA100401) and Special Scientific Research Funds for Central Non-profit Institutes, Chinese Academy of Fishery Sciences (2014C011 and 2013C010).

REFERENCES

Arabidopsis Genome Initiative [AGI]. (2000). Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* 408, 796–815. doi: 10.1038/35048692

Boetzer, M., Henkel, C. V., Jansen, H. J., Butler, D., and Pirovano, W. (2011). Scaffolding pre-assembled contigs using SSPACE. *Bioinformatics* 27, 578–579. doi: 10.1093/bioinformatics/btq683

Bozdag, S., Close, T. J., and Lonardi, S. (2013). A graph-theoretical approach to the selection of the minimum tiling path from a physical map. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 10, 352–360. doi: 10.1109/TCBB.2013.26

Budiman, M. A., Mao, L., Wood, T. C., and Wing, R. A. (2000). A deep-coverage tomato BAC library and prospects toward development of an STC framework for genome sequencing. *Genome Res.* 10, 129–136. doi: 10.1101/gr.10.1.129

Feuillet, C., Leach, J. E., Rogers, J., Schnable, P. S., and Eversole, K. (2011). Crop genome sequencing: lessons and rationales. *Trends Plant Sci.* 16, 77–88. doi: 10.1016/j.tplants.2010.10.005

Gao, S., Sung, W. K., and Nagarajan, N. (2011). Opera: reconstructing optimal genomic scaffolds with high-throughput paired-end sequences. *J. Comput. Biol.* 18, 1681–1691. doi: 10.1089/cmb.2011.0170

Goff, S. A., Ricke, D., Lan, T. H., Presting, G., Wang, R., Dunn, M., et al. (2002). A draft sequence of the rice genome (*Oryza sativa* L. ssp. *japonica*). *Science* 296, 92–100. doi: 10.1126/science.1068275

Gritsenko, A. A., Nijkamp, J. F., Reinders, M. J., and de Ridder, D. (2012). GRASS: a generic algorithm for scaffolding next-generation sequencing assemblies. *Bioinformatics* 28, 1429–1437. doi: 10.1093/bioinformatics/bts175

Hahn, M. W., Zhang, S. V., and Moyle, L. C. (2014). Sequencing, assembling, and correcting draft genomes using recombinant populations. *G3* (*Bethesda*) 4, 669–679. doi: 10.1534/g3.114.010264

Howe, K., Clark, M. D., Torroja, C. F., Torrance, J., Berthelot, C., Muffato, M., et al. (2013). The zebrafish reference genome sequence and its relationship to the human genome. *Nature* 496, 498–503. doi: 10.1038/nature12111

Hunt, M., Newbold, C., Berriman, M., and Otto, T. D. (2014). A comprehensive evaluation of assembly scaffolding tools. *Genome Biol.* 15:R42. doi: 10.1186/gb-2014-15-3-r42

International Chicken Genome Sequencing Consortium [ICGSC]. (2004). Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. *Nature* 432, 695–716. doi: 10.1038/nature03154

International Rice Genome Sequencing Project [IRGSP]. (2005). The map-based sequence of the rice genome. *Nature* 436, 793–800. doi: 10.1038/nature03895

Jaillon, O., Aury, J. M., Brunet, F., Petit, J. L., Stange-Thomann, N., Mauceli, E., et al. (2004). Genome duplication in the teleost fish *Tetraodon nigroviridis* reveals the early vertebrate proto-karyotype. *Nature* 431, 946–957. doi: 10.1038/nature03025

Jiang, Y., Lu, J., Peatman, E., Kucuktas, H., Liu, S., Wang, S., et al. (2011). A pilot study for channel catfish whole genome sequencing and *de novo* assembly. *BMC Genomics* 12:629. doi: 10.1186/1471-2164-12-629

Jiang, Y., Ninwichian, P., Liu, S., Zhang, J., Kucuktas, H., Sun, F., et al. (2013). Generation of physical map contig-specific sequences useful for whole genome sequence scaffolding. *PLoS ONE* 8:e78872. doi: 10.1371/journal.pone.0078872

Kajitani, R., Toshimoto, K., Noguchi, H., Toyoda, A., Ogura, Y., Okuno, M., et al. (2014). Efficient *de novo* assembly of highly heterozygous genomes from whole-genome shotgun short reads. *Genome Res.* doi: 10.1101/gr.170720.113 [Epub ahead of print].

Kasahara, M., Naruse, K., Sasaki, S., Nakatani, Y., Qu, W., Ahsan, B., et al. (2007). The medaka draft genome and insights into vertebrate genome evolution. *Nature* 447, 714–719. doi: 10.1038/nature05846

Kim, S., Park, M., Yeom, S. I., Kim, Y. M., Lee, J. M., Lee, H. A., et al. (2014). Genome sequence of the hot pepper provides insights into the evolution of pungency in Capsicum species. *Nat. Genet.* 46, 270–278. doi: 10.1038/ng.2877

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., et al. (2001). Initial sequencing and analysis of the human genome. *Nature* 409, 860–921. doi: 10.1038/35057062

Larkin, D. M., Everts-van der Wind, A., Rebeiz, M., Schweitzer, P. A., Bachman, S., Green, C., et al. (2003). A cattle-human comparative map built with cattle BAC-ends and human genome sequence. *Genome Res.* 13, 1966–1972. doi: 10.1101/gr.1560203

Lewin, H. A., Larkin, D. M., Pontius, J., and O'Brien, S. J. (2009). Every genome sequence needs a good map. *Genome Res.* 19, 1925–1928. doi: 10.1101/gr.094557.109

Li, R., Fan, W., Tian, G., Zhu, H., He, L., Cai, J., et al. (2009). The sequence and *de novo* assembly of the giant panda genome. *Nature* 463, 311–317. doi: 10.1038/nature08696

Liu, H., Jiang, Y., Wang, S., Ninwichian, P., Somridhivej, B., Xu, P., et al. (2009). Comparative analysis of catfish BAC end sequences with the zebrafish genome. *BMC Genomics* 10:592. doi: 10.1186/1471-2164-10-592

Luo, C., Tsementzi, D., Kyrpides, N., Read, T., and Konstantinidis, K. T. (2012). Direct comparisons of Illumina vs. Roche 454 sequencing technologies on the same microbial community DNA sample. *PLoS ONE* 7:e30087. doi: 10.1371/annotation/64ba358f-a483-46c2-b224-eaa5b9a33939

Mahairas, G. G., Wallace, J. C., Smith, K., Swartzell, S., Holzman, T., Keller, A., et al. (1999). Sequence-tagged connectors: a sequence approach to mapping and scanning the human genome. *Proc. Natl. Acad. Sci. U.S.A.* 96, 9739–9744. doi: 10.1073/pnas.96.17.9739

Mayer, K. F., Waugh, R., Brown, J. W., Schulman, A., Langridge, P., Platzer, M., et al. (2012). A physical, genetic and functional sequence assembly of the barley genome. *Nature* 491, 711–716. doi: 10.1038/nature11543

Messing, J., Bharti, A. K., Karlowski, W. M., Gundlach, H., Kim, H. R., Yu, Y., et al. (2004). Sequence composition and genome organization of maize. *Proc. Natl. Acad. Sci. U.S.A.* 101, 14349–14354. doi: 10.1073/pnas.0406163101

Meyer, A., and Van de Peer, Y. (2005). From 2R to 3R: evidence for a fish-specific genome duplication (FSGD). *Bioessays* 27, 937–945. doi: 10.1002/bies.20293

Moghadam, H. K., Ferguson, M. M., and Danzmann, R. G. (2011). Whole genome duplication: challenges and considerations associated with sequence orthology assignment in Salmoninae. *J. Fish Biol.* 79, 561–574. doi: 10.1111/j.1095-8649.2011.03030.x

Mortazavi, A., Schwarz, E. M., Williams, B., Schaeffer, L., Antoshechkin, I., Wold, B. J., et al. (2010). Scaffolding a *Caenorhabditis* nematode genome with RNA-seq. *Genome Res.* 20, 1740–1747. doi: 10.1101/gr.111021.110

Nagarajan, H., Butler, J. E., Klimes, A., Qiu, Y., Zengler, K., Ward, J., et al. (2010). *De Novo* assembly of the complete genome of an enhanced electricity-producing variant of *Geobacter* sulfurreducens using only short reads. *PLoS ONE* 5:e10922. doi: 10.1371/journal.pone.0010922

Nystedt, B., Street, N. R., Wetterbom, A., Zuccolo, A., Lin, Y. C., Scofield, D. G., et al. (2013). The Norway spruce genome sequence and conifer genome evolution. *Nature* 497, 579–584. doi: 10.1038/nature12211

Quinn, N. L., Levenkova, N., Chow, W., Bouffard, P., Boroevich, K. A., Knight, J. R., et al. (2008). Assessing the feasibility of GS FLX Pyrosequencing for sequencing the Atlantic salmon genome. *BMC Genomics* 9:404. doi: 10.1186/1471-2164-9-404

Ren, C., Lee, M. K., Yan, B., Ding, K., Cox, B., Romanov, M. N., et al. (2003). A BAC-based physical map of the chicken genome. *Genome Res.* 13, 2754–2758. doi: 10.1101/gr.1499303

Siegel, A. F., Trask, B., Roach, J. C., Mahairas, G. G., Hood, L., and van den Engh, G. (1999). Analysis of sequence-tagged-connector strategies for DNA sequencing. *Genome Res.* 9, 297–307. doi: 10.1101/gr.9.3.297

Sierro, N., Battey, J. N., Ouadi, S., Bovet, L., Goepfert, S., Bakaher, N., et al. (2013). Reference genomes and transcriptomes of *Nicotiana sylvestris* and *Nicotiana* tomentosiformis. *Genome Biol.* 14:R60. doi: 10.1186/gb-2013-14-6-r60

Steinke, D., Hoegg, S., Brinkmann, H., and Meyer, A. (2006). Three rounds (1R/2R/3R) of genome duplications and the evolution of the glycolytic pathway in vertebrates. *BMC Biol.* 4:16. doi: 10.1186/1741-7007-4-16

Wang, S., Xu, P., Thorsen, J., Zhu, B., de Jong, P. J., Waldbieser, G., et al. (2007). Characterization of a BAC library from channel catfish *Ictalurus punctatus*: indications of high levels of chromosomal reshuffling among teleost genomes. *Mar. Biotechnol. (NY)* 9, 701–711. doi: 10.1007/s10126-007-9021-5

Waterston, R. H., Lindblad-Toh, K., Birney, E., Rogers, J., Abril, J. F., Agarwal, P., et al. (2002). Initial sequencing and comparative analysis of the mouse genome. *Nature* 420, 520–562. doi: 10.1038/nature01262

Williams, L. J., Tabbaa, D. G., Li, N., Berlin, A. M., Shea, T. P., Maccallum, I., et al. (2012). Paired-end sequencing of Fosmid libraries by Illumina. *Genome Res.* 22, 2241–2249. doi: 10.1101/gr.138925.112

Xu, P., Li, J., Li, Y., Cui, R., Wang, J., Zhang, Y., et al. (2011a). Genomic insight into the common carp (*Cyprinus carpio*) genome by sequencing analysis of BAC-end sequences. *BMC Genomics* 12:188. doi: 10.1186/1471-2164-12-188

Xu, P., Wang, J., Cui, R., Li, Y., Zhao, Z., Ji, P., et al. (2011b). Generation of the first BAC-based physical map of the common carp genome. *BMC Genomics* 12:537. doi: 10.1186/1471-2164-12-537

Xu, P., Wang, S., Liu, L., Peatman, E., Somridhivej, B., Thimmapuram, J., et al. (2006). Channel catfish BAC-end sequences for marker development and assessment of syntenic conservation with other fish species. *Anim. Genet.* 37, 321–326. doi: 10.1111/j.1365-2052.2006.01453.x

Xu, P., Wang, S., Liu, L., Thorsen, J., Kucuktas, H., and Liu, Z. (2007). A BAC-based physical map of the channel catfish genome. *Genomics* 90, 380–388. doi: 10.1016/j.ygeno.2007.05.008

Xu, Q., Chen, L. L., Ruan, X., Chen, D., Zhu, A., Chen, C., et al. (2013). The draft genome of sweet orange (Citrus sinensis). *Nat. Genet.* 45, 59–66. doi: 10.1038/ng.2472

Yu, J., Hu, S., Wang, J., Wong, G. K., Li, S., Liu, B., et al. (2002). A draft sequence of the rice genome (*Oryza sativa* L. ssp. *indica*). *Science* 296, 79–92. doi: 10.1126/science.1068037

Yuan, Q., Liang, F., Hsiao, J., Zismann, V., Benito, M. I., Quackenbush, J., et al. (2000). Anchoring of rice BAC clones to the rice genetic map in silico. *Nucleic Acids Res.* 28, 3636–3641. doi: 10.1093/nar/28.18.3636

Zhang, G., Fang, X., Guo, X., Li, L., Luo, R., Xu, F., et al. (2012). The oyster genome reveals stress adaptation and complexity of shell formation. *Nature* 490, 49–54. doi: 10.1038/nature11413

Zhao, S., Shatsman, S., Ayodeji, B., Geer, K., Tsegaye, G., Krol, M., et al. (2001). Mouse BAC ends quality assessment and sequence analyses. *Genome Res.* 11, 1736–1745. doi: 10.1101/gr.179201

Zimin, A., Stevens, K. A., Crepeau, M. W., Holtz-Morris, A., Koriabine, M., Marcais, G., et al. (2014). Sequencing and assembly of the 22-gb loblolly *Pine* genome. *Genetics* 196, 875–890. doi: 10.1534/genetics.113.159715

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 17 May 2014; paper pending published: 05 June 2014; accepted: 07 July 2014; published online: 22 July 2014.

Citation: Jiang Y, Xu P and Liu Z (2014) Generation of physical map contig-specific sequences. Front. Genet. ***5****:243. doi: 10.3389/fgene.2014.00243*

This article was submitted to Genomic Assay Technology, a section of the journal Frontiers in Genetics.