Part 3. Breeding, genetics and reproduction

Sequencing of the mink genome: plans and perspectives

R. Anistoroaei[1], K. Christensen[1] and B. Benkel[2]
[1]University of Copenhagen, Faculty of Health Sciences, Dept. of Animal and Veterinary Basic Sciences, Division of Animal Genetics and Bioinformatics, Groennegaardsvej 3, 1870 Frederiksberg C,Denmark; [2]Nova Scotia Agricultural College, Department of Plant and Animal Sciences, P.O. Box 550,Truro, B2N 5E3 Nova Scotia, Canada; ran@life.ku.dk

Abstract

The assembled sequence of the whole genome is a must nowadays in any industrially important species. The aim of this research program is to provide the industry and the community with a robust mink genome sequence assembly. It shall be complemented with the identification of widely distributed polymorphic single nucleotide polymorphisms (SNPs) organized into an array designed for undertaking linkage and association studies necessary for mapping, identifying and characterizing genes underlying any specific traits of interest in American mink. Genome assembly based on the sequences of individual clones from the mink Bacterial Artificial Chromosome (BAC) library would result in a draft genome of high quality and accuracy. We explore an alternative approach, called BAC-NGS (from next generation sequencing) that utilizes the BAC library as a solid backbone, complemented with shotgun sequences of the whole genome, for filling gaps for the final assembly. We tested different settings of the new approach in several pilot projects in order to assign the most efficacious approach to sequence assembly. Approximately 15% of the mink genome has so far been generated and assembled by means of this method. This sequencing project represents a long-term investment in the future mink fur industry and farming management, enhancing the development of applications in academic and industrial research.

Keywords: American mink, genome sequencing project, SNPs, BAC-NGS

Introduction

Genomics is one of the corner stones of biology, opening new frontiers in product development and the improvement of animal production while sequence information is becoming essential for understanding the biology of any organism. Whole genome sequence assemblies (WGSA) are essential for understanding the genetics and biology of any organism, since they enable investigations of not only the gene sequences themselves, but also the role of surrounding noncoding DNA and the elucidation of complex/multigenic traits. They also serve as a tool in discovering genomic variation within the species (polymorphisms). Besides the high-quality human genome, assemblies are available for several complex eukaryotic species (e.g. dog-BUILD 3.1, cattle-BUILD 6.1, pig – Sscrofa 9.2; 10.2), but they are commonly known by their users as being incomplete and inaccurate, and in some crucial regions difficult to (re-)sequence. Despite advances in technologies that have accelerated the (re)sequencing complete genomes (Hillier *et al.,* 2008; Wheeler *et al.,* 2008), the need for high-quality assemblies remains (Lewin *et al.,* 2009). Current methods employed for *de novo* genome assemblies are: (1) Classical BAC to BAC method involving Sanger shotgun sequencing of individual BAC clones. This method is time

consuming and very expensive in relation to new sequencing standards. (2) Next Generation Sequencing methods (NGS) (e.g. 454 Roche, Illumina) generate large amounts of sequence data from a genome, avoiding the step of molecular cloning in a vector (e.g. BAC library production). However, genome assemblies generated from NGS have so far fallen short of those obtained with the classical approach. NGS methods are regarded as very effective when re-sequencing a genome for which there is an available draft assembly. Whereas data generated with NGS alone can be readily used for a wide range of research investigations, it has proven difficult to use them to generate high-quality *de novo* genome assemblies of vertebrate genomes, which are large and repeat-rich. Although algorithms have been developed for improving the assembling of *de novo* sequences derived from NGS (Gnerre *et al.,* 2010), without a solid, physically anchored scaffold, the accuracy of the assembly (particularly at segmental duplications or long transposable elements) is compromised. To our knowledge, BAC-based physical maps combined with extensive NGS of BACs have not been exploited for WGSA of any mammalian species.

Additionally, development of specific SNPs is very important for fine mapping of traits and for comparing regions of the genome between cohorts in genome wide association studies and investigations determining genetic diversity and phylogenetic relations. They are highly abundant and randomly distributed throughout the genome and because they are bi-allelic, they are easily assayed. SNP maps are used for identification of genes (aka QTLs) underlying complex traits – areas in which linkage maps have generally proven to be ineffective. SNP arrays and, in some cases, SNP chips (e.g. human – Affymetrix; mouse – Illumina; pig – Illumina) are available for most domesticated species (e.g. dog – Karlsson *et al.,* 2007; horse – Chowdhary and Raudsepp, 2008; cattle – Matukumalli *et al.,* 2009) but not for mink.

Fur animal breeding raises ethical and economic issues, both of which must be addressed at the same time. The most efficient and sustainable way to deal with these issues is by close analysis of information from genomics, proteomics, transcriptomics, non-coding DNA, etc. This information can be obtained by generating a complete, robust, accessible and manageable whole genome sequence assembly (WGSA) supplemented with a fine map of single nucleotide polymorphisms (SNPs) for the American mink (*Neovison vison*). These are the most useful tools researchers from all the biological fields investigating a species can have at their disposal.

This project has been designed to test the novel BAC-NGS hybrid method in order to facilitate the building of a cheap, sufficiently annotated and reliable WGSA for the American mink which ultimately shall be complemented with the identification of widely distributed SNP markers organized into an array (a high resolution genetic map).

Materials and methods

DNA

DNA extraction, purification, libraries preparation and sequencing of the BACs (from the CHORI-231) were performed in different settings both in our laboratories and by outsourcing parts of the pilot-projects. Details of these settings are proprietary. Paired-end libraries were prepared for the Illumina HiSeq2000 platform and reads of 100 n were analysed.

Biocomputing

Gradual *de novo* assembly for sequences of each BAC batch was performed using Velvet (Zerbino and Birney, 2008) and SOAPdenovo (Li *et al.,* 2010) software platforms. They were optimized with respect to processing speed and other settings throughout the development (such as the possibility of connecting contig nodes by the pair ends). GapFiller and PHRAP were employed for the assembly and gap filling at this level. Sets of obtained contigs were organized on a gapped scaffold along the genome. Given that the size of a BAC is 170 kb (Anistoroaei *et al.,* 2011), the number of BACs being sequenced from a specific pool was estimated based on their relative genomic position against the dog genome. Additionally, the existing ZooFISH-based mink/dog/human data (Graphodatsky *et al.,* 2000) and the existing genetic and physical maps (Anistoroaei *et al.,* 2009) facilitated the positioning of the individually assembled contig sequences along the mink chromosomes by linearity (Anistoroaei *et al.,* 2011), indicating breakpoints between mink and human resp. mink and dog.

Results and discussion

The novel strategy *BAC-NGS*, tested in several different pilot projects has enabled the assembly of approximately 15% of the mink genome at various locations as derived from the BAC clones sequencing and assembling. The difference in this BAC-NGS approach compared to NGS *de novo* whole genome sequencing strategies is that the sequencing and *de novo* assembling is performed in parts. By using the CHORI-231 BAC library (Anistoroaei *et al.,* 2011), each part will include a selected number of clones simultaneously NGS analysed at individual iterations, thereby facilitating the working with the data. Another substantial advantage is that the 'per batch' assembling is based on one strand of DNA only (BAC inserts are mono-stranded) thus avoiding the ambiguities given by the heterozygous strands of DNA in a stringent assembling process using genomic DNA as the starting point. Compared to the classic BAC-to-BAC methodology, the method tested here is much cheaper and faster, and the end product will be at least of equal quality. The derived 'per batch' contigs sequences were gradually assembled and integrated. Because the human genome is fully sequenced and comparatively well annotated, we focused on comparison between human and mink. Comparative studies were also performed using the dog genome sequences since is the species phylogenetically closest to mink, that has an assembled genome of an agreeable quality (BUILD 3.1).

The optimal hybrid approach to be utilized for further sequencing of the American mink genome is not yet readily available for publication but the schematic representation of the approach to be embraced is schematized in Figure 1.

Relevance and perspectives of the project

Method and reference

A cheap, resource efficient and at the same time high-quality method for mink WGSA can serve as a model for assemblies and employment in other species. In addition, existing BAC libraries available for several other species could be utilized for WGSA by the BAC-NGS method.

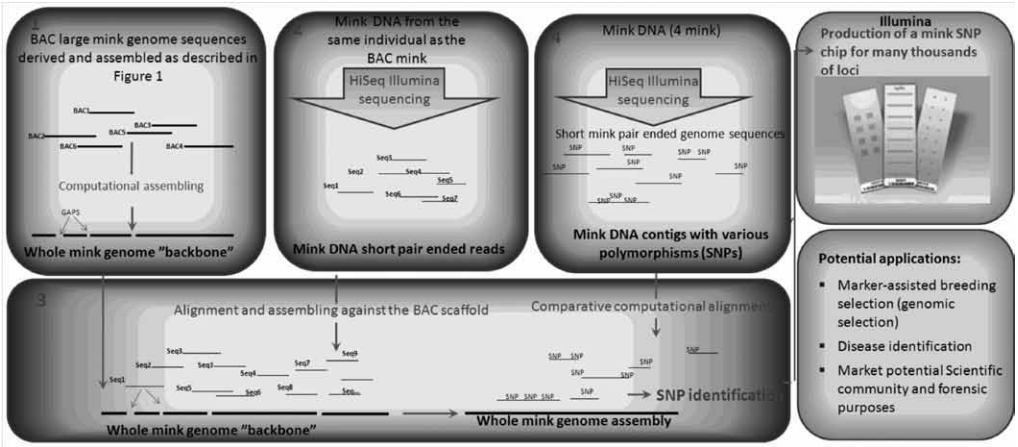

Figure 1. Representation of essential steps (dark frames) taken for producing a WGSA and an array of SNPs for the mink genome. Light frames represent the long term goal of such a project.

Fine association and mapping

An array of SNPs supported by the WGSA is of considerable importance for linkage and genome-wide association studies necessary for identifying genes responsible for mono/poly genic traits and/or diseases and for defining the molecular variation between them.

Characterization of genes

Characterization of genes of interest, their allelic variants and functional investigation will be approached by combined methods utilizing the newly generated genome sequences, which will provide access to the genes for the investigated traits.

Exploring and exploiting the uniqueness of the species

Being able to identify specific and unique genes of mink as well as their complexity would be of interest for biological and medical investigations (e.g. metabolism, immunity), as well as for mink breeding. The mink nomenclature for fur colors differs from that of other animals (Nes *et al.,* 1988), thus understanding its background, would represent an important contribution to explaining the genetics of color.

Serving the industry and animal well-being

WGSA is critically important for positional cloning and identification of the QTLs that were recently identified in the mink genome (Thirstrup *et al.,* 2012). Although positional cloning by interspecific extrapolation is in some cases possible without WGSA, the assembly will greatly accelerate the pace of important research discoveries. It will facilitate implementing sustainable breeding programs and genomic selection. In the farming conditions, traits such as aggressivity

have a direct link to the animal welfare and general stress. They are genetically determined and seem to show variation (Cagan – personal communication). Thus, profiling the specific genes/traits can enable selection of suitable animals that are more adapted. It can also aid ecological studies in protecting wild life. In regard to the investigation of genes of interest, combined strategies can be used: (1) exploitation of the previously established maps substantially supported by the newly generated SNP markers and array to perform fine mapping, association studies and genome scan in informative families to identify regions and/or genes involved in the development of phenotypic variation; (2) exploitation of functional and comparative knowledge about genes and (3) further investigations of the individual genes and their functional background utilizing the WGSA like in all the other eukaryotic large genome projects.

Additional resources

A solid WGSA (with functional annotation) of the mink will enable development of resources for genomic and proteomic research. For example, research on the biological course of Aleutian Disease Virus (ADV), the most vicious disease in mink, and the development of host immunity to infections will be aided greatly by the mink genome sequence information to be generated in this project. It will enable the development of molecular tools (e.g. expression arrays, proteins, antibodies) that are virtually non-existent at present, in an expeditious manner.

Surrogate systems for human experimentation

WGSA can provide investigational potential for disease models such as deafness (*unpublished data*) or Chédiak-Higashi syndrome (Anistoroaei *et al. in press*), which are prevalent on some commercial mink farms.

Forensics

Given the increasing number of robbery cases of furs from the farms, a SNP array (eventually supported on a SNP chip) would serve as an easy to use tool in forensic investigations establishing ownership and origin (Gill, 2001).

Expanding our understanding of evolutionary processes in general and of mink evolution in particular

The mink WGSA will enable the examination of differences in the structures of genes, diversity of gene families, and the accumulation and distribution of repetitive elements among genomes at several levels of evolutionary divergence, and to relate these differences to characteristics of gene expression, function, and epigenetic regulation.

To conclude, the long term prospect of a WGSA project is to support sustainable mink breeding programs leading to improved animal welfare and economy in mink farms globally. It will not only maintain and strengthen the mink research it also benefits genetics in general, with implications in all natural science research areas.

References

Anistoroaei, R., Ansari, S., Farid, A., Benkel, B., Karlskov-Mortensen, P. and Christensen, K., 2009. An extended anchored linkage map and virtual mapping for the American mink genome based on homology to human and dog. Genomics 94(3): 204-210.

Anistoroaei, R, ten Hallers, B., Nefedov, M., Christensen, K. and de Jong P., 2011. Construction of an American mink Bacterial Artificial Chromosome (BAC) library and sequencing candidate genes important for the fur industry BMC Genomics 12: 354.

Anistoroaei, R., Krogh, A.K. and Christensen, K., 2012. A frameshift mutation in the LYST gene is responsible for the Aleutian color and the associated Chédiak-Higashi syndrome in American mink, Animal Genetics, in press.

Chowdhary, B.P. and Raudsepp, T., 2008. The Horse Genome Derby: racing from map to whole genome sequence Chromosome Research 16: 109-127.

Gnerre, S., MacCallum, I., Przybylski, D., Ribeiro, F.J., Burton, J.N., Walker, B.J., Sharpe, T., Hall, G., Shea, T.P., Sykes, S., Berlin, A.M., Aird, D., Costello, M., Daza, R., Williams, L., Nicol, R., Gnirke, A., Nusbaum, C., Lander, E.S. and Jaffe, D.B., 2010. High-quality draft assemblies of mammalian genomes from massively parallel sequence data. PNAS 108: 1513-1518.

Gill, P., 2001. An assessment of the utility of single nucleotide polymorphisms (SNPs) for forensic purpose. Int J Legal Med 114: 204-210.

Graphodatsky, A., Yang, F., Serdukova, N., Perelman, P., Zhdanova, N. and Ferguson-Smith, M., 2000. Dog chromosome-specific paints reveal evolutionary inter- and intrachromosomal rearrangements in the American mink and human. Cytogenet Cell Genet 90: 275-278.

Hameister, H., Klett, C., Bruch, J., Dixkens, C., Vogel, W. and Christensen, K., 1997. Zoo-FISH analysis: the American mink (*Mustela vison*) closely resembles the cat karyotype. Chromos Res 5: 5-11.

Hillier, L.W., Marth, G.T., Quinlan, A.R., Dooling, D., Fewell, G., Barnett, D., Fox, P., Glasscock, J.I., Hickenbotham, M., Huang, W., Magrini, V.J., Richt, R.J., Sander, S.N., Stewart, D.A., Stromberg, M., Tsung, E.F., Wylie, T., Schedl, T., Wilson, R.K. and Mardis, E.R., 2008. Whole-genome sequencing and variant discovery in *C. elegans*. Nat Methods. 5: 183-188.

Karlsson, E.K., Baranowska, I., Wade, C.M., Salmon Hillbertz, N.H.C., Zody, M.C., Anderson, N., Biagi, T.M., Patterson, N., Pielberg, G.R., Kulbokas, E.J., Comstock, K.E., Keller, E.T., Mesirov, J.P., von Euler, H., Kampe, O., Hedhammar, A., Lander, E.S., Andersson, G., Andersson, L. and Lindblad-Toh, K., 2007. Efficient mapping of mendelian traits in dogs through genome-wide association. Nature Genetics 39: 1321-1328.

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.

Li, R, Zhu, H, Ruan, J, Qian, W, Fang, X, Shi, Z, Li, Y, Li, S, Shan, G, Kristiansen, K, Li, S, Yang, H, Wang, J and Wang, J., 2010. *De novo* assembly of human genomes with massively parallel short read sequencing. Genome Res. 20: 265-72.

Matukumalli, L.K., Lawley, C.T., Schnabel, R.D., Taylor, J.F., Allan, M.F., Heaton, M.P., O'Connell, J., Moore, S.S., Smith, T.P.L., Sonstegard, T.S. and Van Tassell, C.P., 2009. Development and characterization of a high density SNP genotyping assay for cattle. PLoS ONE 4(4): e5350.

Nes, N.N., Einarsson E.J. and Lohi O., 1988. Beautiful fur animals – and their colour genetics.Scientifur, Hilleroed, Denmark. 271 pp.

Wheeler, D.A., Srinivasan, M., Egholm, M., Shen, Y., Chen, L., McGuire, A., He, W., Chen, Y.-J., Makhijani, V., Roth, G.T., Gomes, X., Tartaro, K., Niazi, F., Turcotte, C.L., Irzyk, G.P., Lupski, J.R., Chinault, C., Song, X.-z., Liu, Y., Yuan, Y., Nazareth, L., Qin, X., Muzny, D.M., Margulies, M., Weinstock, G.M., Gibbs, R.A. and Rothberg, J.M., 2008. The complete genome of an individual by massively parallel DNA sequencing. Nature 452: 872-876.

Zerbino, D. and Birney, E. 2008. Velvet: Algorithms for de novo short read assembly using Bruijn graphs. Genome Res. 18(5): 821-829.