The draft genome sequence of the American mink (Neovison vison) opens new opportunities of genomic research in mink

Cai, Zexi; Panitz, Frank; Petersen, Bent; Sahana, Goutam; Thomsen, Bo; Bendixen, Christian; Lund, Mogens Sandø; Guldbrandtsen, B

Published in:
Proceedings of the XIth International Scientific Congress in Fur Animal Production

Publication date:
2016

Document Version
Publisher's PDF, also known as Version of record

The draft genome sequence of the American mink *(Neovison vison)* opens new opportunities of genomic research in mink

Z.Cai[1], F. Panitz[2], B. Petersen[3], G. Sahana[1], B. Thomsen[2], C. Bendixen[2], M. S. Lund[1], B. Guldbrandtsen[1]

[1] Center for Quantitative Genetics and Genomics, Department of Molecular Biology and Genetics, Aarhus University, DK-8830 Tjele, Denmark
[2] Department of Molecular Biology and Genetics, Research Centre Foulum, Aarhus University, DK-8830 Tjele, Denmark
[3] Center for Biological Sequence Analysis, Technical University of Denmark, 2800 Kgs. Lyngby, Denmark

zexi.cai@mbg.au.dk

Abstract

The American mink *(Neovison vison)* is a semiaquatic mustelid native to North America. It is an important animal for the fur industry. Although many efforts have been made to locate genes influencing fur quality and color, the lack of a reference genome impedes the search. American mink has the smallest chromosome number among studied *Carnivora* species. Genomic information about American mink is also vital to understand the evolution of *Carnivora*. Hence a reference genome of mink will facilitate genetic improvement of economic traits and will increase our knowledge about the evolution of *Carnivora*.

Here we present the draft genome sequence of American mink. In our study, a male inbred pearl mink was sequenced by Illumina paired-end and mate pair sequencing. The reads were assembled, which lead to 22,419 scaffolds with an N50 (shortest sequence length at 50% of the genome) of 646,304 bp. The assembly constituted 2.4G plus gaps, representing 90% of the estimated genome size. Repeat annotation showed that repeat sequences constitute about 25% of the mink genome. The biggest repeat family was a family of LINEs similar to LINEs found in the dog and ferret genomes. Gene annotation of our draft genome indicated our draft genome contains 87% of 1:1 *Vertebrata* genes (63.5% complete single copy genes, 0.5% duplicated genes and 22% fragment genes). We were able to map on the draft genome all the well-studied genes which are thought to be involved in the coat quality and coat color phenotypes.

Our draft genome has great potential to facilitate genomic research towards improved breeding for high fur quality and will strengthen our understanding of *Carnivora* evolution.

Keywords: NGS, genome assembly, farm animal

Introduction

The American mink *(Neovison vison)* is a semiaquatic species that belongs to the *Carnivora* family. American mink is the only species in the genus *Neovison*. With the release of genome sequences of dog (Lindblad-Toh, Kerstin *et al.* 2005), cat (Pontius *et al.* 2007) and ferret (Peng *et al.* 2014), the genome sequence of American mink will therefore provide additional valuable information into the evolution of *Carnivora*.

For decades the fur industry has used American mink as the major source of fur. The mink fur is an important industry in China, Denmark and Canada etc. (Hansen 2014). As a result, the improvement

in fur quality using genetic markers is a focus of mink breeding research. Since 1996, researchers began to develop SSR markers for mink (O'Connell *et al.* 1996, Fleming *et al.* 1999, Hansen and Jacobsen 1999, Vincent *et al.* 2003, Anistoroaei *et al.* 2006). A linkage map for these markers was released in 2007 (Anistoroaei *et al.* 2007). Subsequently, two updated versions of the linkage map were also released adding new SSR markers (Anistoroaei *et al.* 2009, Anistoroaei *et al.* 2012). Next-generation sequencing (NGS) techniques are also used to generate markers for mink. Using Restriction site-Associated DNA (RAD), 1,256 single nucleotide polymorphism (SNP) markers were identified (Thirstrup *et al.* 2013).

Fur quality is among the most economically important traits in mink production. Therefore, finding genes affecting fur quality is vital for the industry. There was limited effort to identify the QTL for fur quality (Thirstrup *et al.* 2014). A BAC library was constructed and some homology genes for fur quality and color from other mammalian species were analyzed (Anistoroaei *et al.* 2011). The transcriptome of mink was also published, and some color genes were studied using transcriptome data (Christensen and Anistoroaei 2014). But the absence of a mink reference genome hinders the creation of high density maker panels for genotyping mink in order to map genes for economically important traits. Therefore the objective of this study was to construct a mink genome assembly.

Materials and Methods

Genomic data generation and genome assemble

A male pearl American mink *(Neovison vison)* individual from Aarhus University's mink farm was used to generate the genomic sequence. The chosen one is from one of the inbred line. The spleen tissue was used to extract DNA. Extraction of DNA and sequence generation was performed by Aros (http://aros-ab.com/). In order to use the Allpaths-lg pipeline(Butler et al. 2008), we designed one overlap pair-end library (165bp, 100PE) and two long insert mate-pair libraries (3k and 5k,100PE). The total data were 163.2 Gb for pair-end sequence and 184.3 Gb mate-pair sequence. All the data were analyzed using Allpaths-lg to construct the genome assembly.

Down-stream analysis

In order to capture the species-specific repeat sequence from mink, we performed do novo prediction before running RepeatMasker (Smit *et al.* 1996). We used RepeatExplorer (Novák *et al.* 2013) and RepeatModler (Smit and Hubley 2010) with the default parameters to analyze the repeat sequence of mink. For the RepeatExplorer results, we analyzed all the clusters to remove those belonging to non-repeat sequence by blasting(Johnson *et al.* 2008) the contigs against NR (Pruitt *et al.* 2007) database. Finally, we combined results from these two sources and constructed a mink-specific repeat database and used this repeat database to annotate the repeat sequence of the genome. We used BUSCO (Simão *et al.* 2015) to assess the quality of the genome assembly and help annotation.

Results and discussion

Sequencing and assembly

A whole genome sequencing (WGS) strategy was used to sequence and assemble a male brown American mink. A total of 163.2 Gb of next-generation Illumina paired-end reads was generated by sequencing genome shotgun libraries of 165 bp insert size. 184.3 Gb of mate-paired reads were also generated with 3 kb and 5 kb insert size library. The total sequence covered 128.7 fold of the estimated 2.7 Gb genome. All the sequencing reads were assembled by Allpaths-lg (Gnerre et al. 2011), which yielded

Table 1. *The repeat sequence of mink*

Families	Numbers	Length	Percentage
SINEs	993,357	163,046,403 bp	6.71 %
LINEs	555,858	296,246,162 bp	12.19 %
LTR elements	213,839	71,725,044 bp	2.95 %
DNA elements	18,265	32,396,996 bp	1.33 %
Small RNA	919,752	155,347,649 bp	6.39 %
Satellites	6,402	3,167,324 bp	0.13 %
Simple repeats	729,614	28,893,422 bp	1.19 %
Low complexity	134,873	6,823,683 bp	0.28 %
Unclassified	7,651	860,216 bp	0.04 %
Total			24.82 %

Table 2. *The candidate fur quality and color genes*

Gene	Location	Gap	Gene	Location	Gap
FGF5	scaffold2417	780bp	MC2R	scaffold7389	NA
PMEL	scaffold1856	746bp	KITL	scaffold978	2057
MLPH	scaffold4345	NA	KIT	scaffold1621	4084
TYRP1	scaffold4069	1354bp	LYST	scaffold910	3048
MC1R	scaffold1073	NA	RSPO2	scaffold178	3378
KRT71	scaffold819	NA	HLADRB1	scaffold4491	3098
MITF	scaffold53	NA	AGRP	scaffold68	NA
ASIP	scaffold1740	NA	Atoh-1	scaffold810	NA
DEFB103	scaffold1496	NA	ITGB1	scaffold1078	NA
DEFB1	scaffold1496	592	TMIE	scaffold213	120
MC3R	scaffold1804	NA	SLC24A5	scaffold449	NA
TYR	scaffold4579	140			

a 2.43 Gb assemble plus gaps (Table 1). The draft genome consisted of 22,419 scaffolds with N50 of 646,304 bp and largest scaffold was 4.7 Mb.

Down-stream analysis

In order to construct the species-specific repeat library for mink, we used the RepeatModler (Smit and Hubley 2010) and RepeatExplorer (Novák *et al.* 2013) software to analyze repeat families, and then used RepeatMasker (Smit *et al.* 1996) to annotate the genome. Our result showed that 25% of the mink genome was repeat sequence. American mink is thereby a low-repetitive genome, compared to human and dog (Lindblad-Toh, K. *et al.* 2005). The dominant repeat sequence was a family of LINEs similar to LINEs found in the dog and ferret genomes (Peng *et al.* 2014). SINE families were the second largest type of repeat sequence in the mink genome. The detailed composition of different families is listed in *Table 1*. We found three species-specific satellite repeats of mink, which were not reported in the dog and ferret genomes. However, they only covered 0.13% of the genome *(Table 1)*.

To help the annotation of the genome and also to assess the quality of the assembly, we used BUSCO (Simão *et al.* 2015) to analyze the completeness of the *Vertebrata* single-copy orthologs. The result

showed that 87% of 1:1 Vertebrata genes (63.5% complete single copy genes, 0.5% duplicated genes and 22% fragment genes) were present in our assembly. Subsequently, we also mapped on the genome all the well-studied genes which are thought to be involved in coat quality and coat color phenotypes; results are listed in Table 2. Our result capture complete sequence of half of these genes, which complements data from previous research (Anistoroaei *et al.* 2011).

Conclusion

We used NGS technology to generate whole genome sequencing data. The data was used to assemble a reference genome of mink. Our results showed that the assembly is a high quality draft genome. Repeat annotation shows that mink has a low-repetitive genome. Our gene annotation will facilitate the fur quality genes identification. And the SNP calling using reference will provide enough markers to build the high density linkage map and show the great potential in genomic breeding, QTL mapping and GWAS analysis.

References

Anistoroaei R., Farid A., Benkel B., Cirera S. and Christensen K. (2006). Isolation and characterization of 79 microsatellite markers from the American mink *(Mustela vison)*. Animal Genetics **37**(2): 185-188.

Anistoroaei R., Menzorov A., Serov O., Farid A. and Christensen K. (2007). The first linkage map of the American mink *(Mustela vison)*. Anim Genet **38**(4): 384-388.

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., Nielsen V., Markakis M. N., Karlskov-Mortensen P., Jorgensen C. B., Christensen K. and Fredholm M. (2012). A re-assigned American mink *(Neovison vison)* map optimal for genome-wide studies. Gene **511**(1): 66-72.

Butler J., MacCallum I., Kleber M., Shlyakhter I. A., Belmonte M. K., Lander E. S., Nusbaum C. and Jaffe D. B. (2008). "ALLPATHS: de novo assembly of whole-genome shotgun microreads." Genome research **18**(5): 810-820.

Christensen K. and Anistoroaei R. (2014). An American mink *(Neovison vison)* transcriptome. Anim Genet **45**(2): 301-303.

Fleming M. A., Ostrander E. A. and Cook J. A. (1999). Microsatellite markers for American mink *(Mustela vison)* and ermine *(Mustela erminea)*. Molecular Ecology **8**(8): 1352-1355.

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. (2011). High-quality draft assemblies of mammalian genomes from massively parallel sequence data. Proceedings of the National Academy of Sciences of the United States of America **108**(4): 1513-1518.

Hansen H. O. (2014). The global fur industry: trends, globalization and specialization. Journal of Agricultural Science and Technology. A **4**(7A).

Hansen M. M. and Jacobsen L. (1999). Identification of mustelid species: otter (Lutra lutra), American mink *(Mustela vison)* and polecat *(Mustela putorius)*, by analysis of DNA from faecal samples. Journal of Zoology **247**(2): 177-181.

Johnson M., Zaretskaya I., Raytselis Y., Merezhuk Y., McGinnis S. and Madden T. L. (2008). NCBI BLAST: a better web interface. Nucleic acids research 36(suppl 2): W5-W9.

Lindblad-Toh K., Wade C. M., Mikkelsen T. S., Karlsson E. K., Jaffe D. B., Kamal M., Clamp M., Chang J. L., Kulbokas E. J., 3rd, Zody M. C., Mauceli E., Xie X., Breen M., Wayne R. K., Ostrander E. A., Ponting C. P., Galibert F., Smith D. R., DeJong P. J., Kirkness E., Alvarez P., Biagi T., Brockman W., Butler J., Chin C. W., Cook A., Cuff J., Daly M. J., DeCaprio D., Gnerre S., Grabherr M., Kellis M., Kleber M., Bardeleben C., Goodstadt L., Heger A., Hitte C., Kim L., Koepfli K. P., Parker H. G., Pollinger J. P., Searle S. M., Sutter N. B., Thomas R., Webber C., Baldwin J., Abebe A., Abouelleil A., Aftuck L., Ait-Zahra M., Aldredge T., Allen N., An P., Anderson S., Antoine C., Arachchi H., Aslam A., Ayotte L., Bachantsang P., Barry A., Bayul T., Benamara M., Berlin A., Bessette D., Blitshteyn B., Bloom T., Blye J., Boguslavskiy L., Bonnet C., Boukhgalter B., Brown A., Cahill P., Calixte N., Camarata J., Cheshatsang Y., Chu J., Citroen M., Collymore A., Cooke P., Dawoe T., Daza R., Decktor K., DeGray S., Dhargay N., Dooley K., Dooley K., Dorje P., Dorjee K., Dorris L., Duffey N., Dupes A., Egbiremolen O., Elong R., Falk J., Farina A., Faro S., Ferguson D., Ferreira P., Fisher S., FitzGerald M., Foley K., Foley C., Franke A., Friedrich D., Gage D., Garber M., Gearin G., Giannoukos G., Goode T., Goyette A., Graham J., Grandbois E., Gyaltsen K., Hafez N., Hagopian D., Hagos B., Hall J., Healy C., Hegarty R., Honan T., Horn A., Houde N., Hughes L., Hunnicutt L., Husby M., Jester B., Jones C., Kamat A., Kanga B., Kells C., Khazanovich D., Kieu A. C., Kisner P., Kumar M., Lance K., Landers T., Lara M., Lee W., Leger J. P., Lennon N., Leuper L., LeVine S., Liu J., Liu X., Lokyitsang Y., Lokyitsang T., Lui A., Macdonald J., Major J., Marabella R., Maru K., Matthews C., McDonough S., Mehta T., Meldrim J., Melnikov A., Meneus L., Mihalev A., Mihova T., Miller K., Mittelman R., Mlenga V., Mulrain L., Munson G., Navidi A., Naylor J., Nguyen T., Nguyen N., Nguyen C., Nguyen T., Nicol R., Norbu N., Norbu C., Novod N., Nyima T., Olandt P., O'Neill B., O'Neill K., Osman S., Oyono L., Patti C., Perrin D., Phunkhang P., Pierre F., Priest M., Rachupka A., Raghuraman S., Rameau R., Ray V., Raymond C., Rege F., Rise C., Rogers J., Rogov P., Sahalie J., Settipalli S., Sharpe T., Shea T., Sheehan M., Sherpa N., Shi J., Shih D., Sloan J., Smith C., Sparrow T., Stalker J., Stange-Thomann N., Stavropoulos S., Stone C., Stone S., Sykes S., Tchuinga P., Tenzing P., Tesfaye S., Thoulutsang D., Thoulutsang Y., Topham K., Topping I., Tsamla T., Vassiliev H., Venkataraman V., Vo A., Wangchuk T., Wangdi T., Weiand M., Wilkinson J., Wilson A., Yadav S., Yang S., Yang X., Young G., Yu Q., Zainoun J., Zembek L., Zimmer A. and Lander E. S. (2005). "Genome sequence, comparative analysis and haplotype structure of the domestic dog." Nature **438**(7069): 803-819.

Lindblad-Toh K., Wade C. M., Mikkelsen T. S., Karlsson E. K., Jaffe D. B., Kamal M., Clamp M., Chang J. L., Kulbokas E. J. and Zody M. C. (2005). Genome sequence, comparative analysis and haplotype structure of the domestic dog. Nature **438**(7069): 803-819.

Novák P., Neumann P., Pech J., Steinhaisl J. and Macas J. (2013). RepeatExplorer: a Galaxy-based web server for genome-wide characterization of eukaryotic repetitive elements from next-generation sequence reads. Bioinformatics **29**(6): 792-793.

O'Connell M., Wright J. and Farid A. (1996). Development of PCR primers for nine polymorphic American mink *Mustela vison* microsatellite loci. Molecular Ecology **5**(2): 311-312.

Peng X., Alfoldi J., Gori K., Eisfeld A. J., Tyler S. R., Tisoncik-Go J., Brawand D., Law G. L., Skunca N., Hatta M., Gasper D. J., Kelly S. M., Chang J., Thomas M. J., Johnson J., Berlin A. M., Lara M., Russell P., Swofford R., Turner-Maier J., Young S., Hourlier T., Aken B., Searle S., Sun X., Yi Y., Suresh M., Tumpey T. M., Siepel A., Wisely S. M., Dessimoz C., Kawaoka Y., Birren B. W., Lindblad-Toh K., Di Palma F., Engelhardt J. F., Palermo R. E. and Katze M. G. (2014). The draft genome sequence of the ferret *(Mustela putorius furo)* facilitates study of human respiratory disease. Nat Biotechnol **32**(12): 1250-1255.

Pontius J. U., Mullikin J. C., Smith D. R., Team A. S., Lindblad-Toh K., Gnerre S., Clamp M., Chang J., Stephens R. and Neelam B. (2007). Initial sequence and comparative analysis of the cat genome. Genome research **17**(11): 1675-1689.

Pruitt K. D., Tatusova T. and Maglott D. R. (2007). NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic acids research **35** (suppl 1): D61-D65.

Simão F. A., Waterhouse R. M., Ioannidis P., Kriventseva E. V. and Zdobnov E. M. (2015). BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics **31**(19): 3210-3212.

Smit A. and Hubley R. (2010). RepeatModeler Open-1.0. Repeat Masker Website.
Smit A. F., Hubley R. and Green P. (1996). RepeatMasker Open-3.0.

Thirstrup J. P., Pujolar J. M., Larsen P. F., Jensen J., Nielsen R. O. and Pertoldi C. (2013). De novo SNP calling from RAD sequences for a mink *(Neovison vison)* specific genotyping assay. NJF Autumn meeting in Fur Animal Research.

Thirstrup J. P., Anistoroaei R., Guldbrandtsen B., Christensen K., Fredholm M. and Nielsen V. H. (2014). Identifying QTL and genetic correlations between fur quality traits in mink *(Neovison vison)*. Anim Genet **45**(1): 105-110.

Vincent I., Farid A. and Otieno C. (2003). Variability of thirteen microsatellite markers in American mink *(Mustela vison)*. Canadian journal of animal science **83**(3): 597-599.