Genomics Data 10 (2016) 155–157

Contents lists available at ScienceDirect

Genomics Data

journal homepage: www.elsevier.com/locate/gdata

Data in Brief

Application of target capture sequencing of exons and conserved non-coding sequences to 20 inbred rat strains

Minako Yoshihara [a,b], Tetsuya Sato [a,b], Daisuke Saito [a,b], Osamu Ohara [c], Takashi Kuramoto [d,*], Mikita Suyama [a,b,*]

[a] Medical Institute of Bioregulation, Kyushu University, Maidashi 3-1-1, Higashi-ku, Fukuoka 812-8582, Japan
[b] AMED-CREST, Japan Agency for Medical Research and Development, Fukuoka 812-8582, Japan.
[c] Department of Technology Development, Kazusa DNA Research Institute, Kisarazu, Chiba 292-0818, Japan
[d] Institute of Laboratory Animals, Graduate School of Medicine, Kyoto University, Kyoto 606-8501, Japan

ARTICLE INFO

Article history:
Received 28 October 2016
Accepted 9 November 2016
Available online 14 November 2016

ABSTRACT

We report sequence data obtained by our recently devised target capture method TargetEC applied to 20 inbred rat strains. This method encompasses not only all annotated exons but also highly conserved non-coding sequences shared among vertebrates. The total length of the target regions covers 146.8 Mb. On an average, we obtained 31.7× depth of target coverage and identified 154,330 SNVs and 24,368 INDELs for each strain. This corresponds to 470,037 unique SNVs and 68,652 unique INDELs among the 20 strains. The sequence data can be accessed at DDBJ/EMBL/GenBank under accession number PRJDB4648, and the identified variants have been deposited at http://bioinfo.sls.kyushu-u.ac.jp/rat_target_capture/20_strains.vcf.gz.

Specifications [*standardized info for the reader*]	
Organism/cell line/tissue	*Rattus norvegicus* (BDIX/NemOda, BDIX. Cg-Tal/NemOda, BN/SsNSlc, BUF/MNa, DOB/Oda, F344/DuCrlCrlj, F344/Jcl, F344/NSlc, F344/Stm, HTX/Kyo, HWY/Slc, IS/Kyo, IS-*Tlk*/Kyo, KFRS3B/Kyo, LE/Stm, LEC/Tj, NIG-III/Hok, RCS/Kyo, ZF, ZFDM)
Sex	Female and male, see Table 1
Sequencer or array type	Illumina NextSeq 500
Data format	FASTQ and VCF
Experimental factors	Genomic DNA extracted from spleen
Experimental features	Target capture sequencing of exons and conserved non-coding sequences
Consent	Not applicable
Sample source location	Rat strains were provided by the National BioResource Project (NBRP)–Rat (http://www.anim.med.kyoto-u.ac.jp/nbr/).

1. Direct link to deposited data [provide URL below]

http://www.ncbi.nlm.nih.gov/bioproject/?term=PRJDB4648
http://bioinfo.sls.kyushu-u.ac.jp/rat_target_capture/20_strains.vcf.zip

* Corresponding authors.
E-mail addresses: tkuramot@anim.med.kyoto-u.ac.jp (T. Kuramoto), mikita@bioreg.kyushu-u.ac.jp (M. Suyama).

2. Experimental design, materials and methods

Rats are used as animal models of many human diseases, such as cancer and hypertension. Because of its significance in biomedical analyses, the genome sequence of the Brown Norway rat strain was determined as the third complete mammalian genome [1]. The National BioResource Project–Rat (NBRP-Rat) at Kyoto University is one of the largest repositories for rat strains, and currently, >700 strains have been collected and preserved as live animals, embryos, or sperm [2]. Determination of genome sequences for these strains is important not only for understanding genetic causes for various phenotypes but also to augment their value as biological resources.

Whole exome sequencing is an efficient approach to characterize only the exonic portions of a genome, which typically comprise 1%–2% of complete mammalian genomes, and has been successfully used in the identification of relevant genes and their causative mutations in many diseases in humans. Although some non-human exome capture kits exist, there had previously been no such capture probe set for rats. Therefore, we established a target capture kit specifically designed for this rodent species, employing the SeqCap EZ Developer Library (Roche NimbleGen, Madison, WI, USA; design name 140929_RN5_MS_EZ_HX1). In designing our target capture probe set, we included highly conserved non-coding sequences (CNSs) as target regions as well as all annotated exons, covering a total 146.8 Mb of the genome [3]. By applying this target capture method TargetEC (target capture for exons and conserved non-coding

http://dx.doi.org/10.1016/j.gdata.2016.11.010

Table 1
Summary statistics for sequencing and variant calling.

Strain	Sex	Total reads	Read length	Mapped reads after post-processing (%)	Average target depth	SNV (depth ≥ 5×)	INDEL (depth ≥ 5×)
BDIX.Cg-*Tal*/NemOda	Unknown	77,031,192	151	62,133,380 (80.7)	33.0	161,043	25,729
BDIX/NemOda	Female	62,884,340	151	50,668,261 (80.6)	26.2	155,727	24,561
BN/SsNSlc	Male	67,363,478	151	54,385,129 (80.7)	29.4	23,060	5533
BUF/MNa	Male	60,898,020	151	49,603,905 (81.5)	27.2	154,382	24,122
DOB/Oda	Male	68,359,820	151	61,010,641 (89.2)	31.7	196,751	30,148
F344/DuCrlCrlj	Male	73,516,660	151	59,541,186 (81.0)	27.7	152,184	23,890
F344/Jcl	Male	62,994,072	151	50,991,611 (80.9)	26.6	152,141	23,855
F344/NSlc	Male	62,838,170	151	50,726,936 (80.7)	27.5	152,546	23,930
F344/Stm	Male	64,788,908	151	52,984,127 (81.8)	29.1	151,919	23,735
HTX/Kyo	Male	72,484,640	151	64,572,821 (89.1)	33.7	154,418	24,156
HWY/Slc	Male	74,687,034	151	66,579,903 (89.1)	34.6	157,070	24,873
IS/Kyo	Male	79,430,344	151	70,744,396 (89.1)	37.4	187,300	29,120
IS-*Tlk*/Kyo	Male	75,990,092	151	67,761,875 (89.2)	35.8	186,648	28,902
KFRS3B/Kyo	Female	81,643,134	151	72,603,786 (88.9)	35.1	154,292	24,419
LE/Stm	Male	72,300,094	151	58,438,239 (80.8)	31.7	157,488	25,052
LEC/Tj	Unknown	78,990,272	151	70,539,682 (89.3)	37.3	167,547	26,315
NIG-III/Hok	Unknown	78,128,624	151	69,625,354 (89.1)	36.9	164,732	26,195
RCS/Kyo	Male	71,627,648	151	57,894,324 (80.8)	31.6	155,472	24,975
ZF	Male	69,986,466	151	56,655,891 (81.0)	30.2	150,778	23,815
ZFDM	Male	73,535,060	151	59,407,086 (80.8)	31.9	151,101	24,025

sequences) to four rat strains (WTC/Kyo, WTC-*swh*/Kyo, PVG/Seac, and KFRS4/Kyo), we confirmed that TargetEC performs efficiently in the identification of causative mutations, including those present in the non-coding regions [3]. In this study, we further applied TargetEC to 20 additional inbred strains preserved in NBRP-Rat to identify additional variants observed in multiple rat strains. These 20 strains were selected according to the following three categories: disease models derived from selective breeding (BDIX/NemOda, BDIX.Cg-*Tal*/NemOda, BUF/MNa, HTX/Kyo, HWY/Slc, KFRS3B/Kyo, RCS/Kyo, ZF, and ZFDM), those originated from wild populations (BN/SsNSlc, DOB/Oda, IS/Kyo, IS-*Tlk*/Kyo, LE/Stm, LEC/Tj, and NIG-III/Hok), and representative inbred strains (F344/DuCrlCrlj, F344/Jcl, F344/NSlc, and F344/Stm). All animal experimentation protocols were approved by the Institutional Animal Care and Use Committees of Kyoto University and were conducted according to the Regulation on Animal Experimentation at Kyoto University.

Genomic DNA was extracted from spleen samples with standard protocols. Target capture was performed using the standard SeqCap EZ System protocol (Roche NimbleGen). DNA sequencing libraries were prepared using the KAPA HyperPlus Library Preparation Kit (KAPA Biosystems, London, UK) according to the manufacturer's protocol. Sequencing was performed on an Illumina NextSeq 500 platform (Illumina, San Diego, CA, USA) using the High Output Kit (2 × 150 cycles). We obtained 61–82 million reads for each strain (Table 1). Sequence reads were mapped to the rat genome assembly rn5 (RGSC 5.0, March 2012) using BWA (v0.7.4) [4] with the default parameters. SAMtools (v0.1.12a) [5], Picard tools (v1.87) (http://broadinstitute.github.io/picard/), and the Genome Analysis Toolkit (GATK; v2.5.2) [6] were used for post-processing of mapped reads. Variant calling employed the UnifiedGenotyper utility in GATK. We identified 154,330 SNVs and 24,368 INDELs in the target regions, on an average (Table 1). The number of unique SNVs and INDELs among the 20 strains was 470,037 and 68,652, respectively. Sequence data and variants identified for these strains represent valuable resources for further genetic studies in the rat.

Conflict of interest

The authors declare no conflicts of interest.

Acknowledgements

We thank the National BioResource Project–Rat (http://www.anim.med.kyoto-u.ac.jp/nbr/) for providing rat strains. This work was supported in part by the Cooperative Research Project Program of the Medical Institute of Bioregulation, Kyushu University, to OO, and the Genome Information Upgrading Program of the National BioResource Project, Japan Agency for Medical Research and Development, to OO, TK, and MS.

References

[1] R.A. Gibbs, G.M. Weinstock, M.L. Metzker, D.M. Muzny, E.J. Sodergren, S. Scherer, G. Scott, D. Steffen, K.C. Worley, P.E. Burch, G. Okwuonu, S. Hines, L. Lewis, C. DeRamo, O. Delgado, S. Dugan-Rocha, G. Miner, M. Morgan, A. Hawes, R. Gill, Celera, R.A. Holt, M.D. Adams, P.G. Amanatides, H. Baden-Tillson, M. Barnstead, S. Chin, C.A. Evans, S. Ferriera, C. Fosler, A. Glodek, Z. Gu, D. Jennings, C.L. Kraft, T. Nguyen, C.M. Pfannkoch, C. Sitter, G.G. Sutton, J.C. Venter, T. Woodage, D. Smith, H.-M. Lee, E. Gustafson, P. Cahill, A. Kana, L. Doucette-Stamm, K. Weinstock, K. Fechtel, R.B. Weiss, D.M. Dunn, E.D. Green, R.W. Blakesley, G.G. Bouffard, P.J. De Jong, K. Osoegawa, B. Zhu, M. Marra, J. Schein, I. Bosdet, C. Fjell, S. Jones, M. Krzywinski, C. Mathewson, A. Siddiqui, N. Wye, J. McPherson, S. Zhao, C.M. Fraser, J. Shetty, S. Shatsman, K. Geer, Y. Chen, S. Abramzon, W.C. Nierman, P.H. Havlak, R. Chen, K.J. Durbin, A. Egan, Y. Ren, X.-Z. Song, B. Li, Y. Liu, X. Qin, S. Cawley, K.C. Worley, A.J. Cooney, L.M. D'Souza, K. Martin, J.Q. Wu, M.L. Gonzalez-Garay, A.R. Jackson, K.J. Kalafus, M.P. McLeod, A. Milosavljevic, D. Virk, A. Volkov, D.A. Wheeler, Z. Zhang, J.A. Bailey, E.E. Eichler, E. Tuzun, E. Birney, E. Mongin, A. Ureta-Vidal, C. Woodwark, E. Zdobnov, P. Bork, M. Suyama, D. Torrents, M. Alexandersson, B.J. Trask, J.M. Young, H. Huang, H. Wang, H. Xing, S. Daniels, D. Gietzen, J. Schmidt, K. Stevens, U. Vitt, J. Wingrove, F. Camara, M.M. Albà, J.F. Abril, R. Guigo, A. Smit, I. Dubchak, E.M. Rubin, O. Couronne, A. Poliakov, N. Hübner, D. Ganten, C. Goesele, O. Hummel, T. Kreitler, Y.-A. Lee, J. Monti, H. Schulz, H. Zimdahl, H. Himmelbauer, H. Lehrach, H.J. Jacob, S. Bromberg, J. Gullings-Handley, M.I. Jensen-Seaman, A.E. Kwitek, J. Lazar, D. Pasko, P.J. Tonellato, S. Twigger, C.P. Ponting, J.M. Duarte, S. Rice, L. Goodstadt, S.A. Beatson, R.D. Emes, E.E. Winter, C. Webber, P. Brandt, G. Nyakatura, M. Adetobi, F. Chiaromonte, L. Elnitski, P. Eswara, R.C. Hardison, M. Hou, D. Kolbe, K. Makova, W. Miller, A. Nekrutenko, C. Riemer, S. Schwartz, J. Taylor, S. Yang, Y. Zhang, K. Lindpaintner, T.D. Andrews, M. Caccamo, M. Clamp, L. Clarke, V. Curwen, R. Durbin, E. Eyras, S.M. Searle, G.M. Cooper, S. Batzoglou, M. Brudno, A. Sidow, E.A. Stone, J.C. Venter, B.A. Payseur, G. Bourque, C. López-Otín, X.S. Puente, K. Chakrabarti, S. Chatterji, C. Dewey, L. Pachter, N. Bray, V.B. Yap, A. Caspi, G. Tesler, P.A. Pevzner, D. Haussler, K.M. Roskin, R. Baertsch, H. Clawson, T.S. Furey, A.S. Hinrichs, D. Karolchik, W.J. Kent, K.R. Rosenbloom, H. Trumbower, M. Weirauch, D.N. Cooper, P.D. Stenson, B. Ma, M. Brent, M. Arumugam, D. Shteynberg, R.R. Copley, M.S. Taylor, H. Riethman, U. Mudunuri, J. Peterson, M. Guyer, A. Felsenfeld, S. Old, S. Mockrin, F. Collins, Rat genome sequencing project consortium, genome sequence of the Brown Norway rat yields insights into mammalian evolution. Nature 428 (2004) 493–521.

[2] T. Serikawa, T. Mashimo, A. Takizawa, R. Okajima, N. Maedomari, K. Kumafuji, F. Tagami, Y. Neoda, M. Otsuki, S. Nakanishi, K. Yamasaki, B. Voigt, T. Kuramoto, National BioResource Project-Rat and related activities. Exp. Anim. Jpn. Assoc. Lab. Anim. Sci. 58 (2009) 333–341.

[3] M. Yoshihara, D. Saito, T. Sato, O. Ohara, T. Kuramoto, M. Suyama, Design and application of a target capture sequencing of exons and conserved non-coding sequences for the rat. BMC Genomics 17 (2016) 593.

[4] H. Li, R. Durbin, Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25 (2009) 1754–1760.
[5] H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, 1000 Genome Project Data Processing Subgroup, The Sequence Alignment/Map format and SAMtools. Bioinformatics 25 (2009) 2078–2079.
[6] A. McKenna, M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis, A. Kernytsky, K. Garimella, D. Altshuler, S. Gabriel, M. Daly, M.A. DePristo, The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 20 (2010) 1297–1303.