


RESEARCH ARTICLE

Open Access



# Breed-specific reference sequence optimized mapping accuracy of NGS analyses for pigs

Dan Wang<sup>1,2†</sup>, Liu Yang<sup>1†</sup>, Chao Ning<sup>1,2</sup>, Jian-Feng Liu<sup>1</sup> and Xingbo Zhao<sup>1\*</sup> 

## Abstract

**Background:** Reference sequences play a vital role in next-generation sequencing (NGS), impacting mapping quality during genome analyses. However, reference genomes usually do not represent the full range of genetic diversity of a species as a result of geographical divergence and independent demographic events of different populations. For the mitochondrial genome (mitogenome), which occurs in high copy numbers in cells and is strictly maternally inherited, an optimal reference sequence has the potential to make mitogenome alignment both more accurate and more efficient. In this study, we used three different types of reference sequences for mitogenome mapping, i.e., the commonly used reference sequence (CU-ref), the breed-specific reference sequence (BS-ref) and the sample-specific reference sequence (SS-ref), respectively, and compared the accuracy of mitogenome alignment and SNP calling among them, for the purpose of proposing the optimal reference sequence for mitochondrial DNA (mtDNA) analyses of specific populations

**Results:** Four pigs, representing three different breeds, were high-throughput sequenced, subsequently mapping reads to the reference sequences mentioned above, resulting in a largest mapping ratio and a deepest coverage without increased running time when aligning reads to a BS-ref. Next, single nucleotide polymorphism (SNP) calling was carried out by 18 detection strategies with the three tools SAMtools, VarScan and GATK with different parameters, using the bam results mapping to BS-ref. The results showed that all eighteen strategies achieved the same high specificity and sensitivity, which suggested a high accuracy of mitogenome alignment by the BS-ref because of a low requirement for SNP calling tools and parameter choices.

**Conclusions:** This study showed that different reference sequences representing different genetic relationships to sample reads influenced mitogenome alignment, with the breed-specific reference sequences being optimal for mitogenome analyses, which provides a refined processing perspective for NGS data.

**Keywords:** Mitochondrial genome, Mapping, Reference sequence, SNP calling, Pig

\* Correspondence: [zhxbcau@126.com](mailto:zhxbcau@126.com)

<sup>†</sup>Dan Wang and Liu Yang contributed equally to this work.

<sup>1</sup>National Engineering Laboratory for Animal Breeding, Ministry of Agricultural Key Laboratory of Animal Genetics, Breeding and Reproduction, College of Animal Science and Technology, China Agricultural University, Beijing, China

Full list of author information is available at the end of the article



© The Author(s). 2021 **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

## Background

Next-generation sequencing (NGS) technology is characterized by providing millions of DNA sequencing reads at a time, with high efficiency and low costs compared to Sanger sequencing [1]. The mitochondrion contains hundreds of mitochondrial genome copies [2], which are, in contrast to the biparentally inherited nuclear genome, strictly maternally inherited. With NGS data, complete mitogenomes have been obtained for many vertebrate species, including pigs [3–6], chickens [7–9] and cattle [10–12]. It has been reported that the reference sequences used affect the accuracy of genome mapping [13–15]. Reference genomes often cannot represent the full range of genetic diversity of a species as a result of geographical divergence and independent demographic events in different populations. To comprehensively characterize genetic variation, different references for different populations may be necessary. This perspective has fueled the research on pan-genomes, currently focusing on nuclear genomes, and not involving mitogenomes [16, 17]. This lack of studies on the influence of reference sequences on mitogenome alignment requires addressing.

In this study, we explored whether the genetic relationships between reference and sample sequence would impact mapping accuracy of pig mitogenomes. Three breeds of pigs, including one Asian wild boar, two unrelated Diannan small-ear pigs and one Tibetan pig, were sampled and high-throughput sequenced. The sequence data were then analysed by genome alignment and SNP calling. We tested the accuracy of mitogenome alignment and SNP calling based on three kinds of reference sequences, which represent three kinds of genetic relationships to samples. The reference sequences included the commonly used reference sequence (CU-ref), breed-specific reference sequences (BS-ref) and sample-specific reference sequences (SS-ref). In detail, (1) the commonly used reference sequence (CU-ref) refers to a frequently-used sequence from the RefSeq project at NCBI database. For pigs, CU-ref is normally the mitogenome sequence from a Landrace pig (NC\_000845.1) [18–21]. (2) the breed-specific reference sequence (BS-ref), which refers to a sequence of the same breed as the sample, was downloaded from NCBI database. (3) the sample-specific reference sequence (SS-ref), which refers to the consensus sequence, was obtained from the NGS reads of the sample through de novo assembly. Thereinto, for an optimal de novo assembly for the SS-ref sequence, three levels of NGS read sets, namely, all clean read sets, homologous read sets filtered by BLAST, or filtered by BWA mapping, were used and compared in the de novo assembly software SOAPdenovo2 [22] by default parameters with the best k-mer size estimated by KmerGennie [23].

## Results

### Performance of de novo assembly strategies

In order to produce the optimal SS-ref sequences, three de novo assembly strategies were carried out, including the Denovo strategy, the BLAST\_denovo strategy, and the BWA\_denovo strategy. The BLAST\_denovo strategy got a similar result in N50 contig size, consensus length, and genome coverage compared to the Denovo strategy. However, the former strategy yielded no polymorphic site, while the latter created a large number of polymorphisms (193 sites), potentially caused by nuclear mitochondrial sequences (NUMTs). In addition, the third assembly strategy (BWA\_denovo) got the smallest N50 contig size and some polymorphic sites. Therefore, the de novo assembly by homologous sequences filtered by BLAST from NGS data was selected for constructing the SS-ref sequence for each sample. NGS data information is listed in Additional file 1: Table S1, and the assembly quality of each de novo strategy is shown in Table 1.

### Alignment quality by different reference sequences

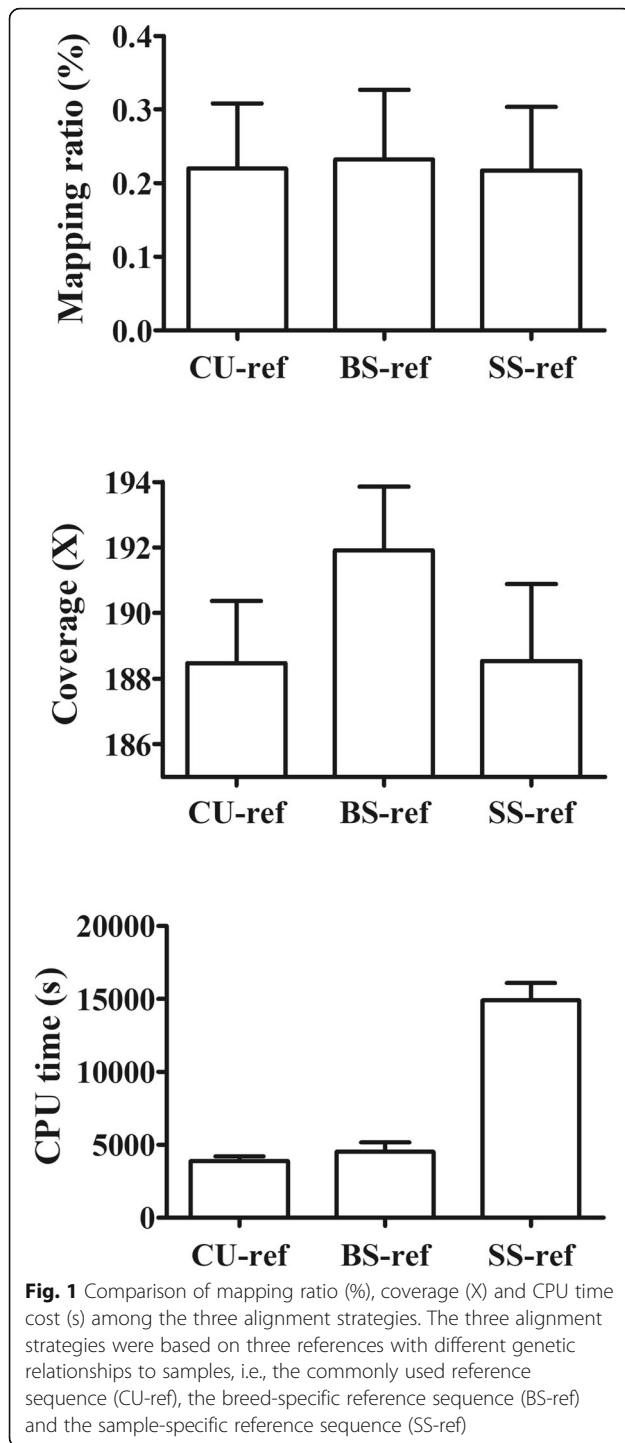
Comparison the mitogenome quality obtained by mapping to different reference sequences showed that alignment against BS-ref yielded a higher mapping ratio and a larger average coverage than CU-ref or SS-ref, while the latter two performed similarly in mapping ratio and average coverage. In terms of time consumption during mitogenome alignment, mapping NGS data to BS-ref or to CU-ref nearly completed at the same time, but faster than to the SS-ref (Fig. 1). Detailed statistics of mapping ratio, coverage and CPU time for mitogenome mapping against the three different types of reference sequences are listed in Additional file 2: Table S2. Moreover, mapping against BS-ref showed a more uniform coverage across the mitogenome than against CU-ref or SS-ref (Fig. 2).

### Performance of SNP calling strategies on mitogenome diversity

As the gold standard, Sanger sequencing data for each specimen were aligned against reference sequence KP765605.1 for sample A1, KM044240.1 for D1 and D2, and KM073256.1 for T1. The number of SNPs was 12 for A1, 7 for D1, 9 for D2, and 10 for T1, which are

**Table 1** Statistics of A1 mitogenome assembly by three strategies

Methods	BLAST_denovo	BWA_denovo	Denovo
<b>best k</b>	31	99	99
<b>N50</b>	1437	112	1867
<b>consensus length</b>	16,596	16,598	16,607
<b>coverage</b>	99.90%	99.91%	99.96%
<b>polymorphic sites</b>	0	2	193



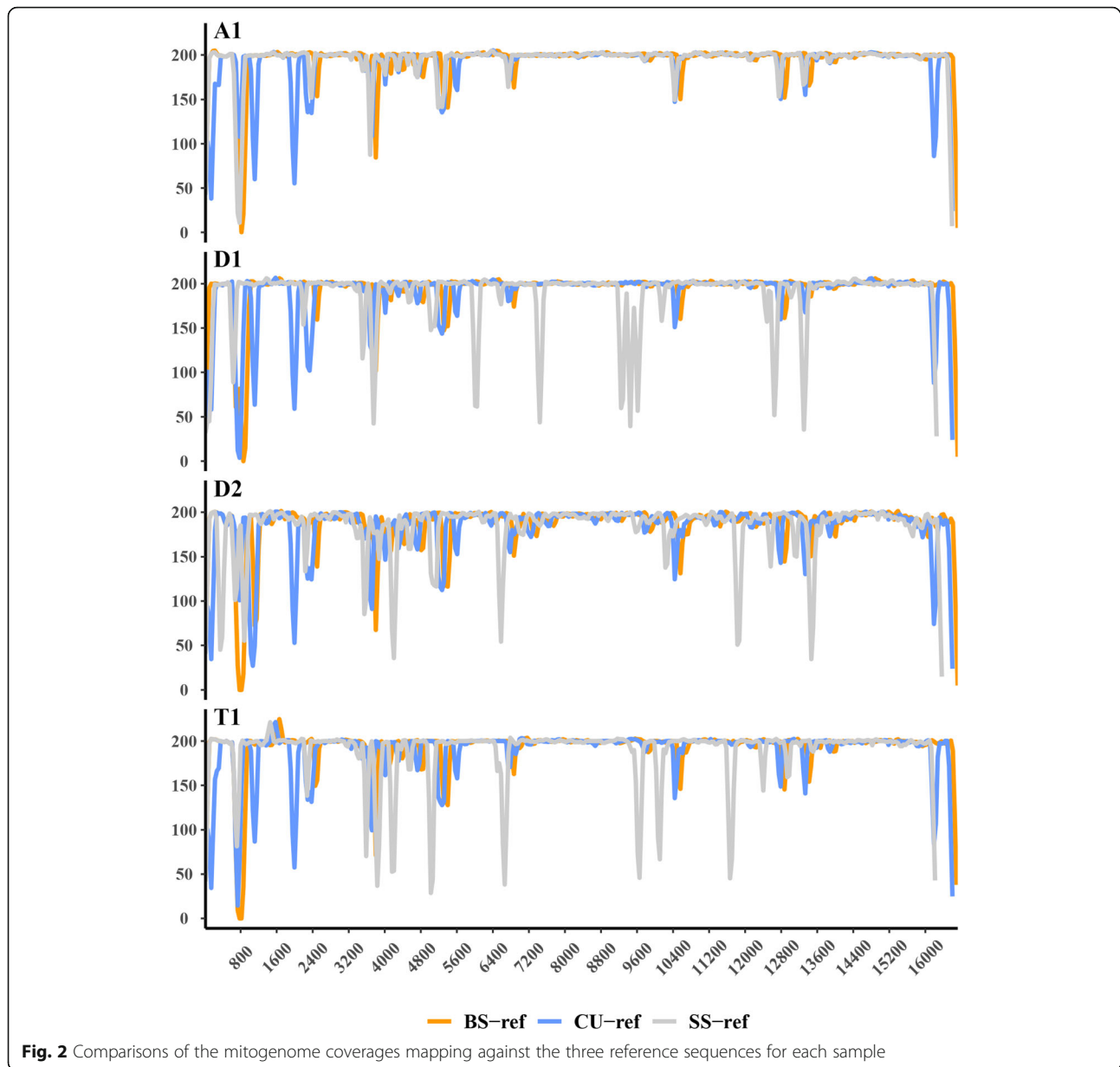
detailed in Additional file 3: Table S3. A total of 18 SNP calling strategies, performed with three variant callers, i.e., SAMtools 1.3.1 [24], VarScan 2.3.9 [25] and GATK 3.7 [26] with different parameter combinations (detailed in Table 2), were carried out to call SNPs using the bam files resulted from mitogenome alignments against the BS-ref sequences. All the SNPs we found were homogeneous substitutions compared to BS-ref. Analysis of the

concordance between the SNP calling results from NGS and Sanger data revealed that all the eighteen strategies detected all true SNPs, with zero false positives and zero false negatives (see Additional file 4: Table S4).

## Discussion

In order to generate accurate genome sequences from NGS data, many studies have explored optimal alignment strategies [13–15, 27–29]. Both de novo and reference-based approaches were used in mitogenome reconstructions of *Clarias batrachus* from NGS data, resulting in consensus sequences with different lengths [13]. Moreover, different reference sequences led to different mapping performances. Liu et al. found that when the sample-specific sequence, i.e. a sequence with the same genotype sequence as the sample, was used as mapping reference in NGS analyses of HBV (Hepatitis B Virus), mapping accuracy and variant calling were optimized compared to the other four HBV sequences from the GenBank database commonly used as reference [14]. In general, a reference sequence belonging to the same species as the sample population, i.e., a reference sequence referred to as CU-ref sequence in this study, is the choice for sequence alignment [30–32]. However, reference genomes often cannot represent the full range of genetic diversity as a result of geographical divergence and independent demographic events of different populations. To comprehensively characterize genetic variation, different references for different populations are necessary. This perspective has motivated the research on pan-genomes, currently focusing on nuclear genomes, and not involving mitogenomes [16, 17].

In this study, we compared the mitogenome alignment quality obtained by mapping to three types of reference sequences, and proposed the optimal reference sequence for mtDNA analyses of specific populations. As Figs. 1 and 2 shown, alignment against BS-ref performed better in mapping ratio, average coverage and mitogenome coverage uniformity than against the other two kinds of reference sequences, and in the NGS data mapping process both BS-ref and CU-ref using have an advantage over SS-ref regarding computing time. Thus, the comparison above shows that the mapping strategy based on BS-ref showed a slightly better performance than the other two. This result was consistent with previous studies [27, 33]. Lee et al. revealed that the references with a closer genetic relationship to investigate *Mycobacterium tuberculosis* samples showed the highest proportion of reads that successfully aligned, which influenced the detection of mitochondrial SNP in later step [33]. To deeply analyze the basic reason in different mapping efficiency, we aligned the three kinds of reference sequences (CU-ref, BS-ref and SS-ref sequences) by MEGA7 [34], and the results were detailed in Additional file 5: Table



S5. More than 200 polymorphic loci existed between CU-refs and the other two reference sequences. Polymorphisms between BS-ref and SS-ref were low-level, ranging from 9 to 38. The polymorphic difference among the reference sequences showed just right the difference of genetic relationships between references and samples, consistent with the statements of the three references. The reference sequence difference was a basis of difference in mapping efficiency, which devoted the importance of reference sequences for specific population during mitogenome alignment.

Mitogenome SNP detection is important, for example for functional annotation. A total of 18 different SNP calling strategies using three software programs with different

options were compared, and led to the same SNP results in terms of true SNPs, false positives and false negatives. The results showed that the mitogenome obtained from the breed-specific alignment had a low requirement for variation calling tools and parameter choices, which indicated that mapping to the breed-specific reference sequence contributed to an accurate mitogenome. Therefore, breed-specific reference sequences were useful for mitogenome alignment of high reliability, and also worked well for the later detection of mitogenome variation. Though the mitogenome mapping against the BS-ref showed a slightly better performance in the mapping process, its performance in the SNP calling process was a plus, especially useful for NGS data of low depth [35], for

**Table 2** Summary of the different software packages used to call SNPs

Name	Software	Command	Option
BCF	SAMtools; BCFtools	mpileup; call	Default; -c
BCF_B	SAMtools; BCFtools	mpileup; call	-B; -c
BCF_E	SAMtools; BCFtools	mpileup; call	-E; -c
HAP	GATK	HaplotypeCaller	Default
UNI	GATK	UnifiedGenotyper	Default
UNI1	GATK	UnifiedGenotyper	-ploidy 1
VAR_B01	SAMtools; VarScan	mpileup; mpileup2snp	-B; --min-var-freq 0.01
VAR_B10	SAMtools; VarScan	mpileup; mpileup2snp	-B; --min-var-freq 0.10
VAR_B25	SAMtools; VarScan	mpileup; mpileup2snp	-B; --min-var-freq 0.25
VAR_B50	SAMtools; VarScan	mpileup; mpileup2snp	-B; --min-var-freq 0.50
VAR_E01	SAMtools; VarScan	mpileup; mpileup2snp	-E; --min-var-freq 0.01
VAR_E10	SAMtools; VarScan	mpileup; mpileup2snp	-E; --min-var-freq 0.10
VAR_E25	SAMtools; VarScan	mpileup; mpileup2snp	-E; --min-var-freq 0.25
VAR_E50	SAMtools; VarScan	mpileup; mpileup2snp	-E; --min-var-freq 0.50
VAR01	SAMtools; VarScan	mpileup; mpileup2snp	Default; --min-var-freq 0.01
VAR10	SAMtools; VarScan	mpileup; mpileup2snp	Default; --min-var-freq 0.10
VAR25	SAMtools; VarScan	mpileup; mpileup2snp	Default; --min-var-freq 0.25
VAR50	SAMtools; VarScan	mpileup; mpileup2snp	Default; --min-var-freq 0.50

example for the study of ancient DNA, by reducing erroneous base incorporations.

In addition, the de novo assembly comparison showed that the homologous read sets filtered by BLAST from clean data were suitable. The BLAST algorithm [36], proposed by Altschul et al. in 1990, is now the most widely used search tool for homologous sequences in nucleotide databases. Its tolerance for mismatches and gaps is better than that of BWA, which only identifies extremely stringent sequence similarities [37]. Inadequately, in this study, the de novo assemblies of the specimens were not complete containing some gaps, which might result in a slightly poor performance in the alignments to SS-ref sequences.

## Conclusions

In this study, different kinds of reference sequences, representing different genetic relationships to the investigated samples, were used in mitogenome analyses based on NGS data. The breed-specific sequence gave an optimal performance due to its high accuracy both in mitogenome mapping and SNP calling. Overall, this study underscored the importance of reference sequence choice in mitogenome research.

## Methods

### Animal ethics statement

All experimental pigs were maintained according to the guidelines of the experimental animal management of China Agricultural University. Animal management and

experimental protocols complied with the guidelines approved by the Institutional Animal Care and Use Ethics Committee (IACUC) at China Agricultural University. After the study, the pigs were still living in the original environment.

### DNA sequencing

Ear tissues from four pigs of three breeds were collected, including an Asian wild boar (A1), two unrelated Dinnan small-ear pigs (D1 and D2) and a Tibetan pig (T1). The pigs were all female except D1, and the ear samples were collected in their early adult life. Total DNA was extracted using the QIAamp DNA Investigator kit (QIAGEN, Hilden, Germany) following the manufacturer's instructions. DNA quality was evaluated by spectrophotometry and agarose gel electrophoresis. DNA templates were ultrasonically sheared using a Covaris E220 (Covaris, Woburn, USA), and were converted into DNA libraries following the NEBNext Ultra DNA Library preparation protocol. Multiple Ampure Bead XP cleanups (Beckman Coulter, Brea, CA, USA) were conducted to remove any adapter dimers that might have developed. Quality and concentration of libraries were determined on an Agilent Bioanalyzer 2100 (Agilent Technologies, Santa Clara, CA). Subsequently, the quality-controlled genomic library for each sample was PE100 sequenced using the Illumina HiSeq 2000 sequencing system.

The traditional sequencing approach, Sanger sequencing, was also performed on the samples to represent



the gold standard in variant detection. The mitogenome was PCR-amplified with 16 primer pairs used in a previous study [21]. Amplicons were bi-directionally sequenced using the BigDye Terminator version 3.1 technology on an ABI 3730 system (Applied Biosystems, Foster City, CA). Mitogenomes were analysed with the software packages MEGA6 [38] and DnaSP v5 [39].

#### Quality control

Read quality was assessed using FastQC focusing on base quality scores and sequence length. Adapters and low-quality bases were removed with Clip&Merge. Reads shorter than 35 bp and a Phred quality score lower than 20 were removed. Next, forward and reverse reads were merged into single sequences if they overlapped by at least 8 bp. The above tools were used in an integrated pipeline, EAGER [40]. The filtered NGS data after the above steps were then used for downstream analyses.

#### Alignment to different reference sequences

Mitogenome mapping was performed with BWA [41] with default parameters for the commands “aln” and “samse”. Three types of reference sequences regarding the genetic relationship between the reference sequences and the sample were used as follows.

- (1) The commonly used reference sequence (CU-ref), which referred to a frequently-used sequence from the RefSeq project in the NCBI database. Here NC\_000845.1 was used, which is the mitogenome sequence from a Landrace [18–21].
- (2) The breed-specific reference sequence (BS-ref), which referred to a sequence of the same breed as the sample, and was downloaded from the NCBI database. For the Asian wild boar, KP765605.1 was used as BS-ref, which is from a Changbai mountains wild boar and 16,720 bp in length; for the Diannan small-ear pigs, KM044240.1 was used, which is a complete mitogenome of 16,720 bp obtained from a Diannan small-ear pig in Yunnan Province; and for the Tibetan pig, KM073256.1 was used, which is a Tibetan complete mitogenome of 16,710 bp.
- (3) The sample-specific reference sequence (SS-ref), which referred to the consensus sequence obtained from the NGS reads of the sample through de novo assembly.

The BAM files obtained from BWA were filtered for sequences with a mapping quality of at least 30. Duplicate reads that showed identical start and end coordinates were removed using DeDup. These tools were also integrated into the EAGER-pipeline [40].

The quality of mitogenome mapping was assessed by the mapping ratio, average coverage and run time. The mapping ratio refers to the ratio of reads mappable to the mitochondrial reference to all clean reads. Average coverage refers to the number of times the mitogenome is sequenced. Runtime refers to CPU time consumption during the mapping processing, instead of elapsed time, including waiting for input/output operations or entering low-power mode.

#### De novo assembly for SS-ref construction

To produce the optimal SS-ref, three modified de novo assembly strategies were compared based on the NGS data from A1. They were different in the NGS read sets, including all clean read sets, homologous read sets filtered by BLAST, or BWA mapping. De novo assembly was performed using SOAPdenovo2 [22] with default parameters with the best k-mer size estimated by KmerGennie [23]. The detailed assembly information was as follows.

- (1) “Denovo”: the de novo assembly directly from all clean reads [23, 35]. All clean data from NGS were put into SOAPdenovo2 and contigs assembled. Then these contigs were aligned to NC\_000845.1 using MEGA6, and a consensus inferred [35].
- (2) “BLAST\_denovo”: the de novo assembly by homologous read sets filtered from clean data by BLAST. Clean data were filtered against a reference panel composed of all complete *Sus Scrofa* mitogenome sequences (219) downloaded from the NCBI database by the BLAST tool with the blastn command, and then these sets were put into SOAPdenovo2 for de novo assembly.
- (3) “BWA\_denovo”: the de novo assembly by homologous read sets filtered from clean data by BWA. Clean reads were mapped against the above-mentioned reference panel to filter homologous sequences of each sample by BWA, and then these sequences were assembled by SOAPdenovo2.

To assess the three de novo assembly strategies, the indicators including the N50, consensus length, coverage and sequence polymorphism resulted from each strategy were measured.

#### SNP calling of mitochondrial genomes

Three variation callers, i.e., SAMtools 1.3.1 [24], VarScan 2.3.9 [25] and GATK 3.7 [26], were applied with different parameter combinations detailed in Table 2 to the bam files resulting from the mitogenome alignments. These parameters were selected to ensure comparability among different callers. The

minimum base quality required to consider a base for calling was set to 30.

The performance of SNP calling was evaluated using the overall genotype concordance by comparing the NGS results with the Sanger data, with the assumption that the Sanger sequencing gave the correct calling [42–44]. Only positions where a Sanger sequence was available were kept, and SNPs concordant between Sanger and NGS data for each individual were considered as true SNPs, while discrepancies were considered as errors. When NGS data identified an alternate homozygote not observed by Sanger sequence, it was considered as a false positive. Accordingly, when NGS data did not see an alternate homozygote found with Sanger sequence, it was considered as a false negative. The number of true SNPs, false positives and false negatives were analysed.

#### Abbreviations

NGS: Next-generation sequencing; Mitogenome: Mitochondrial genome; MtDNA: Mitochondrial DNA; SNP: Single nucleotide polymorphism; NUMTS: Nuclear mitochondrial sequences; HBV: Hepatitis B virus

#### Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s12864-021-08030-1>.

**Additional file 1: Table S1.** Sample information on whole-genome sequencing data.

**Additional file 2: Table S2.** Statistics of mapping ratio, coverage and CPU time cost for mitogenome mapping of NGS data.

**Additional file 3: Table S3.** SNP information resulting from Sanger sequences aligning to the breed-specific references.

**Additional file 4: Table S4.** The results of SNP calling for NGS data.

**Additional file 5: Table S5.** Polymorphic loci among the three kinds of reference sequences.

#### Acknowledgements

The authors appreciated Prof. Michael Hofreiter, University of Potsdam, for proofreading and polishing the manuscript.

#### Authors' contributions

X.Z. designed the study. D.W. and J.L. collected the samples and provided data. D.W., L.Y. and C.N. analysed data, and X.Z., D.W., L.Y., C.N. and J.L. wrote the manuscript. All authors read and approved the final manuscript.

#### Funding

Sample collection and sequencing were funded by the National Natural Science Foundation of China-Deutsche Forschungsgemeinschaft (31961133031). Data analyses were funded by the National Natural Science Foundation of China (32102526 and 32002172), China Postdoctoral Science Foundation (2020M682217), Shandong Provincial Postdoctoral Program for Innovative Talent, and Shandong Provincial Natural Science Foundation (ZR2020QC176 and ZR2020QC175).

#### Availability of data and materials

The datasets analysed during the current study are available in the NCBI repository under the accession number PRJNA378496 including SRA544899 (A1), SRA544150 (D1), SRA544170 (D2) and SRA544142 (T1).

#### Declarations

##### Ethics approval and consent to participate

All experimental pigs were maintained according to the guidelines of the experimental animal management of China Agricultural University. The animal management and the experimental protocols complied with the guidelines approved by the Institutional Animal Care and Use Ethics Committee (IACUC) at the China Agricultural University. After the study, the pigs were still living in the original environment.

##### Consent for publication

Not applicable.

##### Competing interests

The authors declare that they have no competing interests.

##### Author details

<sup>1</sup>National Engineering Laboratory for Animal Breeding, Ministry of Agricultural Key Laboratory of Animal Genetics, Breeding and Reproduction, College of Animal Science and Technology, China Agricultural University, Beijing, China. <sup>2</sup>College of Animal Science and Technology, Shandong Agricultural University, Tai'an, China.

Received: 7 October 2020 Accepted: 22 September 2021

Published online: 12 October 2021

#### References

- Hasan MS, Wu X, Zhang L. Performance evaluation of indel calling tools using real short-read data. *Hum Genomics*. 2015;9(1):20. <https://doi.org/10.1186/s40246-015-0042-2>.
- Footo AD, Morin PA, Durban JW, Pitman RL, Wade P, Willerslev E, et al. Positive selection on the killer whale mitogenome. *Biol Lett*. 2011;7(1):116–8. <https://doi.org/10.1098/rsbl.2010.0638>.
- Tan Y, Shi KZ, Wang J, Du CL ZCP, Zhang X, et al. The complete mitochondrial genome of Kele pig (*Sus scrofa*) using next-generation deep sequencing. *Conserv Genet Resour*. 2018;10(2):195–9. <https://doi.org/10.1007/s12686-017-0797-y>.
- Singh AP, Jadav KK, Kumar D, Rajput N, Srivastav AB, Sarkhel BC. Complete mitochondrial genome sequencing of central Indian domestic pig. *Mitochondrial DNA Part B*. 2016;1(1):949–50. <https://doi.org/10.1080/23802359.2016.1197077>.
- Lord E, Collins C, deFrance S, LeFebvre MJ, Matisoo-Smith E. Complete mitogenomes of ancient Caribbean Guinea pigs (*Cavia porcellus*). *J Archaeol Sci Rep*. 2018;17:678–88. <https://doi.org/10.1016/j.jasrep.2017.12.004>.
- Mao H, Zhao G, Guan Y, Guo X, Lamaocao Z. The complete mitochondrial genome of Juema pig (*Suina: Suidae*). *Conserv Genet Resour*. 2018;10(3):1–3. <https://doi.org/10.1007/s12686-017-0865-3>.
- Kwak W, Song K-D, Oh J-D, Heo K-N, Lee J-H, Lee WK, et al. Uncovering genomic features and maternal origin of Korean native chicken by whole genome sequencing. *PLoS One*. 2014;9(12):e114763. <https://doi.org/10.1371/journal.pone.0114763>.
- Suwannapoom C, Wu Y-J, Chen X, Adeola AC, Chen J, Wang W-Z. Complete mitochondrial genome of the Thai red Junglefowl (*Gallus gallus*) and phylogenetic analysis. *Zool Res*. 2018;39(2):127–9. <https://doi.org/10.24272/j.issn.2095-8137.2017.028>.
- Huang X-H, Li G-M, Chen X, Wu Y-J, Li W-N, Zhong F-S, et al. Identification of a novel mtDNA lineage B3 in chicken (*Gallus gallus domesticus*). *Zool Res*. 2017;38(4):208–10. <https://doi.org/10.24272/j.issn.2095-8137.2017.039>.
- Guo X, Ding X, Wu X, Bao P, Xiong L, Yan P, et al. Complete mitochondrial genome of Anxi cattle (*Bos taurus*). *Conserv Genet Resour*. 2018;10(3):393–5. <https://doi.org/10.1007/s12686-017-0833-y>.
- Guo X, Pei J, Xiong L, Bao P, Zhu Y, Wangdui B, et al. The complete mitochondrial genome of Shigaste humped cattle (*Bos taurus*). *Conserv Genet Resour*. 2018;10(4):789–91. <https://doi.org/10.1007/s12686-017-0931-x>.
- Pramod RK, Velayutham D, Zachariah A, Zachariah A, Dhinoth Kumar B, Iype S, et al. Complete mitogenome reveals genetic divergence and phylogenetic relationships among Indian cattle (*Bos indicus*) breeds. *Anim Biotechnol*. 2019;30:1–14.

13. Kushwaha B, Kumar R, Agarwal S, Pandey M, Nagpure NS, Singh M, et al. Assembly and variation analyses of *Clarias batrachus* mitogenome retrieved from WGS data and its phylogenetic relationship with other catfishes. *Meta Gene*. 2015;5:105–14. <https://doi.org/10.1016/j.mgene.2015.06.004>.
14. Liu WC, Lin CP, Cheng CP, Ho CH, Lan KL, Cheng JH, et al. Aligning to the sample-specific reference sequence to optimize the accuracy of next-generation sequencing analysis for hepatitis B virus. *Hepatol Int*. 2016;10(1):147–57. <https://doi.org/10.1007/s12072-015-9645-x>.
15. Yuan S, Qin Z. Read-mapping using personalized diploid reference genome for RNA sequencing data reduced bias for detecting allele-specific expression. 2012 IEEE International Conference on Bioinformatics and Biomedicine Workshops: IEEE; 2012. p. 718–24.
16. Gerdol M, Moreira R, Cruz F, Gómez-Garrido J, Vlasova A, Rosani U, et al. Massive gene presence-absence variation shapes an open pangenome in the Mediterranean mussel. *Genome Biol*. 2020;21(1):1–21.
17. Li M, Chen L, Tian S, Lin Y, Tang Q, Zhou X, et al. Comprehensive variation discovery and recovery of missing sequence in the pig genome using multiple de novo assemblies. *Genome Res*. 2017;27(5):865–74. <https://doi.org/10.1101/gr.207456.116>.
18. O'Leary NA, Wright MW, Brister JR, Ciuffo S, Haddad D, McVeigh R, et al. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res*. 2015;44(D1):D733–D45. <https://doi.org/10.1093/nar/gkv1189>.
19. Liu H, Shi W, Wang D, Zhao X. Association analysis of mitochondrial DNA polymorphisms with oocyte number in pigs. *Reprod Fertil Dev*. 2019;31(4):805–9. <https://doi.org/10.1071/RD18219>.
20. Liu H, Wang J, Wang D, Kong M, Ning C, Zhang X, et al. Cybrid model supports mitochondrial genetic effect on pig litter size. *Front Genet*. 2020;11:579382. <https://doi.org/10.3389/fgene.2020.579382>.
21. Wang D, Ning C, Xiang H, Zheng X, Kong M, Yin T, et al. Polymorphism of mitochondrial tRNA genes associated with the number of pigs born alive. *J Anim Sci Biotechnol*. 2018;9(1):86. <https://doi.org/10.1186/s40104-018-0299-0>.
22. Luo R, Liu B, Xie Y, Li Z, Huang W, Yuan J, et al. SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *Gigascience*. 2012;1(1):2047–217X-1-18. <https://doi.org/10.1186/2047-217X-1-18>.
23. Chikhi R, Medvedev P. Informed and automated k-mer size selection for genome assembly. *Bioinformatics*. 2014;30(1):31–7. <https://doi.org/10.1093/bioinformatics/btt310>.
24. Li H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics*. 2011;27(21):2987–93. <https://doi.org/10.1093/bioinformatics/btr509>.
25. Koboldt DC, Zhang Q, Larson DE, Shen D, McLellan MD, Lin L, et al. VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res*. 2012;22(3):568–76. <https://doi.org/10.1101/gr.129684.111>.
26. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytzky A, et al. The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res*. 2010;20(9):1297–303. <https://doi.org/10.1101/gr.107524.110>.
27. Machado D, Lyra M, Grant T. Mitogenome assembly from genomic multiplex libraries: comparison of strategies and novel mitogenomes for five species of frogs. *Mol Ecol Resour*. 2016;16(3):686–93. <https://doi.org/10.1111/1755-0998.12492>.
28. Dubchak I, Poliakov A, Kislyuk A, Brudno M. Multiple whole-genome alignments without a reference organism. *Genome Res*. 2009;19(4):682–9. <https://doi.org/10.1101/gr.081778.108>.
29. Angiuoli SV, Salzberg SL. Mugsy: fast multiple alignment of closely related whole genomes. *Bioinformatics*. 2016;27(3):334.
30. Wang D, Xiang H, Ning C, Liu H, Liu J-F, Zhao X. Mitochondrial DNA enrichment reduced NUMT contamination in porcine NGS analyses. *Brief Bioinform*. 2020;21(4):1368–77. <https://doi.org/10.1093/bib/bbz060>.
31. Charre C, Ginevra C, Sabatier M, Regue H, Destras G, Brun S, et al. Evaluation of NGS-based approaches for SARS-CoV-2 whole genome characterisation. *Virus Evol*. 2020;6(2):veaa075.
32. Chen X, Wang D, Xiang H, Dun W, Brahi DO, Yin T, et al. Mitochondrial DNA T7719G in tRNA-Lys gene affects litter size in small-tailed Han sheep. *J Anim Sci Biotechnol*. 2017;8(1):31. <https://doi.org/10.1186/s40104-017-0160-x>.
33. Lee RS, Behr MA. Does choice matter? Reference-based alignment for molecular epidemiology of tuberculosis. *J Clin Microbiol*. 2016;54(7):1891–5. <https://doi.org/10.1128/JCM.00364-16>.
34. Kumar S, Stecher G, Tamura K. MEGA7: molecular evolutionary genetics analysis version 7.0 for bigger datasets. *Mol Biol Evol*. 2016;33(7):1870–4. <https://doi.org/10.1093/molbev/msw054>.
35. Cronn R, Liston A, Parks M, Gernandt DS, Shen R, Mockler T. Multiplex sequencing of plant chloroplast genomes using Solexa sequencing-by-synthesis technology. *Nucleic Acids Res*. 2008;36(19):e122-e.
36. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *J Mol Biol*. 1990;215(3):403–10. [https://doi.org/10.1016/S0022-836\(05\)80360-2](https://doi.org/10.1016/S0022-836(05)80360-2).
37. Niu B, Zhu Z, Fu L, Wu S, Li W. FR-HIT, a very fast program to recruit metagenomic reads to homologous reference genomes. *Bioinformatics*. 2011;27(12):1704–5. <https://doi.org/10.1093/bioinformatics/btr252>.
38. Tamura K, Stecher G, Peterson D, Filipiński A, Kumar S. MEGA6: molecular evolutionary genetics analysis version 6.0. *Mol Biol Evol*. 2013;30(12):2725–9. <https://doi.org/10.1093/molbev/mst197>.
39. Librado P, Rozas J. DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. *Bioinformatics*. 2009;25(11):1451–2. <https://doi.org/10.1093/bioinformatics/btp187>.
40. Peltzer A, Jäger G, Herbig A, Seitz A, Knip C, Krause J, et al. EAGER: efficient ancient genome reconstruction. *Genome Biol*. 2016;17(1):60. <https://doi.org/10.1186/s13059-016-0918-z>.
41. Li H, Durbin R. Fast and accurate short read alignment with burrows-wheeler transform. *Bioinformatics*. 2009;25(14):1754–60. <https://doi.org/10.1093/bioinformatics/btp324>.
42. Scarcelli N, Mariac C, Couvreur T, Faye A, Richard D, Sabot F, et al. Intra-individual polymorphism in chloroplasts from NGS data: where does it come from and how to handle it? *Mol Ecol Resour*. 2016;16(2):434–45. <https://doi.org/10.1111/1755-0998.12462>.
43. Alexander A, Steel D, Slikas B, Hoekzema K, Carraher C, Parks M, et al. Low diversity in the mitogenome of sperm whales revealed by next-generation sequencing. *Genome Biol Evol*. 2013;5(1):113–29. <https://doi.org/10.1093/gbe/evs126>.
44. Elsensohn M, Leblay N, Dimassi S, Campan-Fournier A, Labalme A, Roucher-Boulez F, et al. Statistical method to compare massive parallel sequencing pipelines. *BMC Bioinformatics*. 2017;18(1):1–11. <https://doi.org/10.1186/s12859-017-1552-9>.

## Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

**Ready to submit your research? Choose BMC and benefit from:**

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

**At BMC, research is always in progress.**

Learn more [biomedcentral.com/submissions](https://biomedcentral.com/submissions)

