ISSN: 2162-4453 Volume 7, Number 1, March 2017

Scientific Research Publishing

Open Journal of Genetics

ISSN: 2162-4453

www.scirp.org/journal/ojgen

Journal Editorial Board

ISSN: 2162-4453 (Print), 2162-4461 (Online)
http://www.scirp.org/journal/ojgen

Editor-in-Chief

Editorial Board

Scientific Research Publishing

Open Journal of Genetics, 2017, 7, 1-67
http://www.scirp.org/journal/ojgen
ISSN Online: 2162-4461
ISSN Print: 2162-4453

Table of Contents

Volume 7 Number 1 March 2017

Open Journal of Genetics (OJGen)
Journal Information

SUBSCRIPTIONS

The *Open Journal of Genetics* (Online at Scientific Research Publishing, www.SciRP.org) is published quarterly by Scientific Research Publishing, Inc., USA.

Subscription rates:
Print: $79 per issue.
To subscribe, please contact Journals Subscriptions Department, E-mail: sub@scirp.org

SERVICES

Advertisements
Advertisement Sales Department, E-mail: service@scirp.org

Reprints (minimum quantity 100 copies)
Reprints Co-ordinator, Scientific Research Publishing, Inc., USA.
E-mail: sub@scirp.org

COPYRIGHT

PRODUCTION INFORMATION

For manuscripts that have been accepted for publication, please contact:
E-mail: ojgen@scirp.org

Scientific Research Publishing

Open Journal of Genetics, 2017, 7, 1-8
http://www.scirp.org/journal/ojgen
ISSN Online: 2162-4461
ISSN Print: 2162-4453

Multiple z-Score Based Method for Noninvasive Prenatal Test Using Cell-Free DNA in Maternal Plasma

Hyuk Jung Kwon, Amit Goyal, Heesu Im, Kichan Lee, Seon Young Yun, Yoon Hee Kim, Sungjong Lee, Mi-Gyeong Lee, Hyuna Lee, Reena Garg, Boram Park, Soyoung Choi, Joungsu Joo, Jin-Sik Bae, Min-Jeong Kim, Min Seob Lee, Sunghoon Lee*

EONE-DIAGNOMICS Genome Center Co. Ltd., Inchon, Korea
Email: *shlee@edgc.com

How to cite this paper: Kwon, H.J., Goyal, A., Im, H., Lee, K., Yun, S.Y., Kim, Y.H., Lee, S., Lee, M.-G., Lee, H., Garg, R., Park, B., Choi, S., Joo, J., Bae, J.-S., Kim, M.-J., Lee, M.S. and Lee, S. (2017) Multiple z-Score Based Method for Noninvasive Prenatal Test Using Cell-Free DNA in Maternal Plasma. *Open Journal of Genetics*, **7**, 1-8.
https://doi.org/10.4236/ojgen.2017.71001

Received: January 1, 2017
Accepted: February 4, 2017
Published: February 7, 2017

Abstract

Objective: To improve the detecting accuracy of chromosomal aneuploidy of fetus by non-invasive prenatal testing (NIPT) using next generation sequencing data of pregnant women's cell-free DNA. **Methods:** We proposed the multi-Z method which uses 21 z-scores for each autosomal chromosome to detect aneuploidy of the chromosome, while the conventional NIPT method uses only one z-score. To do this, mapped read numbers of a certain chromosome were normalized by those of the other 21 chromosomes. Average and standard deviation (SD), which are used for calculating z-score of each sample, were obtained with normalized values between all autosomal chromosomes of control samples. In this way, multiple z-scores can be calculated for 21 autosomal chromosomes except oneself. **Results:** Multi-Z method showed 100% sensitivity and specificity for 187 samples sequenced to 3 M reads while the conventional NIPT method showed 95.1% specificity. Similarly, for 216 samples sequenced to 1 M reads, Multi-Z method showed 100% sensitivity and 95.6% specificity and the conventional NIPT method showed a result of 75.1% specificity. **Conclusion:** Multi-Z method showed higher accuracy and robust results than the conventional method even at low coverage reads.

Keywords

Cell-Free DNA, z-Score, Multiple Thresholds, Coefficient of Variance, Noninvasive Prenatal Testing, NIPT

1. Introduction

The most common chromosomal aneuploidy for a new born infant is Trisomy 21. The overall occurrence of trisomy 21 is around 0.001%, but the risk increases

up to 0.02% for women above 45 years old [1] [2] [3]. Traditional methods of prenatal screening for fetal aneuploidy have a high miscarriage risk since it involves invasive sampling. Recent technology advancement in Next Generation Sequencing (NGS) and Bioinformatics led to a novel Non-Invasive Prenatal Test (NIPT) method to analyze fetus aneuploidy using cell-free DNA (cfDNA) in the plasma of pregnant women. This NIPT has been shown to be both highly sensitive and highly specific across numerous studies [4]. The whole chromosome analysis uses massive parallel sequencing data and applies statistical normalization of each chromosome read count. Sequence reads are mapped to the human reference genome and quantified according to their genomic locus. After normalizing the read count, one z-score per chromosome was calculated to determine fetal aneuploidy [5]. Most of published NIPT studies rely on the z-score which represents the quantitative variations of the chromosome of interest and they show the results as positive or negative by checking if the z-score exceeds the predefined threshold [6].

Even though these methods become highly accurate, they still have a 0.1% possibility of reporting false positives and need enough read count to score the high sensitivity [7] [8] [9].

While many NIPT methods have been developed and introduced, studies are under way to increase accuracy. Sunshin Kim *et al*. [10], for example, introduced a new algorithm based on selecting reference samples adaptively using CV according to the shared ranges of GC content and DNA reads fraction. They showed reliable results within GC (0.424 ± 0.001) for 7.4 ± 2.1 million raw reads, but the insufficient, yet large sample size for selecting reference samples is a concerning issue.

In order to save sequencing cost and time, some research is being conducted with less reads, for example, Lau, T.K. *et al*. [11], reported the clinical performance of NIPT based on low-coverage whole-genome sequencing as 0.1× on average with approximately 300 bp which produces the minimal amount of unique sequencing reads less than 3.5 million.

In this study, we devised a new algorithm which uses multiple z-scores to determine the fetal aneuploidy. Multi-Z algorithm shows 100% sensitivity and specificity for 3 M-reads samples, and 100% sensitivity and 95.6% specificity for 1 M-reads samples.

2. Method

2.1. cfDNA Sequencing

About 10 mL of blood was collected from 216 pregnancies into a cfDNA Vangenes Cell Free DNA Tubes (Vangenes, US) and centrifuged at 1900 g for 15 min at RT. The plasma was transferred into 1.5 mL tubes and then centrifuged at 16,000 g for 15 min at RT. The separated plasma was transferred to 5 mL Cryogenic Tube and stored at −40˚C. cfDNA was isolated from 2 mL plasma by using the QIAsymphony DSPvirus/Pathogen midi kit (Qiagen, Germany) according to the manufacturer's instructions. Ion Proton sequencing libraries were prepared

by using cfDNA samples (<100 ng) according to the manufacturer's instructions (Life Technology, US). Ion PI™ Chip kit v3 was used to yield an average 7 million and 1 million sequencing reads for nucleotide.

2.2. Data Analysis

The DNA fragments were mapped to the human reference genome sequence (hg19) by using BWA 0.7.10 [12] and duplicated DNA reads were removed by using Picard 1.81 [13]. Finally, uniquely aligned counts for each chromosome were calculated by Samtools 0.1.18 [14].

We generated artificial samples consist of 1 M and 3 M sequence reads by randomly selecting 1 million and 3 million reads from the samples sequenced up to 7 M reads, respectively. In this study, we calculated the multiple z-score for each chromosome of interest in two steps. First dividing the read count of the chromosome of interest to the all remaining autosomal chromosome one by one to get normalized read count as shown in **Figure 1**, and then, z-score calculation using Equation (1), by utilizing the average and SD obtained from sample dataset which are composed of 200 normal samples. For example, if chr13 is our chromosome of interest, read count of chromosome 13 is divided by chr1, chr2, chr3, etc. except the chr13. So, it will yield 21 normalized read count for the chromosome 13. Now using the average and SD for each normalized read count in sample dataset, we developed the multiple z-scores for the chromosome 13, *i.e.* $Zscore_{13,1}$, $Zscore_{13,2}$, $Zscore_{13,3}$, etc. Similarly, multiple z-scores were developed for each of the autosomal chromosome.

We also calculated the CV through the calculated SD in sample dataset and it is used to apply the order of z-score thresholds gradually lowest to highest, e.g. chr7 is the lowest and chr12, chr14, chr9, chr11, followed in ascending order.

Multiple z-scores are calculated for normal samples using the reference mean and SD as follows:

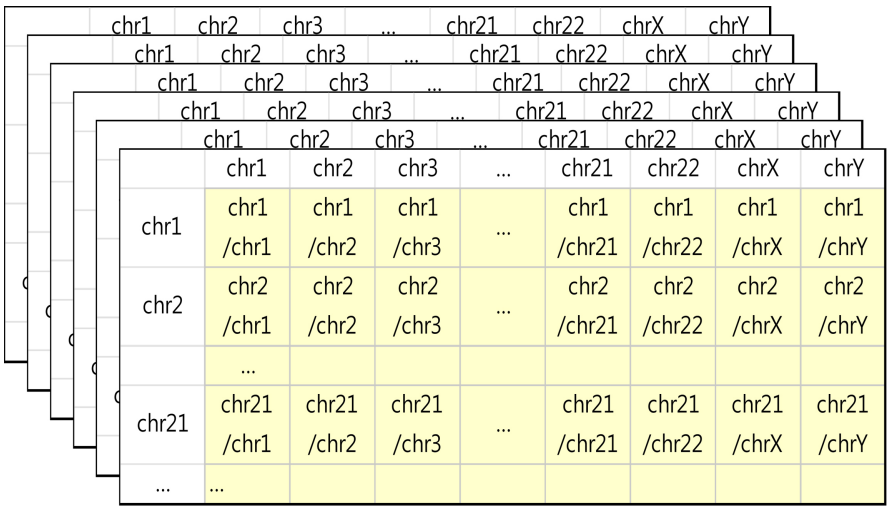

Figure 1. Normalization between chromosomes. Normalized value between two chromosomes is calculated by dividing the value of interested chromosome by that of each chromosome.

$$\text{Zscore}_{i,j} = \frac{\left(\text{ratioof}\left(\frac{\text{chri}}{\text{chrj}}\right)_{\text{normal}} - \text{mean}\left(\frac{\text{chri}}{\text{chrj}}\right)_{\text{reference}}\right)}{\text{SD}\left(\frac{\text{chri}}{\text{chrj}}\right)_{\text{reference}}} \quad (1)$$

The equation is a multiple z-scores calculation. In this equation, i is for chromosome of interest and j is for other chromosomes to normalize.

Now, Multi-Z algorithm classifies negative versus positive case for test samples by applying the smallest z-score of a certain chromosome. Since we used 70 samples as thresholds in this paper, there are 70 z-scores for a chromosome of interest and the smallest z-score among them is chosen as an applicable threshold. Repeatedly, Multi-Z algorithm applies the next threshold of another chromosome according to the ascending order of CV.

3. Results

The algorithm was tested on a cohort of 216 pregnant women including seven T21 and we found the Multi-Z algorithm produced reliable results with higher sensitivity and specificity. Up to 70 control samples which had undergone the amniocentesis were used for determining thresholds. We generated 70 randomly selected samples from seven trisomy samples to make robust thresholds by increasing number of control samples.

As for 3 M-reads samples, the amount of unique sequencing reads is around 2 million and the average length of sequenced read is around 160 bp that can be calculated approximately 0.1× low coverage depth. The uniquely mapped read for 1 M samples is 0.7 M and the coverage depth is 0.035× as same calculation as 3 M-reads.

In this paper, we used the term 3 M-reads, 1 M-reads as the amount of sequenced reads

In Figure 2, the results of Multi-Z for 187 normal samples randomly generated at 3 M read length (black dots) with 70 control samples (white dots) show 100% specificity at only seven thresholds as listed in Table 1. The red-dashed line means the least z-score of interested chromosome and this line classifies the normal samples as trisomy or normal while repeating the chromosomes in interest. The number of normal samples is decreased as we repeat applying the algorithm and all the normal samples, finally, are categorized as normal after applying chromosome 6 which is 7th threshold.

In Figure 3, we added 29 samples which were experimentally sequenced at 1 million reads, represented by red dots. Seven of these samples among 216 samples were judged to be trisomy by using eight thresholds without any false positives, while nine false positive samples were found in 209 normal samples resulting in 95.6% specificity as listed Table 1.

In Figure 4, we found the conventional NIPT detection algorithm shows lower specificity compared to the multi-dimensional detecting algorithm. In this figure, we can find the only one threshold as red-dashed line which is the smallest z-score of 70 control samples. There are 9 falsely detected samples as trisomy

Table 1. Accuracy comparison between the conventional NIPT and Multi-Z.

Method	#Sample	#TP	#FP	#TN	#FN	Sensitivity	Specificity
conventional NIPT							
3 M-reads	187	N/A	9	178	N/A	N/A	95.1
1 M-reads	216	7	52	157	0	100%	75.1
Multi-Z							
3 M-reads	187	N/A	0	187	N/A	N/A	100%
1 M-reads	216	7	9	200	0	100%	95.6%

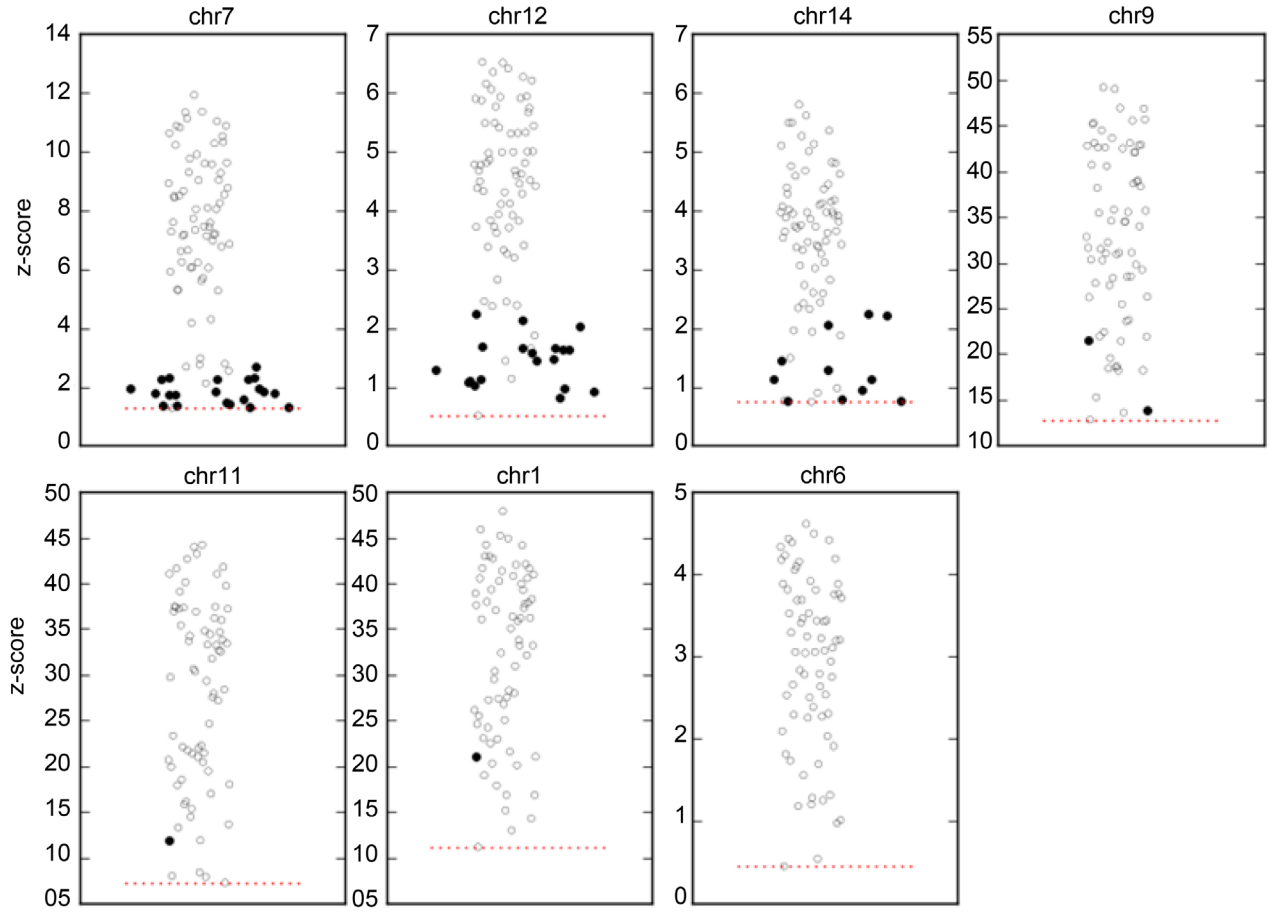

Figure 2. 3 M-reads by Multi-Z. Every normal sample (close black dots) is classified as normal by using seven thresholds of control sample (open black dots). Red dashed line in each figure represents the smallest z-score for applied threshold of control samples. Samples remained above red dashed line are considered as trisomy.

for 3 M-reads samples which yield 95.1% Specificity as listed in Table 1. As for 1 M-reads samples, we found 52 false positive samples and it shows the worse specificity as 75.1% in Table 1. This result shows the conventional NIPT method is not applicable for low reads samples.

4. Discussion

We compared the sensitivity and the specificity between the conventional NIPT

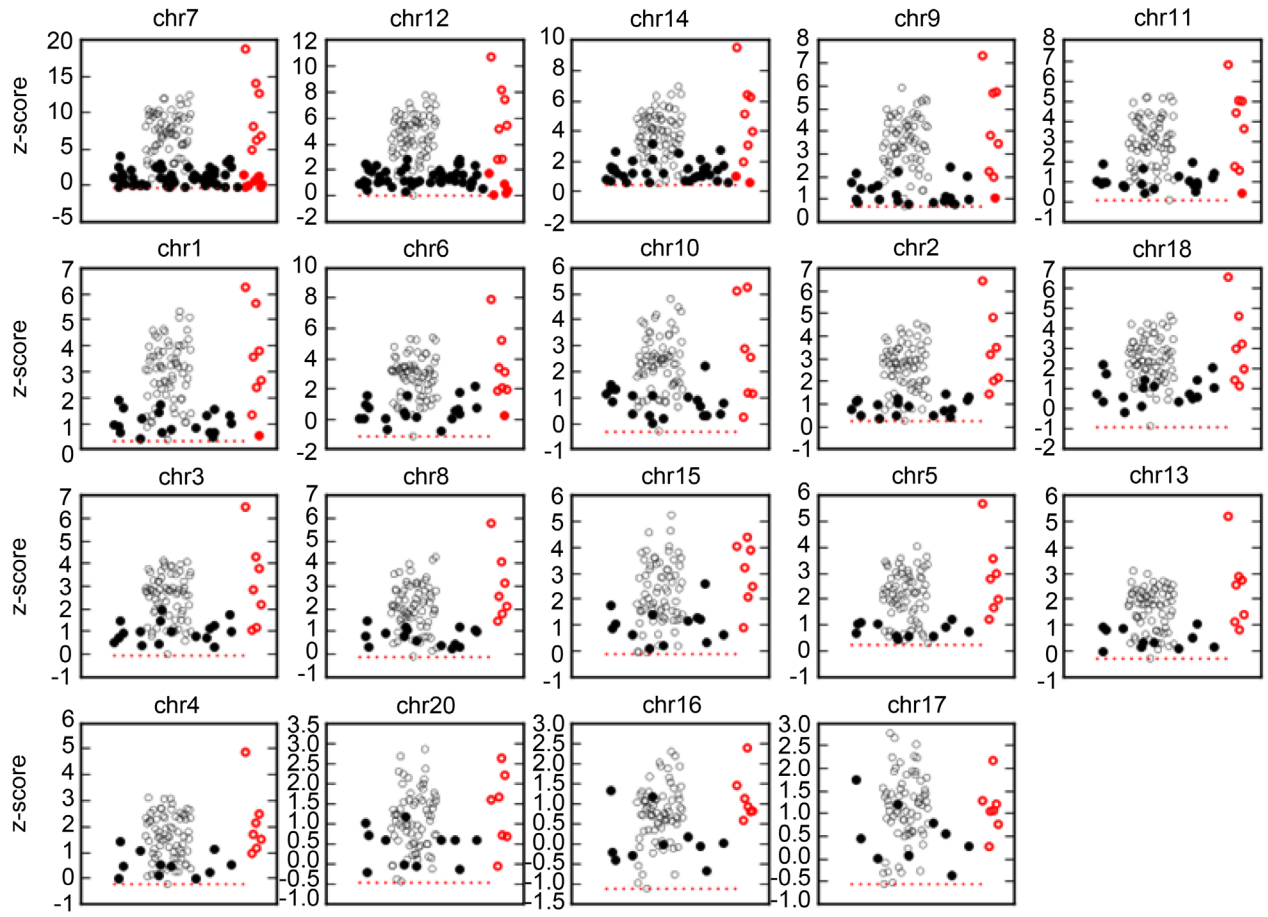

Figure 3. 1 M-reads by Multi-Z. Seven of 1M sequencing samples (red dots) are judged as trisomy, and 9 normal samples (close black dots) are defined as trisomy (False Positive) by using 19 z-scores. Red dashed line in each figure represents the smallest z-score for applied threshold of control samples.

and Multi-Z algorithm for 1 M-reads and 3 M-reads. The results showed that Multi-Z algorithm can be used for low coverage cfDNA NIPT samples with higher specificity. When applying single Z-score threshold, it is often difficult to distinguish between the anueploid and the euploid, especially when the z-score is close to borderline, besides it may cause false positives call for some samples. We observed that the use of Multi Z-score approach for such ambiguous samples, using the correlation between specific chromosomes, showed a higher z-score and the same tendency was observed for all our test samples. Therefore, we concluded that more thresholds can be used through correlation with other chromosomes, and confirmed that it reduced the possibility of false positives. Also, we found the proper order of applying thresholds according to the CV.

Since the Multi-Z can be applied to low reads samples, it has potential competitive advantages such as reduction of experiment cost and the rapid analysis in the emerging NIPT market. A major concern in this study was to determine the precise number of thresholds to call the aneuploidy, *i.e.* how many z-score thresholds would be effective to detect Trisomy 21 and how to decide borderline range with selected thresholds. If the number of applied thresholds are increased, it may cause false negative and *vice versa*. Although, this method can

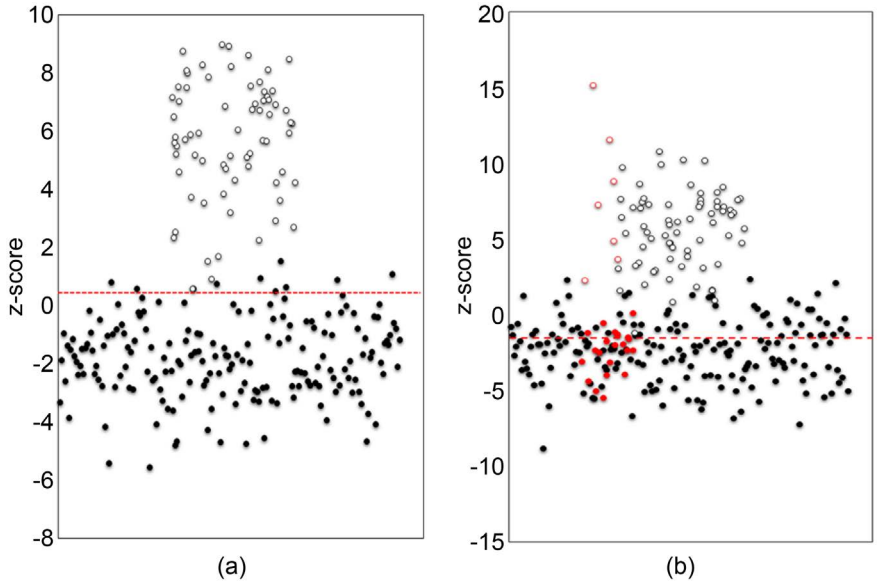

Figure 4. Accuracy of previous NIPT method at low read sequencing data. Nine and 52 normal samples were detected falsely as aneuploidy in (a) 3 M-reads data and (b) 1 M-reads data, respectively. Open black circles in (a) and (b) represent the trisomy sample's z-scores, and the red dashed line represents the chosen threshold among all z-scores. Closed black circle represents normal samples' z-score. red circles in (b) represent the samples sequenced at 1 million, and open red circles represent seven trisomy samples. Red dashed line represents the threshold which is the smallest z-score among control samples and we assume the samples above the red line as trisomy.

detect all the chromosome 21 aneuploidies accurately in our current dataset, a large number of samples would be required to validate our approach further.

5. Conclusion

In conclusion, we confirmed that the Multi-Z method provides an optimal way to reduce false positives of T21 for low coverage samples compared to the conventional NIPT algorithm. It is expected that detecting T13, T18 and sex chromosome aneuploidy can be applied in the same way after the research on T21 is completed.

References

[1] Savva, G.M., Morris, J.K., Mutton, D.E. and Alberman, E. (2006) Maternal Age-Specific Fetal Loss Rates in Down Syndrome Pregnancies. *Prenatal Diagnosis*, **26**, 499-504. https://doi.org/10.1002/pd.1443

[2] Morris, J.K., Mutton, D.E. and Alberman, E. (2002) Revised Estimates of the Maternal Age Specific Live Birth Prevalence of Down's Syndrome. *Journal of Medical Screening*, **9**, 2-6. https://doi.org/10.1136/jms.9.1.2

[3] Haddow, J.E. (1990) Prenatal Screening for Open Neural Tube Defects, Down's Syndrome, and Other Major Fetal Disorders. *Seminars in Perinatology*, **14**, 488-503.

[4] Bianchi, D.W., Parker, R.L., Wentworth, J., Madankumar, R., Saffer, C., Das, A.F., Craig, J.A., Chudova, D.I., Devers, P.L., Jones, K.W., Oliver, K., Rava, R.P., Sehnert, A.J. and Group, C.S. (2014) DNA Sequencing versus Standard Prenatal Aneuploidy Screening. *New England Journal of Medicine*, **370**, 799-808.

https://doi.org/10.1056/NEJMoa1311037

[5] Fan, H.C., Blumenfeld, Y.J., Chitkara, U., Hudgins, L. and Quake, S.R. (2008) Noninvasive Diagnosis of Fetal Aneuploidy by Shotgun Sequencing DNA from Maternal Blood. *Proceedings of the National Academy of Sciences of the USA*, **105**, 16266-16271. https://doi.org/10.1073/pnas.0808319105

[6] Jiang, F., Ren, J., Chen, F., Zhou, Y., Xie, J., Dan, S., Su, Y., Xie, J., Yin, B., Su, W., Zhang, H., Wang, W., Chai, X., Lin, L., Guo, H., Li, Q., Li, P., Yuan, Y., Pan, X., Li, Y., Liu, L., Chen, H., Xuan, Z., Chen, S., Zhang, C., Zhang, H., Tian, Z., Zhang, Z., Jiang, H., Zhao, L., Zheng, W., Li, S., Li, Y., Wang, J., Wang, J. and Zhang, X. (2012) Noninvasive Fetal Trisomy (NIFTY) Test: An Advanced Noninvasive Prenatal Diagnosis Methodology for Fetal Autosomal and Sex Chromosomal Aneuploidies. *BMC Medical Genomics*, **5**, 57. https://doi.org/10.1186/1755-8794-5-57

[7] Lo, Y.M., Corbetta, N., Chamberlain, P.F., Rai, V., Sargent, I.L., Redman, C.W. and Wainscoat, J.S. (1997) Presence of Fetal DNA in Maternal Plasma and Serum. *Lancet*, **350**, 485-487. https://doi.org/10.1016/S0140-6736(97)02174-0

[8] Chiu, R.W., Chan, K.C., Gao, Y., Lau, V.Y., Zheng, W., Leung, T.Y., Foo, C.H., Xie, B., Tsui, N.B., Lun, F.M., Zee, B.C., Lau, T.K., Cantor, C.R. and Lo, Y.M. (2008) Noninvasive Prenatal Diagnosis of Fetal Chromosomal Aneuploidy by Massively Parallel Genomic Sequencing of DNA in Maternal Plasma. *Proceedings of the National Academy of Sciences of the USA*, **105**, 20458-20463. https://doi.org/10.1073/pnas.0810641105

[9] Fairbrother, G., Johnson, S., Musci, T.J. and Song, K. (2013) Clinical Experience of Noninvasive Prenatal Testing with Cell-Free DNA for Fetal Trisomies 21, 18, and 13, in a General Screening Population. *Prenatal Diagnosis*, **33**, 580-583. https://doi.org/10.1002/pd.4092

[10] Kim, S., Jung, H., Han, S.H., Lee, S., Kwon, J., Kim, M.G., Chu, H., Han, K., Kwak, H., Park, S., Joo, H.J., An, M., Ha, J., Lee, K., Kim, B.C., Zheng, H., Zhu, X., Chen, H. and Bhak, J. (2016) An Adaptive Detection Method for Fetal Chromosomal Aneuploidy Using Cell-Free DNA from 447 Korean Women. *BMC Medical Genomics*, **9**, 61. https://doi.org/10.1186/s12920-016-0222-5

[11] Lau, T.K., Cheung, S.W., Lo, P.S., Pursley, A.N., Chan, M.K., Jiang, F., Zhang, H., Wang, W., Jong, L.F., Yuen, O.K., Chan, H.Y., Chan, W.S. and Choy, K.W. (2014) Non-Invasive Prenatal Testing for Fetal Chromosomal Abnormalities by Low-Coverage Whole-Genome Sequencing of Maternal Plasma DNA: Review of 1982 Consecutive Cases in a Single Center. *Ultrasound in Obstetrics & Gynecology*, **43**, 254-264. https://doi.org/10.1002/uog.13277

[12] Li, H. and Durbin, R. (2009) Fast and Accurate Short Read Alignment with Burrows-Wheeler Transform. *Bioinformatics*, **25**, 1754-1760. https://doi.org/10.1093/bioinformatics/btp324

[13] Picard Tools. Broad Institute. http://broadinstitute.github.io/picard/

[14] Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R. and Genome Project Data Processing, S. (2009) The Sequence Alignment/Map Format and SAMtools. *Bioinformatics*, **25**, 2078-2079. https://doi.org/10.1093/bioinformatics/btp352

Scientific Research Publishing

Open Journal of Genetics, 2017, 7, 9-19
http://www.scirp.org/journal/ojgen
ISSN Online: 2162-4461
ISSN Print: 2162-4453

Ultra-Fast Next Generation Human Genome Sequencing Data Processing Using DRAGEN™ Bio-IT Processor for Precision Medicine

Amit Goyal, Hyuk Jung Kwon, Kichan Lee, Reena Garg, Seon Young Yun, Yoon Hee Kim, Sunghoon Lee, Min Seob Lee*

EONE-DIAGNOMICS Genome Center Co. Ltd., Inchon, Korea
Email: *a.goyal@edgc.com

How to cite this paper: Goyal, A., Kwon, H.J., Lee, K., Garg, R., Yun, S.Y., Kim, Y.H., Lee, S. and Lee, M.S. (2017) Ultra-Fast Next Generation Human Genome Sequencing Data Processing Using DRAGEN™ Bio-IT Processor for Precision Medicine. *Open Journal of Genetics*, **7**, 9-19.
https://doi.org/10.4236/ojgen.2017.71002

Received: February 14, 2017
Accepted: March 6, 2017
Published: March 9, 2017

Abstract

Slow speed of the Next-Generation sequencing data analysis, compared to the latest high throughput sequencers such as HiSeq X system, using the current industry standard genome analysis pipeline, has been the major factor of data backlog which limits the real-time use of genomic data for precision medicine. This study demonstrates the DRAGEN Bio-IT Processor as a potential candidate to remove the "Big Data Bottleneck". DRAGEN™ accomplished the variant calling, for ~40× coverage WGS data in as low as ~30 minutes using a single command, achieving the over 50-fold data analysis speed while maintaining the similar or better variant calling accuracy than the standard GATK Best Practices workflow. This systematic comparison provides the faster and efficient NGS data analysis alternative to NGS-based healthcare industries and research institutes to meet the requirement for precision medicine based healthcare.

Keywords

NGS Data Analysis, BWA-GATK, DRAGEN Bio-IT Processor, Genomics, INDEL, Mapping

1. Introduction

With the emergence of the 2nd generation high throughput Next Generation Sequencing (NGS) platforms as well as accurate and consistent identification of the genomic variants, the use of the personal genome sequencing information for the diagnostic and prognostic purpose has become the reality [1] [2]. Furthermore, fast sequencing turnaround time and roughly $1000 NGS whole genome cost is encouraging more institutes and individuals to opt for NGS based

personalized medicine [3]-[8]. However, the "Big Data Bottleneck" is still the largest obstacle to use the NGS-based precision medicine in the real time disease and healthcare management. For instance, high throughput NGS HiSeq X Ten System has around 18,000 humans' whole genome sequencing capacity at 30× genome coverage annually which translates into just ~30 - 40-minute turnaround time for each genome sequencing. The most commonly used Genome Analysis Toolkit (GATK) best practice pipelines requires several hours to several days to analyze one human whole genome sequencing data, depending on the available processors. At commercial level, NGS-based data analysis time can be reduced significantly using the clusters of hundreds or thousands of CPUs. Also, several cloud-based solutions, such as GenomePilot by Appistry [9], etc., to accelerate NGS-data analysis platform to speed-up the analysis has been introduced. However, this conventional cluster approach requires expensive computer system, maintenance and monitoring. Similarly, cloud-based platforms require massive data upload and download which is a limitation/burden for many research institutes and small to medium scale companies, especially in low bandwidth supported countries. Overall, the data processing strategy is not suitable for real time guidance/management of many medical diseases/conditions such as tolerance and rejection monitoring in organ transplant recipients, etc. Considering the routine use of the NGS-based diagnostic and prognostic in clinical setting, the need for the fast turnaround time, easy operation and accurate NGS-based data analysis platform has become prominent.

In this study, we assessed the performance of the world's first bioinformatics processor DRAGEN Bio-IT Processor [10] [11] which is designed to analyze the NGS data. The DRAGEN (Dynamic Read Analysis for Genomics) Processor uses a field-programmable gate array (FPGA), implemented on a PCIe card embedded in a pre-configured server, to provide hardware-accelerated implementations of genome pipeline algorithms, such as BCL conversion, compression, mapping, alignment, sorting, duplicate marking and haplotype variant calling.

This study was carried out in two steps. First, run time performance of the DRAGEN Bio-IT Genome pipelines with the most commonly used GATK's best-practice guidelines were analyzed for the 2 replicates of NA12878 Whole Genome Sequencing (WGS) dataset. Second, the variant calling efficiencies of the two pipelines were evaluated by comparing variants with the GIABv2.19 high confidence (truth) call-set [12] [13]. These studies demonstrate that the employment of the DRAGEN Bio-IT processor decreased the WGS NGS-data analysis time to just ~40 minute while achieving the equivalent or better genotype variant calling accuracy than the standard GATK Best Practices workflow.

2. Methods

2.1. Sequence Data-Set and GIAB Validation Call-Set

Two WGS replicates of the Coriell Cell Repository NA12878 reference sample NA12878 were downloaded from the Garvan NA12878 HiSeqX datasets [18]. These datasets have been sequenced on the Illumina HiSeq X platform using the

Illumina's TruSeq Nano kit using 350 bp inserts. Each dataset contains over 120 GB of fastq data yield, with > 87% bases with quality > Q30. These replicates are sequenced to assess the reproducibility and has been provided freely for research purpose by the The Garvan Institute of Medical Research, DNA nexus and AllSeq.

As a gold standard practice to validate the variant calling platform's performance, the high confidence reference variant calls for the 1000 Genome project individual (sample NA12878), published by the Genome in a Bottle (GIAB) consortium [12] using hg19 coordinates, were utilized. The highly confident variant call-set in the Variant Call Format (NISTIntegratedCalls_14datasets_131103_allcall_UGHapMerge_HetHomVarPASS_VQSRv2.19_2mindatasets_5minYesNoRatio_all_nouncert_excludesimplerep_excludesegdups_excludedecoy_excludeRepSeqSTRs_noCNVs.vcf.gz, GIAB v2.19) as well as the high confidence genomic region file (union13callableMQonlymerged_addcert_nouncert_excludesimplerep_excludesegdups_excludedecoy_excludeRepSeqSTRs_noCNVs_v2.19_2mindatasets_5minYesNoRatio.bed.gz) were downloaded for the validation purpose.

2.2. GATK Best Practices Workflow

GATK Best Practices workflow is used most commonly to analyze the genomic data. The complete best practice pipeline [19] can be basically divided into two phase. First, preprocessing the raw data which includes, alignment the raw fastq data to the hg19 reference genome using mapping by BWA (version 0.7.12-r1039) [20], sorting by samtools (version 1.2 using htslib 1.2.1) [21], MarkDuplicate and addRG by steps using picard-tools (version 1.119), and Base Recalibration using GATK (version 3.6-0-g89b7209) [19]. Second, Variants calling using GATK HaplotypeCaller. This study followed the GATK best practice workflow recommended commands and arguments at each step which were executed on 48 core (using-nt and -nct arguments) the Intel Xeon E5-2697v2 12C server with 2.7 GHz processors,128 GB RAM and 3.2 TB capacity SSD running on CentOS 6.6.

2.3. DRAGEN Bio-IT Processor and DRAGEN Genome Pipelines

Unlike the traditional Genome analysis workflows, DRAGEN Bio-IT processor is the hardware accelerated platform which comes equipped with a custom Peripheral Component Interconnect Express (PCIe) board with a field-programmable gate array (FPGA) which has been bundled with two 24 core Intel Xeon E5-2697v2 12C, 2.7 GHz processors with 128 GB RAM and 3.2 TB capacity SSD running on CentOS 6.6. DRAGEN system is supplied with DRAGEN Genome Pipeline which utilizes the DRAGEN Bio-It Platform with the improved and highly optimized mapping, aligning, sorting, duplicate marking and haplotype variant calling algorithms 11. A single DRAGEN run for WGS data, from fastq files to vcf files, can be completed in just one simple command. Also, DRAGEN can be run to output the intermediate alignment BAM file to be used with other variant caller (alignment mode) and vice versa. More details about the DRAGEN

Bio-IT Platform and DRAGEN Genome Pipeline has been recently published recently by Miller NA *et al*. [10] [11].

DRAGEN Command

"dragen--num-threads 48-r/path/to/reference_Dir/--output-directory/path/to/Output_Dir/--output-file-prefix PREFIX-1 Sequence_R1.fastq-2 Sequence_R2.fastq--enable-variant-caller true--vc-reference/path/to/hg19.fa--vc-sample-name SampleID--enable-duplicate-marking true--remove-duplicates true—enable-bam-indexing true--enable-map-align-output true--intermediate-results-dir/staging/tmp".

2.4. Performance Assessment of the Two Variant Calling Pipelines

This study utilized below mentioned two WGS data analysis pipelines to process the dataset consist of two replicate of the NA12878.

Pipeline 1. DRAGEN Alignment and DRAGEN Variant Caller (DRAGEN Genome Pipeline)

Pipeline 2 GATK Best Practices workflow (BWA alignment, BAM file pre-processing and HaplotypeCaller).

All the analyses were performed on the server equipped with two 48 core Intel Xeon E5-2697v2 12C, 2.7 GHz processors with 128 GB RAM and 3.2 TB capacity SSD running on CentOS 6.6. Variant called using the pipelines were compared with the GIAB variants truth-set. For WGS dataset, a subset of variants in the GIAB high confidence genomic region bed file was extracted for each pipeline and compared with the GIAB's high confident variant call-set to assess the performance.

To draw receiver operating characteristic (ROC) curve and calculate the sensitivity and specificity of SNPs and INDELs, we defined the true positive (TP), true negative (TN), false positive (FP), and false negative (FN) variants as follows:

TP: Correctly called ALT genotype which is also listed in GIAB truth-set.

TN: Correctly called REF genotype which is also not listed in GIAB-truth-set

FP: Incorrectly called ALT genotype which is not listed in GIAB truth-set.

FN: Incorrectly missed ALT genotype which is listed in GIAB truth-set.

$$\text{Sensitivity} = \text{TP}/(\text{TP}+\text{FN}) \quad (1)$$

$$\text{Specificity} = \text{TP}/(\text{TP}+\text{FP}) \quad (2)$$

$$\text{Accuracy} = (\text{TP}+\text{TN})/(\text{TP}+\text{TN}+\text{FP}+\text{FN}) \quad (3)$$

3. Results

3.1. Research Scheme

Figure 1 shows the research scheme to assess the variant calling pipelines performance for the whole genome sequencing data. This research employed two genome analysis pipelines, *i.e.* the DRAGEN genome pipelines and GATK Best Practice Pipelines, to assess the read-alignment and variant calling accuracy (as

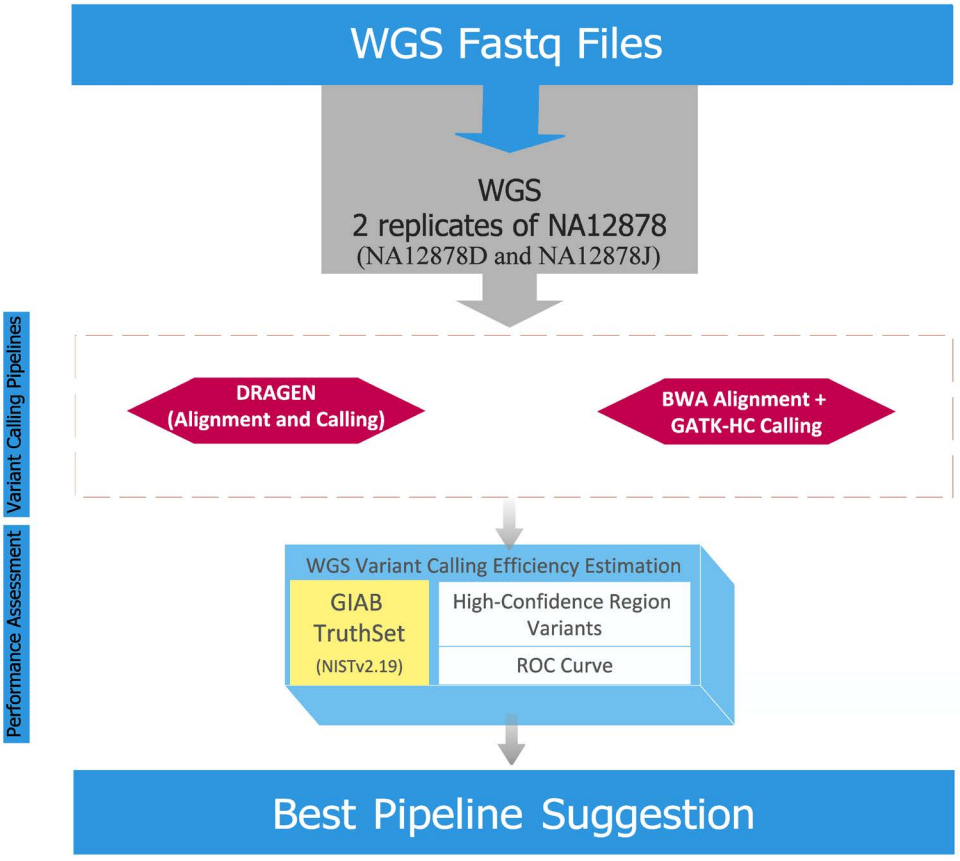

Figure 1. Scheme of assessment of NGS data analysis pipelines. Flow chart shows the steps to assess the variant calling performance of the various pipelines using DRAGEN and GATK-best practices guidelines.

described in Method section). Two replicates of the NA12878 WGS sample, labelled as NA12878D and NA12878J18 with the coverage of 39.00× and 38.65× respectively, were used to assess the consistency and reproducibility of the variant calling workflows. GIAB high confidence truth-set, along with the high confidence genomic region bed file, was used to assessed the performance of the both variant calling pipelines and suggest the best pipeline to the NGS-based researcher. More details about the sequence dataset and validation call-set can be found in Method section.

3.2. Runtime Performance of the Genome Analysis Pipelines

One of the main object of this study is the Run-time assessment of the DRAGEN against the GATK Best Practice pipelines. Run time matrices were divided into, mapping time (*i.e.* fastq to Bam file generation time) and variant calling time (Bam to VCF generation). Table 1 lists the run time matrices for both the pipelines. DRAGEN alignment includes the sorting, duplicate marking, ReadGroup information adding, etc. GATK best practice pipeline includes the mapping by BWA and preprocessing by samtools, Picard-tools, GATK, etc. All the command utilized the multithreaded option with maximum of 48 core, except the Picard-tools which utilized single core.

Table 1. Performance comparison: run time assessment of variant calling pipelines.

Dataset	Sample	Analysis Step	Dragen	BWA + HC
WGS	NA12878D	FastQ2BAM	00:18:38	23:18:32
		Bam2VCF	00:23:17	9:13:19
		FastQ2VCF	00:37:53	32:31:51
	NA12878J	FastQ2BAM	00:19:21	23:24:08
		Bam2VCF	00:24:42	09:31:12
		FastQ2VCF	00:40:15	32:55:20

Here, table lists the run time profile of the two pipelines measured on the 2 replicates of NA12878 WGS dataset. For each pipeline, run-time for individual step, *i.e.* Mapping/Alignment (FastQ2BAM), Variant Calling (Bam2VCF) and complete pipeline run (FastQ2VCF), is listed.

As listed in Table 1, DRAGEN competed alignment and BAM preprocessing for the NA12878D dataset in ~18 minute while GATK best practice pipeline took over 23 hours for the same (Figure 2). Likewise, variant calling using the GATK HaplotypeCaller completed in over 9 hours while DRAGEN Haplotype aware variant caller took just 23 minutes. All over, DRAGEN NGS data run was completed in ~37 minutes while the GATK tools over 32 hours. A similar run-time was obtained while analyzing the another WGS dataset (NA12878J). In a nutshell, around 50-fold NGS data processing speed can be obtained by utilized the DRAGEN Genome Pipeline as compared to the GATK best practice recommendations.

3.3. Variant Calling Accuracy of the WGS Variant Calling Pipelines

Variant calling accuracy of the two pipelines were assessed against the standard GIAB high confidence region truth-set (v2.19). For this, both the WGS data analysis pipelines were executed for the two replicates of NA12878 WGS data. For each pipeline, a high-confident subset of variants, in the GIAB high confidence genome region (bed file), were selected and compared with the GIAB truth-set to calculate the sensitivity, specificity and accuracy of each pipelines.

Both the pipelines showed ~99% and ~90% variant calling sensitivities for SNPs and INDELs, respectively while maintaining over 98% detection specificity in both the cases. As listed in Table 2 and shown in Figure 3, DRAGEN Genome pipelines showed slightly higher SNP detection accuracy than GATK best practice workflow. On the other hand, GATK best practice pipelines showed high INDEL detection accuracy for NA12878D dataset while DRAGEN pipeline for other (NA12878J) dataset. Further, as shown in Figure 4, ROC curve of SNPs and INDELs for DRAGEN genome pipeline showed high variant detection sensitivity at low False Positive Rate, but gradually with the in-crease of false positive hits the curve become similar (or lag behind) to that of the GATK HaplotypeCaller. Overall, DRAGEN Genome Pipelines showed the comparable or better variant calling accuracy than the GATK best practice workflow.

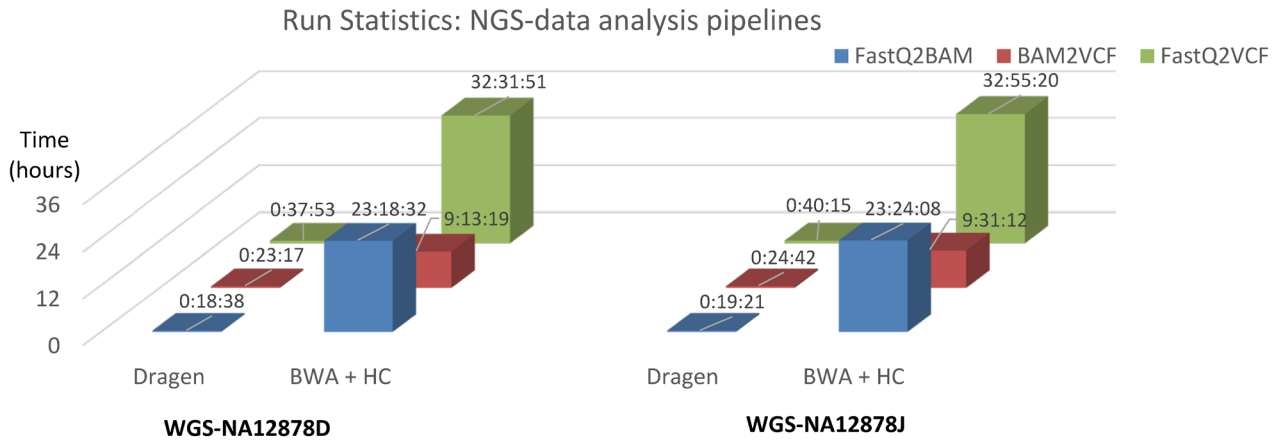

Figure 2. Genome analysis pipelines run-profile statistics. The figure shows the run-profile statistics for each steps of the NGS-data analysis, *i.e.* the Alignment step (FastQ2BAM), Variant Calling step (BAM2VCF) and total run-time (FastQ2VCF) for each dataset, for the two NGS data analysis pipelines in this study.

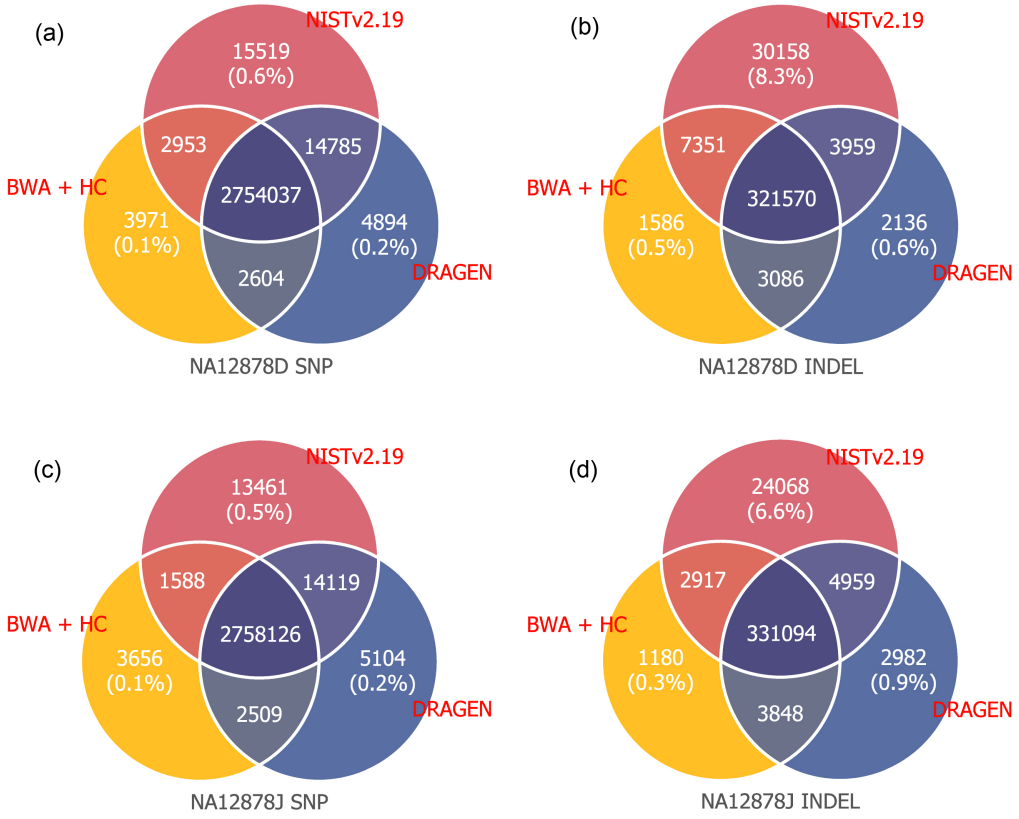

Figure 3. Variant calling performance assessment for WGS dataset. The figure shows the performance assessment of the genome analysis pipelines against the GIAB truth-set for the NA12878 sample. The Venn diagram shows the concordant SNPs (a) and (c) and the INDELs call (b) and (d) obtained by two pipelines against the GIAB truth-set.

4. Discussion

The major focus on the NGS-data analysis workflow is how to speed-up the analysis time without sacrificing the variant calling accuracy to utilize the NGS-based diagnosis more effectively, especially in a real-time disease management,

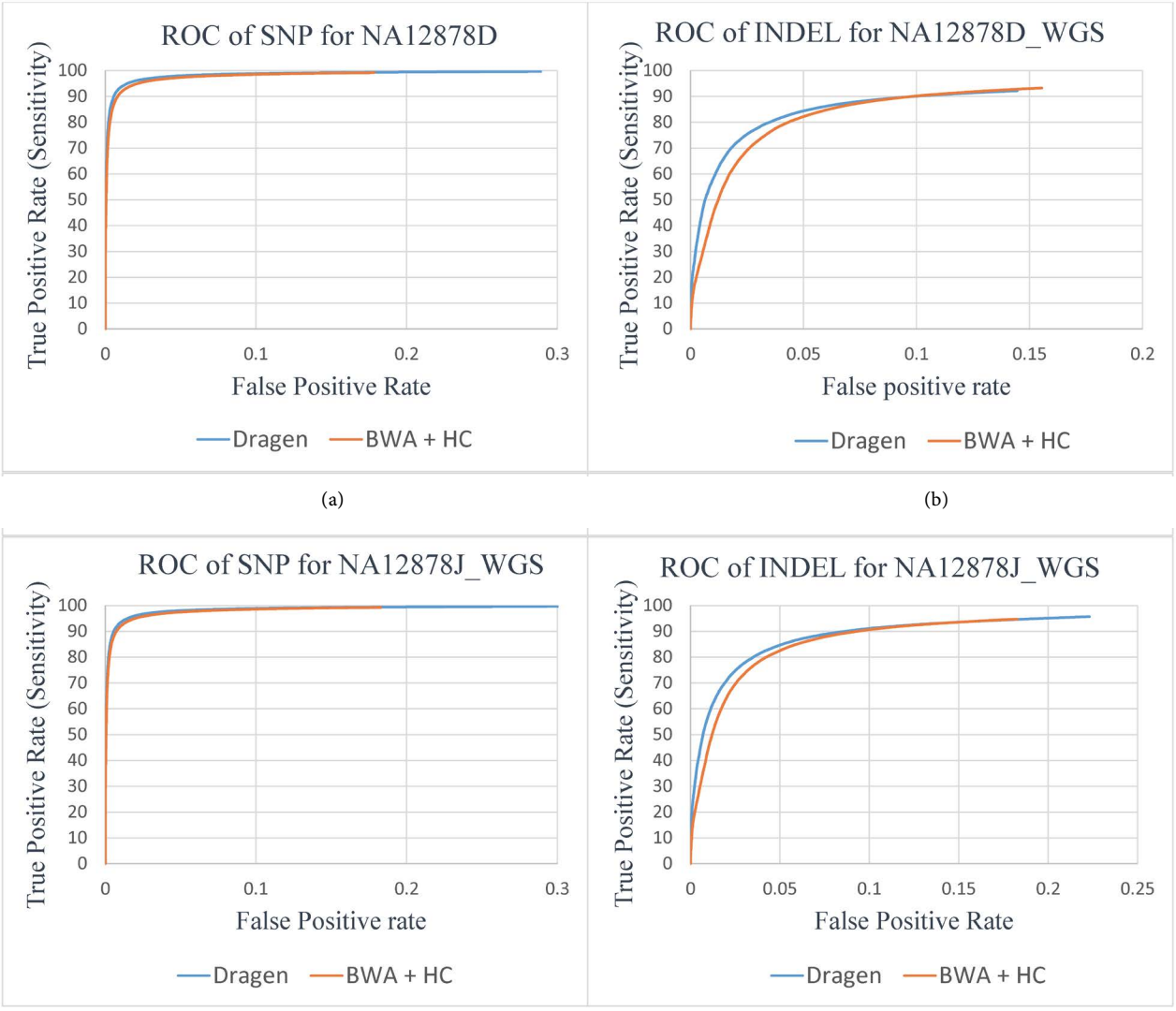

Figure 4. ROC curves of SNPs and INDELs for WGS dataset. ROC curves showing sensitivity vs. false positive rate for two replicates of the whole genome, (a) and (c) SNPs and (b) and (d) INDELs, for the NA12878 data set. Variant quality and true positive/false positive variants were identified as described in the Online Methods section.

Table 2. Performance comparison: accuracy of the variants calling pipelines.

Pipeline	#SNP†	#FP SNPs	Sensitivity (%)	Specificity (%)	Accuracy	#INDEL§	#FPINDELs	Sensitivity (%)	Specificity (%)	Accuracy (%)
					WGS-NA12878D					
DRAGEN	2,776,320	2488	99.33	99.72	**99.07**	330,751	1076	89.66	98.42	88.39
BWA + HC	2,763,565	1433	98.91	99.76	98.68	333,596	457	90.59	98.59	**89.45**
					WGS-NA12878J					
DRAGEN	2,779,858	2457	99.46	99.72	**99.18**	342,883	1799	92.56	98.01	**90.85**
BWA + HC	2,765,879	1163	99.01	99.77	98.79	339,039	357	92.20	98.51	90.74

Here, table lists the variant calling sensitivity and specificity profile of the 2 pipelines measured on the 2 replicated of NA12878 WGS dataset. For each pipeline, total number of SNP/INDEL, false positives, Sensitivity, Specificity and Accuracy is listed. Definition and formula of FP, sensitivity, specificity and accuracy is described in Method section. For SNP calling, DRAGEN Pipeline is shown to be more efficient than the BWA + HC pipelines. Similarly, DRAGEN pipeline shows comparable or better INDEL calling accuracy than GATK best practice workflow.

outbreaks of infectious disease and disaster situations, etc. This study assessed the analysis speed of sequencing data and variant calling accuracy of two genome analysis pipeline. The results showed that the ~40× coverage human WGS data processing using the DRAGEN Bio-IT Genome Pipelines can be completed in less than 40 minutes while obtaining the comparable accuracy with the standard GATK best practice workflow.

One of the main objects of this study is to identify the fast, accurate and efficient genomic analysis solution which can deal with the high computing demand in the era of massive NGS data analysis. Utilization of the DRAGEN Bio-IT processor with DRAGEN Genome Pipeline can provide an efficient solution to the "Big-data bottleneck" since it can complete the standard human whole genome sequencing data analysis (fastq to vcf) in less than 40 minutes. The DRAGEN system processing time is sufficient to support ~30 - 40 minutes sequencing time for a single WGS sample in currently available high throughput sequencer. Therefore, one DRAGEN-system is enough to analyze the raw data generated from the high throughput sequencing system such as Illumina HiSeq X 10 sequencing center.

In the recent time, several modifications of the GATK best practice pipelines have been published, e.g. Churchill [14], SpeedSeq [15], etc. Churchill pipeline claims to accomplish the 30× WGS sample in ~11 hours on a 48-core single CPU or ~1 hour 50 minutes on Ohio Supercomputer Center's Glenn Cluster (768 cores over 96 nodes). Similarly, SpeedSeq claims 13-hour run-time for 50× NA12878 WGS using default software parameters and a single 16-core server (allowing 32 threads) with 128 GB of RAM. Even though, this study doesn't compare the DRAGEN Genome pipeline's speed and accuracy with such pipelines, but ~40 minutes WGS data analysis time is much less than above mentioned pipelines which makes the DRAGEN system highly promising at industrial scale.

One important observation in our study is that DRAGEN Genome Pipelines is highly sensitive at low False Positive Rate. As shown in ROC curve of SNPs and INDELs for WGS dataset in the **Figure 4**, with the increase in the variant calling sensitivity (over 92% for SNPs and over 80% for INDEL case) the false positive hits increased significantly which reduces the overall DRAGEN variant callers' accuracy. For example, as shown in **Table 2**, NA12878D and NA12878J samples have 1% and 0.5% lower INDEL calling specificity than the GATK HaplotypeCaller, respectively. Accurate detection of INDEL from the NGS-data has been challenging due to the varying size and difficulty to map to the correct position in the genome (especially in the case of longer INDEL), etc. [16] [17]. These are the well-known issues which are caused by the technical limitation of NGS-platforms and analysis workflows. In the current study, we only compared the result of INDEL calling from the available resource/software without additional INDEL detection accuracy improvement.

Altogether, this study focused on demonstrating the proficiency and comparison of DRAGEN Bio-IT software and DRAGEN Genome Pipelines with tradi-

tional approaches. These results implicate that the DRAGEN system can be used as a single platform to analyze the genomic data accurately in quicker time at industrial scale. We expect, this research will help the scientist to make an informed choice to set-up a new (or modify the existing) genome analysis platform in their laboratory and/or institute.

References

[1] Hayden, E.C. (2014) Technology: The $1,000 Genome. *Nature*, **507**, 295-295.

[2] Watson, M. (2014) Illuminating the Future of DNA Sequencing. *Genome Biology*, **15**, 108-108. https://doi.org/10.1186/gb4165

[3] Petric, R.C., Pop, L.-A., Jurj, A., Raduly, L., Dumitrascu, D., Dragos, N. and Neagoe, I.B. (2015) Next Generation Sequencing Applications for Breast Cancer Research. *Clujul Medical*, **88**, 278-287. https://doi.org/10.15386/cjmed-486

[4] George, A. (2015) UK BRCA Mutation Testing in Patients with Ovarian Cancer. *British Journal of Cancer*, **113**, S17-S21. https://doi.org/10.1038/bjc.2015.396

[5] Vivante, A. and Hildebrandt, F. (2016) Exploring the Genetic Basis of Early-Onset Chronic Kidney Disease. *Nature Reviews Nephrology*, **12**, 133-146. https://doi.org/10.1038/nrneph.2015.205

[6] Zutt, R., van Egmond, M.E., Elting, J.W., van Laar, P.J., Brouwer, O.F., Sival, D.A., Kremer, H.P., de Koning, T.J. and Tijssen, M.A. (2015) A Novel Diagnostic Approach to Patients with Myoclonus. *Nature Reviews Neurology*, **11**, 687-697. https://doi.org/10.1038/nrneurol.2015.198

[7] Hoyle, J.C., Isfort, M.C., Roggenbuck, J., Arnold, D. and Hoyle, C. (2015) The Genetics of Charcot-Marie-Tooth Disease: Current Trends and Future Implications for Diagnosis and Management. *Application of Clinical Genetics*, **8**, 235-243.

[8] Ono, S., Lam, S., Nagahara, M. and Hoon, D. (2015) Circulating microRNA Biomarkers as Liquid Biopsy for Cancer Patients: Pros and Cons of Current Assays. *Journal of Clinical Medicine*, **4**, 1890-1907. https://doi.org/10.3390/jcm4101890

[9] Appistry. AppistryGenomePilot™. http://www.appistry.com/genomepilot/

[10] DRAGEN. Edicogenome. http://www.edicogenome.com/

[11] Miller, N.A., Farrow, E.G., Gibson, M., Willig, L.K., Twist, G., Yoo, B., Marrs, T., Corder, S., Krivohlavek, L., Walter, A., Petrikin, J.E., Saunders, C.J., Thiffault, I., Soden, S.E., Smith, L.D., Dinwiddie, D.L., Herd, S., Cakici, J.A., Catreux, S., Ruehle, M. and Kingsmore, S.F. (2015) A 26-Hour System of Highly Sensitive Whole Genome Sequencing for Emergency Management of Genetic Diseases. *Genome Medicine*, **7**, 100. https://doi.org/10.1186/s13073-015-0221-8

[12] Zook, J.M., Chapman, B., Wang, J., Mittelman, D., Hofmann, O., Hide, W. and Salit, M. (2014) Integrating Human Sequence Data Sets Provides a Resource of Benchmark SNP and Indel Genotype Calls. *Nature Biotechnology*, **32**, 246-251. https://doi.org/10.1038/nbt.2835

[13] Hwang, S., Kim, E., Lee, I. and Marcotte, E.M. (2015) Systematic Comparison of Variant Calling Pipelines Using Gold Standard Personal Exome Variants. *Scientific Reports*, **5**, 17875-17875. https://doi.org/10.1038/srep17875

[14] Kelly, B.J., Fitch, J.R., Hu, Y., Corsmeier, D.J., Zhong, H., Wetzel, A.N., Nordquist, R.D., Newsom, D.L. and White, P. (2015) Churchill: An Ultra-Fast, Deterministic, Highly Scalable and Balanced Parallelization Strategy for the Discovery of Human Genetic Variation in Clinical and Population-Scale Genomics. *Genome Biology*, **16**, 6. https://doi.org/10.1186/s13059-014-0577-x

[15] Chiang, C., Layer, R.M., Faust, G.G., Lindberg, M.R., Rose, D.B., Garrison, E.P., Marth, G.T., Quinlan, A.R. and Hall, I.M. (2015) SpeedSeq: Ultra-Fast Personal Genome Analysis and Interpretation. *Nature Methods*, **12**, 966-968. https://doi.org/10.1038/nmeth.3505

[16] Jiang, Y., Turinsky, A.L. and Brudno, M. (2015) The Missing Indels: An Estimate of Indel Variation in a Human Genome and Analysis of Factors That Impede Detection. *Nucleic Acids Research*, **43**, 7217-7228. https://doi.org/10.1093/nar/gkv677

[17] Fang, H., Wu, Y., Narzisi, G., O'Rawe, J.A., Barrón, L.T.J., Rosenbaum, J., Ronemus, M., Iossifov, I., Schatz, M.C. and Lyon, G.J. (2014) Reducing INDEL Calling Errors in Whole Genome and Exome Sequencing Data. *Genome Medicine*, **6**, 89-89. https://doi.org/10.1186/s13073-014-0089-z

[18] HiSeq X-Ten Test Data. https://dnanexus-rnd.s3.amazonaws.com/NA12878-xten.html

[19] Van der Auwera, G.A., Carneiro, M.O., Hartl, C., Poplin, R., del Angel, G., Levy-Moonshine, A., Jordan, T., Shakir, K., Roazen, D., Thibault, J., Banks, E., Garimella, K.V., Altshuler, D., Gabriel, S. and DePristo, M.A. (2013) From fastQ Data to High-Confidence Variant Calls: The Genome Analysis Toolkit Best Practices Pipeline. *Current Protocols in Bioinformatics*, **43**, 1-33. https://doi.org/10.1002/0471250953.bi1110s43

[20] Li, H. and Durbin, R. (2010) Fast and Accurate Long-Read Alignment with Burrows-Wheeler Transform. *Bioinformatics*, **26**, 589-595. https://doi.org/10.1093/bioinformatics/btp698

[21] Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R. and Genome Project Data Processing Subgroup (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, **25**, 2078-2079. https://doi.org/10.1093/bioinformatics/btp352

Scientific Research Publishing

Open Journal of Genetics, 2017, 7, 20-39
http://www.scirp.org/journal/ojgen
ISSN Online: 2162-4461
ISSN Print: 2162-4453

Why Do We Care for Old Parents? Evolutionary Genetic Model of Elderly Caring

Takahiro Miyo

2-17-4 Misumi-cho, Higashimurayama-shi, Tokyo 189-0023, Japan
Email: takahiro_miyo@hotmail.com

How to cite this paper: Miyo, T. (2017) Why Do We Care for Old Parents? Evolutionary Genetic Model of Elderly Caring. *Open Journal of Genetics*, **7**, 20-39.
https://doi.org/10.4236/ojgen.2017.71003

Received: January 21, 2017
Accepted: March 26, 2017
Published: March 29, 2017

Abstract

From the standpoint of evolution, caring for old parents may be maladaptive, because they have ceased reproduction, so that the benefit for inclusive fitness may not be expected. Then why do we care for old parents? In this study, the evolution of care for the elderly was examined, by using an evolutionary genetic model, in which pleiotropic constraints between behaviors expressed in different social contexts among family members were assumed. It was suggested that establishing a solid relationship with parents during infancy should be selectively favorable, even though old parents have to be cared for in the future, but that caring for old parents may be excluded from the population if this behavior imposes high costs on reproduction of the younger generation. Despite the diminishing numbers of individuals within the population, care for the elderly may not be readily selected against, but at the same time this may not contribute to the rate of increase in population size. The significance of discussing the behavior of elderly caring from the standpoint of evolutionary genetics was emphasized.

Keywords

Elderly Caring, Evolutionary Genetic Model, Infancy, Pleiotropic Constraint

1. Introduction

Senescence, which is defined as the decline in a wide range of bodily functions with age, is an almost universal phenomenon among organisms [1]. In spite of its universality, the problem posed by this apparently deleterious phenomenon is to explain why organisms had evolved not to continue living and reproducing forever but to senesce and die ultimately. Evolutionary biologists have proposed solid theories to explain how senescence had been established and maintained by natural selection [1]. Among them, Medawar [2] [3] suggested the critical con-

cept for solving this problem that the force of natural selection on deleterious genes declines with age. That is, as organisms age, they produce offspring, so that there would be essentially no effects of deleterious genes on fitness if these genes are expressed in fully late ages, because they have already produced enough offspring. Therefore, the later the deleterious genes are expressed, the weaker the force of natural selection. As a result, deleterious genes expressed late in life would not be eliminated from the population, and the population tends to accumulate these deleterious mutations, resulting in senescence and ultimate death of organisms as an inevitable consequence of evolutionary processes within the population.

While Medawar suggested the above evolutionary mechanism of senescence more than 50 years ago, he also raised the alarm about population aging in the last paragraph of the same paper, which is now a serious problem in such countries as Japan. He mentioned as follows:

> "The moral is that the problem of doing something about old age becomes slowly but progressively more urgent. Something must be done, if it is not to be said that killing people painlessly at the age of seventy is, after all, a *real* kindness" (italicized by Medawar himself) [2].

The above alarm statement by Medawar seems somewhat radical, but it is not a fantasy at all in such countries as Japan, where various problems resulting from population aging are and will be serious matters. In fact, at some nursing facilities for the elderly, it is not unusual that one sometimes hears old users saying "It is boring," "I feel lonely," "I want to die quickly," and so on (the author's care-worker experience).

The elderly over 65 years old in Japan are 33,920,000 people at the point of 2015, and the proportion of the elderly over 65 years old in the total population are 26.7% [4]. However, it is predicted that this proportion would keep rising from now on and reach 39.9% in 2060, which means one out of 2.5 people would be over 65 years old, while the whole Japanese population would keep decreasing. The number of elderly people who need nursing care is expected to increase rapidly, and in particular, it is predicted that patients over 65 years old suffering from dementia would reach about 7,000,000 people in 2025 (one out of 5 people over 65 years old) [4]. Despite the fact that not only elderly people but also old people requiring nursing care and dementia patients are increasing rapidly, there is the fear of serious lack of the care staffs in the future who should care for the elderly [5]. In addition, people in the state of so-called "double care", who need to carry out caring for their old parents as well as their children simultaneously, reached more than 250,000 people, about 168,000 women and about 85,000 men [6]. If caring for elderly people is predicted to be a serious problem in Japan, would such a situation, in which elderly caring is cut off, be realized, as Medawar warned?

Care for offspring by parents itself is not altruistic, because the care would increase in their own fitness [7]; however, caring for old parents by their offspring

would be altruistic, because old parents should have ceased reproduction, so that it is unlikely that the care for the elderly would result in the benefit of fitness from the standpoint of inclusive fitness theory [8]. Because all psychological mechanisms causing costly altruistic behaviors should exist through population genetic processes, including natural selection [9], if we care for old parents despite the fact that elderly caring costs very much, it is a kind of paradox in which not only proximate but also evolutionary mechanisms causing elderly caring need to be examined. In order to investigate the evolution of elderly caring, an approach using evolutionary genetic models was adopted, because evolution of a trait must be ultimately understood in terms of changes in gene frequency or its equivalent, even though it may be an invisible psychological trait [10].

Most behaviors are widely believed to be multifactorial [11], and many human psychological traits are influenced by genetic factors [12]. Like other complex human behaviors, elderly caring must also consist of many components, both physical and psychological, each of which must itself consist of many components, which may be influenced by genetic and environmental factors. For example, there would be little doubt that some psychological traits, including attachment, empathy, and altruism between parents and offspring, more or less contribute to and play roles in caring for old parents. These psychological traits have in fact been suggested to have significant genetic components, based on twin studies [13] [14] [15] [16]. Thus elderly caring can be regarded as a quantitative trait, which is influenced by genetic as well as environmental factors, and the evolution of elderly caring would rather be investigated by using evolutionary genetic models.

Since an individual plays a different role in his/her family (*i.e.*, offspring, parent, grandparent, etc.) through his/her life course, social behaviors among family members would be desirable to be treated as an integrated and developmental process [17]. Based on considerations using the theoretical models and empirical data, Lynch [17] suggested that if the types of behaviors which are expressed at different social contexts are genetically correlated, there would be some constraints in social interactions which can evolve among family members, and that such genetic constraints in behaviors should be common in natural populations. We human beings establish attachment (strong affectional bonds which result in close proximity to particular others) with our parents since infancy, and the attachment with parents is likely to continue through the whole life, "from the cradle to the grave" [18] [19]. Because we cannot survive without care by others during infancy, a genetic bias for developing solid attachment with care-givers, mainly parents, would increase the survival probabilities of infants and therefore fitness of parents per se [20]. In addition, because the attachment is likely to influence the strategies of offspring for survival and reproduction in the physical and social environments [21] [22], it should play important roles in our whole lives in the evolutionary and adaptive terms. It would be healthy that parents and children in the solid attachment relationships would reverse their roles with time, when parents are getting older or disabled [19]. Therefore, it is necessary

to examine whether establishing the solid attachment relationships with parents during infancy, such that children have to care for old parents in the future, is selectively favorable, even though there would not be payoff in fitness in the elderly caring. Thus, in this study, an evolutionary genetic model was developed, assuming pleiotropic constraints between interactions in the care for infants by parents and interactions in the care for old parents by children. In addition, because elderly caring is inevitably related to age relationships between family members (e.g., parents and children, grandparents and parents, and so on), evolutionary genetic models should take into account the age-structure of the population. Therefore, following a theory of kin selection in age-structured populations [23], the selective advantage of the parent-offspring interactions involving elderly caring and the consequence of diminishing population size on it, which such countries as Japan are now or will be faced with, were examined in order to respond to the alarm proposition by Medawar [2]. The approach by Charlesworth and Charnov [23] is to examine the condition that a rare allele A_2, which affects behavior and whose effects are assumed to be expressed in the heterozygous state, can invade a random mating population, which is initially fixed with the other allele A_1 [1]. In this study, two main questions were examined using simple numerical models for hypothetical populations: 1) whether a rare dominant allele causing high levels of care for old parents as well as infants (high care gene) can invade a population that is initially composed of an relatively indifferent genotype, whose net reproduction rate was set to be one, and 2) whether a rare dominant allele causing an indifferent phenotype (indifference gene) can invade a decreasing population that is initially composed of a genotype providing high levels of care for old parents as well as infants, which might be the very situation that the countries suffering from population aging and diminishing population size, such as Japan, is being faced with.

2. Methods

2.1. Vital Statistics for the Hypothetical Population

Vital statistics for the hypothetical population suffering from population aging and diminishing population size were created, using population statistics of Japan available on the official website of National Institute of Population and Social Security Research, Japan (http://www.ipss.go.jp/syoushika/tohkei/Popular/Popular2016.asp?chap=0). Population statistics used were numbers of Japanese females by 5-year age-group from 2001 to 2012, and age-specific fecundity data for females by 5-year age-group from 2001 to 2007, which were transformed into the age-specific survival rates ($P(x)$) and fecundities ($m(x)$) for 5-year time-intervals. For example, using numbers of females by 5-year age-group in 2001 and 2006, $P(x)$ for the 5-year interval was calculated. $m(x)$ was calculated by multiplying the age-specific fecundities for each year by five. Thus seven sets of $P(x)$ and $m(x)$ for 5-year intervals were created for these years. After averaged over the seven years, one set of basic vital statistics were obtained for the hypothetical populations in this

study (the upper Figure 1).

It would be better to note that my intention is not to predict the precise fate of the Japanese population, but to discuss the evolution of elderly caring in such populations suffering from population aging and diminishing population size as the Japanese population. Therefore, although there were some discrepancies, including some $P(x)$ values larger than one which may be due to immigration, figures taken from the above website were used without any correction, because these discrepancies were minor in effect and did not change my conclusion.

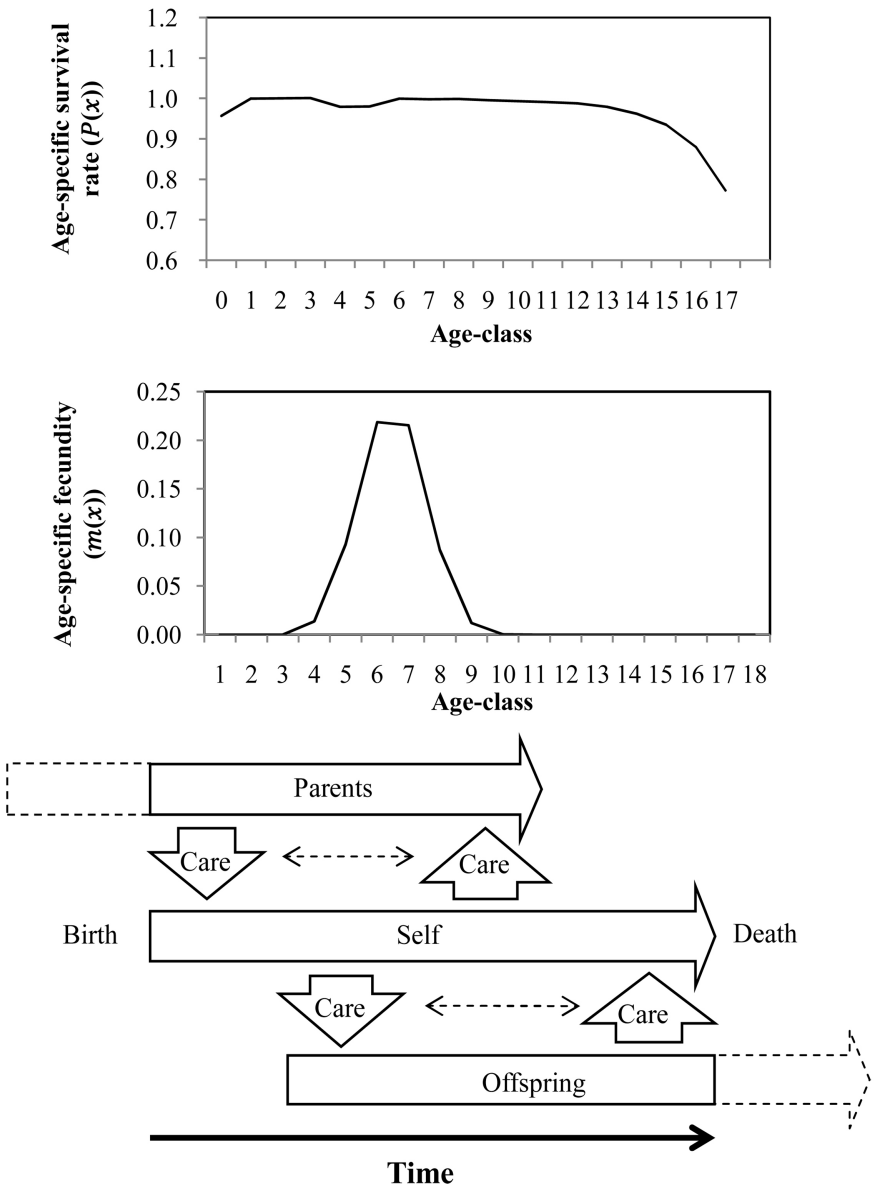

Figure 1. Basic vital statistics used in the evolutionary genetic model (upper) and a schematic diagram depicting interactions among family members (lower). $P(0)$ is the probability that a female child born at some point during a 5-year interval will live to the end of the interval; $P(1)$ is the probability that a female aged between 0 and 4 years will survive for 5 years, and so on. Likewise, $m(4)$ is the number of daughters a female aged between 15 and 19 years is expected to bear during 5 years, and so on.

2.2. Effects of Age-Specific Changes on Fitness

In populations suffering from population aging and diminishing population size, such as Japan, consequences of positive or negative changes in individual vital statistics on population increase would be especially important. The relationships between changes in individual vital statistics and corresponding changes in population growth were determined, by using the intrinsic rate of increase (r) as a measure of population growth and conducting partial differentiation of r with respect to $P(x)$ or $m(x)$ [1] [24].

The relationship between these vital statistics and r, known as the Euler-Lotka equation, is

$$\sum_{x=b}^{d} \mathrm{e}^{-rx} l(x) m(x) = 1 \tag{1}$$

[25], in which b and d represent a lower and an upper limit to the age of reproduction, respectively, and $l(x) = P(0)P(1)P(2)\cdots P(x-1)$, the probability that a female survives from birth to age x. Applying implicit differentiation to Equation (1), the relationship between a change in $\ln P(x)$ and a corresponding change in r is obtained as follows

$$\frac{\partial r}{\partial \ln P(x)} = \frac{\sum_{y=x+1}^{d} \mathrm{e}^{-ry} l(y) m(y)}{\sum_{y=b}^{d} y \mathrm{e}^{-ry} l(y) m(y)}. \tag{2}$$

Likewise, the relationship between a change in $m(x)$ and a corresponding change in r is

$$\frac{\partial r}{\partial m(x)} = \frac{\mathrm{e}^{-rx} l(x)}{\sum_{y=b}^{d} y \mathrm{e}^{-ry} l(y) m(y)}. \tag{3}$$

By obtaining these partial derivatives, we can find the effects of small changes in $\ln P(x)$ or $m(x)$ on r [1]. Because r is a measure of population growth, here we can consider the impact of the changes in $\ln P(x)$ or $m(x)$ on population increase and decrease.

For the basic vital statistics used in this study (the upper **Figure 1**), the intrinsic rate of increase, generation time, and net reproduction rate in the time-unit of 5 years were calculated by the Newton-Raphson iteration, using FORTRAN 77 program [26]. Then values of $\partial r / \partial \ln P(x)$ and $\partial r / \partial m(x)$ for each age were calculated using Equations (2) and (3).

2.3. Evolutionary Genetic Model of Elderly Caring

Based on Charlesworth and Charnov [23], whether or not a rare variant causing some behavioral trait can increase in frequency within a population fixed initially with the other variant was examined. Although the behavior of elderly caring is likely to be a complex quantitative trait, to which various factors, genetic as well as environmental, should contribute, it is necessary to incorporate the age-structure as an inevitable factor into a model, so that it would be too complex to deal with, if elderly caring is considered to be a quantitative trait. Therefore, the age-structure of the population is more emphasized, and a simple evolutionary genetic model, assuming two variants, one causing high levels of care

for old parents as well as infants (high care gene) and the other causing relatively low levels of care (indifference gene), was adopted.

2.3.1. Invasion of a High Care Gene

It was assumed that a rare dominant autosomal allele causing high levels of care for the elderly as well as infants (*A*) was introduced into a random mating population fixed initially with a relatively indifferent allele (*a*), in which the initial frequency among individuals of age-class 4, starting reproduction, was set to be 1:100 (*Aa*:*aa*). Under this condition, combinations of mates and constitutions of family members were greatly simplified, so that the complex behavior of elderly caring could be examined more readily (**Figure 2**). It was assumed that there were genetic constraints between the types of behavior among family members in different social contexts, so that individuals possessing the dominant high care allele would establish a solid relationship between parents and offspring, in which parents would provide high levels of care for offspring, resulting in high survival rates during infancy, while offspring would provide high levels of care for old parents, resulting in high survival rates during old ages (the lower **Figure 1**). The genotype possessing the rare dominant allele should be *Aa*, and its mate should be *aa*, because individuals with two *A* alleles and mating between *Aa* individuals should be so rare that they could be ignored. Under these conditions, combinations of mates which should be taken into account were *Aa* × *aa* and *aa* × *aa*, and it would be expected that the former family would produce offspring whose genotype is *Aa* or *aa* (1:1), while the latter family would produce offspring whose genotype is only *aa*. Because one parent of the former family possesses *Aa*, offspring in this family would receive high levels of care during infancy; on the

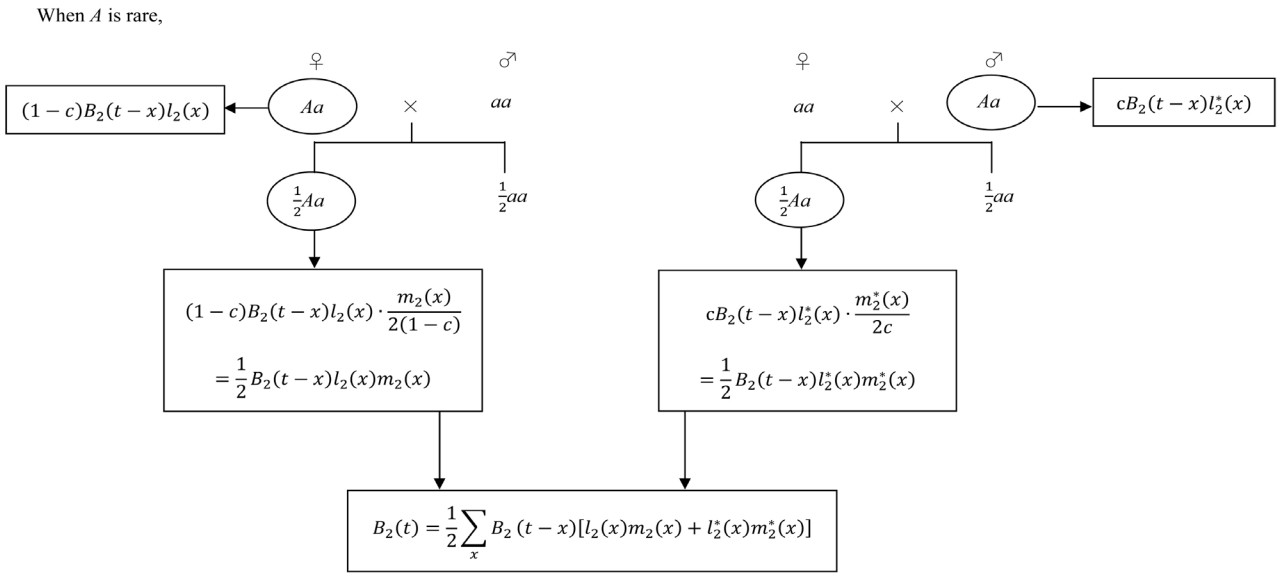

Figure 2. A schematic diagram depicting how to calculate the total number of *Aa* newborns produced at time t ($B_2(t)$), when a dominant *A* gene is rare. Here, c is the frequency of males among *Aa* newborns, $l_2(x)$ is the probability that an *Aa* female survives from birth to age x, and $m_2(x)$ is the number of female offspring which an *Aa* female of age x is expected to bear. $l_2^*(x)$ and $m_2^*(x)$ are the similar vital statistics for an *Aa* male. In this study, females and males were assumed to have the same vital statistics (for more details, see the text and [23]).

other hand, parents would receive high levels of care later in life because half the offspring are expected to possess *Aa*. Therefore, through parent-offspring interactions, *Aa* individuals were expected to have higher survival rates during infancy as well as later ages than *aa* individuals, at the level of individual genotypes. In this study, the basic $P(x)$ calculated from the population statistics of Japan (the upper **Figure 1**) was used for $P(x)$ of *Aa* individuals. For indifferent individuals, $P(x)$ values during age-classes 0 - 2 (birth to 9 years old) and age-classes 15 - 17 (70 to 84 years old) were assumed to be 5%, 10% or 20% lower than those of the high care genotype. In addition, the actual $P(x)$ values of Taiwan females in 1906, more than 100 years ago, were used for reference, which were calculated from the life table in [24].

$m(x)$ values were set to be those which made the net reproduction rate of the relatively indifferent genotype unity, by multiplying the basic $m(x)$ values (the upper **Figure 1**) by constants. Under this condition, if the high care phenotype was selectively favorable, the relative frequency of the high care gene would be increased. If disadvantageous, its relative frequency would be decreased [1]. In addition, the state of so-called "double care" is likely to impose double burdens on people who care for old parents as well as children simultaneously [6]. Under these circumstances, care for old parents could impose much cost on offspring, and if there is no elderly caring, offspring could make an additional effort to reproduce. Thus negative effects of high levels of care for the elderly on reproduction were examined by assigning to the high care genotype the $m(x)$ values which were reduced by multiplying constants. Under these conditions, female and male vital statistics were assumed to be the same for each genotype. After calculating the trajectories of the numbers of *Aa* and *aa* newborns (**Figure 2**), the relative frequencies of *A* allele (p_2) were approximately estimated as

$$p_2 = (\text{number of } Aa \text{ newborns}) / 2(\text{number of } aa \text{ newborns}).$$

When the gene frequencies are sufficiently increased, these genes are not rare any more, so that a variety of combinations of genotypes among mates and among family members cannot be ignored. Under this situation, interactions between parents and offspring would be expected to be much more complex than the above situation. Whether or not the high care gene eventually reached fixation within the population was examined, using a discrete generation model, in which selection coefficients were determined based on the intrinsic rate of increase of the high care genotype [27].

2.3.2. Invasion of an Indifferent Gene

It was assumed that a rare dominant autosomal allele causing a relatively indifferent behavior (*B*) was introduced into a random mating population fixed initially with a high care allele (*b*), in which the initial frequency among individuals of age-class 4, starting reproduction, was set to be 1:100 (*Bb*:*bb*). The procedure of the analysis was essentially the same as above, but here the factors of population aging and diminishing population size were taken into account, which Japan and other countries are now and will be faced with, so that the basic $m(x)$ values calculated from the recent population statistics of Japan were used

(the upper Figure 1). Therefore, it matters whether high care individuals prohibit indifferent individuals from increasing or relatively indifferent individuals increase in frequency and eventually exclude high care individuals, while the population size is decreasing. In addition, taking the cost of elderly caring on reproduction of offspring into account, negative effects of high levels of care for the elderly on reproduction were examined by assigning to the relatively indifferent genotype the $m(x)$ values which were increased by multiplying constants.

2.4. Correlation in Survival Rate between the Immature and the Elderly

In this study, pleiotropic constraints in interactions affecting survival rates through life courses between parents and offspring were assumed. Because there is no research concerning the relationship between the level of care for infants and the level of care for old parents, to my knowledge, this assumption cannot be examined directly. However, if the pleiotropic constraints exist in interactions between parents and offspring, there should exist a positive correlation between survival rates during the immature stage and those later in life. Therefore, the correlation between survival rates in the immature ages and those later in life was examined. To do this, I used data from the 2015 revision of the World Population Prospects, obtained from the website of the United Nations (UN), Population Division [28]. The data I used included age-specific fertility rates, female population by 5-year age-group, and feminity ratio in age 0 - 1, for country-groups, classified by income levels, based on 2014 Gross National Income *per capita* from the World Bank, from 1990 to 2015. For each country-group (high-, upper-middle-, lower-middle-, and low-income), five sets of $P(x)$ and $m(x)$ for 5-year intervals were created from the data in the same way as the above hypothetical population. After age-classes were divided into three groups (0 - 14, 15 - 64, and 65+ years old), survival rates were averaged over the age-classes. The correlation coefficient in survival rate between age-group 0 - 14 and age-group 65+ was estimated. In addition, multiple regression analyses of survival rates of one age-group on survival rates of the other two age-groups were conducted, because each age-group should interact with other age-groups, so that survival rates of one age-group may be affected by those of the other age-groups.

Some $P(x)$ values, especially for the high-income country-group, were more than unity, probably due to international migration. Although the net migration rates for these country-groups during the time period used in this study ranged from +3.7 to −1.3 per 1000 population per year [28], any adjustment was not conducted for the population data from the UN, in order to avoid making arbitrary decision, so that survival rates at face value were used to obtain insights into the interactions affecting survival rates through life courses between parents and offspring.

3. Results

3.1. Effects of Age-Specific Changes on Population Growth

For the basic vital statistics of the hypothetical population used in this study (the

upper Figure 1), which is suffering from population aging and diminishing population size, the intrinsic rate of increase, generation time, and net reproduction rate in the 5-year time-unit were estimated as −0.081, 6.56 (about 32.8 years), and 0.590, respectively.

The relationship between changes in $\ln P(x)$ or $m(x)$ and corresponding changes in the intrinsic rate of increase (r) for the hypothetical basic population is shown in Figure 3. In the case of $\partial r/\partial \ln P(x)$, the maximum value of 0.152 was attained in age-classes 0 to 3 (0 - 14 years old), but the value then decreased with age and reached zero in age-class 10 (45 - 49 years old). In the case of $\partial r/\partial m(x)$, the minimum value of 0.158 in age-class 1 (0 - 4 years old) increased gradually and reached the maximum value of 0.436 in age-class 16 (75 - 79 years old), and then decreased afterward.

3.2. Invasion of a High Care Gene

The results of the evolutionary genetic analyses of elderly caring, in which a rare high care allele (A) was introduced into a population initially fixed with a relatively indifferent allele (a), are indicated in Figure 4(a). Within the population initially occupied by relatively indifferent individuals, whose survival rates during infancy and later in life were reduced by 5% to 20%, individuals possessing the rare high care allele increased in frequency, if they had the same $m(x)$ as the indifferent. The larger the reductions in survival rate of indifferent individuals, the larger the rates of increase of the high care allele within the population.

However, the rates of increase in the frequencies of the high care allele were decreased, if high care for the elderly influenced reproduction negatively, and

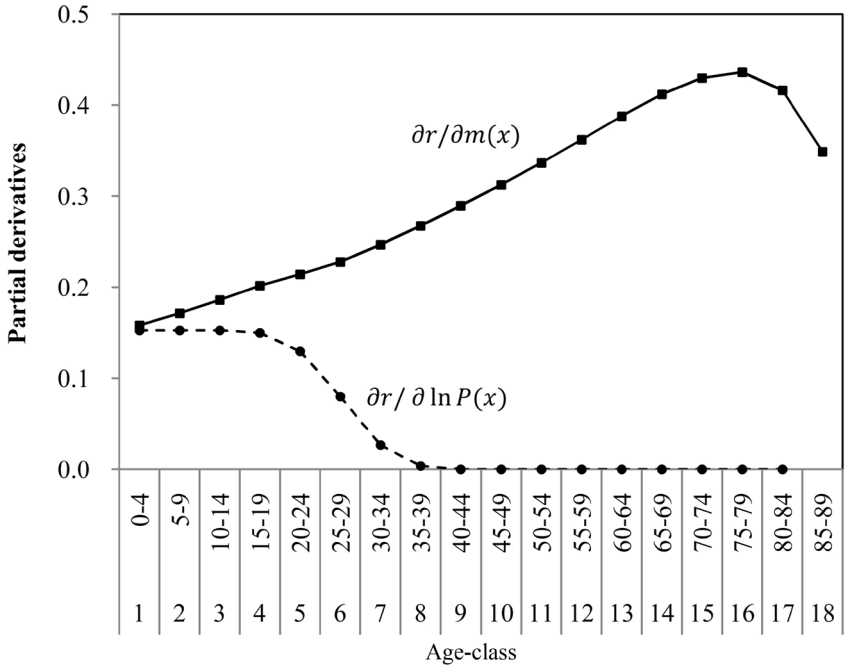

Figure 3. The relationship between changes in $\ln P(x)$ or $m(x)$ and corresponding changes in the intrinsic rate of increase, r.

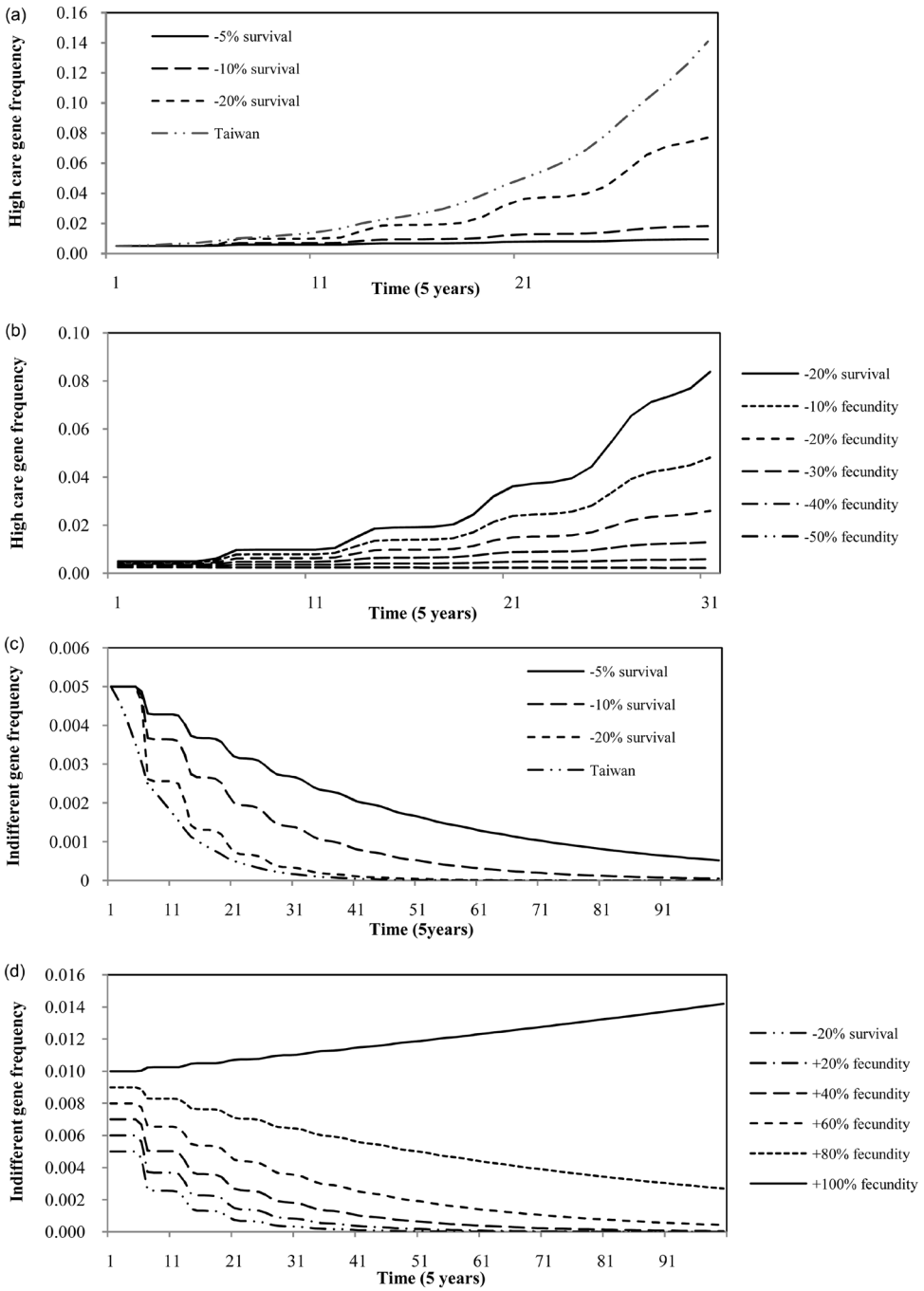

Figure 4. Changes in frequency of a rare dominant autosomal allele which affects the behavior of carriers of the gene. ((a) (b)) The cases, in which a rare high care gene was introduced into a stationary population occupied by relatively indifferent individuals, whose survival rates during infancy and later in life were reduced by 5% to 20% and whose net reproduction rate was set to be one. (a) Age-specific fecundities were the same between the genotypes. (b) Age-specific fecundities for the high care genotype were reduced by 10% to 50% due to high levels of care for the elderly within the −20% indifferent population. ((c) (d)) The cases, in which a rare relatively indifferent gene causing 5% to 20% reduction in survival rates during infancy and later in life was introduced into a decreasing population occupied by high care individuals. (c) Age-specific fecundities were the same between the genotypes. (d) Age-specific fecundities for the −20% indifferent genotype were increased by 20% to 100% due to low levels of care for the elderly. The results using actual age-specific survival rates for Taiwan females in 1906 are also shown in (a) and (c) for reference.

the high care allele was excluded from the population, if the negative effects on reproduction were sufficiently large. The changes in fecundity of the high care genotype, resulting in the intrinsic rates of increase equivalent to those of the relatively indifferent genotypes, are shown in Table 1. For example, the high care allele within the initial population fixed with the indifferent allele conferring −20% survival can increase in frequency, even though it brought about a 40% reduction in fecundity (the initial gene frequency of 0.0030 would indeed reach 0.0048 after 20 time-units (100 years); Figure 4(b)). However, it cannot, if $m(x)$ was reduced by 50% (the initial gene frequency 0.0025 would indeed decrease to 0.0023 after 20 time-units).

Whether or not such rare alleles that increased initially within the population eventually reached fixation was verified, based on a discrete generation model. The results obtained by using the discrete generation model were consistent with those obtained by using the evolutionary genetic model taking into account the age-related interactions between family members, so that it was expected that the rare allele which increased initially within the population would increase until it reaches fixation (data not shown).

3.3. Invasion of an Indifferent Gene

The results of the cases where a rare relatively indifferent allele (B) was introduced into a decreasing population fixed with a high care allele (b) are shown in Figure 4(c). While the population was decreasing, the rare indifferent gene which reduced $P(x)$ during infancy and later in life by 5% to 20% cannot increase in frequency within the high care population, if both genotypes possessed the same $m(x)$.

In addition, the rates of decrease in the frequencies of the relatively indifferent allele were decreased, if low levels of care for the elderly had positive consequences on reproduction, and the indifferent allele would exclude the high care allele eventually from the population, if the positive consequences of low care on reproduction were sufficiently large. For example, the relatively indifferent allele which reduced $P(x)$ during infancy and later in life by 20% could not increase in frequency even if $m(x)$ increased by 1.8 times, but its frequency would finally

Table 1. The changes in fecundity of the invading genotypes which result in the intrinsic rates of increase equivalent to those of the initial populations.

invader	Initial population	Reduction in survival rates of indifferent individuals			
		−5%	−10%	−20%	Taiwan 1906
High care	Stationary indifferent[a]	−14.26%	−27.10%	−48.80%	−53.41%
Indifferent	Decreasing high care[b]	+16.63%	+37.17%	+95.31%	+116.44%

[a]A rare dominant allele causing high levels of care for both offspring and old parents is assumed to be introduced into a stationary population fixed with an indifferent allele. [b]A rare dominant allele causing relative indifference is assumed to be introduced into a decreasing population fixed with a high care allele.

increase if $m(x)$ increased by 2 times (Figure 4(d) and Table 1).

3.4. Correlation in Survival Rate between the Immature and the Elderly

There was a significant positive correlation in survival rate between the immature (0 - 14 years old) and the elderly (more than 65 years old) (Figure 5). The correlation coefficient was 0.930 (n = 20, P < 0.01) when survival rates of each year were treated separately, and 0.994 (n = 4, P < 0.01) when survival rates for each country-group were averaged over years. These survival rates also seemed to be associated with income levels from high to low.

Partial regression coefficients of age-group survival rate variables are shown in Table 2. The survival rate of the middle stage (15 - 64 years old) contributed

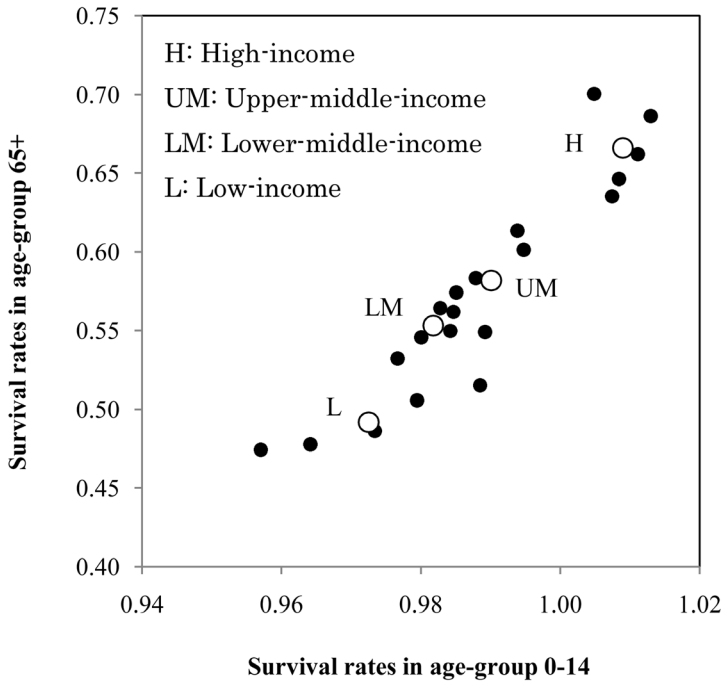

Figure 5. The correlation in survival rate between the immature (0 - 14 years old) and the elderly (more than 65 years old), estimated by using the population data for country-groups classified based on income levels [28]. Each black dot represents survival rates of each separate year for each country-group, and each larger circle represents survival rates averaged over years for each country-group.

Table 2. Partial regression coefficients of age-group survival rate variables.

Dependent variable	Intercept	Independent variables		
		AGE1	AGE2	AGE3
AGE1	0.4658 ± 0.1682*	—	0.5089 ± 0.2121*	0.0510 ± 0.0681
AGE2	0.3597 ± 0.1802	0.4970 ± 0.2072*	—	0.2066 ± 0.0466***
AGE3	−2.5625 ± 0.3426***	0.6264 ± 0.8359	2.5962 ± 0.5853***	—

*P < 0.05, ***P < 0.001. AGE1 represents the survival rate of the immature stage (0 - 14 years old), AGE2 the middle stage (15 - 64 years old), and AGE3 the elder stage (65+ years old).

significantly to the survival rates of both the immature and elder stages. On the other hand, the survival rate of the middle stage was significantly contributed by the survival rates of both the immature and elder stages. In the case of the survival rate of the elderly as a dependent variable, the intercept (−2.5625, $P < 0.001$) was remarkably negative, and a contribution to the survival rate from the middle stage (2.5962, $P < 0.001$) was remarkably large; on the other hand, the intercept (0.4658, $P < 0.05$) and a contribution to the survival rate from the middle stage (0.5089, $P < 0.05$) were significantly positive, when the survival rate of the immature stage was a dependent variable.

4. Discussion

The intrinsic rate of increase (r) and the net reproduction rate in the 5-year time-unit, estimated by using the basic vital statistics for the hypothetical population (the upper **Figure 1**), were −0.081 and 0.590, respectively. Because r needs to be more than zero and the net reproduction rate more than unity in order for the population to be maintained, this hypothetical population is considered a decreasing population, as expected. In general, the relationship between changes in $m(x)$ and corresponding changes in r for an increasing population tends to decrease with age; however, it is possible that the relationship is not necessarily monotonic and could increase with age in the case of a decreasing population like this hypothetical population (**Figure 3**; [1]). Under the conditions of population aging and diminishing population size, which countries such as Japan are or will be faced with, the relationship between changes in $m(x)$ and corresponding changes in r increased with age, so that benefits contributed by old people to $m(x)$ of younger individuals would have more positive effects on population increase than those contributed by any other young individual, if the amount of benefits is the same. At the same time, when elderly caring imposes costs on $m(x)$ of younger individuals, it is suggested that costs by old people would have more negative effects on population increase than costs by any other young individual, if the amount of costs is the same. Therefore, if there are contributions from each age to young individuals, the positive contributions from older ages would be most efficient for population increase. On the contrary, it is suggested that decrease of the population such as Japan would get worse, if young individuals continue to receive costs of elderly caring.

In this study, the selective advantage of high levels of care for the elderly as well as infants was examined by using an evolutionary genetic model assuming two alleles, one causing high levels of care for old parents as well as infants and the other causing a relatively indifferent behavior. In the population fixed with the indifferent allele which reduces $P(x)$ during infancy and later in life by 5 to 20%, the rare high care allele initially increased in frequency, if $m(x)$ was the same between the genotypes (**Figure 4(a)**). Therefore, it is suggested that the establishment of a solid relationship with parents during infancy, which increases survival rates during that period, would be selectively favorable, even if it is necessary to take care of old parents in the future, from which benefits in inclusive

fitness will not be expected. Because trajectories were below that for the Taiwan population in 1906, about 100 years ago (Figure 4(a), Taiwan), the assumption of 5% to 20% reduction in survival rates during infancy and later in life for relatively indifferent genotypes seems not to be unrealistic figures. However, the rates of increase in the frequencies of the high care allele were decreased, if high levels of elderly caring imposed costs on reproduction, and the high care allele was excluded from the population, if costs of care were sufficiently large (Figure 4(b) and Table 1). Therefore, it is suggested that high levels of care for the elderly as well as infants should not necessarily have absolute advantages, which may be excluded from a population if high levels of care for the elderly impose high costs on reproduction. However, in this situation, even though high levels of care for the elderly are matters of great concern, high levels of care for infants may also be excluded from the population due to pleiotropic constraints.

In addition, the case in which a rare relatively indifferent allele was introduced into a decreasing population fixed with a high care allele was examined, in order to discuss the possibility that we evolve to cut off caring for old parents, if elderly caring becomes a serious matter due to deterioration of population aging and diminishing population size, which is near the situation Japan is now or will be faced with. Although the population was decreasing, the rare relatively indifferent allele causing reduction in $P(x)$ during infancy and later in life by 5% to 20% could not invade the high care population, if $m(x)$ was the same between the genotypes (Figure 4(c)). Therefore, it is suggested that the genotype which establishes a solid relationship with parents and needs to take care of old parents altruistically in the future is still be selectively advantageous than the relatively indifferent genotype, even if the population does not have prospects for increase. As in the case of the high care gene invasion, the assumption of reduction in $P(x)$ during infancy and later in life by 5% to 20% seems not to be unrealistic, because trajectories of the indifferent gene exclusion were above that for Taiwan population in 1906, which was rather near the trajectory for the case of 20% reduction (Figure 4(c), Taiwan). However, the rates of decrease in the frequencies of the relatively indifferent allele were decreased, if indifferent individuals performed low levels of care for old parents and increased reproductive efforts, and the indifferent allele would exclude the high care genotype eventually from the population, if low levels of care resulted in sufficient increases in reproduction (Figure 4(d) and Table 1). In this situation, however, the indifferent genotype was necessary to increase $m(x)$ very much. For example, the indifferent genotype reducing $P(x)$ during infancy and later in life by 20% needed to increase $m(x)$ by about 100% in order to invade the population fixed with the high care allele. Therefore, selective advantages in establishing a solid relationship with parents are considered to be huge, even if old parents needs to be cared for altruistically in the future, which may impose even costs on reproduction. At the same time, the results that the indifferent allele was excluded from the population even though $m(x)$ of the genotype was increased so much may indicate that the behavior of care for old parents would not be readily excluded from the population

even though the population is decreasing, so that prohibiting population decrease would not be an easy task because this behavior would not be easily excluded from the population.

In this study, pleiotropic constraints in interactions between parents and offspring through the whole life were assumed in the evolutionary genetic model. In the analyses using the data from the UN [28], the highly significant correlation in survival probability between the immature and the elderly was in fact estimated (**Figure 5**). Although survival rates would be affected by many factors, the assumption made in the model was supported at least superficially. On the other hand, the variation in survival rates seemed to also be related with income levels at the same time; the higher the income levels of country-groups, the higher the survival rates of both age-groups. Although any adjustment was not conducted in using the data from the UN [28], this correlation is clear and seems reasonable from the standpoint of our expectations. Thus it is possible that the state of poverty or richness might affect the quality of the care for the young and the elderly; that is, environmental factors, such as poverty and richness, might influence the quality of the care. This may be true, but the question in this study in the first place is why we care for old parents, even though they have ceased reproduction and care for them costs very much. Therefore, the poverty-richness hypothesis does not answer the question why we care for old parents under the wealthy condition, even though the gain of inclusive fitness should not be expected. If we care for the elderly under the wealthy condition, but cannot under the poor condition, it is likely that we have the nature to care for old parents, because it is rather natural that we cannot care for the elderly under the poor condition, even though we hope to care for them. Thus the results of this analysis might suggest the predisposed association between care for old parents and care for infants, supporting the assumption made in the evolutionary genetic model of the elderly caring.

The results of the multiple regression analyses indicate significant positive contributions from the middle age-group to the immature and the elder age-groups, and significant positive contributions from the immature and the elder age-groups to the middle (**Table 2**). Because survival rates among age-groups were also highly correlated with each other, caution should be taken, but the results did not necessarily indicate the same tendency among partial regression coefficients. Especially, when the survival rate of the elder age-group was a dependent variable, the contribution of the middle age-group was remarkably positive, and the intercept of the regression model when both independent variables were zero was remarkably negative, suggesting that survival rates in the elder stage may be largely dependent on the contributions from offspring generation, and that it may be difficult for the elderly to survive without their contribution. Therefore, the results of the multiple regression analyses may imply the existence of the care for the elderly. On the other hand, when the survival rate of the immature age-group was a dependent variable, the intercept of the regression model was significantly positive. Because infants cannot survive by themselves,

the intercept, which should indicate the survival rate when contributions from the other age-groups are zero, may be expected to be near zero. Therefore, the significant positive intercept may imply the actual existence of contributions from a parental generation and a positive bias for survival of the immature stage when parental care is provided, which Bowl by [20] assumed. Overall, these results obtained by using population data from the UN [28] may be considered to support the assumption made in the evolutionary genetic model at least superficially.

5. Conclusions

From the standpoint of evolution, caring for old parents is considered to be a maladaptive behavior, because they have ceased to reproduce, so that this behavior is unlikely to produce benefits in inclusive fitness. In order for such a behavioral trait to be maintained within a population, some large benefits should exist in interactions between parents and offspring at other life stages. Thus in this study, I focused on the simple fact that we can survive only if someone (usually parents) does care for us, when we are infants, and discussed the evolutionary mechanism of elderly caring, using the evolutionary genetic model assuming pleiotropic constraints between care for infants by parents and care for old parents by offspring.

One of the most important consequences of population decrease, especially for human being, would be the increase of impacts from elder ages on population increase, and the ultimate population extinction would be the most important [1]. Therefore, deterioration of population aging and diminishing population size would result in serious social crisis. However, from the results obtained in this study, it may be concluded that the behavior of caring for old parents should not be excluded from a population until social situations are in serious crisis (establishing a solid relationship with parents and receiving care from them during infancy would have such a large selective advantage even if we have to care for old parents in the future), so that I believe it is unlikely that such a situation as in the Medawar's alarming statement or the movie "The Ballad of Narayama" (directed by Keisuke Kinoshita, the theme of which is a Japanese folk tale of a poor farm village, in which old people who have reached 70 years old must be carried on their offspring's shoulder to the mountain called Narayama and left there to die) will come about under current conditions. Nevertheless, this is not necessarily absolute, and it might be possible that the behavior of caring for old parents could be excluded from a population, if the serious social crisis comes about due to population aging and diminishing population size. However, the important point which should be emphasized is that behaviors which should evolve are not only abandonment of old parents but also abandonment of high levels of care for infants as well as positive interactions during other life stages, because elderly caring should be a behavioral trait which is not maintained within a population, if there are not pleiotropic constraints in interactions between parents and offspring. In other words, the behavior of elderly caring,

which costs very much, should not be maintained within a population unless it entails high selective advantages during infancy and/or other life stages. Therefore, it is unlikely that abandonment of old parents will evolve in spite of the fact that people have received high levels of care from their parents in their childhood. When the serious social crisis in elderly caring has come about, it should be considered, not simply that it is better to abandon elderly caring, but that effects of abandonment of old parents may be brought about in various aspects of interactions between parents and offspring. In this sense, the following statement Medawar made is very suggestive:

> "Those who argue that our concern is with the preservation of life in infancy and youth, so that pediatrics must forever take precedence of what people are beginning to call 'gerontology', fail to recognize that the outcome of pediatrics is to preserve the young for an old age that is grudged them. There is no *sense* in that sort of discrimination" (italicized by Medawar himself) [2].

The world in which the elderly can be easily abandoned may also be the world in which lives of infants are in serious danger. In fact, among the Ache, ancestral hunter-gatherers living in Paraguay, South America, infanticides as well as geronticides were common [29] [30]. From the perspectives of people living at present, neither world would be happy. However, it is also true that abuses of children as well as the elderly are frequently reported in our society. Is it pessimistic to consider this situation as that the frequency of the relatively indifferent allele is increasing in our society?

My care-worker experience at nursing facilities for the elderly and the Medawar's alarming statement quoted in Introduction motivated me to write this article and pose the serious matters of population aging and diminishing population size, which such countries as Japan are now or will be faced with. It is anticipated that the proportion of the elderly over 65 years old would reach about 40% of the total population in Japan in 2060, which is about 1.5 times as large as its current value [4]. It might be possible for us to imagine how serious this situation would be, if we consider a little about the situation, in which the number of old users in a nursing facility increases by 1.5 times under current care-work conditions. However, the real seriousness of the situation may not be imagined, because many events beyond our presumptions happen every day even at present. Although the evolutionary genetic model of elderly caring in this study is a simple numerical model for the hypothetical population suffering from population aging and diminishing population size, and there would be room for considering further details, such as the extent to which interactions between parents and offspring are pleiotropically constrained, I proposed the evolutionary mechanism of elderly caring in order to discuss this serious problem from various perspectives, other than social or political sciences. After discussing the evolutionary mechanisms of senescence, Medawar mentioned.

> "… there was *some* truth amidst a good deal of what we can now see to be

> nonsense, and… it would stir up his successors to think up a more polished and cogent explanation" (italicized by Medawar himself) [3].

I hope there is some truth in the model used in this study, which could result in valuable discussion and constructive criticisms for the future generations.

Acknowledgements

I thank Professor Brian Charlesworth (Univ. of Edinburgh) and Emeritus Professors Yuzuru Oguma (Univ. of Tsukuba) and Kenichi Aoki (Univ. of Tokyo) for valuable comments and suggestions on the manuscript. I am grateful to Dr. John Cartwright (Univ. of Chester) for his suggestions on this research at the early stage and to Mr. Takuji Miyo and Ms. Sachiko Miyo for valuable discussion on elderly caring. Finally, I would like to express my sincere appreciation to users and staffs at nursing facilities where I worked as a care-worker to provide me with invaluable opportunities for thinking about elderly caring.

References

[1] Charlesworth, B. (1994) Evolution in Age-Structured Populations. 2nd Edition, Cambridge University Press, Cambridge. https://doi.org/10.1017/CBO9780511525711

[2] Medawar, P.B. (1946) Old Age and Natural Death. *Modern Quarterly*, **1**, 30-56. (Reprinted in Medawar, P.B. (1981) The Uniqueness of the Individual. 2nd Revised Edition, Dover, New York, 1-27)

[3] Medawar, P.B. (1952) An Unsolved Problem of Biology. H. K. Lewis, London. (Reprinted in Medawar, P.B. (1981) The Uniqueness of the Individual. 2nd Revised Edition, Dover, New York, 28-54)

[4] Cabinet Office, Government of Japan (2016) Annual Report on the Aging Society. (In Japanese) http://www8.cao.go.jp/kourei/whitepaper/w-2016/gaiyou/28pdf_indexg.html

[5] Aoki, M. (2016) In Graying Japan, Caregivers in Short Supply. The Japan Times, 28 June, 3.

[6] Otake, T. (2016) Burden of "Double Care" on the Rise. The Japan Times, 7 May, 3.

[7] Charlesworth, B. (1978) Some Models of the Evolution of Altruistic Behaviour between Siblings. *Journal of Theoretical Biology*, **72**, 297-319.

[8] Hamilton, W.D. (1964) The Genetical Evolution of Social Behaviour. *Journal of Theoretical Biology*, **7**, 1-16.

[9] Wilson, D.S. (2015) Does Altruism Exist? Culture, Genes, and the Welfare of Others. Yale University Press, New Haven.

[10] Aoki, K. (1984) A Quantitative Genetic Model of Two-Policy Games between Relatives. *Journal of Theoretical Biology*, **109**, 111-126.

[11] Crow, J.F. and Aoki, K. (1982) Group Selection for a Polygenic Behavioral Trait: A Differential Proliferation Model. *Proceedings of the National Academy of Sciences of the United States of America*, **79**, 2628-2631. https://doi.org/10.1073/pnas.79.8.2628

[12] Bouchard, T.J. (2004) Genetic Influence on Human Psychological Traits: A Survey. *Current Directions in Psychological Science*, **13**, 148-151. https://doi.org/10.1111/j.0963-7214.2004.00295.x

[13] Davis, M.H., Luce, C. and Kraus, S.J. (1994) The Heritability of Characteristics As-

sociated with Dispositional Empathy. *Journal of Personality*, **62**, 369-391. https://doi.org/10.1111/j.1467-6494.1994.tb00302.x

[14] Finkel, D., Wille, D.E. and Matheny, A.P. (1998) Preliminary Results from a Twin Study of Infant-Caregiver Attachment. *Behavior Genetics*, **28**, 1-8. https://doi.org/10.1023/A:1021448429653

[15] Rushton, J.P., Fulker, D.W., Neale, M.C., Nias, D.K.B. and Eysenck, H.J. (1986) Altruism and Aggression: The Heritability of Individual Differences. *Journal of Personality and Social Psychology*, **50**, 1192-1198. https://doi.org/10.1037/0022-3514.50.6.1192

[16] Zahn-Waxler, C., Robinson, J.L. and Emde, R.N. (1992) The Development of Empathy in Twins. *Developmental Psychology*, **28**, 1038-1047. https://doi.org/10.1037/0012-1649.28.6.1038

[17] Lynch, M. (1987) Evolution of Intrafamilial Interactions. *Proceedings of the National Academy of Sciences of the United States of America*, **84**, 8507-8511. https://doi.org/10.1073/pnas.84.23.8507

[18] Bowlby, J. (1977) The Making and Breaking of Affectional Bonds. I. Aetiology and Psychopathology in the Light of Attachment Theory. *British Journal of Psychiatry*, **130**, 201-210. https://doi.org/10.1192/bjp.130.3.201

[19] Ainsworth, M.D.S. (1985) Attachments across the Life Span. *Bulletin of the New York Academy of Medicine*, **61**, 792-812.

[20] Bowlby, J. (1981) Psychoanalysis as a Natural Science. *International Review of Psycho-Analysis*, **8**, 243-256. https://doi.org/10.1111/j.1468-2273.1981.tb01318.x

[21] Chisholm, J.S. (1996) The Evolutionary Ecology of Attachment Organization. *Human Nature*, **7**, 1-38.

[22] Simpson, J.A. and Belsky, J. (2008) Attachment Theory within a Modern Evolutionary Flamework. In: Cassidy, J. and Shaver, P.R., Eds., *Handbook of Attachment: Theory, Research, and Clinical Applications*, 2nd Edition, Guilford Press, New York, 131-157.

[23] Charlesworth, B. and Charnov, E.L. (1981) Kin Selection in Age-Structured Populations. *Journal of Theoretical Biology*, **88**, 103-119.

[24] Hamilton, W.D. (1966) The Moulding of Senescence by Natural Selection. *Journal of Theoretical Biology*, **12**, 12-45.

[25] Charlesworth, B. and Charlesworth, D. (2010) Elements of Evolutionary Genetics. Roberts and Company Publishers, Greenwood Village.

[26] Miyo, T., Oguma, Y. and Charlesworth, B. (2003) The Comparison of Intrinsic Rates of Increase among Chromosome-Substituted Lines Resistant and Susceptible to Organophosphate Insecticides in *Drosophila melanogaster*. *Genes and Genetic Systems*, **78**, 373-382. https://doi.org/10.1266/ggs.78.373

[27] Charlesworth, B. and Charlesworth, D. (1973) The Measurement of Fitness and Mutation Rate in Human Populations. *Annals of Human Genetics*, **37**, 175-187. https://doi.org/10.1111/j.1469-1809.1973.tb01825.x

[28] United Nations, Department of Economic and Social Affairs, Population Division (2015) World Population Prospects: The 2015 Revision. https://esa.un.org/unpd/wpp/

[29] Hill, K. and Hurtado, A.M. (1996) Ache Life History: The Ecology and Demography of a Foraging People. Aldine Transaction, New Brunswick.

[30] Mace, R. (2000) Evolutionary Ecology of Human Life History. *Animal Behaviour*, **59**, 1-10.

Scientific Research Publishing

Open Journal of Genetics, 2017, 7, 40-49
http://www.scirp.org/journal/ojgen
ISSN Online: 2162-4461
ISSN Print: 2162-4453

Increase Data Characters to Construct the Molecular Phylogeny of the *Drosophila auraria* Species Complex

Lu Gan[1*], Gaodong Li[1*], Wenhao Li[1], Qingtao Zeng[1], Yong Yang[1,2#]

[1]College of Life Science, Hubei University, Wuhan, China
[2]Hubei Collaborative Innovation Center for Green Transformation of Bio-Resources, Wuhan, China
Email: #yangyong@hubu.edu.cn

How to cite this paper: Gan, L., Li, G.D., Li, W.H., Zeng, Q.T. and Yang, Y. (2017) Increase Data Characters to Construct the Molecular Phylogeny of the *Drosophila auraria* Species Complex. *Open Journal of Genetics*, **7**, 40-49.
https://doi.org/10.4236/ojgen.2017.71004

Received: December 12, 2016
Accepted: March 27, 2017
Published: March 30, 2017

Abstract

Previous phylogenetic analyses of the *auraria* species complex have led to conflicting hypotheses concerning their relationship; therefore the addition of new sequence data is necessary to discover the phylogeny of this species complex. Here we present new data derived from 22 genes to reconstruct the phylogeny of the *auraria* species complex. A variety of statistical tests, as well as maximum likelihood mapping analysis, were performed to estimate data quality, suggesting that all genes had a high degree of contribution to resolve the phylogeny. Individual locus was analyzed using maximum likelihood (ML), and the concatenated dataset (21,882 bp) were analyzed using partitioned maximum likelihood (ML) and Bayesian analyses. Separated analysis produced various phylogenetic relationships. Phylogenetic topologies from ML and Bayesian analysis based on concatenated dataset show that *D. subauraria* was well supported as the first species by separated analysis, concatenated dataset analysis, and some previous analysis, then followed by *D. auraria* and *D. biauraria*, *D. quadraria* and *D. triauraria*. The close relationships of *D. quadraria* and *D. triauraria* were consistent with most previous studies. The phylogenetic position of the *D. auraria* and *D. biauraria* will be resolved by more data sets.

Keywords

Drosophila auraria Species Complex, Phylogenetic Reconstruction, Multiple Genes

1. Introduction

The members of the *auraria* species complex in which ordinarily five members were involved (*D. auraria*, *D. biauraria*, *D. subauraria*, *D. quadraria* and *D. tri-*

*These authors contributed equally to this work.

auraria) [1] [2] were considered as perfective model for reproductive isolation, flight activities, ability of diapauses, courtship songs and cold tolerance[3]. Recently, the phylogeny of the *auraria* species complex was studied based on various data, specifically, DNA sequence data. However, all analyses brought conflicting phylogenetic hypotheses [4]-[18] (**Figure 1**). The cause of the conflicting hypotheses is not known. All previous studies on the phylogeny of this species complex are based on different sample sizes or genetic markers. Differences in the number of taxa and the number of genes can have an effect on phylogenetic accuracy [19]. In many previous phylogenetic treatments of this species complex, representatives of only 4 species or less were included [4] [7] [13] [14] [15] [16] [17] [20]. Incomplete or insufficient taxon sampling has led to major inconsistencies in phylogenetic reconstructions [20] [21] [22] [23] [24]. On the other hand, differing sets of genetic markers were selected in previous studies, the most previous investigations were based on no more than two genetic markers [9] [11] [12] [15] [16] [17], phylogenetic hypotheses deduced from small amounts of sequence data would be incongruent or pool support [25]. Moreover, highly conserved genetic markers were involved in some analyses [3] [7] [13] [14] but some authors suggested that fast-evolving DNA regions were prior to analysis the molecular phylogenies of closely related species [26]. Although the phylogenetic relationships of these five members were deduced from 17 loci [18], the hypothesis that "increasing sampling outside the group may decrease accuracy" [27] may have applied; therefore, Yang (2012) did not resolve this complex problem. Many investigations suggested that maximizing gene numbers was advantageous to resolve complex phylogeny [28] [29]. Consequently, it

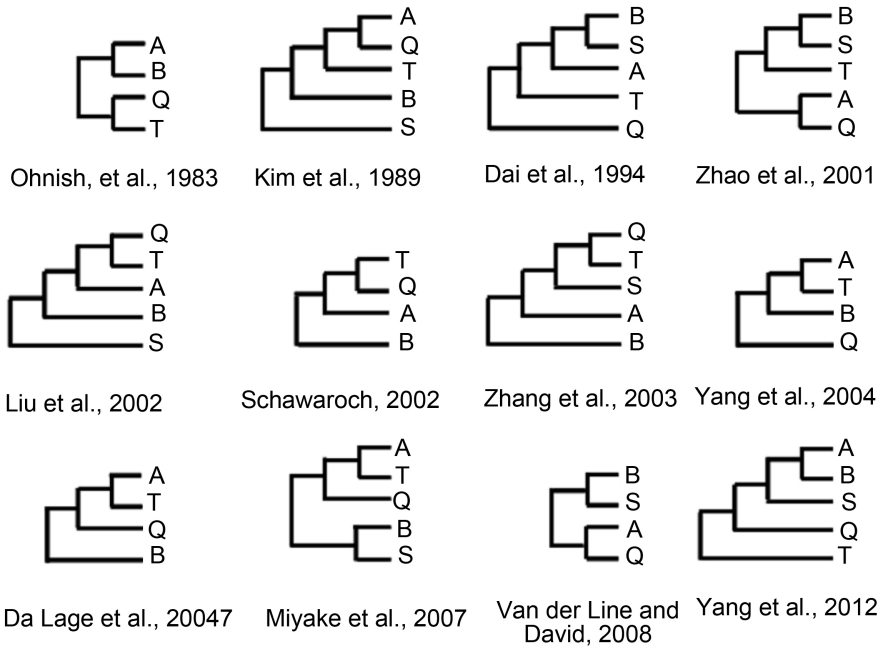

Figure 1. The diagram of the phylogenetic relationships of *auraria* species complex based on different data sets. (A, B, Q, S and T are *D. auraria, D. biauraria, D. subauraria, D. quadraria* and *D. triauraria*, respectively.

was advantageous to reconstruct the phylogeny of the five species based on increasing gene sampling sizes.

Finally, in this study, 22 genes segments were first used to reanalyze the phylogenetic relationships of *D. auraria*, *D. biauraria*, *D. subauraria*, *D. quadraria* and *D. triauraria*. These loci included partial genomic sequences of mitochondrial genes: cytochrome oxidase subunit I (COI), cytochrome oxidase subunit II (*COII*), mitochondrial genes ND1 (*ND*1) and ND4 (*ND*4); and nuclear ribosomal sequences: 28S rDNA (28*S*), internal transcribed spacer of nuclear ribosomal DNA (ITS including ITS1, 5.8S, 2S, and ITS2), and nuclear genes: amylase (*amy*), a paralogue of the amylase genes (*amr*), sn-glycerol-3-phosphate dehydrogenase (*gpdh*), histone 2 spacers (*h*2*s*), Dopa decarboxylase (*ddc*), extra sexcombs (*esc*), hunchback (*hb*), extron 2, 3, 4 of alcohol dehydrogenase gene (*adh*234), nucleoporin 96 - 98 gene (*nup*), membrane protein (patched) gene (*ptc*), and Xenopus Cdc6 (*cdc*), genes for odorant-binding protein 57d, odorant-binding protein 57e (*odo*), multidrug-resistance associated protein 1-(*mrp*1), wingless (*wgl*), intron1 of bab gene (*bab*1), endophilin B (*endoB*).

2. Materials and Methods

2.1. The Study Taxa and Sequences Data

The sequences of *COI*, *COII*, *ND*1, *ND*4, 28*S*, *ITS*, *amy*, *amryel*, *gpdh*, *h*2*s*, *ddc*, *esc*, *hb*, *adh*234, *nup*, *ptc*, *Cdc*6 were download from GneBank (GenBank accession numbers were listed in Yang *et al*., 2012). Sequences of odorant-binding protein 57d and e (*odo*), multidrug-resistance associated protein 1(*mrp*1), wingless (*wgl*), intron1 of *bab* gene (*bab*1) and endophilin B (*endoB*) were newly presented in this study. The detail information is given in Table 1. *D. melanogaster* was selected as the out group. PCR conditions and primers are listed in Table 2.

2.2. Sequence Alignment and Statistical Tests

Alignment of multiple DNA sequences was performed with MUSCLE for each gene [30]. The base composition, variable sites, and average genetic *p*-distance among all taxa were calculated by MEGA 4 [31]. The degree of nucleotide substitution saturation for each gene was tested using DAMBE 4.5.47 software [32]. A test for homogeneity of base frequencies across taxa was conducted using

Table 1. Experimental species name and GenBank accession numbers.

Species	GenBank Accession Number				
	odo	*mrp*1	*wgl*	*bab*1	*endoB*
D. auraria	EU835204	HQ850387	DQ778962	EU835204	YY971326
D. biauraria	AY465281	AE154622	HN546526	AB235812	JN974392
D. quadraria	AY465282	HQ850397	HN546524	AB235813	JN974393
D. subauraria	AY465283	AE154623	HN546523	AB235814	YY971325
D. triauraria	AY465284	HQ850403	HN546527	AB235815	JN974394
D. melanogaster	AE014297	AE014134	AE014134	AE014296	AE013599

Table 2. PCR conditions and primers.

Gene segment and length (bp)	Forward primers (5'-3')	Reverse primers (5'-3')
odorant-binding protein 57d and e (~1070 bp)	OdF1: CTTTTGAATTACATTGCCGTA OdF2: GCTATAAGCACGCGGATT OdF3: TTCCGTCGTCTTCAATCCCT	OdR1: AATCCGCGTGCTTATAGC OdR2: AGGGATTGAAGACGACGGAA OdR3: CATCCAGATATTTGAAGCGA
multidrug-resistance associated protein 1 (~1070 bp)	MrpF1: TTATGCGGTTCCCAGT MrpF2: GGAATGCCGCGACAGACCAA MrpF3: GCTGGGACCCTCTGTGCT	MrpR1: TTGGTCTGTCGCGGCATTC MrpR2: AGCACAGAGGGTCCCAGC MrpR3: TAGCTTCGAGAAGCAAGT
wingless (~1070 bp)	WF1: GCTGGATGCGACTGGCAA WF2: GGTCGCAAACATAATAGGT	WR1: ACCTATTATGTTTGCGACC WR2: GGCGCATCGCTCCACCACCA
endophilin B (~1070 bp)	EbF1: GGAGGCGGGTACCACGA EbF2: GCCGCTGCGCAAGTTCCT	EbR1: AGGAACTTGCGCAGCGGC EbR2: ACTACAAGCAGTGCGGCGA
intron1 of *bab* gene (~1070 bp)	BabF1: CACATAAAAATCAGCAACA	BabR1: TGCCGGACGCATGCTGCAAC

PAUP 4.0 beta 10 [33].

2.3. Nucleotide Evolutionary Model Selection, Phylogenetic Analysis

For separate analysis, maximum likelihood (ML) trees for each locus were constructed in PAUP*v.4.0b10 [33] with the best nucleotide substitution model as determined by the Akaike Information Criterion (AIC). The concatenated dataset was divided into 22 partitions representing 25 genes, the best-known likelihood (BKL) tree for concatenated dataset was inferred after conducting 1000 RAxML runs using the f-d option for thorough searching and bootstrap replicates were performed in the multithread compiled version of RAxML-7.04. And Bayesian analysis running in MrBayes-3.1.2 [34] with 1,000,000 MCMC generations using the substitution model and parameters deduced from Model Test 3.06 [35].

2.4. Alternative Phylogenetic Hypotheses Test

SH test using CONSEL version 0.1 [36] were performed to test the statistical support of most of the previous hypotheses (**Figure 1**) and the hypotheses deduced from separate analysis. The BKL tree as optimal likelihood tree was modified using TreeView [37] to produce phylogenetic trees representing the alternative hypotheses.

3. Results

3.1. Sequence Alignment and Statistical Tests

Aligned sequences for the individual gene regions varied from 334 to 2455 bp in length, and the variation and parsimony informative sites were quite different among all genes, *bab*1 and 28S contain the highest and lowest number of parsimony informative sites, respectively (**Table 3**). Most of the average *p*-distances among the taxa were lower than 10% (18 out of 22); the *mrp*1 and *bab*1 have

Table 3. The characters of the 22 genes across 5 of the *auraria* species complex.

Characteristics	*28S*	*adh*	*amy*	*amr*	*bab1*	*cdc*	*COI*	*COII*	*ddc*	*endoB*	*esc*	*gpd*	*h2s*	*hb*	*mrp1*	*ITS*	*ND1*	*ND4*	*odo*	*nup*	*ptc*	*wgl*	22 genes
Length (bp)	*343*	668	705	1732	2455	1469	407	362	564	1117	360	298	497	504	2029	1425	939	1338	*2103*	492	1328	781	21,882
*P**	0	5	29	29	282	41	16	10	0	88	0	12	3	2	176	32	26	12	72	18	44	60	747
*A**	2.2	1.4	8.6	7.4	14.4	11	4.5	6.2	5.3	8.6	4.7	5.8	7.6	4.8	13.8	9.2	4.0	3.0	2.2	11.5	6.4	12.2	8.2

P*: Parsimony information sites; A*: Average genetic p-distance (%).

larger values, whereas 28*S*, *adh*234, and *odo* have very small values (all lower than 2.2%). The test for substitution saturation [32] show that all gene regions have no substitution saturation. The sequences of all fragments show homogeneity of base frequencies ($P \geq 0.05$).

3.2. Phylogenetic Analysis

The topologies of the trees deduced from the concatenated dataset under the ML and Bayesian analysis were completely identical (**Figure 2**), the five species in *auraria* species complex consisted of three lineages, the *D. subauraria* is the first species, then *D. auraria* and *D. biauraria*, *D. quadraria* and *D. triauraria.* The percent bootstrap support in ML analysis and posterior probabilities in Bayesian analysis all are 100 and 1.0, respectively. Maximum likelihood (ML) trees for each locus constructed in PAUP*v.4.0b10 [33] were different (**Figure 3**).

3.3. Alternative Hypotheses Test

All ML trees from each gene completely supported the *melanogaster* species group comprised of three monophyletic lineages: the *ananassae* subgroup, the *montium* subgroup, and the *melanogaster* subgroup plus oriental subgroups; however genes differed in the relationships among these groups. The *montium subgroup* was supported as the sister taxon of all remaining members of the *melanogaster* group by 6 of the 17 genes. The close relationships of the *melanogaster*, *suzukii*, and *takahashii* subgroups were supported by 4 genes. All 17 genes supported *suzukii* and *takahashii* as the sister lineages, and 7 genes supported the monophyly clade of *ficusphila*, *eugracilis*, *elegans*, and *rhopaloa* subgroups. Five of the 17 genes accepted the paraphyly of the *suzukii* subgroup, in which *D. lucipennis* is the sister species of *D. elegans* (see **Table 3**, and supplemental material). The p-values of all alternative phylogenetic hypotheses (**Figure 1**), except the hypotheses (p-values = 0.016) suggested by van der Linde and Houle (2008), are significantly lower than 0.005.

4. Discussion

Quality Evaluation for All the Representative Genes

Data quality is crucial for phylogenetic analysis, especially, based on DNA data-

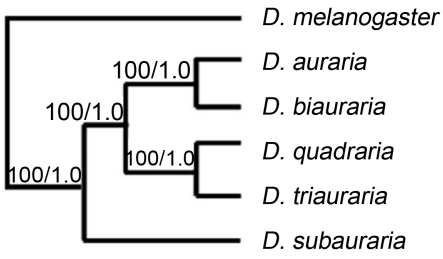

Figure 2. The phylogenies of *auraria* complex were deduced from the concatenated dataset under the ML and Bayesian analysis. the number on the branch refer to bootstrap support in ML analysis and posterior probabilities in Bayesian analysis. *D. melanogaster* was out group.

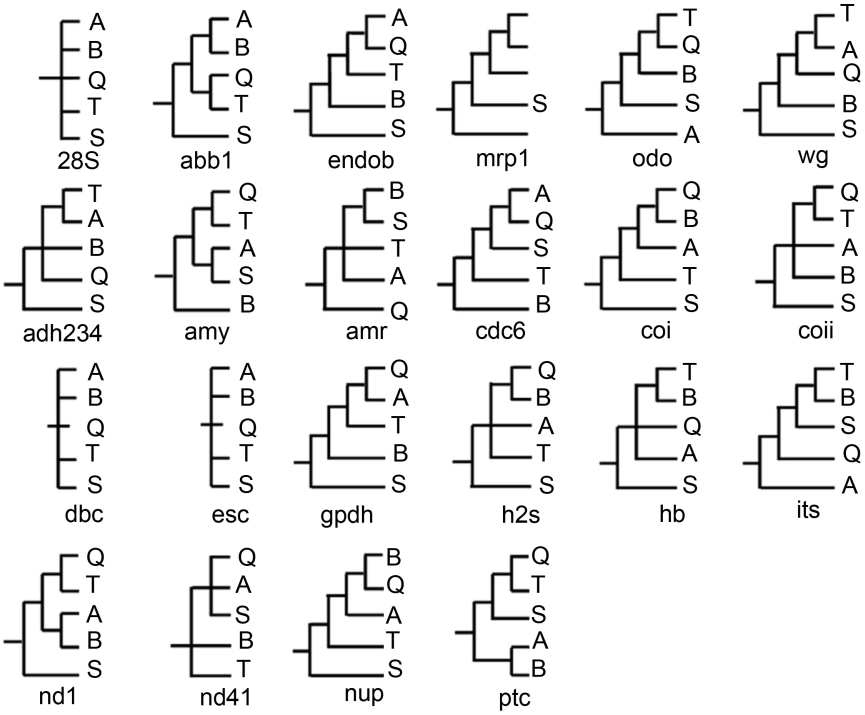

Figure 3. Phylogenetic relationships of the *auraria* species complex were constructed based on different single locus.

sets; improper data which conceal conflicting evidence could lead to incorrect phylogenetic tree topologies [38]. All data in this study included various parsimony information sites, and the average genetic p-distance was close to 10% (**Table 1**), except for 28S (4.0%) and the four mtDNA (COI = 8.6%, COII = 8.1%, ND1 = 7.8%, and ND4 = 7.9%). 28S is 343 bp in length, a small part of the 28S complete sequence; the low genetic p-distance could come from arbitrary selection, hence, the conservative regions of this gene was selecte D. The mtDNA was traditionally considered as conservative genes; these kinds of genes were effective for discovering "higher-level" phylogenies. The likelihood mapping (**Table 3**) results also indicated that these four genes have significant contributions to resolve the phylogeny of the *melanogaster* species group (COI = 80.0%, COII =

81.5%, ND1 = 88.8%, and ND4 = 89.9%).

According to the test method described by Xia *et al*. (2003), most of the genes showed no saturation (**Table 2**) except for ITS and nup, with little saturation. Although saturation could bring noise into the phylogenetic analysis, these two genes included useful phylogenetic information, and the values of the average genetic p-distance suggest that these two genes have "fast" evolutionary rates. Thollesson (1999), however, suggested that fast genes also included phylogenetic signal and Scholler (1994) demonstrated divergent genes, e.g., ITS, are appropriate genetic markers to discover the phylogeny of the melanogaster species group. The likelihood mapping also showed these two genes have considerable amounts of phylogenetic signal. Therefore, these genes were included to avoid loss of phylogenetic signal.

The concatenated dataset has a large amount of useful parsimony information (parsimony sites, 4365), and the average genetic p-distance (14.0%), together with the likelihood mapping values (resolved quarters, 99.5%), indicated that the combined data has provided sufficient phylogenetic signals to resolve the phylogenetic relationships of the melanogaster species group.

In the present study, the phylogenetic relationships of the auraria species complex were reconstructed based on 22 gene segments (**Figure 2**). The phylogenetic tree indicated that *D. subauraria* is the first species, then *D. auraria and D. biauraria*, *D. quadraria* and *D. triauraria*, which were supported with high bootstrap values and posterior probabilities (100 and 1.0, respectively). *D. subauraria* was absent in the phylogenetic analysis before it was found in Japan [39], but based on the RAPD data, it was the first species in this species complex. The phylogenetic position ofthe *D. subauraria* as the first species was also supported by some previous analysis [3] [5] [8] [10] and halfofthe separated analysis (11 genes of the 22 genes. On the other hand, no evidences show that *D. subauraria* could produce fertility offspring with other member of this species complex. All the above mentioned evidences suggested that the *D. subauraria* was the earlier divergent species. *D. quadraria* was always considered as the first species before *D. subauraria* was taken into phylogenetic analysis [8], however, *D. quadraria* and *D. triauraria* could produce cross-fertilize generations [2], which indicated that there was no reproductive isolation between these two species, therefore, these two species were ever treated as the same species [1] [2]. Moreover, some previous analysis based on two-dimensional electrophores and DNA data sets [9] [15] supported the close relationships of *D. quadraria* and *D. triauraria*, which were also supported by sixes gene data sets in the analysis (abb1, edo, ND1, ptc, amy, and COII). The phylogenetic relationships of *D. auraria* and *D. biauraria* were only supported by three gene data sets in separated analysis (ptc, abb1, and mrp1) and some previous evidences [4] [18]. From all the analysis, it could be draw a primary conclusion that *D. subauraria* was the first species, and *D. quadraria* and *D. triauraria* were the close relative two species, but the phylogenetic positions of *D. auraria* and *D. biauraria* should be resolved basedon more data sets. However, in the future analysis of the phylogenetic relationships

of these species, it is advantageous to consider add more DNA sequences data. From all the separated and most previous analysis, it was obvious that any signal gene data was limited to discover the phylogenetic relationships of this species complex because almost all the previous analysis and separated analysis were rejected by analysis based on the concatenated dataset. Especially, some authors emphasized that phylogenetic hypotheses.

5. Conclusions

1) The phylogenetic relationship of Auraria species complex were analyzed based on different data, *D. subauraria* is the first species, *D. auraria and D. biauraria*, *D. quadraria* and *D. triauraria* as the second clusters.

2) To discuss the phylogenetic relationship of relative species, it was advantageous to concatenate large dataset than single locus.

Acknowledgements

This work was supported by the Department science of Hubei Province (2012FFB00304) and the Foundation of Wuhan City Technology Bureau (2013070104010013). We thank the Drosophila Genetic Resource Center and the Bloomington Drosophila center for supplying the flies used in this study.

References

[1] Bock, I.R. and Wheeler, M.R. (1972) The *Drosophila melanogaster* Species Group Studies in Genetics VII. Univ. Texas Publ., 7213.

[2] Kimura, M.T. (1987) Habitat Differentiation and Speciation in the *Drosophila auraria* Species-Complex (Diptera, Drosophilidae). *Japanese Journal of Entomology*, **55**, 429-436.

[3] Miyake, H. and Watada, M. (2007) Molecular phylogeny of the *Drosophila auraria* Species Complex and Allied Species of Japan Based on Nuclear and Mitochondrial DNA Sequences. *Genes & Genetic Systems*, **82**, 77-88. https://doi.org/10.1266/ggs.82.77

[4] Ohnishi, S., Kim, K.-W. and Watanabe, T.K. (1983) Biochemical Phylogeny of the *Drosophila montium* Species Subgroup. *The Japanese Journal of Genetics*, **58**, 141-151. https://doi.org/10.1266/jjg.58.141

[5] Kim, B.K., Watanabe, T.K. and Kitagawa, O. (1989) Evolutionary Genetics of the *Drosophila montium* Subgroup. I. Reproductive Isolations and the Phylogeny. *The Japanese Journal of Genetics*, **64**, 177-190. https://doi.org/10.1266/jjg.64.177

[6] Dai, Z. (1994) Study on Evolutionary Genetics of *Drosophila auraria* Species Complex—Cladistic Analysis and Phonetic Analysis. *Journal of Genetics and Genomics*, **21**, 436-440.

[7] Goto, S.G. and Kimura, M. (2001) Phylogenetic Utility of Mitochondrial COI and COI Inuclear Gpdh Genes in Drosophila. *Molecular Phylogenetics and Evolution*, **18**, 404-422.

[8] Zhao, Z.-M. (2001) Genetic Differentiation within *Drosophila auraria* Species Complex Revealed by Random Amplified Polymorphic DNA (RAPD). *Acta Zoologica Sinica*, **47**, 625-631.

[9] Liu, Z.-M. (2002) Preliminary Studies on the Thr-Gly Region of the Period Gene in

the *Drosophila auraria* Species Complex. *Zoological Research*, **23**, 1-6.

[10] Lu, J., Lu, J., Chen, H.-X., Zhang, W.-X. and Dai, Z.-H. (2002) Molecular Phylogeny of *Drosophila auraria* Species Complex. *Journal of Genetics and Genomics*, **29**, 39-49.

[11] Zhang, Z. and Inomata, N. (2003) Phylogeny and the Evolution of the *Amylase* Multigenes in the *Drosophila montium* Species Subgroup. *Journal of Molecular Evolution*, **56**, 121-130. https://doi.org/10.1007/s00239-002-2384-3

[12] Yang, Y., Zhang, Y.P., Qian, Y.H. and Zeng, Q.T. (2004) Phylogenetic Relationships of the *Drosophila melanogaster* Species Group Deduced from Spacer Regions of Histone Gene H2A. *Molecular Phylogenetics and Evolution*, **30**, 336-343.

[13] Mou, S.L., Zeng, Q.T., Yang, Y., Qian, Y.H. and Hu, G.A. (2005) Phylogeny of Melanogster Species Group Inferred from ND4L and ND4 Genes. *Zoological Research*, **26**, 344-349.

[14] Lewis, R.L., Beckenbach, A.T. and Mooers, A.O. (2005) The Phylogeny of the Subgroups within the Melanogaster Species Group: Likelihood Tests on COI and COII Sequences and a Bayesian Estimate of Phylogeny. *Molecular Phylogenetics and Evolution*, **37**, 15-24.

[15] Da Lage, J.L., Kergoat, G.J., Maczkowiak, F., Silvain, J.-F., Cariou, M.-L. and Lachaise, D. (2007) A Phylogeny of Drosophilidae Using the *Amyrel* Gene: Questioning the *Drosophila melanogaster* Species Group Boundaries. *Journal of Zoological Systematics and Evolutionary Research*, **45**, 47-63. https://doi.org/10.1111/j.1439-0469.2006.00389.x

[16] Van der Linde, K. and Houle, D. (2008) A Supertree Analysis and Literature Review of the Genus *Drosophila* and Closely Related Genera (Diptera, Drosophilidae). *Insect Systematics & Evolution*, **39**, 241-267. https://doi.org/10.1163/187631208788784237

[17] Van der Linde, K., Houle, D., Spicer, G.S. and Steppan, S. (2010) A Supermatrix-Based Molecular Phylogeny of the Family Drosophilidae. *Genetics Research*, **92**, 25-38. https://doi.org/10.1017/S001667231000008X

[18] Yang, Y., Hou, Z.-C., Qian, Y.-H., Kang, H. and Zeng, Q.-T. (2012) Increasing the Data Size to Accurately Reconstruct the Phylogenetic Relationships between Nine Subgroups of the *Drosophila melanogaster* Species Group (Drosophilidae, Diptera). *Molecular Phylogenetics and Evolution*, **62**, 214-223.

[19] Rokas, A. and Carroll, S.B. (2005) More Genes or More Taxa? The Relative Contribution of Gene Number and Taxon Number to Phylogenetic Accuracy. *Molecular Biology and Evolution*, **22**, 1337-1344. https://doi.org/10.1093/molbev/msi121

[20] Schawaroch, V.A. (2002) Phylogeny of a Paradigm Lineage: The *Drosophila melanogaster* Species Group (Diptera: Drosophilidae). *Biological Journal of the Linnean Society*, **76**, 21-37. https://doi.org/10.1111/j.1095-8312.2002.tb01711.x

[21] Hillis, D.M., Pollock, D., Mcguire, J.A. and Zwickl, D.J. (2003) Is Sparse Taxon Sampling a Problem for Phylogenetic Inference? *Systematic Biology*, **52**, 124-126. https://doi.org/10.1080/10635150390132911

[22] Rosenberg, M.S. and Kumar, S. (2001) Incomplete Taxon Sampling Is Not a Problem for Phylogenetic Inference. *Proceedings of the National Academy of Sciences of the United States of America*, **98**, 10751-10756. https://doi.org/10.1073/pnas.191248498

[23] Pollock, D.D., Zwickl, D.J., Mcguire, J.A. and Hillis, D.M. (2002) Increased Taxon Sampling Is Advantageous for Phylogenetic Inference. *Systematic Biology*, **51**, 664-671. https://doi.org/10.1080/10635150290102357

[24] Zwickl, D.J., Hillis, D.M. and Crandall, K. (2002) Increased Taxon Sampling Greatly Reduces Phylogenetic Error. *Systematic Biology*, **51**, 588-598. https://doi.org/10.1080/10635150290102339

[25] Kopp, A. and True, J.R. (2002) Phylogeny of the Oriental *Drosophila melanogaster* Species Group: A Multilocus Reconstruction. *Systematic Biology*, **51**, 786-805. https://doi.org/10.1080/10635150290102410

[26] Schlötterer, C. and Hauser, M.T. (1994) Comparative Evolutionary Analysis of rDNA ITS Regions in Drosophila. *Molecular Biology and Evolution*, **11**, 513-522.

[27] Rannala, B., Huelsenbeck, J.P., Yang, Z., Nielsen, R. and Cannatella, D. (1998) Taxon Sampling and the Accuracy of Large Phylogenies. *Systematic Biology*, **47**, 702-710. https://doi.org/10.1080/106351598260680

[28] Bapteste, E., Brinkmann, H., Lee, J. A., Moore, D. V., Sensen, C. W., Gordon, P., Durufle, L., Gaasterland, T., Lopez, P., Muller, M. and Philippe, H. (2002) The Analysis of 100 Genes Supports the Grouping of Three Highly Divergent Amoebae: *Dictyostelium*, *Entamoeba*, and *Mastigamoeba*. *Proceedings of the National Academy of Sciences of the United States of America*, **99**, 1414-1419. https://doi.org/10.1073/pnas.032662799

[29] Wolf, A.M., Conaway, M.R., Crowther, J.Q., Hazen, K.Y., Nadler, J.L., Oneida, B. and Bovbjerg, V.E. (2004) Translating Lifestyle Intervention to Practice in Obese Patients with Type 2 Diabetes: Improving Control with Activity and Nutrition (ICAN) Study. *Diabetes Care*, **27**, 1570-1576. https://doi.org/10.2337/diacare.27.7.1570

[30] Edgar, R.C. (2004) MUSCLE: Multiple Sequence Alignment with High Accuracy and High Throughput. *Nucleic Acids Research*, **32**, 1792-1797. https://doi.org/10.1093/nar/gkh340

[31] Tamura, K., Dudley, J., Nei, M. and Kumar, S. (2007) MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) Software Version 4.0. *Molecular Biology and Evolution*, **24**, 1596-1599. https://doi.org/10.1093/molbev/msm092

[32] Xia, X., Xie, Z., Salemi, M., Chen, L. and Wang, Y. (2003) An Index of Substitution Saturation and Its Application. *Molecular Phylogenetics and Evolution*, **26**, 1-7.

[33] Swofford, D.L. (2002) PAUP*: Phylogenetic Analysis Using Parsimony (and Other Methods) Version 4. Associates, Sinauer, Sunderland, MA.

[34] Ronquist, F. and Huelsenbeck, J.P. (2003) MrBayes 3: Bayesian Phylogenetic Inference under Mixed Models. *Bioinformatics*, **19**, 1572-1574. https://doi.org/10.1093/bioinformatics/btg180

[35] Posada, D. and Crandall, K. (1998) MODELTEST: Testing the Model of DNA Substitution. *Bioinformatics*, **14**, 817-818. https://doi.org/10.1093/bioinformatics/14.9.817

[36] Shimodaira, H. and Hasegawa, M. (2001) CONSEL: For Assessing the Confidence of Phylogenetic Tree Selection. *Bioinformatics*, **17**, 1246-1247. https://doi.org/10.1093/bioinformatics/17.12.1246

[37] Page, R.D. (1996) TreeView: An Application to Display Phylogenetic Trees on Personal Computers. *Computer Applications in the Biosciences*, **12**, 357-358.

[38] Kimura, M. (1983) The Neutral Theory of Molecular Evolution. Cambridge University Press, Cambridge. https://doi.org/10.1017/CBO9780511623486

[39] Ohnishi, S. and Watanabe, T.K. (1984) Systematics of the *Drosophila montium* Species Subgroup: A Biochemical Approach. *Zoological Science*, **1**, 801-807.

Scientific Research Publishing

Open Journal of Genetics, 2017, 7, 50-61
http://www.scirp.org/journal/ojgen
ISSN Online: 2162-4461
ISSN Print: 2162-4453

Y-Chromosomal Profile and Mitochondrial DNA of the Chevalier Bayard (1476?-1524)

Gérard Lucotte*, Alexandra Bouin Wilkinson

Institute of Molecular Anthropology, Paris, France
Email: *lucotte@hotmail.com

How to cite this paper: Lucotte, G. and Wilkinson, A.B. (2017) Y-Chromosomal Profile and Mitochondrial DNA of the Chevalier Bayard (1476?-1524). *Open Journal of Genetics*, **7**, 50-61.
https://doi.org/10.4236/ojgen.2017.71005

Received: February 23, 2017
Accepted: March 27, 2017
Published: March 30, 2017

Abstract

Objective: We report the results of Y-chromosomal profile and mtDNA (mitochondrial DNA) of the Chevalier Bayard (1476?-1524). **Methods:** His genomic DNA was extracted from a tooth of his mandible. His Y-STRs profile was obtained using the AmFirst identifier PCR amplification kit. The mtDNA genomic sequence intervals for *HVR*1 and *HVR*2 were amplified by PCR, with specific primers. **Results**: We obtained the complete STR (Short Tandem Repeats) profile, based on fourteen STRs (DYS19, DYS385.a, DYS389.I and .b, DYS390, DYS391, DYS392, DYS393, DYS438, DYS439, DYS448, DYS456 and DYS458 and Y-GATA-H4). The deduced Y-STRs profile corresponds to the sub-clade S21 of the major European haplogroup **R1b-M269** (the "Germanic" haplotype). There are six mutations (16093C, 16211T and 16519C in the *HVR*1 sequence, 263G, 309.1C and 315.1C in the *HVR*2 sequence) in the mtDNA of Bayard. The 263G mutation determines the **H** mtDNA haplogroup and the 16211T suggests the **H5** sub-clade of the **H** haplogroup (a sub-clade found at >8% frequency in France, at the periphery of the Alpine arch region). This sub-clade **H5** (subsequently assimilated to the **H10e** haplotype) is that (with a perfect match) of a modern living male related (to 32 generations) to the Bayard maternal ascendance. The Bayard mtDNA haplotype was found once only in a database of 100 South-German mtDNA control sequences. **Conclusions**: The resulting **R1b-M269** Y haplogroup established confirms the Germanic origin of the Bayard ancestors, suggested by genealogic studies concerning his paternal ascendance. The result concerning the mtDNA **H10e** haplotype found in the modern living male related to Bayard by matrilinear ascendance establishes that the DNA tooth is well of him, with a 99% of chance.

Keywords

The Chevalier Bayard, Tooth, Genomic DNA, Y-STRs Profile, mtDNA Mutations

1. Introduction

The Chevalier Bayard (1476?-1524)—named Pierre (III) du Terrail—is well known in French history as "le chevalier sans peur et sans reproche". He is unanimously considered as the last true knight in shining armor, the last flower of the late Middle Ages, and the epitome of chivalry before the modern world took over [1].

As a soldier, Bayard was one of the most skilful commanders of his time. He served under three successive French kings: Charles VIII (with whom he participated to the conquest of the kingdom of Naples), Louis XII (with whom he acted to the conquest of Genoa) and François Ier. On the accession of François Ier in 1515, Bayard was made lieutenant-general (governor) of Dauphiné (a French region at this time); after the victory of Marignan, to which his valour largely contributed, he had the honour of being conferred knighthood from his youthful sovereign.

When the war broke out again between François Ier and Charles Quint (the Holy Roman Emperor), Bayard held Mézières against an army of 3500 men, and after six weeks compelled the imperial generals to raise the siege. This stubborn resistance saved central France from invasion. The parliament thanked Bayard as the saviour of his country and the king made him (in 1521) a knight of the order of Saint Michel.

In 1524, Bayard was sent to Italy with the Admiral Bonnivet, who had been defeated at Robecco and wounded in a combat. During the retreat of the French army Bayard repulsed the foremost pursuers, but in guarding the rear at the passage of the Sesia, was mortally wounded at Rovasenda (April 30, 1524) by an arquebuse ball which pierced his armor.

We do not know exactly when Bayard was born (between 1473 and 1476) but he was very probably born at Château Bayard (near Pontcharra, Isère), in Dauphiné.

Bayard was the second son (the third-born child of a kinship of eight children) of Aymon du Terrail (1458-1490), the second Lord de Bayard, and Hélène Alleman de Laval (born before 1436-died after 1504); we ignore the exact date of their marriage.

Bayard's cranium is at present kept in the Dauphiné Museum of Grenoble. We have reconstituted the whole Bayard cranium (Lucotte, unpublished) from his upper, lateral and posterior parts, his mandible and most of his osseous facial part. Detailed examination of the reconstituted cranium establishes that it is the skull of a Caucasian male, aged from 45 to 50 years; some observed particularities of the mandible (brachygnathy, elevated corpus, squared and non-protruding chin) correspond to those depicted on the portraits of Bayard.

A molar tooth was extracted from this mandible. Genomic DNA obtained from this tooth permits us to study the Y-chromosomal profile and the mtDNA (mitochondrial DNA) of Bayard, in a similar way to that recently explored [2] for King Richard III of England.

2. The Sample

We extracted a tooth from the mandible articulated to the cranium. This incomplete cranium, presumed as the Bayard's skull, is stored in the Dauphiné Museum (Grenoble). This tooth (Figure 1) is the first molar located at the left side of the mandible (the tooth number 36 according to the Nomenclature Dentaire Internationale). The second root basis of that tooth was saved, and the interior of the canal of the crown was abrased with a dentist drill. The recuperated powder was sterilized, and then used for DNA extraction.

3. Methods

The dentine powder was washed with 15% HCl, rinsed with UV-treated ddH_2O, and dried under an UV lamp for 15 min.

The sterilized powder was introduced in 15 ml tubes (Costar), and DNA was extracted according to a modified silica-based protocol [3]. Briefly, 2 ml of an undiluted commercial guanidine thiocyanate solution (DNAzol®) was added to the tube and incubated at room temperature for 3 days; after that, the supernatant was passed through a silica column (QIAquick®, Qiagen).

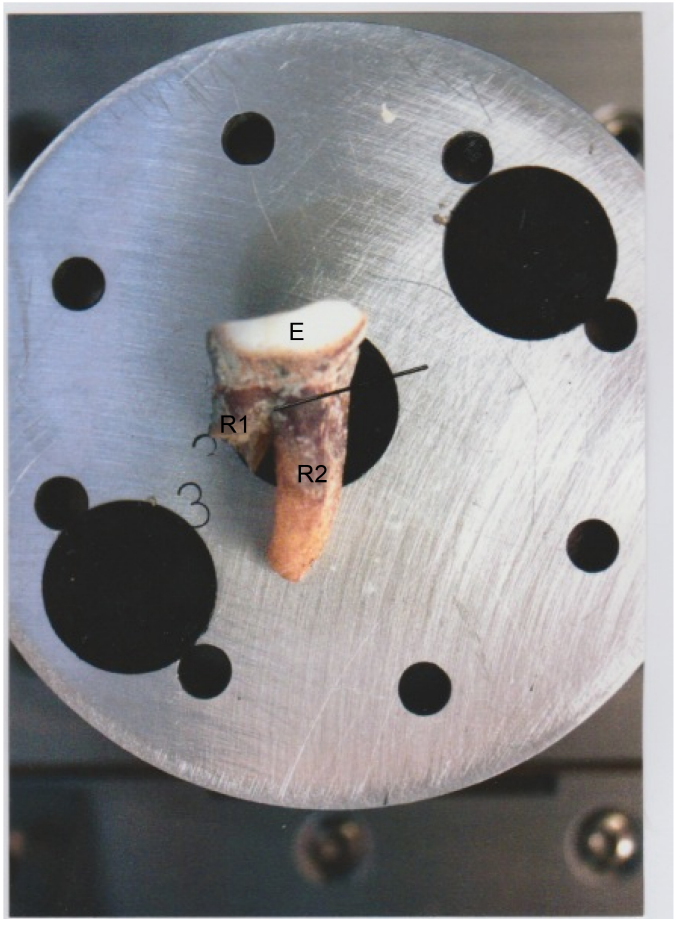

Figure 1. The tooth extracted from the mandible (lingual view). E: enamel, R1: the first (broken) tooth' root; R2: the second (sawed along the line) tooth root.

All staff involved in the sampling wore protective clothing, sterile gloves and facemasks, to prevent exogenous contamination. DNA extraction and purification were performed according to our previously published protocol [4], in a dedicated laboratory.

We amplified from the genomic DNA extracted 14 Y-chromosomal short tandem repeats (Y-STRs) by using the AmFirst Identifier PCR amplification kit (Amp FIRSTLY filer™, Applied Biosystems), according to the instructions given by the Company; this amplification kit is specially adapted to the study of ancient DNA (a-DNA). The fifteen STRs studied are the followings: DYS19 (= DYS394), DYS385.a, DYS389.I and .b (DYS389.b = DYS389.II *minus* DYS389.I), DYS390 (=DYS708), DYS391, DYS392, DYS393 (=DYS395), DYS438, DYS439 (=Y-GATA-A4), DYS448, DYS456, DYS458, and DYS635 (=Y-GATA-C4); Y-GATA-H4 was detected in an independent PCR (Polymerase Chain) reaction. To detect the long STR alleles, we proceeded to two successive essays, with various degrees of stringency.

We predict, starting from the allele values, the corresponding Y haplogroup using the Whit Atey's Haplogroup Predictor [5].

From the genomic DNA extracted, we studied also the mtDNA. The mtDNA genomic sequence intervals for *HVR*1 and *HVR*2 (Hypervariable regions 1 and 2) were amplified by PCR with primers F15971 and R16610 and with primers L15 and H484, respectively. For each PCR, the DNA extract from the tooth root specimen was amplified by PCR in a 12.5 µl reaction mixture: 2 mM $MgCl_2$, 50 mM KCl, 10 mM Tris/HCl pH = 9, 0.1% Triton X-100, 0.2 mM each dNTP, 0.1 µM each primer, and 2.5 U of DNA polymerase (Ampli Taq Gold; Applied Biosystems). The amplification was carried out with an initial denaturation step at 95˚C for 6 min., followed by 35 cycles at 95˚C for 1 min., 55˚C for 1 min., and 72˚C for 1 min.

PCR products were purified from agarose gel (QIA-Quick PCR purification kit; Qiagen). Both strands of all the amplified mtDNA fragments eluted from agarose gel slides were directy sequenced (Big Dye Terminator Cycle Sequencing kit; Applied Biosystems) and separated (ABI PRISM 3130 Genetic Analyzer; Applied Biosystem).

The sequences obtained were aligned against the Revised Cambridge Reference Sequence [6], to identify the presence of polymorphic sites. Seqscape software (Applied Biosystems) and Clustal analysis were used for pairwise alignment.

The laboratory performed DNA typing under strict precautions, following previously published criteria for ancient DNA authentication [7].

4. Results

A quantity of about 100 ng of a total genomic DNA was obtained from the dental powder.

Preliminary experiments on the genomic DNA extracted established that it contains sequences of the amelogenin human gene, which show two peaks cor-

responding to the two X and Y chromosomes. Consequently, the individual under study is truly a XY male.

Table 1 gives the allelic Y-STRs profile obtained. All these allele values were confirmed in a second PCR essay; we were not able to obtain the allele value corresponding to DYS635 (probably the longest).

The predicted Y haplogroup using the Haplogroup Predictor is **I-M223**.

We obtained DNA sequences (from 16,025 to 16,555 and from 67 to 369, respectively) of the *HVR*1 (16,024 16,569) and *HVR*2 (1-576) segments of the mtDNA. Three mutations (16093C, 16221T and 16519C) are present in the *HVR*1 sequence, and three also (263G, 309.1C and 315.1C) in the *HVR*2 sequence. The same results were obtained in a replication study.

5. Discussion

In the present study we obtain, with genomic DNA extracted from one of his tooth, the Y-STRs profile (coming from his father) and the mtDNA sequence (coming from his mother) of Bayard.

The predicted Y-haplogroup **I-M223** is equivalent to **I2a2a**, previously known as **I2b1** of haplogroup I [8]. In fact one of us (G.L.) established that this haplogroup corresponds to the sub-clade S21/U106 of the major haplogroup **R1b-M269** [9]. In current European populations, this sub-clade shows a peak of S21 frequencies centred on Germany and surrounding areas (Figure 2); because of these particularities, it was named the "Germanic" Y-haplotype.

Table 2 summarizes what we know about the real paternal Bayard ancestry [10]. Bayard's father, Aymon du Terrail, was the second Lord de Bayard. His father, Pierre said "le Jeune" (1421-1460) was the first Lord de Bayard; his

Table 1. Y-chromosomal STR data analysis.

Numbers	Y-STRs	Allele values
1	DYS19	16
2	DYS385.a	14
3	DYS389.I	14
4	DYS389.b	18
5	DYS390	24
6	DYS391	10
7	DYS392	12
8	DYS393	14
9	DYS438	10
10	DYS439	12
11	DYS448	20
12	DYS456	15
13	DYS458	16
14	DYS635	?
15	Y-GATA-H4	11

Table 2. Bayard's paternal ancestry.

Generations	Bayard paternal ancestry	Names	Birth- and death rates	Titles
0	Chevalier Bayard	Pierre (III) du Terrail	1476?-1524	third Lord de Bayard
1	Bayard's father	Aymon (or Amon) du Terrail	1458-1490	second Lord de Bayard
2	Bayard's grand-father	Pierre II (junior) du Terrail	1421-1460	first Lord de Bayard
3	Bayard's aïeul	Pierre I (senior) Terrail	1387-1433/34	
4	Bayard's bisaïeul	Pierre Terrail	?-1387	

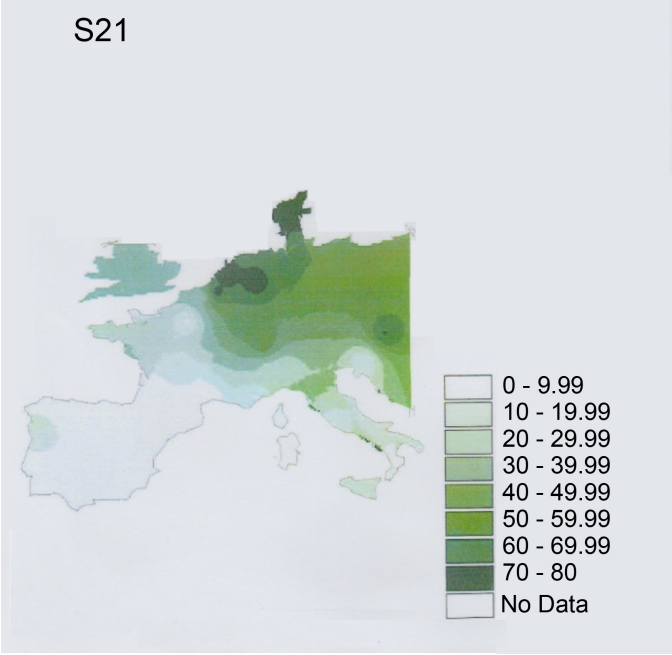

Figure 2. Isofrequency map of R-S21 in West-Europe (from Lucotte, 2015). Isofrequency lines indicate the artificial limits (of the areas with various nuances of green) between decreasing S21 values from the peak.

grand-father, Pierre said "le Vieux", was born in 1387 and died in 1433/34. We do not know the birth date of the ancestor Pierre Terrail, but he probably died in 1387. Nothing is known with certainty concerning Bayard's remote paternal ancestors.

Some past biographers spread various legendary accounts about Bayard's paternal ancestry [11]. But it is generally admitted that the House of de Terrail ("Terra alii" means stranger, in latin language) is very ancient in Dauphiné, and that they come from Germany ("at the time when the Emperors possessed the Dauphiné"). This explains why Bayard's Y-STRs profile corresponds to the Germanic Y-haplotype S21/U106.

The only-known Bayard descendant is an (illegitimate) daughter: Jeanne (1501-

1580), who married in 1525 (one year after Bayard's death) to François II de Bocsozel (1483-1532); they have four sons: Pierre, Jehan, Piraud and Soffrey, but the break in the Bayard paternal transmission line (with Jeanne) do not permit us to compare their Y-STRs profiles (and those of their further male descendants) to that of Bayard.

Concerning now Bayard's mtDNA haplogroup, the *HVR*2 sequence mutation 263G defines the mtDNA haplogroup **H**; it is the most commonly found mtDNA haplogroup in Europe [12]. The *HVR*1 sequence mutation 16221T indicates (because of the absences of both 456T and 16304C mutations) the **H5** sub-clade of **H** [13].

The Eupedia map [14] of sub-clade **H5** shows a remarkable concentration (that can attain >8%) of **H5** frequencies (in West-Europe) at the near-periphery of the Alpine arch (in Slovenia, Austria, Switzerland and in the south-eastern part of France).

Bayard inherited his mtDNA haplotype from his mother Hélène Alleman de Laval. Now the Dauphiné region-from which Hélène came from-is located at the Western part of the Alpine arch periphery (an area with the maximal concentration of **H5** frequencies). But, as for other European nobility [2], the female mobility of Hélène's family tends to be higher than the general population.

J. C. Parisot de Bayard, who funded this study, was able to identify a modern descendant of Bayard's family (**Figure 3**). This subject (named P. R.) is a living male individual born near the town of Annecy (in Savoy), 32 generations removed from Bayard (from his mother Hélène Alleman du Terrail) on the female line.

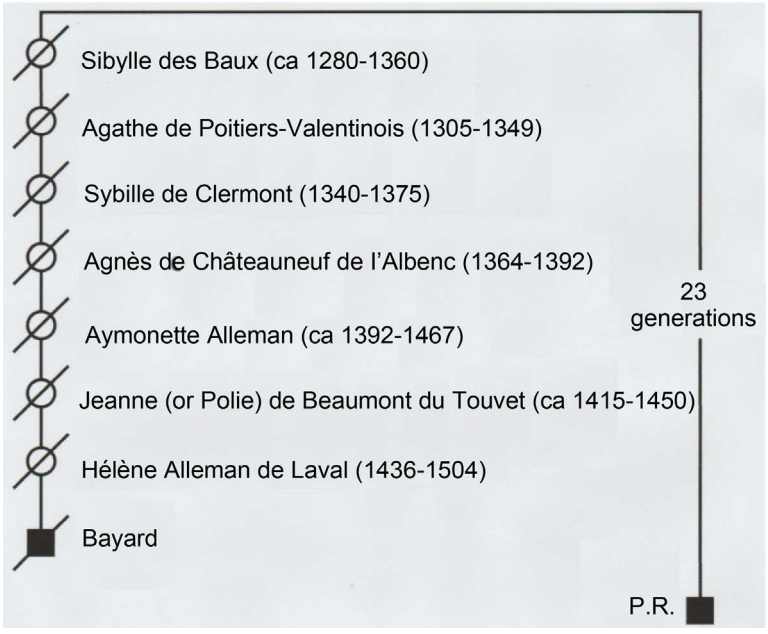

Figure 3. Genealogical links between Bayard and a modern-day relative (P. R.) who participated in this study. Genealogical information links P. R. (who descended from Bayard's mother) and Bayard through a female-only line (ca: circa).

The mtDNA analysis (repeated twice) of P.R. shows a perfect match (he had the three *HVR*1 mutations 16093C, 16221T and 16519C, and the three other *HVR*2 mutations 263G, 309.1C and 315.1C) between his mtDNA haplotype and that of Bayard; this is consistent with this subject being a matrilineal relative-in a similar form that it was found in [2]-at the genealogical time depth considered.

In contrast to false paternity, false-maternity is much less likely [2]; but historical records of female-line lineages are usually more difficult to track over multiple generations, due to the change of women's surnames after marriage. In the case reported here, the family tree concerns noble families (often better recorded) tracing down from Bayard to the beginning of the 13th century.

Because of the maximal value of >8% observed frequency of the sub-clade **H5** in the geographic region considered, we considered at first that there is less than about 10% of probability that the mtDNA match between P.R. and Bayard could have occurred by chance. But that is certainly an overestimate, because Bayard's mtDNA haplotype (Table 3) is only one of those constituting haplogroup **H5** considered as a whole. To note that the attribution adopted here to consider Bayard's mtDNA haplogroup as being of the sub-clade **H** (usually determined by the two supplementary mutations 16304C and 456T) is entirely due to the presence in it of the 16221T mutation; this mutation is generally considered, together with the 16085T and 16106A mutations, as a typical "Alpine mutation" [15].

Next we investigated more precisely the probability that Bayard's mtDNA haplogroup had occurred by chance, by tempting to find it in three databases of complete mtDNA control sequences concerning subjects originating from geographic regions located in the North periphery of the Alpine arch. These three databases are: 1/a first database [16] concerning 104 Slovenians and 144 Bosnians; 2/a second database [17] concerning 273 unrelated West-Eurasians from Austria; 3/a third database [18] concerning one hundred of samples that were collected from native German speakers in the middle of Southern Germany (in the region of Ulm city, that is located between Lake Constance and the Swabian Alps).

The only perfect match we found is for the UL1A8 subject of South-Germany (Table 4) of the third database; he is nomenclatured under the mtDNA haplogroup **H*** (the paragroup). So, for this geographic region , we found Bayard's

Table 3. Bayard's mtDNA haplotype.

HVR sequences	Mutations	Comments
1	16093C	this variant appeared in about 5% of most haplogroups, but most commonly in **K**
	16221T	indicates the **H5** sub-clade of **H**
	16519C	corresponds to a hotspot; it appears in almost every haplogoup, in over half of them
2	263G	determinates the **H** haplogroup
	309.1C	is one of the most recurrent mutation
	315.1C	most members of **H** also have this mutation

Table 4. Results of the research to find the Bayard mtDNA haplotype in three databases of mtDNA control sequences.

Populations	Samples	Bayard mtDNA haplotype						Attributed haplogroups	%
		16093C	16221T	16519C	263G	309.1C	315.1C		
1. Slovenia and Bosnia	Number 45 in the list	+	+	?	+	+	+	H	1/248
2. Switzerland	F1E3	+	-	+	+	+	+	H*	1/273
3. South-Germany	UL1A8	+	+	+	+	+	+	H*	1/100

+indicates presence of the mutation; −: their absence; ?: uncertain.

mtDNA haplotype at a frequency of 1/100.

The subject number 47 (a Slovenian) of the first database is of a mtDNA haplogroup very similar to that of Bayard's (but he had 150T as a supplementary mutation in the *HVR*2 sequence); the reason why he had not 16519C is that the second primer of the PCR reaction used for *HVR*1 sequence covers until 16400 only. The nearest mtDNA haplotype we found in the second database is that of the Swiss f1E3 subject (but he had 16271C as a supplementary mutation in the *HVR*1 sequence); the fact that he had not 16221T is possibly related to the event that this variant could be a phantom mutation [19], that is a systematic artefacts generated in the course of the sequencing process itself.

A synthetic study [20] concerned an analysis of 1350 mtDNA haplotypes belonging to **H**, originating from Central Europe: Austria (**H** samples = 973), Germany (=31), Hungary (=71), Macedonia (=100), Romania (=124), and from Dubaï (=51). The prevalence of sub-clades **H1**, **H5**, **H6** and **H10** (defined by the 16093C and 16221T mutations), **H13**, **H14**, **H15**, **H16**, **H17** and **H21** was consistent across Europe; the **H10** sub-clade particularly concerns 44 samples (3.3% of the total), but manifests hardly and genetic heterogeneity.

Today [21] the 16221T mutation defines the **H10e** type. Various sub-types of **H10e** are characterized by some specific mutations (Table 5). We had the opportunity to examine 56 mitogenomes (complete mtDNA sequences) belonging to **H10e** for the presence of the Bayard mtDNA haplotype: three **H10e1** sequences (among 9), four **H10e1a** (among 6), three **H10e2** (among 4) and one **H10e3a** (among 2) contained it.

Among the thirty five mitogenomes examined, twelve of them (Table 6) are of the Bayard mtDNA haplogroup: three subjects from Denmark, three from UK, two from England, one from Wales, one from France (he is not P.R.) and two from USA (who revendicate remote Anglo-Saxon ancestries).

6. Conclusion

In conclusion, it is the first time that the DNA of the Chevalier Bayard is studied (for paternal and maternal ancestries). His Y-STRs profile shows that he belongs to the Germanic S21/U106 Y haplogroup sub-clade. His mtDNA haplogroup,

Table 5. Characteristic mutations of the **H10e** haplogroup type and of four sub-types found in the collection of 56 mitogenomes bearing the Bayard mtDNA haplotype.

Type	Mutations	Numbers found
H1Oe	16221T	35
Sub-types		
H10e1	13830C	9
H10e1a	16266T	6
H10e2	14602G	4
H10e3a	961C	2

Table 6. Characterisations of the twelve **H10e** mitogenomes (on 35) of the Bayard mtDNA haplotype.

Numbers (on 12)	Numbers (on 35)	GenBank ID	Geographic origins
1	7	KF161474	Denmark
2	26	JX153333	Denmark
3	29	KF161060	Denmark
4	11	JQ701809	UK
5	21	JQ705324	UK
6	24	JQ705702	UK
7	14	JQ703082	England
8	18	JQ704209	England
9	17	JQ704082	Wales
10	5	HQ662520	France
11	3	GU569076	USA
12	4	HM101252	USA

found again in a living male individual removed from 32 generations on the female line of Bayard ancestry, is of the **H5** sub-clade further precised as being the **H10e** type. The Bayard mtDNA haplotype is found at an approximate percentage of 1/100 in the geographic region located at the periphery of the Alpine arch. We now move towards phenotypical DNA markers concerning his skin, eyes and hair pigmentation and his nose and chin forms, in order to compare them to the corresponding characters observable on Bayard's portraits.

Acknowledgements

We thank T. Tomasset (UST of Compiègne) for the photograph of the tooth; R. Rottenberg (Rosny-sous-Bois) for the obtaining of the dentine powder of the interior of the tooth; F. Dieterlen (Geneva, Switzerland) for his help to construct the Y-SNP S21 isofrequency map. P. R. (Annecy) accepted to participate anonymously to the study. The Prefect J. C. Parisot de Bayard (Peols) provided a detailed genealogy linking P.R. to Bayard (his remote ancestor). We acknowledge J. P. Jospin, Chief Curator at the Dauphiné Museum, who permitted to one of us (G. L.) to have access to the cranium. Thanks to Pr A. Torroni (Pavia Uni-

versity, Italy) who informed us about the most recently progress concerning 56 mitogenomic sequences belonging to **H10e**.

Funding

The present work was realized with the financial help of J. C. Parisot de Bayard and of "les Amis de Bayard" Association.

Conflict of Interest

The authors declare no conflict of interest.

References

[1] Champier S. (1525) The Whole Life of the Brave Chevalier Bayard. Gilbert de Villiers Editor, Lyon.

[2] King, T.E., Gonzalez Fortes, G., Balareque, P., Thomas, M.G., Balding, D., Maisano Delser, P., *et al.* (2014) Identification of the Remains of King Richard III. *Nature Communications*, **5**, Article No. 5631. https://doi.org/10.1038/ncomms6631

[3] Höss, M. and Pääbo, S. (1993) DNA Extraction from Pleistocene Bones by a Silica-Based Purification Method. *Nucleic Acids Research*, **21**, 3913-3914. https://doi.org/10.1093/nar/21.16.3913

[4] Lucotte, G. (2010) A Rare Variant of the mtDNA HVS1 Sequence in the Hairs of Napoléon's Family. *Investigative Genetics*, **1**, 7-10. https://doi.org/10.1186/2041-2223-1-7

[5] Athey, T.W. (2006) Haplogroup Prediction from Y-STR Values Using a Bayesian Allele Frequency Approach. *Journal of Genetic Genealogy*, **2**, 34-39.

[6] Andrews, R.M., Kubacha, I., Chinnery, P.F., Lightowlers, R.N., Turnbull, D.M. and Howel, N. (1999) Reanalysis and Revision of the Cambridge reference Sequence for Human Mitochondrial DNA. *Nature Genetics*, **23**, 147. https://doi.org/10.1038/13779

[7] Hofreiter, M., Serre, D., Poinar, H.N., Kuch, M. and Pääbo, S. (2001) Ancient DNA. *Nature Reviews Genetics*, **2**, 353-359. https://doi.org/10.1038/35072071

[8] Athey, T.W. and Nordtvedt, K. (2005) Resolving the Placement of Haplogroup I-M223 in the Y-Chromosome Phylogenetic Tree. *Journal of Genetic Genealogy*, **1**, 54-55.

[9] Lucotte, G. (2015) The Major Y-Chromosome Haplogroup R1b-M269 in West-Europe, Subdivided by the Three SNPs S21/U106, S145/L21 and S28/U152, Shows a Clear Pattern of Geographic Differentiation. *Advances in Anthropology*, **5**, 22-30. https://doi.org/10.4236/aa.2015.51003

[10] Verdier, R. (2007) Des Terrail à Bayard, in "Cross Stories of the Chevalier Bayard". University Press of Grenoble Editor, 41-54.

[11] Parisot de Bayard, J.C. (2016) Genealogy of the Magnificent Chevalier Bayard. Christian Editor, Paris.

[12] Roostalu, U., Kutuev, I., Loogväli, E.L., Metspalu, E., Tambets, K. and Reidla, M. (2006) Origin and Expansion of Haplogroup H, the Dominant Human Mitochondrial DNA Lineage in West Eurasia: The Near Eastern and Caucasian Perspective. *Molecular Biology and Evolution*, **24**, 436-448. https://doi.org/10.1093/molbev/msl173

[13] List of Mitochondrial DNA (mtDNA) Haplogroups and Subclades with Their De-

fining Mutations. http://www.eupedia.com/europe/Haplogroup_H_mtDNA.shtml

[14] Eupedia: haplogroup H (mtDNA), see Eupedia map of mtDNA haplogroup H5.

[15] Malyarchuck, B.A. and Derenko, M.V. (1999) Molecular Instability of the Mitochondrial Haplogroup T Sequences at Nucleotide Positions 16292 and 16296. *Annals of Human Genetics*, **63**, 489-497. https://doi.org/10.1046/j.1469-1809.1999.6360489.x

[16] Malyarchuk, B.A., Grzybowski, T., Dorenko, M.V., Czarny, J., Drobnic, K. and Miscicka-Sliwka, D. (2003) Mitochondrial DNA Variability in Bosnians and Slovenians. *Annals of Human Genetics*, **67**, 412-425. https://doi.org/10.1046/j.1469-1809.2003.00042.x

[17] Bandstätter, A., Niederstätter, H., Pavlic, M., Grubwieser, P. and Parson, W. (2007) Generating Population Data for the EMPOP Database—An Overview of the mtDNA Sequencing and Data Evaluation Processes Considering 273 Austrian Control Region Sequences as Example. *Forensic Science International*, **166**, 164-175. https://doi.org/10.1016/j.forsciint.2006.05.006

[18] Bandstätter, A., Klein, R., Duftner, N., Wiegand, P. and Parson, W. (2006) Application of a Quasi-Median Network Analysis for the Visualization of Character Conflicts to a Population Sample of Mitochondrial DNA Control Region Sequences from Southern Germany (Ulm). *International Journal of Legal Medicine*, **120**, 310-314. https://doi.org/10.1007/s00414-006-0114-x

[19] Bandelt, H.J., Quintana-Murci, L., Salas, A. and Macaulay, V. (2002) The Fingerprint of Phantom Mutations in Mitochondrial DNA Data. *American Journal of Human Genetics*, **71**, 1150-1160. https://doi.org/10.1086/344397

[20] Brandstäter, A., Zimmermann, B., Wagner, J., Göbel, T., Röck, A.W., Salas, A., *et al.* (2008) Timing and Deciphering Mitochondrial DNA Macro-Haplogroup R0 Variability in Central Europe and Middle East. *BMC Evolutionary Biology*, **8**, 191. https://doi.org/10.1186/1471-2148-8-191

[21] Phylo Tree. Org-mtDNA Tree Build 17.18 Feb 2016: Subtree RO.

Scientific Research Publishing

Open Journal of Genetics, 2017, 7, 62-67
http://www.scirp.org/journal/ojgen
ISSN Online: 2162-4461
ISSN Print: 2162-4453

Methylenetetrahydrofolate Reductase (MTHFR) Gene Mutations in Patients with Idiopathic Scoliosis: A Clinical Chart Review

Mark W. Morningstar[1*], Megan N. Strauchman[1], Clayton J. Stitzel[2], Brian Dovorany[3], Aatif Siddiqui[4]

[1]Natural Wellness & Pain Relief Center, Grand Blanc, MI, USA
[2]Lancaster Spinal Health Center, Lititz, PA, USA
[3]Posture & Spine Care Center, Green Bay, WI, USA
[4]Esprit Wellness, New York, NY, USA
Email: *drmorningstar@nwprc.com

How to cite this paper: Morningstar, M.W., Strauchman, M.N., Stitzel, C.J., Dovorany, B. and Siddiqui, A. (2017) Methylenetetrahydrofolate Reductase (MTHFR) Gene Mutations in Patients with Idiopathic Scoliosis: A Clinical Chart Review. *Open Journal of Genetics*, **7**, 62-67.
https://doi.org/10.4236/ojgen.2017.71006

Received: January 9, 2017
Accepted: March 27, 2017
Published: March 30, 2017

Abstract

The effects of genetic variations of methylenetetrahydrofolate reductase (MTHFR) enzyme activity have been the discussion of many research papers. It has been associated with multiple neurological sequelae, and has been implicated in other chronic diseases. Although many genetic influences on the development and/or progression of idiopathic scoliosis have been reported, there has been no report of any relationship between MTHFR mutations and idiopathic scoliosis. This paper compared two groups of patients who received MTHFR genetic testing. One group had a history of idiopathic scoliosis, while the other served as a control group. The scoliosis group showed a positive MTHFR mutation in 23 out of 44 patients, while the control group showed 11/44 ($P < 0.01$). Given the increased incidence of MTHFR defects in scoliosis patients, this study warrants further investigation into how MTHFR variations may trigger the development or progression of idiopathic scoliosis.

Keywords

Genetics, MTHFR, Polymorphism, Methylation, Scoliosis, Spine

1. Introduction

Recent medical literature has seen demonstrated a substantially increased interest in genetic testing for predicting various diseases, as well as to evaluate the functionality of various metabolic pathways. For example, one of the most studied single nucleotide polymorphisms (SNPs) is the methylenetetrahydrofolate reductase (MTHFR) enzyme. MTHFR gene variations have been associated with

ADHD [1], autism [2], cataract [3], colon cancer [4], glioma or meningioma [5], methotrexate toxicity [6], and migraines [7].

Idiopathic scoliosis is a curvature of the spine involving a 3-dimensional displacement measuring at least 10 degrees [8]. Evidence [9] suggests that there are multiple potential associations between specific genetic [10], neurological [11], and/or endocrine [12] variations that may lead to the cause, or progression of, idiopathic scoliosis. Although reduction in MTHFR activity has been linked to other health conditions, it has not been studied in patients with idiopathic scoliosis. Given that methylation is responsible for many enzymatic conversions along several hormone pathways, including the melatonin pathway [13], it is postulated that alterations of the genes responsible for encoding MTHFR activity may somehow be associated with scoliosis. This is in light of the fact that melatonin deficiency [14] and melatonin signaling abnormalities [15] have been previously implicated in scoliosis etiology.

Single nucleotide polymorphisms of the 677CT and 1298AC alleles cause a decrease in MTHFR enzyme activity. [16]. A homozygous mutation (677TT) of the 677CT allele has been shown to result in a decrease in MTHFR enzyme activity by 60% [17], while a heterozygous mutation of both alleles can also cause a 50% - 60% reduction in MTHFR activity. [18]. The purpose of this paper was to evaluate the presence of any MTHFR gene variants in a group of patients with idiopathic scoliosis (IS) patients compared to patients who did not have idiopathic scoliosis.

2. Materials and Methods

Patient charts at a private medical clinic were reviewed for those patients who presented with scoliosis. These patients had been given an option to receive genetic testing. Testing was performed via blood, saliva collection or buccal swab, depending upon other concomitant lab work ordered. To minimize specimen collection per patient, the specimen required for other concurrent lab studies was also selected for genetic analysis. Genetic testing was focused on the status of the MTHFR genes, specifically looking at 677CT and 1298AC. For purposes of this study, only the charts of those patients with idiopathic scoliosis were selected. Patients with a history of neuromuscular, syndromic, or congenital scoliosis were excluded. Past treatment history in patients with idiopathic scoliosis was irrelevant and not considered. Based upon these criteria, a total of 44 patient charts were consecutively selected.

Once these patient charts had been identified, an additional set of 44 patient charts was selected. These charts contained information on patients who presented to the same medical clinic, also had genetic testing performed, but did not have any type of scoliosis in their history. This group of charts served as the control group. Once all files were selected, patients whose files were chosen gave their written informed consent to use their non-identifying information. For both groups, the MTHFR gene results were analyzed. Patients would be considered positive for an MTHFR variation if they had one of the following results:

1) a double mutation of either the 677 or 1298 allele (hereafter referred to as homozygous positive), or 2) a single mutation in both alleles (hereafter referred to as heterozygous positive). These results were obtained for all patients in each group, and then compared quantitatively and qualitatively to the other group. Patients in both groups whose charts were selected subsequently provided written permission to use their non-identifying lab results and demographics.

3. Results

When comparing the total sample size of each group to one another, the scoliosis group was positive for an MTHFR mutation in 23 out of 44 cases (52%). The control group was positive in 11 of 44 cases (25%). Figure 1 provides an illustration of this data.

It was also of interest to evaluate the incidence of MTHFR mutations when considering ethnicity. In the scoliosis group, 3 of the 23 positive patients were African-American, 1 was Hispanic, and 1 was Indian. The remaining 18 positive patients were Caucasian. Given a total of 37 total Caucasian scoliosis patients, this makes the incidence of MTHFR defect among this ethnicity to be 49%. In the control group, 8 Caucasians, 2 African-Americans, and 1 Hispanic patient were MTHFR positive, while 2 Indian patients and 1 Hispanic patient were negative. The remaining 38 patients were Caucasian, resulting in an incidence of 21%. Table 1 provides a summary of these results.

With a 99% confidence interval, the data were compared using independent t-tests. These results are shown in Table 2. When examining specific genotypes of each gene, the 677CT/TT genotypes were not statistically different between patient groups. However, the 1298AC/CC genotypes were significantly higher in the scoliosis groups compared to the control group ($P < 0.01$). The heterozygous

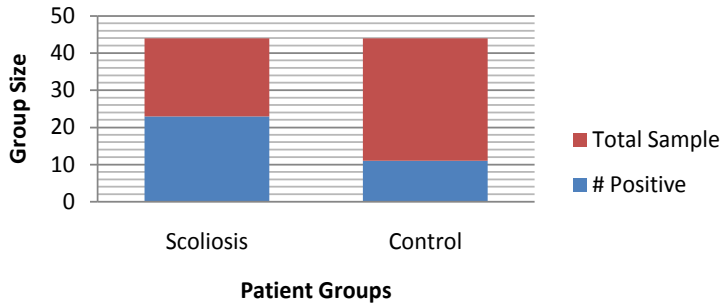

Figure 1. Frequency of heterozygous and homozygous positive genotypes. Positive genotypes in scoliosis were statistically significant at $P < 0.01$ (0.002924).

Table 1. Frequency of 677CT genotypes in IS and control groups.

677CT	CC (normal)	CT (heterozygosity)	TT (homozygosity)
Scoliosis	26	13*	5*
Control	31	9	4

*P = 0.112335.

Table 2. Frequency of 1298AC genotypes in IS and control groups.

1298AC	AA (normal)	AC (heterozygosity)	CC (homozygosity)
Scoliosis	25	13*	6*
Control	36	6	2

*Combined occurrence was statistically significant at $P < 0.01$ (0.007257).

mutation occurred over twice as often, and the homozygous mutations occurred three times more than in the control group.

4. Discussion

The data demonstrate that MTFHR polymorphisms are found significantly more frequently in patients with idiopathic scoliosis when compared with non-scoliotic patients. It is unlikely, however, that MTHFR variants are causative for scoliosis, given that a notable minority of non-scoliotic patients also carry these polymorphisms. This is the first investigation into the MTHFR genetic variation of the A1298C allele in patients with idiopathic scoliosis. Genetic variations of the A1298C allele may reduce tetrahydrobiopterin (BH4), which serves as a cofactor for phenylalanine hydroxylase (PAH), tyrosine hydroxylase (TH), and tryptophan hydroxylase (TPH) [19]. These enzymes are required to produce serotonin and the catecholamines. Morningstar *et al.* [20] have previous identified neurotransmitter imbalances in a cohort of patients with adolescent idiopathic scoliosis compared to age-matched controls. Moreover, when these neurotransmitter imbalances were corrected, scoliosis correction was improved compared to scoliotic controls [21].

It is possible that there could be other genetic variants occurring simultaneously in patients with idiopathic scoliosis, thus increasing the chance of developing the spinal deformity. Alternatively, there may also be environmental factors at play affecting idiopathic scoliosis patients that are not affecting non-scoliotics. The non-scoliotics may also not be exposed to the same environmental factors contributing to the IS onset.

Limitations

This study does not account for the downstream influences of MTHFR variants on neurotransmitter metabolism. For example, methylation is required to convert norepinephrine into epinephrine [22], and is also important in the negative feedback loop in serotonin metabolism [23]. Morningstar et al have previously described the role of these neurotransmitters in reflexive postural control [24], which is often altered in the majority of patients with IS [25]. However, it is likely that there are different gradations of downstream activity, possibly dependent upon the patient's current lifestyle and dietary habits, and local environmental factors. Therefore, investigation into these aspects in this patient population is warranted.

5. Conclusion

When comparing two small clinical samples of patient charts, 52% of patients

with a history of idiopathic scoliosis tested positive for an MTHFR genetic variation while only 25% of non-scoliotics tested positive. Given the impact that the methylation cycle has on neurotransmitter metabolism, this study may help to foster further investigation into this impact relative to neuromotor control of postural reflex pathways. It is currently unknown how their interaction may be associated with the development or progression of idiopathic scoliosis.

References

[1] Gokcen, C., Kocak, N. and Pekgor, A. (2011) Methylenetetrahydrofolate Reductase Gene Polymorphisms in Children with Attention Deficit Hyperactivity Disorder. *International Journal of Medical Sciences*, 8.

[2] Boris, M., Goldblatt, A., Galanko, J. and James, S.J. (2004) Association of MTHFR Gene Variants with Autism. *Journal of the American Physicians and Surgeons*, **9**, 106-108.

[3] Wang, X., Qiao, C., Wei, L., Han, Y., Cui, N., Huang, Z., Li, Z., Zheng, F. and Yan, M. (2015) Associations of Polymorphisms in MTHFR Gene with the Risk of Age-Related Cataract in Chinese Han Population: A Genotype-Phenotype Analysis. *PLoS ONE*, **10**, e0145581. https://doi.org/10.1371/journal.pone.0145581

[4] Cao, H.-X., Gao, C.-M., Takezaki, T., Wu, J.-Z., Ding, J.-H., Liu, Y.-T., Li, S.-P., *et al*. (2008) Genetic Polymorphisms of Methylenetetrahydrofolate Reductase and Susceptibility to Colorectal Cancer. *Asian Pacific Journal of Cancer Prevention*, **9**, 203-208.

[5] Bethke, L., Webb, E., Murray, A., Schoemaker, M., Feychting, M., Lonn, S., Ahlbom, A., *et al*. (2008) Functional Polymorphisms in Folate Metabolism Genes Influence the Risk of Meningioma and Glioma. *Cancer Epidemiology, Biomarkers & Prevention*, **17**, 1195-1202. https://doi.org/10.1158/1055-9965.EPI-07-2733

[6] Fisher, M.C. and Cronstein, B.N. (2009) Meta-Analysis of Methylenetetrahydrofolate Reductase (MTHFR) Polymorphisms Affecting Methotrexate Toxicity. *Journal of Rheumatology*, **36**, 539-545. https://doi.org/10.3899/jrheum.080576

[7] Liu, A., Menon, S., Colson, N.J., Quinlan, S., Cox, H., Peterson, M., Tiang, T., *et al*. Analysis of the MTHFR C677T Variant with Migraine Phenotypes. *BMC Research Notes*, **3**, 213. https://doi.org/10.1186/1756-0500-3-213

[8] Miller, N.H. (2011) Idiopathic Scoliosis: Cracking the Genetic Code and What Does It Mean? *Journal of Pediatric Orthopaedics*, **31**, S49-S52. https://doi.org/10.1097/BPO.0b013e318202bfe2

[9] Porter, R.W. (2001) The Pathogenesis of Idiopathic Scoliosis: Uncoupled Neuro-Osseous Growth? *European Spine Journal*, **10**, 473-481. https://doi.org/10.1007/s005860100311

[10] Moreau, A., Wang, D.S., *et al*. (2004) Melatonin Signaling Dysfunction in Adolescent Idiopathic Scoliosis. *Spine*, **29**, 1772-1781. https://doi.org/10.1097/01.BRS.0000134567.52303.1A

[11] Burwell, R.G., Clark, E.M., Dangerfield, P.H. and Moulton, A. (2016) Adolescent Idiopathic Scoliosis (AIS): A Multi-Factorial Cascade Concept for Pathogenesis and Embryonic Origin. *Scoliosis and Spinal Disorders*, **11**, 8. https://doi.org/10.1186/s13013-016-0063-1

[12] Wajchenberg, M., Astur, N., Kanas, M. and Martins, D.E. (2016) Adolescent Idiopathic Scoliosis: Current Concepts on Neurological and Muscular Etiologies. *Scoliosis and Spinal Disorders*, **11**, 4. https://doi.org/10.1186/s13013-016-0066-y

[13] Lee, H.Y., Byeon, Y., Lee, K., Lee, H.J. and Back, K. (2014) Cloning of Arabidopsis Serotonin N-Acetyltransferase and Its Role with Caffeic Acid O-Methyltransferase in the Biosynthesis of Melatonin *in Vitro* Despite Their Different Subcellular Localizations. *Journal of Pineal Research*, **57**, 418-426. https://doi.org/10.1111/jpi.12181

[14] Machida, M., Dubousset, J., Yamada, T. and Kimura, J. (2009) Serum Melatonin Levels in Adolescent Idiopathic Scoliosis Prediction and Prevention for Curve Progression—A Prospective Study. *Journal of Pineal Research*, **46**, 344-348. https://doi.org/10.1111/j.1600-079X.2009.00669.x

[15] Man, G.C., Wang, W.W., Yim, A.P., Wong, J.H., Ng, T.B., Lam, T.P., Lee, S.K., Ng, B.K., Wang, C.C., Qiu, Y. and Cheng, C.Y. (2014) A Review of Pinealectomy-Induced Melatonin-Deficient Animal Models for the Study of Etiopathogenesis of Adolescent Idiopathic Scoliosis. *International Journal of Molecular Sciences*, **15**, 16484-16499. https://doi.org/10.3390/ijms150916484

[16] Chango, A., Boisson, F., Barbe, F., *et al*. (2000) The Effect of 677CT and 1298AC Mutations on Plasma Homocysteine and 5,10-Methylenetetrahydrofolate Reductase Activity in Healthy Subjects. *British Journal of Nutrition*, **83**, 593-596. https://doi.org/10.1017/S0007114500000751

[17] Weisberg, I., Tran, P., Christensen, B., Sibani, S. and Rozen, R. (1998) A Second Genetic Polymorphism in Methyltetrahydrofolate Reductase (MTHFR) Associated with Decreased Enzyme Activity. *Molecular Genetics and Metabolism*, **64**, 159-172. https://doi.org/10.1006/mgme.1998.2714

[18] Rady, P.L., Szucs, S., Grady, J., *et al*. (2002) Genetic Polymorphisms of Methylenetetrahydrofolate Reductase (MTHFR) and Methionine Synthasereductase (MTRR) in Ethnic Populations in Texas: A Report of a Novel MTHFR Polymorphic Site, G1793A. *American Journal of Medical Genetics*, **107**, 162-168. https://doi.org/10.1002/ajmg.10122

[19] Werner, E.R., Blau, N. and Thony, B. (2011) Tetrahydrobiopterin: Biochemistry and Pathophysiology. *Biochemical Journal*, **438**, 397-414. https://doi.org/10.1042/BJ20110293

[20] Morningstar, M. (2013) Neurotransmitter Patterns in Patients with Adolescent Idiopathic Scoliosis (AIS). *Scoliosis*, **8**, O1. https://doi.org/10.1186/1748-7161-8-S2-O1

[21] Morningstar, M.W., Siddiqui, A., Dovorany, B. and Stitzel, C.S. (2014) Can Neurotransmitter Status Affect the Results of Exercise-Based Scoliosis Treatment? Results of a Controlled Comparative Chart Review. *Alternative & Integrative Medicine*, **3**, 177.

[22] Pohorecky, L.A., Zigmond, M., Karten, H. and Wurtman, R.J. (1969) Enzymatic Conversion of Norepinephrine to Epinephrine by the Brain. *Journal of Pharmacology and Experimental Therapeutics*, **165**, 190-195.

[23] Hensler, J. (2006) Serotonin. In: Siegel, G.J., Albers, R.W., Brady, S.T. and Price, D.L., Eds., *Basic Neurochemistry*: *Molecular*, *Cellular*, *and Medical Aspects*, Elsevier Academic Press, Burlington.

[24] Morningstar, M.W. (2016) Neurotransmitter Status and Idiopathic Scoliosis: A Commentary on Pathways, Testing, Clinical Utility, and Treatment. *Current Pediatric Research*, **20**, 14-19.

[25] Pialasse, J.P., Mercier, P., Descarreaux, M. and Simoneau, M. (2016) Sensorimotor Control Impairment in Young Adults with Idiopathic Scoliosis Compared with Healthy Controls. *Journal of Manipulative and Physiological Therapeutics*, **39**, 473-479. https://doi.org/10.1016/j.jmpt.2016.06.001

Call for Papers

Open Journal of Genetics

ISSN: 2162-4453 (Print), 2162-4461 (Online)
http://www.scirp.org/journal/ojgen

Editor-in-Chief

Editorial Board

Open Journal of Genetics (OJGen) is an international journal dedicated to the latest advancements of Genetics. The goal of this journal is to provide a platform for scientists and academicians all over the world to promote, share, and discuss various new issues and developments in different areas of Genetics. All manuscripts must be prepared in English, and are subject to a rigorous and fair peer-review process. Accepted papers will immediately appear online followed by printed hard copy. The journal publishes original papers including but not limited to the following fields:

- Classical and Developmental Genetics
- Conservation and Ecological Genetics (including Behavioural Genetics)
- Evolutionary and Population Genetics (except Human)
- Genetic Engineering
- Genetics of Plants and Animals in Agronomy (including Quantitative Genetics)
- Genomics
- Human and Medical Genetics (including Quantitative Genetics, Population Genetics, Psychiatric Genetics)
- Immunogenetics
- Microbial Genetics
- Molecular Genetics (including Gene Characterization and Polymorphisms, Control of Gene Expression and Epigenetics)
- Technical Tips and Improvement

We are also interested in: 1) Short Reports—2-5 page papers where an author can either present an idea with theoretical background but has not yet completed the research needed for a complete paper or preliminary data; 2) Book Reviews—Comments and critiques.

Website and E-Mail

http://www.scirp.org/journal/ojgen E-mail: ojgen@scirp.org

What is SCIRP?

Scientific Research Publishing (SCIRP) is one of the largest Open Access journal publishers. It is currently publishing more than 200 open access, online, peer-reviewed journals covering a wide range of academic disciplines. SCIRP serves the worldwide academic communities and contributes to the progress and application of science with its publication.

What is Open Access?

All original research papers published by SCIRP are made freely and permanently accessible online immediately upon publication. To be able to provide open access journals, SCIRP defrays operation costs from authors and subscription charges only for its printed version. Open access publishing allows an immediate, worldwide, barrier-free, open access to the full text of research papers, which is in the best interests of the scientific community.

- High visibility for maximum global exposure with open access publishing model
- Rigorous peer review of research papers
- Prompt faster publication with less cost
- Guaranteed targeted, multidisciplinary audience

Scientific Research Publishing

Website: http://www.scirp.org

Subscription: sub@scirp.org

Advertisement: service@scirp.org