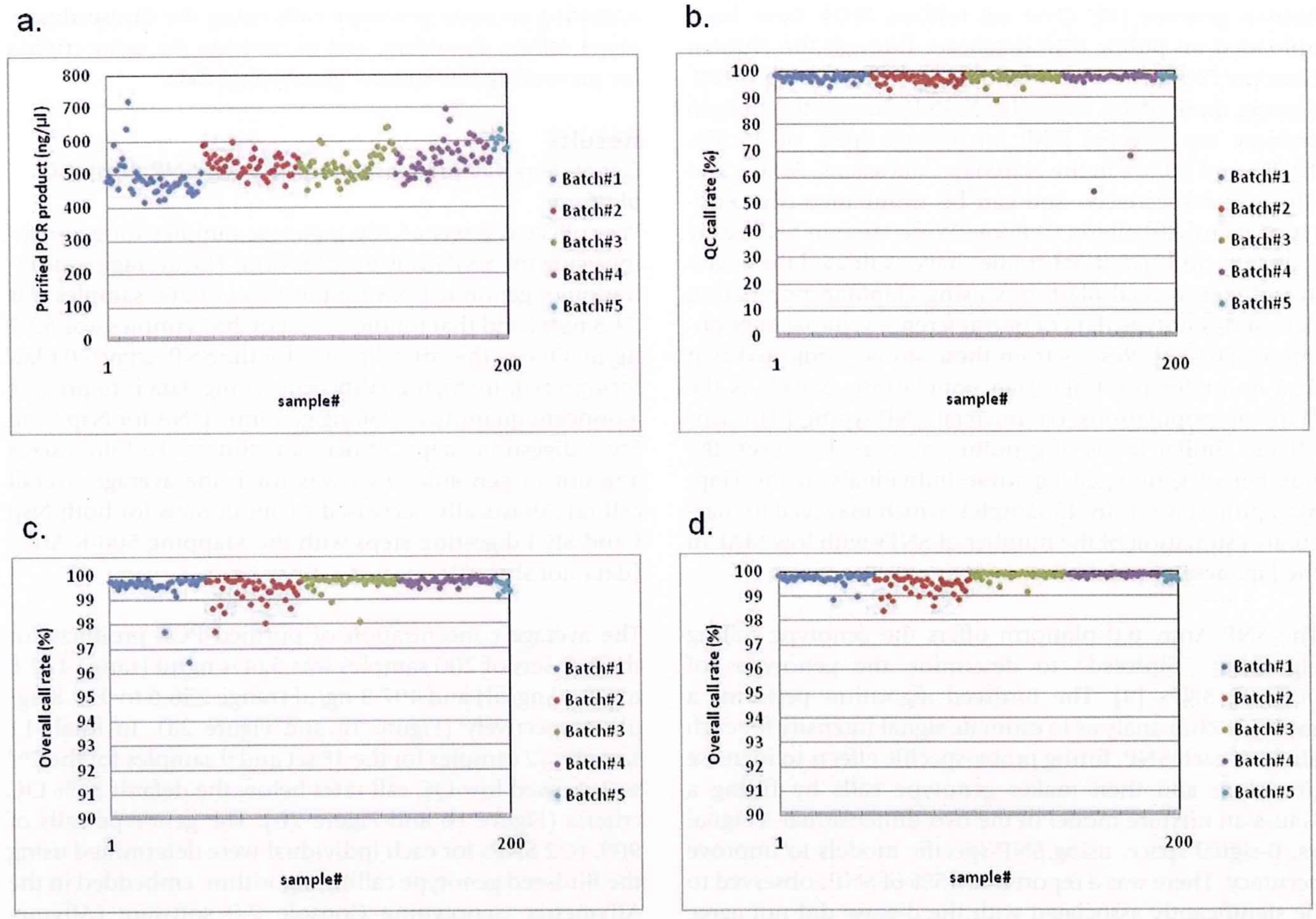


Figure 1
Genotyping results of the 1st set of 200 samples using the SNP Array 6.0 platform. Colours are based on every 48 samples analyzed simultaneously as a batch. a. Concentration of purified PCR products for each sample. b. QC call rate for each sample. c. Overall call rate for each sample, as determined by the Birdseed algorithm using total 198 samples that passed the default 86% QC criteria. d. Overall call rate for each sample, as determined by the Birdseed algorithm using samples in the same batch.

(Figure 2d). The concentration of purified PCR products from batch #1 drastically fluctuated among the 48 samples (Figure 2a). The CV (standard deviation/average) of the purified PCR product concentration for batch #1 was much higher than that for any other batches from the two sets of 200 samples (Figure 3). The CV of the purified PCR product concentration is a new indicator to assess experimental quality for each of the running batches, and may remove the experimental errors occurring on the running batches prior to hybridization on the GeneChip arrays.

For the 48 samples from batch #1 of the 2nd set, the intact genomic DNA could not be detected clearly when the samples were electrophoresed on 1.0% agarose gels (Figure 4). Therefore, these genomic DNAs for batch #1 of the 2nd set may have degraded due to repetitive freezing and thawing, which led to low-quality genotyping results. Preparation of the exact amount of intact genomic DNA is considered to be one of the crucial points for the SNP array 6.0 platform.

In order to assess the performance of the SNP Array 6.0 platform and the Birdseed algorithm, we mainly used genotyping data obtained from the 1st set of 200 samples because the 2nd set contained samples in poor condition.

Genotype calling accuracy with "Birdseed" algorithm

The genotype calling accuracy of the Birdseed algorithm was considered to be improved as the sample number for determining genotype calls increased. We determined 909,622 genotype calls for 12 samples among 198 samples with over 86% QC criteria, and used these genotype calls as a reference. We also determined the genotype calls of the same 12 samples under 6 different sample sizes,

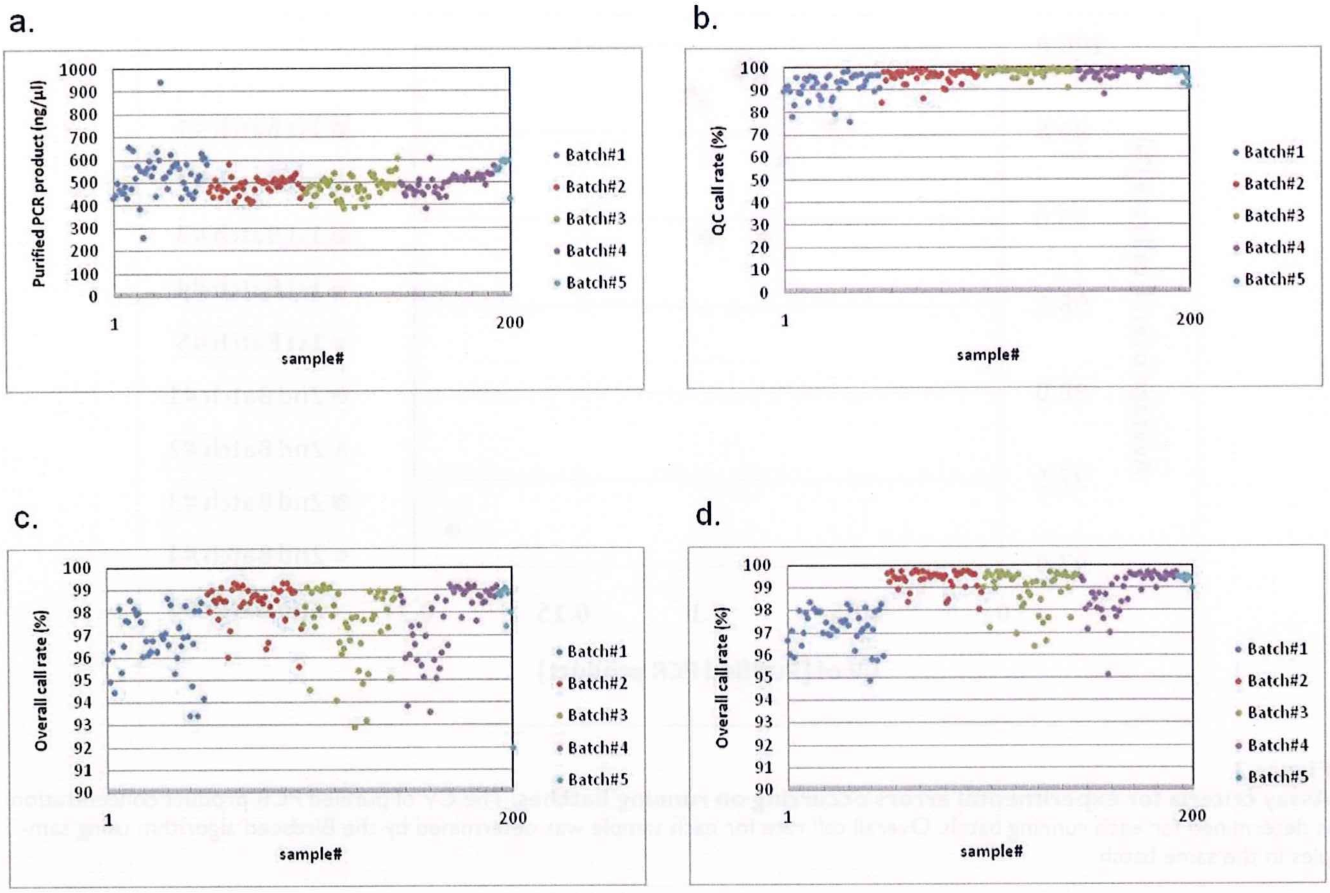


Figure 2

Genotyping results of 2nd set of 200 samples using the SNP Array 6.0 platform. a. Concentration of purified PCR products for each sample. b. QC call rate for each sample. c. Overall call rate for each sample, as determined by the Birdseed algorithm using a total of 191 samples that passed the default 86% QC criteria. d. Overall call rate for each sample, as determined by the Birdseed algorithm using samples in the same batch.

using 12 samples, 24 samples, 36 samples, 48 samples, 72 samples and 96 samples. To investigate the genotype calling accuracy of the Birdseed algorithm, we compared the genotype calls determined under 6 different sample sizes to the reference genotype calls for each of the 12 samples. We prepared 4 sets of 12 samples from a batch of 48 samples (Batch #3) and performed the genotype call comparison for each set of 12 samples. Figure 5 shows the average overall call rate and the average concordance rate for each set of the 12 samples. The average overall call rate for 4 sets of the 12 samples, which were determined with 12 samples, 24 samples, 36 samples, 48 samples, 72 samples, 96 samples and 198 samples, were 99.84%, 99.86%, 99.84%, 99.83%, 99.79%, 99.75% and 99.71%, respectively. The average concordance rate for the 4 sets of the 12 samples under 6 different sample sizes were 99.47%, 99.75%, 99.80%, 99.84%, 99.86% and 99.87%, respectively. Here, "No Calls" was excluded from the concordance calculation.

Our results showed that the average overall call rate of the 12 samples was almost constant when the genotype calls were determined with fewer than 48 samples; however, it gradually decreased as the sample number increased from 48 to 198, which showed a negative correlation with a P value of 0.0053. In contrast, the concordance rate gradually increased as the sample number increased, which showed a positive correlation with a P value of 0.0115.

Removing low-quality samples by adjusting QC criteria

Our results showed that the average overall call rate gradually decreased as the sample number increased, presumably due to low-quality samples included in the genotype calling with the Birdseed algorithm. Indeed, there was one sample which had an overall call rate lower than 97% among the 198 samples with over 86% QC call rate. Therefore, we applied more stringent QC criteria to remove the low-quality samples, because a linear relationship was observed between QC call rate and overall call

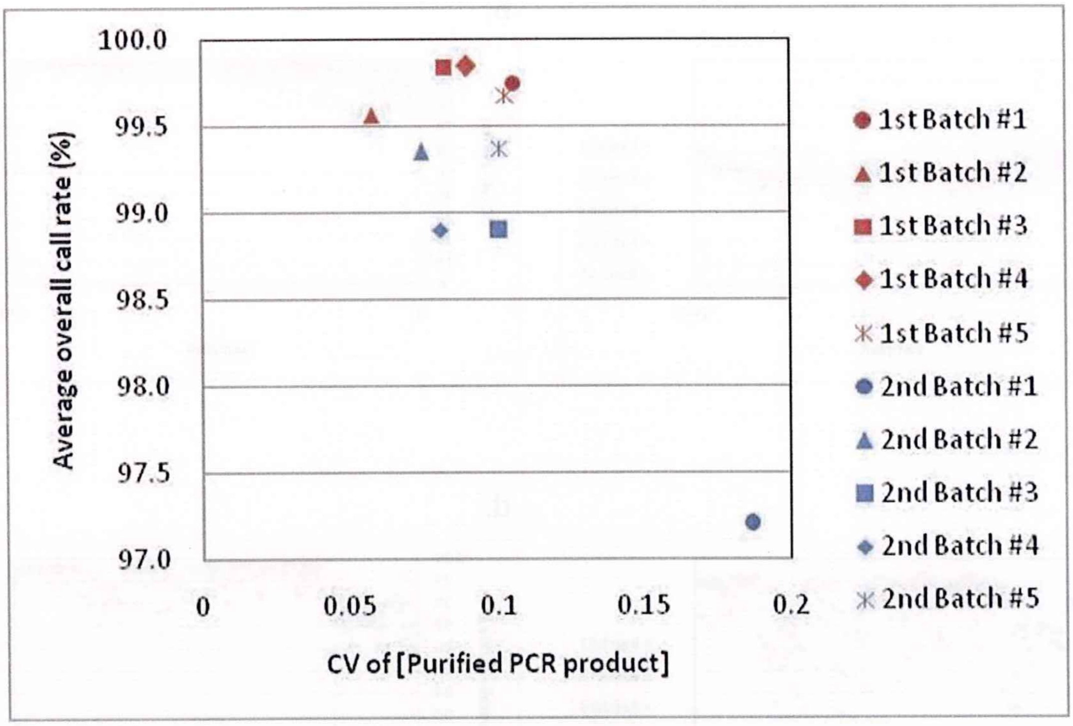


Figure 3
Assay criteria for experimental errors occurring on running batches. The CV of purified PCR product concentration is determined for each running batch. Overall call rate for each sample was determined by the Birdseed algorithm using samples in the same batch.

rate (Figure 6a). When we applied 95% QC criteria, 189 samples passed the QC criteria and the average overall call rate improved from 99.58 to 99.65%. By comparing the overall call rate determined under the 95% QC criteria with that under the default criteria, 187 of 189 samples improved by an average of 0.018% in overall call rate; however, the remaining two samples showed decreased overall call rate (by 0.76% and 0.12%) (Figure 6b). These two samples were considered as outliers on the genotype calling with the Birdseed algorithm and had to be removed. We repeated the removal of samples until none had a lower overall call rate than that determined under the default criteria. A total of 184 samples had an overall call rate that improved over the one determined under the default criteria, with an average change of 0.035%. The average overall call rate for the 184 samples was 99.71%, which was 0.13% higher than the default QC criteria (Figure 6c).

Number of SNPs available for GWAS in the Japanese population

The genotype calls of 909,622 SNPs were determined with 184 samples after sample filtering with adjusted QC criteria. However, these genotype calls still included inaccurate SNPs, which could lead to inflation of false positives, presumably due to systematically miss-called SNPs. Therefore, SNP filtering was considered to be important for a reliable and accurate set of genotype calls that avoid false association signals and false negative signals, allowing rapid identification of disease susceptibility genetic factors. We reported that the poorly behaving SNPs were effectively eliminated with the SNP filtering parameters; MAF > 5% or 1%, HWE p-value > 0.001 and SNP call rate > 95% [14]. Here, SNP call rate was defined for each SNP as the number of successfully genotyped samples divided by the number of total samples genotyped.

Among a total of 909,622 SNPs genotyped using 184 samples, 590,248 SNPs passed the three SNP filtering criteria with MAF > 5%, HWE p-value > 0.001 and SNP call rate > 95%, while 661,559 SNPs passed with MAF > 1%, HWE p-value > 0.001, and SNP call rate > 95%. A total of 180,859 SNPs were observed to be monomorphic in the Japanese population.

Discussion

The emerging SNP typing technologies have enabled genome-wide association studies to be conducted with hundreds of thousands of genotyped SNPs. According to Affymetrix, the SNP Array 6.0 platform can genotype over

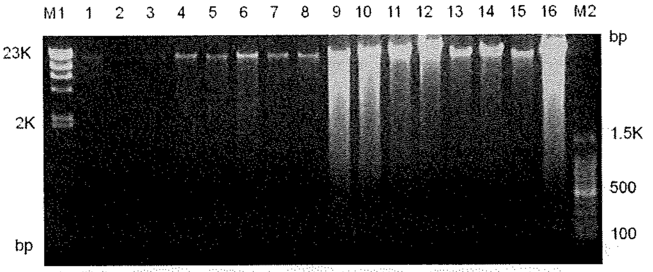


Figure 4
Agarose gel electrophoresis pattern showing genomic DNA from batch #1 of the 2nd set (lanes 1–8) and batch #2 of the 2nd set (lanes 9–16). Fifty nanograms of genomic DNA for each of the sample was electrophoresed on 1.0% agarose gels. M1 and M2 indicate lambda DNA digested with Hind III and 100-bp DNA ladder marker, respectively.

900 K SNP markers across the human genome with an overall call rate of at least 97%, over 99.7% concordant with the HapMap genotypes, and the Mendelian inheritance consistency for 10 Trios of greater than 99.9% when performing analysis under the default 86% QC criteria. To evaluate the SNP 6.0 Array platform and the Birdseed genotype calling algorithm, we genotyped two sets of 200 non-HapMap Japanese samples using the SNP Array 6.0 platform.

When we applied the default 86% QC criteria, 2 samples out of the 1st set of 200 samples were excluded and the average overall call rate was 99.58%. There was one sample with an overall call rate of lower than 97% among the 198 samples. Here, we found a linear relationship between QC call rate and overall call rate. Therefore, we applied stringent QC criteria of over 95% in order to remove the low-quality samples and found that the average overall call rate for 189 samples passing the stringent QC criteria improved to 99.65%. Among the 189 samples, 187 samples had higher overall call rates than those determined under the default QC criteria; however, the remaining two samples showed lower overall call rates (by 0.76% and 0.12%). When we repeated the removal of samples until none had a lower overall call rate than the one determined under the default criteria, none of the remaining 184 samples with an overall call rate lower than 97%. The average overall call rate of 184 samples was thus improved to 99.71%. The decay of average overall call rate may be caused by some samples that pass the QC criteria, but still have a low overall call rate. We can thus improve overall call rate by removing these samples and adjusting the QC criteria.

One of the crucial points for the SNP array 6.0 platform is to prepare the exact amount of intact genomic DNA. A 10-fold excess amount of genomic DNA decreased the overall call rate of each sample to by about 80% and another study revealed that samples with less than 50 ng/μl genomic DNA show low overall call rates [15]. Therefore, we checked the concentration and condition of genomic DNA with the NanoDrop quatitation and agarose gel electrophoresis. The SNP array 6.0 platform has three check points to assess experimental errors prior to hybridization on GeneChip arrays. Here, we found that the CV of the purified PCR product concentration was another critical indicator prior to hybridization in assessing the performance of each running batches. We suggest that samples with a CV value over 0.15 are excluded from the remainder of the assay.

The genotype calling accuracy of the Birdseed algorithm was assessed by comparing the 909,622 genotype calls of 12 samples from among198 samples with over 86% QC criteria, to those of 12 samples determined with six different sample sizes; 12 samples, 24 samples, 36 samples, 48 samples, 72 samples and 96 samples. The concordance rate gradually increased as the number of samples increased. The average concordance rate was almost constant over 99.8%, when the genotype calls were determined with over 48 samples using the Birdseed algorithm. However, the average overall call rate of the 12 samples gradually decreased as the sample number increased from 48 to 198. We could explain the reasons why the overall call rate decreases, and why the concordance rate increases for these 12 samples in a grouping of samples greater than 48 by means of characteristic properties of the Birdseed algorithm and minor allele frequency of each SNP. When the sample number was smaller than 48, all of three clusters designating AA, AB and BB genotypes were rarely observed for the SNPs with low MAF. In such cases, the Birdseed algorithm would determine the genotype as a single cluster, however, would ambiguously genotype as AA, BB and AB (tend to miss-genotype). Therefore, high call rate and low concordance were observed with the sample number smaller than 48. In contrast, when the sample number was greater than 48, two or three clusters would be observed for many SNPs. For these SNPs, the Birdseed algorithm could determine the outlying samples from each cluster as "No Calls", leading to low call rate and high concordance.

We can accurately determine the genotype calls with high overall call rates by determining the genotype calls with more than 48 samples, after removing low-quality samples by adjusting the QC criteria. Our results showed that the SNP Array 6.0 platform reached the expected level reported by the manufacturer, with an average overall call rate of over 99.5% and an average concordance rate of

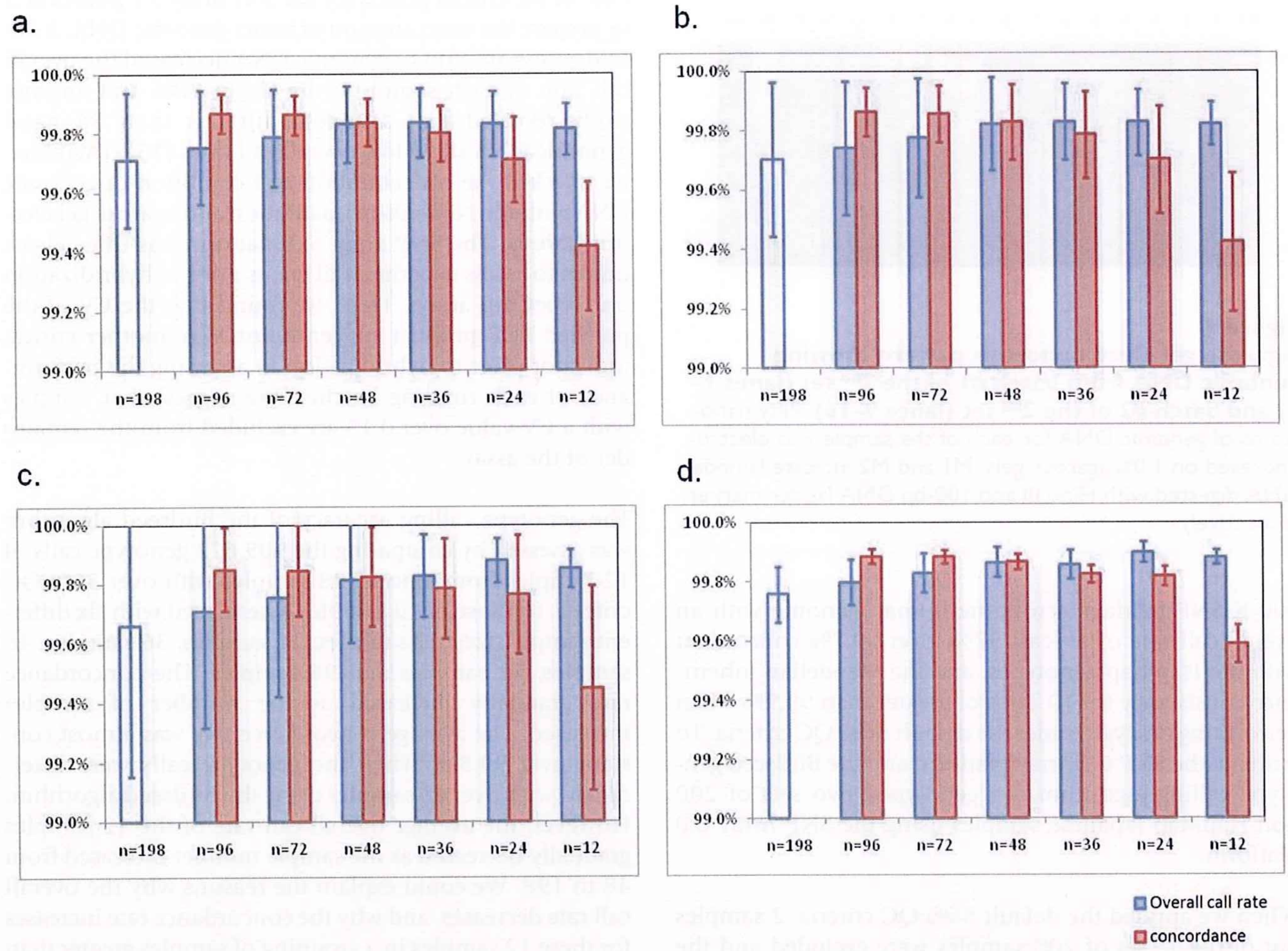


Figure 5
Genotype calling accuracy with Birdseed algorithm. a-d. Genotype calls determined using 198 samples with over 86% QC criteria were used as a reference. The average overall call rate for the 4 sets of the 12 samples were determined with 7 different sample sizes; 12 samples, 24 samples, 36 samples, 48 samples, 72 samples, 96 samples and 198 samples. The average concordance rates for the 4 sets of 12 samples were determined by comparison with the reference genotype calls. A negative correlation with a P value of 0.0053 and a positive correlation with a P value of 0.0115 were shown for overall call rate and concordance rate by fitting the power-law distribution to the data with least-squares approximation.

over 99.8%. However, about 20% of a total of 909,622 SNPs were found to be monomorphic in the Japanese population, which is due to SNP selection methods. The SNPs assayed on the SNP Array 6.0 platform were mainly selected as observed with high MAF in the Caucasian population. Among a total of 909,622 SNPs genotyped using the SNP Array 6.0 platform with 184 Japanese samples, 590,248 SNPs passed three SNP filtering criteria; MAF > 5%, HWE p-value > 0.001 and SNP call rate > 95%. Although the exact number of SNPs within the human genome remains under discussion, it has been reported that the genome coverage of the JPT + CHB population in the Phase II HapMap data was 66% using the Mapping 500 K Array set [10]. The genome coverage of the SNP array 6.0 platform was estimated using the same calculation and was revealed to be 75% with the 590,248 SNPs in the Japanese population.

Conclusion

The current Affymetrix SNP Array 6.0 platform enables the genotyping of over 900 K SNPs with high overall call rate (over 99.5%) and high concordance rate (over 99.8%). The number of SNPs available for GWAS in the Japanese population was revealed to be over 660 K SNPs, all of which passed the three SNP filtering criteria; MAF > 1%, HWE p-value > 0.001 and SNP call rate > 95%. GWAS

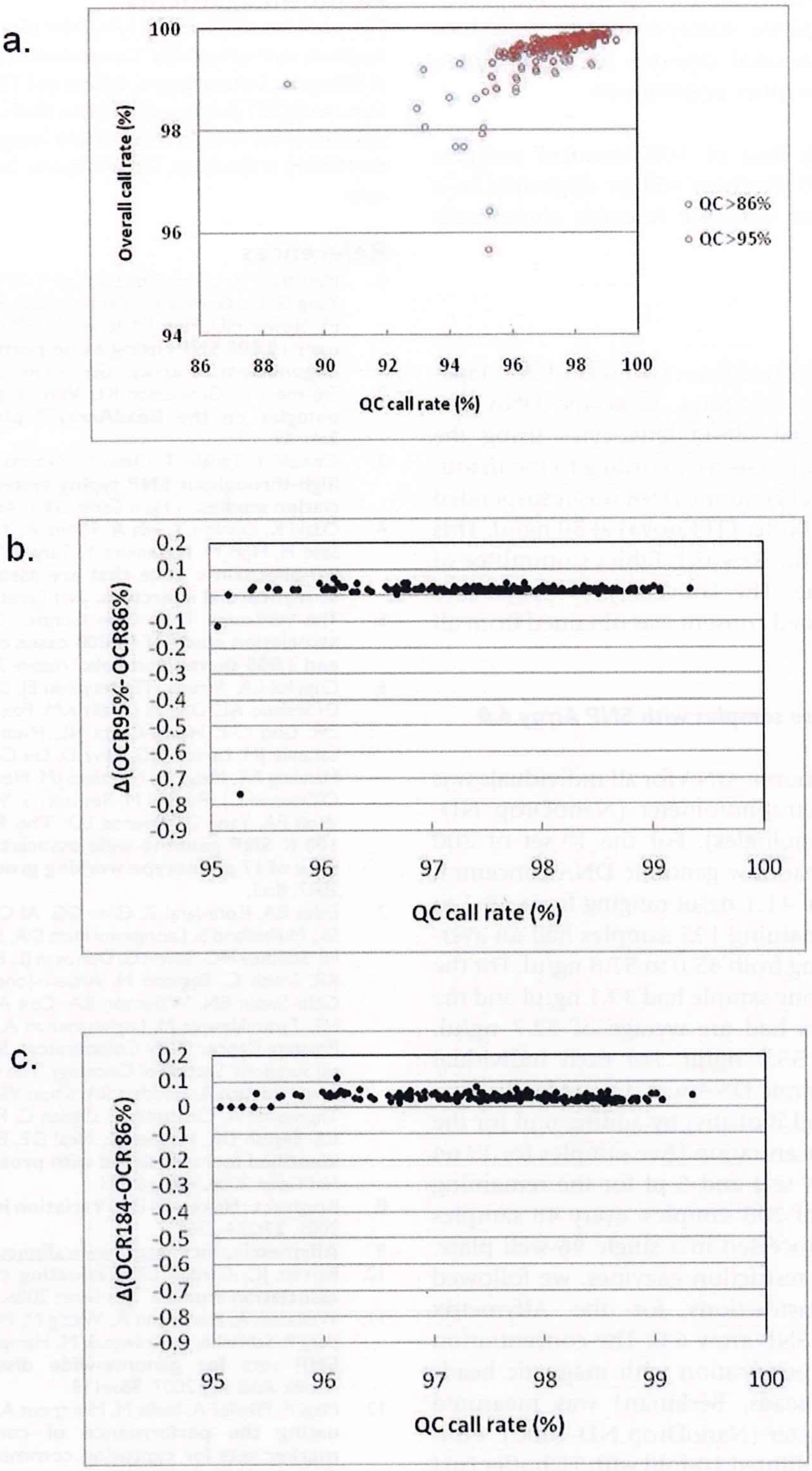


Figure 6

Removal of low-quality samples by adjusting QC criteria. Overall call rate for each sample was determined using total samples that passed the QC criteria. a. Overall call rate and QC call rate for each sample plotted with QC criteria > 86% and > 95%. b. Overall call rate (OCR) determined with 86% QC criteria compared with that determined with 95% QC criteria. c. Overall call rate (OCR) determined with 86% QC criteria compared with that determined using 184 samples.

using the SNP Array 6.0 platform has considerable potential in identifying candidate susceptibility or resistance genetic loci for multifactorial diseases in the Japanese population, as well as in other populations.

Finally, the genotyping data of 400 Japanese samples using the SNP array 6.0 platform will be deposited in a public database to share with the research community [16].

Methods

Study sample

Blood samples were obtained from two sets of 200 Japanese individuals in two institutes. Genomic DNA was extracted from peripheral blood leukocytes using the QIAamp Blood Mini Kit (Qiagen) according to the manufacturer's instructions. All genomic DNA was resuspended with Reduced EDTA TE Buffer (TEKnova) at 50 ng/μl. This study was approved by the Research Ethics Committee of the Faculty of Medicine, The University of Tokyo and Tokai University. Informed consent was obtained from all participants.

Genotyping 400 Japanese samples with SNP Array 6.0 platform

The concentration of genomic DNA for all individuals was measured using a spectrophotometer (NanoDrop ND-1000, NanoDrop Technologies). For the 1st set of 200 samples, five samples had low genomic DNA concentrations with an average of 41.1 ng/μl ranging from 38.2 to 44.5 ng/μl, and the remaining 195 samples had an average of 54.8 ng/μl, ranging from 45.0 to 57.8 ng/μl. For the 2nd set of 200 samples, one sample had 39.1 ng/μl and the remaining 199 samples had an average of 52.7 ng/μl, ranging from 45.0 to 55.9 ng/μl. For each individual assayed, 250 ng of genomic DNA was digested with Sty I and Nsp I (New England BioLabs) by adding 6 μl for the 6 samples with low concentration (five samples for 1st set and one sample for 2nd set) and 5 μl for the remaining samples. For two sets of 200 samples, every 48 samples were simultaneously processed in a single 96-well plate. After the reaction with restriction enzymes, we followed the manufacturer's instructions for the Affymetrix Genome-wide Human SNP array 6.0. The concentration of PCR products after purification with magnetic beads (Agencourt Magnetic Beads, Beckman) was measured using a spectrophotometer (NanoDrop ND-1000). Purified PCR products were diluted 10-fold with TE buffer (pH 8.0) (WAKO) in order to have a suitable concentration for the spectrophotometer to measure. The genotype calls of each individual were determined by the Birdseed version 1 genotype calling algorithm, embedded in the software Affymetrix Genotyping Console 2.0 (Affymetrix). The number of samples used to determine the genotype calls varied depending on the examination.

Acknowledgements

This work was supported by KAKENHI (Grant-in-Aid for Scientific Research) on Priority Areas "Comprehensive Genomics" from the Ministry of Education, Culture, Sports, Science and Technology of Japan, by a grant from the CREST program of the Japan Science and Technology Agency, Japan and by the contract research fund "Integrated Database Project" from the Ministry of Education, Culture, Sports, Science, and Technology of Japan.

References

1. Matsuzaki H, Loi H, Dong S, Tsai Y-Y, Fang J, Law J, Di X, Liu W-M, Yang G, Liu G, Huang J, Kennedy GC, Ryder TB, Marcus GA, Walsh PS, Shriver MD, Puck JM, Jones KW, Mei R: **Parallel genotyping of over 10,000 SNPs using a one-primer assay on a high-density oligonucleotide array.** *Genome Res* 2004, **14:**414-425.
2. Steemers FJ, Gunderson KL: **Whole genome genotyping technologies on the BeadArray™ platform.** *Biotechnol J* 2007, **2:**41-49.
3. Ohnishi Y, Tanaka T, Ozaki K, Yamada R, Suzuki H, Nakamura Y: **A high-throughput SNP typing system for genome-wide association studies.** *J Hum Genet* 2001, **46:**471-477.
4. Ozaki K, Ohnishi Y, Iida A, Sekine A, Yamada R, Tsunoda T, Sato H, Sato H, Hori M, Nakamura Y, Tanaka T: **Functional SNPs in the lymphotoxin-α gene that are associated with susceptibility to myocardial infarction.** *Nat Genet* 2002, **32:**650-654.
5. The Wellcome Trust Case Control Consortium: **Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls.** *Nature* 2007, **447:**661-678.
6. Cupples LA, Arruda HT, Benjamin EJ, D'Agostino RB Sr, Demissie S, DeStefano AL, Dupuis J, Falls KM, Fox GS, Gottlib DJ, Govindaraju DR, Guo C-Y, Heard-Costa NL, Hwang S-J, Kathiresan S, Kiel DP, Laramie JM, Larson MG, Levy D, Liu C-Y, Lunetta KL, Mailman MD, Manning AK, Meigs JB, Murabito JM, Newton-Cheh C, O'Connor GT, O'Donnell CJ, Pandey M, Seshadri S, Vasan RS, Wang ZY, Wilk JB, Wolf PA, Yang Q, Atwood LD: **The Framingham Heart Study 100 K SNP genome-wide association study resource: overview of 17 phenotype working group reports.** *BMC Med Genet* 2007, **8:**s1.
7. Eeles RA, Kote-Jarai Z, Giles GG, Al Olama AA, Guy M, Jugurnauth SK, Mulholland S, Leongamornlert DA, Edwards SM, Morrison J, Field HI, Southey MC, Severi G, Donovan JL, Hamdy FC, Dearnaly DP, Muir KR, Smith C, Bagnato M, Ardern-Jones AT, Hall AL, O'Brien LT, Gehr-Swain BN, Wilkinson RA, Cox A, Lewis S, Brown PM, Jhavar SG, Tymrakiewicz M, Lophatananon A, Bryant SL, The UK Genetic Prostate Cancer Study Collaborators, British Association of Urological Surgeons' Section of Oncology, The UK ProtecT Study Collaborators, Horwich A, Huddart RA, Khoo VS, Parker CC, Woodhouse CJ, Thompson A, Christmas T, Ogden C, Fisher C, Jamieson C, Cooper CS, English DR, Hopper JL, Neal DE, Easton DF: **Multiplex newly identified loci associated with prostate cancer susceptibility.** *Nat Genet* 2008, **40:**316-321.
8. Kruglyak L, Nickerson DA: **Variation is the spice of life.** *Nat Genet* 2001, **27:**234-236.
9. **Affymetrix, Inc** [http://www.affymetrix.com/index.affx]
10. Barrett JC, Cardon LR: **Evaluating coverage of genome-wide association studies.** *Nat Genet* 2006, **38:**659-662.
11. Wollstein A, Herrmann A, Wittig M, Mothnagel M, Franke A, Nürnberg P, Schreiber S, Krawczak M, Hampe J: **Efficacy assessment of SNP sets for genome-wide disease association studies.** *Nucleic Acids Res* 2007, **35:**e113.
12. Mägi R, Pfeufer A, Nelis M, Montpetit A, Metspalu A, Remm M: **Evaluating the performance of commercial whole-genome marker sets for capturing common genetic variation.** *BMC Genomics* 2007, **8:**159-166.
13. Hua J, Craig DW, Brun M, Webster J, Zismann V, Tembe W, Joshipura K, Huentelman MJ, Dougherty ER, Stephan DA: **SNiPer-HD: improved genotype calling accuracy by an expectation-maximization algorithm for high-density SNP arrays.** *Bioinformatics* 2007, **23:**57-63.
14. Miyagawa T, Nishida N, Ohashi J, Kimura R, Fujimoto A, Kawashima M, Koike A, Sasaki T, Tanii H, Otowa T, Momose Y, Nakahara Y, Gotoh J, Okazaki Y, Tsuji S, Tokunaga K: **Appropriate data cleaning methods for genome-wide association study.** *J Hum Genet* 2008 in press.

15. Woo JG, Sun G, Haverbusch M, Indugula S, Martin L, Broderick JP, Deka R, Woo D: **Quality assessment of buccal versus blood genomic DNA using the Affymetrix 500 K GeneChip.** *BMC Genet* 2007, **8**:79-83.

16. **Ministry of Education, Culture, Sports, Science, and Technology (MEXT) Integrated Database Project** [http://lifesciencedb.mext.go.jp/en/]

Distribution and Effects of Nonsense Polymorphisms in Human Genes

Yumi Yamaguchi-Kabata[1¤a], Makoto K. Shimada[1,2¤b], Yosuke Hayakawa[1,2], Shinsei Minoshima[3], Ranajit Chakraborty[4], Takashi Gojobori[1,5], Tadashi Imanishi[1]*

1 Biological Information Research Center, National Institute of Advanced Industrial Science and Technology, Tokyo, Japan, **2** Japan Biological Information Research Center, Japan Biological Informatics Consortium, Tokyo, Japan, **3** Hamamatsu University School of Medicine, Hamamatsu, Shizuoka, Japan, **4** Center for Genome Information, University of Cincinnati, Cincinnati, Ohio, United States of America, **5** Center for Information Biology and DNA Data Bank of Japan, National Institute of Genetics, Mishima, Shizuoka, Japan

Abstract

Background: A great amount of data has been accumulated on genetic variations in the human genome, but we still do not know much about how the genetic variations affect gene function. In particular, little is known about the distribution of nonsense polymorphisms in human genes despite their drastic effects on gene products.

Methodology/Principal Findings: To detect polymorphisms affecting gene function, we analyzed all publicly available polymorphisms in a database for single nucleotide polymorphisms (dbSNP build 125) located in the exons of 36,712 known and predicted protein-coding genes that were defined in an annotation project of all human genes and transcripts (H-InvDB ver3.8). We found a total of 252,555 single nucleotide polymorphisms (SNPs) and 8,479 insertion and deletions in the representative transcripts in these genes. The SNPs located in ORFs include 40,484 synonymous and 53,754 nonsynonymous SNPs, and 1,258 SNPs that were predicted to be nonsense SNPs or read-through SNPs. We estimated the density of nonsense SNPs to be 0.85×10^{-3} per site, which is lower than that of nonsynonymous SNPs (2.1×10^{-3} per site). On average, nonsense SNPs were located 250 codons upstream of the original termination codon, with the substitution occurring most frequently at the first codon position. Of the nonsense SNPs, 581 were predicted to cause nonsense-mediated decay (NMD) of transcripts that would prevent translation. We found that nonsense SNPs causing NMD were more common in genes involving kinase activity and transport. The remaining 602 nonsense SNPs are predicted to produce truncated polypeptides, with an average truncation of 75 amino acids. In addition, 110 read-through SNPs at termination codons were detected.

Conclusion/Significance: Our comprehensive exploration of nonsense polymorphisms showed that nonsense SNPs exist at a lower density than nonsynonymous SNPs, suggesting that nonsense mutations have more severe effects than amino acid changes. The correspondence of nonsense SNPs to known pathological variants suggests that phenotypic effects of nonsense SNPs have been reported for only a small fraction of nonsense SNPs, and that nonsense SNPs causing NMD are more likely to be involved in phenotypic variations. These nonsense SNPs may include pathological variants that have not yet been reported. These data are available from Transcript View of H-InvDB and VarySysDB (http://h-invitational.jp/varygene/).

Citation: Yamaguchi-Kabata Y, Shimada MK, Hayakawa Y, Minoshima S, Chakraborty R, et al. (2008) Distribution and Effects of Nonsense Polymorphisms in Human Genes. PLoS ONE 3(10): e3393. doi:10.1371/journal.pone.0003393

Editor: Alan Christoffels, University of Western Cape, South Africa

Received March 6, 2008; **Accepted** September 3, 2008; **Published** October 14, 2008

Funding: This research is financially supported by the Ministry of Economy, Trade and Industry of Japan (METI), the Ministry of Education, Culture, Sports, Science and Technology of Japan (MEXT) and the Japan Biological Informatics Consortium (JBIC).

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: t.imanishi@aist.go.jp

¤a Current address: Center for Genomic Medicine, RIKEN, Yokohama City, Kanagawa, Japan
¤b Current address: Institute for Comprehensive Medical Science, Fujita Health University, Toyoake, Aichi, Japan

Introduction

Genetic variations in the human genome are maintained by a balance of mutation, selection and random genetic drift. Some of the polymorphisms cause phenotypic variations and diseases. Therefore, many studies have attempted to identify causative variants of genetic diseases and the relationships between genetic variations and phenotypic effects. Genetic variations within linked loci are inherited to the same gamete. Based on the linkage of genetic variations, loci that contain disease-causing genes have been mapped by using polymorphic markers. At present, about 14 million clusters of genetic polymorphisms have been identified in the human genome [1]. On average, two haploid genomes are estimated to differ by one single nucleotide polymorphism (SNP) in every 1200–1500 bp [2]. SNPs have been recently used to conduct genome-wide association studies to find genomic regions that are susceptible to diseases and phenotypic variations [3,4,5,6]. In this approach, usually, causative polymorphisms for diseases or phenotypic variations are identified after the identification of susceptible genomic regions by using SNP markers. Such SNPs are called landmark SNPs, and the indirect relationships between polymorphisms and phenotypic variations were examined to identify genomic regions where causative genes are located.

Another approach in finding pathological variants is to extract polymorphisms that alter amino acids in functional genes or affect gene expression or splicing, using a comprehensive set of functional elements of the human genome. Several studies have analyzed nonsynonymous SNPs to predict pathological variants [7,8,9, 10,11,12,13,14]. A large number of nonsynonymous SNPs also have been examined for associations with diseases[15,16].

Although many pathological mutations have been identified [17,18], the number of such variants is small compared to the number of known polymorphisms, and it is still unclear which polymorphisms have biological effects. In a study of consanguineous marriage [19], it was estimated that each person has deleterious alleles that are equivalent to a few lethal genes. Gene-centric SNP surveys have shown that the ratio of nonsynonymous to synonymous SNPs is significantly higher in the low frequency class than in the common frequency class [20,21,22]. These results suggest that a large fraction of the low frequency nonsynonymous SNPs are deleterious. To understand the molecular basis of the effects of human genetic variations on phenotypic variations, a prediction analysis of possible effects of polymorphisms on gene function in all human genes appears to be needed.

In this study, to detect polymorphisms affecting gene function, we analyzed all publicly available polymorphisms in the Single Nucleotide Polymorphism Database (dbSNP) (build 125) in the exons of all 36,712 protein-coding genes that were defined in an annotation project of all human genes and transcripts (H-InvDB ver3.8)[23,24]. In summary with representative transcripts (one transcript from one gene), we detected 53,754 nonsynonymous SNPs and 1,417 SNPs causing changes between amino acids and stop codons. Among possible point mutations in ORFs, nonsense mutations cause the most drastic changes of gene products. In fact, several reports have shown that nonsense mutations cause genetic diseases [25,26,27,28]. Truncation of a polypeptide by a premature termination codon causes a drastic change in the gene product. Furthermore, it is known that a nonsense mutation can cause decay of mRNA resulting in the absence of the gene product. This process, called 'nonsense-mediated decay (NMD)' limits the synthesis of abnormal proteins[29,30,31]. On the other hand, the loss of a termination codon in a transcript also appears to cause decay of mRNA (referred to as non-stop decay) and thus to prevent translation[32,33]. In spite of the severe effects of nonsense mutations, the distribution of nonsense SNPs in human genes is little understood. In this study, we examined the density of nonsense SNPs in human genes, and showed that nonsense SNPs exist at a lower density than nonsynonymous SNPs, possibly due to the more severe effects of premature stop codons than amino acid changes. About a half of nonsense SNPs are predicted to cause NMD. The correspondence between known pathological variants and nonsense SNPs suggests that nonsense SNPs causing NMD are more likely to be involved in phenotypic variations.

Results

Selection and classification of polymorphisms in exon regions

We analyzed 9,235,997 polymorphisms (dbSNP build 125) in the human genome with exon positions and predicted ORFs that were revealed in our annotation project of human genes (H-InvDB) (Figure 1). In all of the 36,712 protein-coding loci in the genome, we detected 252,555 SNPs and 8,479 insertions and deletions (indels) that exist in exon regions of the representative transcript (one transcript from one gene) (Table 1). The polymorphisms in the exon regions were further classified according to the predicted ORFs. We detected 96,164 SNPs within the ORFs, 51,881 SNPs in the 5′UTR regions and 104,510 SNPs in the 3′UTR regions. Among the SNPs in the ORFs, 40,484 were synonymous and 53,754 were nonsynonymous (Further analyses of nonsynonymous SNPs are described in Results S1.). Most of the indels were detected in the UTR regions. The ORF regions contained 1,258 SNPs that cause changes between amino acids and stop codons (Table S1). Of the 1,258 SNPs, 1,183 SNPs were regarded as nonsense SNPs, while 75 were found to have stop codons as ancestral alleles. We also detected 247 SNPs at termination codon sites, 88 of which were synonymous. The remaining 159 SNPs were changes between stop codons and amino acids. After checking ancestral alleles, 110 of the 159 SNPs were inferred to be read-through SNPs, while the other 49 were inferred to changes to stop codons.

Distribution of polymorphisms in exon regions

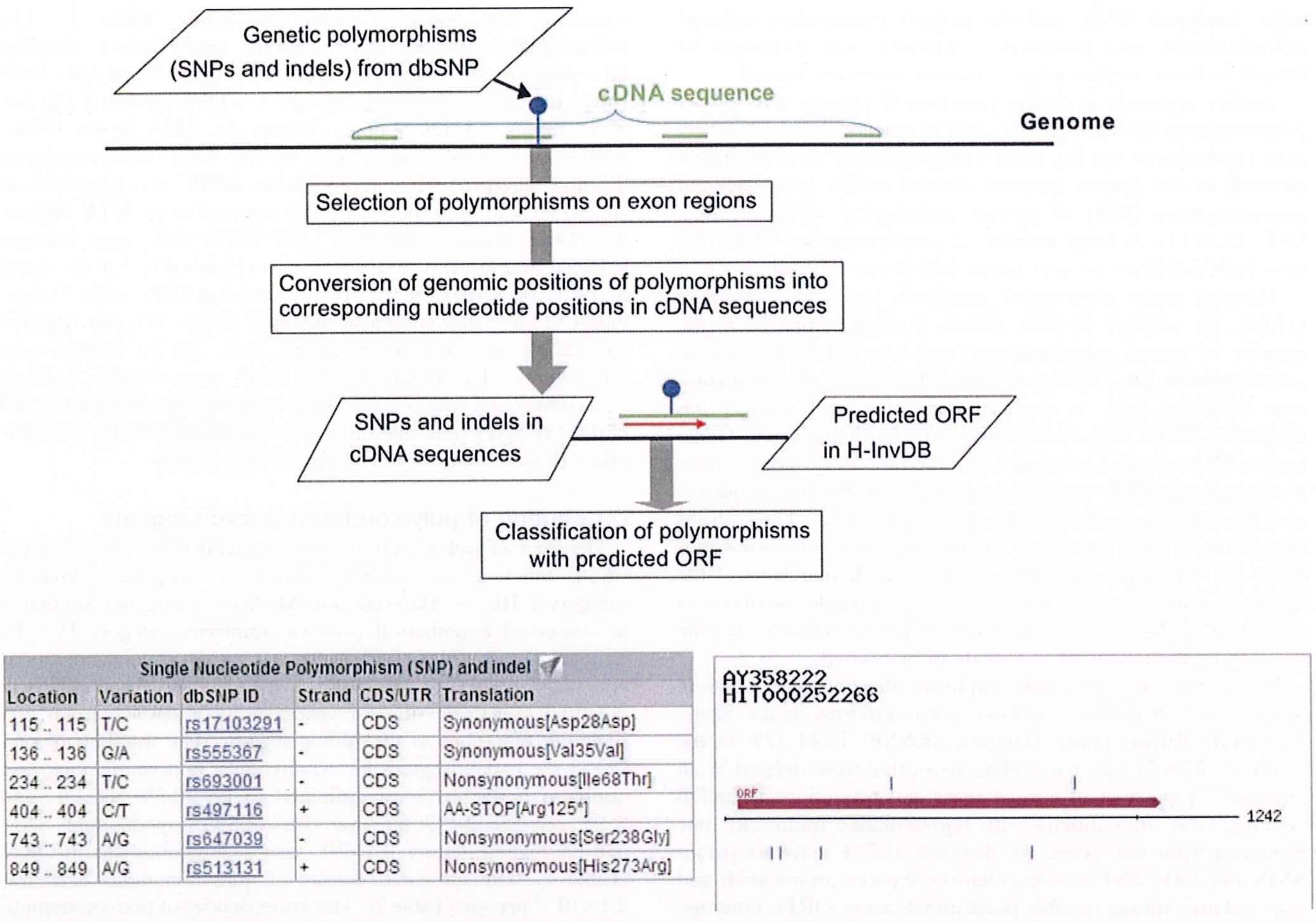
Densities of polymorphisms were estimated for 23,717 genes whose functions are clearly defined or suggested (similarity category I–III, see Materials and Methods) and genes annotated as conserved hypothetical proteins (similarity category IV). To estimate the densities of SNPs for synonymous, nonsynonymous and nonsense SNPs in the ORFs, we calculated the numbers of potential nucleotide sites for synonymous, nonsynonymous and nonsense mutations in the coding regions. The fractions of sites (%) in the coding regions for synonymous, nonsynonymous, and nonsense mutations were estimated to be 28.5%, 68.1%, and 3.4%, respectively. Of the three types of SNPs, synonymous SNPs had the highest density, 4.1×10^{-3} per synonymous site, in ORFs (Table 2). The estimated density of nonsynonymous SNP was 2.1×10^{-3} per site (Table 2). The lower density of nonsynonymous SNPs compared with synonymous SNPs (51%) is due to the functional constraint of amino acid changes, and is in agreement with previous studies [20,22,34]. However, the ratio of the numbers of nonsynonymous SNPs to synonymous SNPs per site is higher in this study compared with previous studies (32–34%) [20,21,22], which they focused on specific populations. The higher ratio of nonsynonymous SNPs in this study may be due to the fact that our study is based on pooled data from various populations world wide. This study includes many nonsynonymous SNPs that exist in relatively lower frequencies and are likely to be more population-specific in comparison to synonymous SNPs [20].

Among random nucleotide mutations in ORFs, 3.4% would be expected to be nonsense mutations; however, the distribution of nonsense SNPs has not been evaluated or reported. The density of nonsense SNPs was estimated to be 0.85×10^{-3} per site (Table 3), which is only 21% of the density of synonymous SNPs, and 40% of the density of nonsynonymous SNPs. The reason for the lowest density of nonsense SNPs may be that premature stop codons have more severe effects than amino acid changes.

In the exons of the 36,712 loci, 8479 indels were detected, and 1,532 of them were found in ORFs. Among the latter, 1,331 are expected to cause frame shifts, resulting in drastic changes of proteins. The density of indels in ORFs was much lower than in the UTR regions (Table 4). The lower density of indels in the 5′UTRs than in the 3′UTRs suggests that functional constraint for insertions and deletions is higher in the 5′UTR regions than in the 3′UTR regions.

Nonsense SNPs

We examined the patterns and the positions of the nonsense SNPs. There are 23 possible ways to change codons into stop


Genetic polymorphisms (SNPs and indels) from dbSNP
cDNA sequence
Genome
Selection of polymorphisms on exon regions
Conversion of genomic positions of polymorphisms into corresponding nucleotide positions in cDNA sequences
SNPs and indels in cDNA sequences
Predicted ORF in H-InvDB
Classification of polymorphisms with predicted ORF
AY358222
HIT000252266
ORF
1242

Single Nucleotide Polymorphism (SNP) and indel

Location	Variation	dbSNP ID	Strand	CDS/UTR	Translation
115 .. 115	T/C	rs17103291	-	CDS	Synonymous[Asp28Asp]
136 .. 136	G/A	rs555367	-	CDS	Synonymous[Val35Val]
234 .. 234	T/C	rs693001	-	CDS	Nonsynonymous[Ile68Thr]
404 .. 404	C/T	rs497116	+	CDS	AA-STOP[Arg125*]
743 .. 743	A/G	rs647039	-	CDS	Nonsynonymous[Ser238Gly]
849 .. 849	A/G	rs513131	+	CDS	Nonsynonymous[His273Arg]

Figure 1. Analysis of polymorphisms with gene structure. Top: Scheme of analysis pipeline of polymorphisms with gene structure. Bottom: Screen shots taken from 'Transcript View' in H-InvDB that show classified SNPs and their positions (blue bars) in the *CASP12* gene.
doi:10.1371/journal.pone.0003393.g001

codons (nine, seven and seven for the first, second and third positions, respectively), and all 23 were found (Table 5). Nonsense SNPs were more frequent at the first codon position than at the second and third positions ($p<0.005$, chi-square test). The most frequent type of nonsense mutation is the change from CGA to TGA (Table 5), which is a transitional change at CpG mutation hotspots [35]. However, it is notable that there were frequent transversional mutations such as GAA to TAA and GAG to TAG. Our analyses of nonsense polymorphisms revealed that changes between hydrophilic amino acids and termination codons by nucleotide changes at the first codon positions were very frequent.

We examined the positions of 1,183 nonsense polymorphisms in the coding regions. On average, nonsense SNPs were located at 250 codons upstream of the original termination codons. To predict whether a nonsense mutation causes nonsense-mediated decay (NMD) of mRNA, we examined the locations of nonsense SNPs in the exon-intron structure of the genes (Table 6). As a result, of the 1183 nonsense SNPs, 581 were predicted to cause NMD, and thus to prevent translation. The other 602 cases of nonsense SNPs were predicted to result in truncated proteins. For the cases that truncated proteins are produced, the average truncation was estimated to be 75 amino acids.

Table 1. SNPs and indels in exon, intron and other genomic regions.

	Exon	Intron	Other genomic regions
SNPs	249,182	3,332,537	5,209,127
Indels	9,742	185,761	249,648

Polymorphisms mapped on single positions were analyzed with 36,712 protein-coding genes.
doi:10.1371/journal.pone.0003393.t001

To see which of these nonsense SNPs were known pathological mutations, we compared them with allelic variants in the Online Mendelian Inheritance in Man (OMIM) database. Only eight of 1,183 nonsense SNPs (rs17602729 in *AMPD1*, rs283413 in *ADH1C*, rs10250779 in *PGAM2*, rs17215500 in *KCNQ1*, rs497116 in *CASP12*, rs2228325 in *ACTN3*, rs3092891 in *RB1* and rs28989186 in *BUB1B*) matched the variants in the OMIM database that are known variants with phenotypic variations (Table 7). This low value suggests that the biological effects of most nonsense SNPs have not yet been reported. Interestingly, each of the eight cases that matched known pathological variants was predicted to cause NMD (Table 7).

SNPs that cause read-though of the original termination codon

Among the 247 SNPs at termination codon sites, 119 SNP-mRNA pairs were found to be read-through mutations. If the allele having the stop codon is the ancestral type, the SNP is

Table 2. Classified SNPs in exon regions.

Region	Effects on translation	Genes in category I-IV[a]	All protein-coding genes[b]
5′UTR		23454 [3.3×10^{-3}/site][c]	51881
ORF	Total	85233 [2.7×10^{-3}/site]	96164
	Synonymous	37484 [4.1×10^{-3} /site]	40484
	Nonsynonymous	46261 [2.1×10^{-3} /site]	53754
	AA↔Ter[d]	938	1258
	Unclassified[e]	398	421
Stop codon	Total	152	247
	Synonymous	63	88
	Ter↔AA[d]	89	159
3′UTR		69691 [3.3×10^{-3}/site]	104510
Total		178378	252555

[a]Representative transcripts in 23,717 genes whose function were defined or suggested (similarity category I–III) and genes annotated as conserved hypothetical proteins (similarity category IV).

[b]Representative transcripts in all protein-coding genes (36,712) including genes in similarity category I–IV plus similarity category V–VII (hypothetical protein, hypothetical short protein, and pseudogene candidate, respectively).

[c]Densities of polymorphisms are shown in brackets as average number of polymorphisms per site. The average lengths of the 5′UTR, ORF and 3′UTR regions in 23717 genes were 303.9 bp, 1343.5 bp, and 877.6 bp, respectively. The densities of SNPs for synonymous, nonsynonymous and nonsense SNPs in ORFs were calculated based on the numbers of potential nucleotide sites for synonymous, nonsynonymous and nonsense mutations in coding regions. The density of nonsense SNPs is shown in Table 3.

[d]SNPs causing changes between amino acids and stop codons.

doi:10.1371/journal.pone.0003393.t002

Table 3. SNPs causing changes between amino acids and stop codons.

Region	Effects on translation	Genes in category I-IV[a]	All protein-coding genes[a]
ORF	Nonsense	910 [0.85×10^{-3}/site][d]	1183
	Read-through[b]	28	75
Stop codon	Read-through	67	110
	Nonsense[c]	22	49

[a]These two gene sets are the same as Table 2.

[b]Possible read-through SNPs in which alleles coding stop codons were ancestral type. This may be due to existence of shorter ORFs in the ancestral population.

[c]Possible nonsense SNPs in which alleles coding stop codons were derived alleles. This may be due to existence of longer ORFs in the ancestral population.

[d]The densities of nonsense SNPs in ORFs were calculated based on the numbers of potential nucleotide sites for nonsense mutations in coding regions.

doi:10.1371/journal.pone.0003393.t003

Table 4. Insertions and deletions in exon regions.

	Genes in category I-IV[a]	All protein-coding genes[a]
5′UTR	785 [0.11×10^{-3}][b]	2005
ORF	1120 [0.035×10^{-3}]	1532
3′UTR	3323 [0.16×10^{-3}]	4942
Total	5225[c]	8479

[a]These two gene sets are the same as Table 2.

[b]Densities of polymorphisms are shown in brackets as average number of polymorphisms per site.

[c]Three indels were located on both of ORF and UTR.

doi:10.1371/journal.pone.0003393.t004

Table 5. Frequency of each type of codon change for nonsense SNPs.

	TAA		TAG		TGA		Total
	Aaa→Taa	33	Aag→Tag	31	Aga→Tga	20	
1st	**Caa→Taa**	**62**	**Cag→Tag**	**162**	**Cga→Tga**	**203**	748*
	Gaa→Taa	80	Gag→Tag	125	Gga→Tga	32	
	tCa→tAa	27	tCg→tAg	19	tCa→tGa	25	
2nd			**tGg→tAg**	**80**			200
	tTa→tAa	18	tTg→tAg	18	tTa→tGa	13	
	taC→taA	25	taC→taG	25	tgC→tgA	22	
3rd					**tgG→tgA**	**85**	235
	taT→taA	19	taT→taG	27	tgT→tgA	32	
Total		264		487		432	1183

Bold letters show nucleotide changes by transition.

*$P<0.005$ by chi-square test.

doi:10.1371/journal.pone.0003393.t005

regarded as a change causing elongation of the polypeptide. However, an extended polypeptide would be expected only if there is an additional termination codon downstream. For 108 SNP-mRNA pairs, an additional termination codon was found in the 3′UTR region. The average extension was estimated to be 29 amino acids. Interestingly, we found five SNP-mRNA pairs that have no stop codons in the 3′UTR at all (The remaining six SNP-mRNA pairs do not have 3′UTR regions). For example, the T-to-C substitution (rs15941) in the *DDR2* gene (X74764) is predicted to be a read-through mutation (from TAG to CGA), and the transcript has no other stop codon in the 3′UTR region. The frequency of this SNP is unknown (it is monomorphic in the four populations in HapMap project [4]). However, if this polymorphism really exists, transcripts having this read-through mutation would not produce a protein. Another example is the T-to-C substitution (rs17850833) in the *MFSD3* gene (CR620962), which causes a change from TGA to CGA resulting in a change to arginine.

Functional bias of genes having nonsense SNPs

To see whether there is any functional bias in genes having nonsense SNPs, we examined the frequent biological terms in the genes having nonsense SNPs. We classified the genes having nonsense SNPs into two categories: genes with nonsense SNPs that are predicted to cause NMD and genes with nonsense SNPs that are not predicted to cause NMD. For genes having nonsense SNPs that would cause NMD (Table 8), the molecular functions that are most overrepresented included phosphorylation, ATP binding, iron/calcium ion binding, nucleotide/RNA binding and transporter activity. The localization of these genes was also biased to the cell membrane and the proteinaceous extracellular matrix. On the other hand, the genes having nonsense SNPs predicted to not cause NMD showed less bias in biological function (Table 9).

Table 6. Nonsense SNPs and prediction of NMD.

	Predicted to cause NMD[a]	Not for NMD[b]	Total
Known pathological variants	8[c]	0	8
Other nonsense SNPs	573	602	1175
Total	581	602	1183

[a] This prediction is based on that mRNA would be destroyed if a stop codon occurs in the 5′ side of the boundary, which is 50–55 nucleotides upstream from the 3′ end of the second to last exon. Here, the nonsense SNPs located in the 5′ side of the boundary, which was set at 50 nucleotides upstream from the 3′ end of the second to last exon, were predicted to cause NMD.

[b] This number includes SNPs in genes consisting of only one exon.

[c] P = 0.0033 by Fisher's exact test.

doi:10.1371/journal.pone.0003393.t006

Table 7. Nonsense SNPs with known pathological effects.

Acc#	Chr	Gene symbol	SNP	Variation	OMIM	Biological effects
M60092	1	*AMPD1*	rs17602729	Gln12Ter	102770	AMPD deficiency
M12272	4	*ADH1C*	rs283413	Gly78Ter	103730	Parkinson disease
BC073741	7	*PGAM2*	rs10250779	Trp78Ter	261670	Myopathy
AF000571	11	*KCNQ1*	rs17215500	Arg518Ter	607542	Long QT syndrome 1
AY358222	11	*CASP12*	rs497116	Arg125Ter	608633	Sepsis susceptibility
M86407	11	*ACTN3*	rs2228325	Arg577Ter	102574	Athletic performance
L41870	13	*RB1*	rs3092891	Arg445Ter	180200	Bilateral retinoblastoma
AF068760	15	*BUB1B*	rs28989186	Arg194Ter	602860	Premature chromatid separation trait and mosaic variegated aneuploidy syndrome

doi:10.1371/journal.pone.0003393.t007

Table 8. Functional bias of genes having nonsense SNPs causing NMD.

Top level	Gene Ontology no.	Gene Ontology	Observed gene no.[a]	Expected gene no.[b]	Ratio of enrichment	P value[c]
Biological process	0006118	electron transport	15	4.23	3.55	5.03×10^{-5}
	0006468	protein amino acid phosphorylation	16	7.28	2.20	4.98×10^{-3}
Cellular component	0016020	membrane	41	22.55	1.82	5.57×10^{-4}
	0005578	proteinaceous extracellular matrix	8	1.21	6.62	2.17×10^{-6}
Molecular function	0005524	ATP binding	35	17.15	2.04	1.79×10^{-4}
	0004713	protein tyrosine kinase activity	16	6.46	2.48	1.56×10^{-3}
	0004674	protein serine/threonine kinase activity	16	6.78	2.36	2.51×10^{-3}
	0000166	nucleotide binding	14	5.61	2.50	2.79×10^{-3}
	0004672	protein kinase activity	16	7.15	2.24	4.21×10^{-3}
	0003723	RNA binding	10	3.11	3.22	1.82×10^{-3}
	0005506	iron ion binding	8	2.00	4.00	1.32×10^{-3}
	0005509	calcium ion binding	16	7.65	2.09	7.89×10^{-3}
	0005215	transporter activity	10	3.44	2.91	3.76×10^{-3}
	0016491	oxidoreductase activity	11	4.24	2.59	5.76×10^{-3}
	0003779	actin binding	6	1.27	4.74	2.24×10^{-3}
	0004759	carboxylesterase activity	5	0.24	20.44	4.19×10^{-6}

[a] Number of genes with a molecular function in the 581 genes in which nonsense SNPs causing NMD were found.

[b] Expected number of genes that have a biological function in a sample of 581 genes, assuming a proportion of genes with a molecular function in all human genes.

[c] Enrichment of a biological term in the genes for nonsense SNPs was statistically evaluated as a upper probability in a hypergeometric distribution.

doi:10.1371/journal.pone.0003393.t008

Table 9. Functional bias of genes having nonsense SNPs not causing NMD.

Top level	Gene Ontology no.	Gene Ontology	Observed gene no.[a]	Expected gene no.[b]	Ratio of enrichment	P value[c]
Biological process	0007156	homophilic cell adhesion	6	1.42	4.23	3.05×10^{-3}
	0006310	DNA recombination	3	0.19	15.50	8.25×10^{-4}
	0006414	translational elongation	3	0.34	8.85	4.48×10^{-3}
	0042254	ribosome biogenesis and assembly	2	0.15	13.77	8.68×10^{-3}
Cellular component	0005853	eukaryotic translation elongation factor 1 complex	2	0.13	15.50	6.82×10^{-3}
Molecular function	0004194	pepsin A activity	2	0.18	11.27	1.30×10^{-2}
	0003746	translation elongation factor activity	2	0.29	6.89	3.35×10^{-2}

[a]Number of genes with a molecular function in the 602 genes in which nonsense SNPs causing NMD were found.
[b]Expected number of genes that have a biological function in a sample of 602 genes, assuming a proportion of genes with a molecular function in all human genes.
[c]Enrichment of a biological term in the genes for nonsense SNPs was statistically evaluated as a upper probability in a hypergeometric distribution.

doi:10.1371/journal.pone.0003393.t009

Discussion

In this study, we conducted an extensive analysis of human genome polymorphisms with a comprehensive catalogue of human genes, and detected more than 50,000 polymorphisms that affect proteins. The distribution of polymorphisms showed different densities of polymorphisms among the 5′UTR, ORF and 3′UTR. The density of SNPs was lower in ORFs than in the 5′UTR and 3′UTR. The density of synonymous SNPs in the ORFs was higher than the densities of SNPs in the UTR regions. The reduction in density of SNPs in the UTR regions is consistent that there are functional constraints on nucleotide changes in UTRs related to the transcriptional and translational efficiency[22]. The density of nonsynonymous SNPs was much lower than the densities of other types of SNPs, possibly due to that the nucleotide changes with alteration of amino acids changes are under strong negative selection [36]. It was not known how nonsense SNPs are distributed in protein-coding regions. Here we showed that the density of nonsense SNPs is much lower than that of nonsynonymous SNPs. Although the biological effects of nonsense mutations appear to vary widely depending on their positions and the genes, the low density of nonsense SNPs that we found suggests that nonsense mutations have more disadvantageous effects than nonsynonymous mutations.

While nonsense mutations that cause NMD result in 'loss of function', nonsense mutations that do not cause NMD produce truncated proteins which could have the dominant effects. The proportion of predicted nonsense SNPs causing NMD in this study is in agreement with a previous study which showed that dbSNP (build 125) has 1301 nonsense SNPs, about half of which were predicted to result in NMD [37]. In order to understand the biological effects of nonsense SNPs, it is important to know whether they do or do not cause NMD, because premature stop codons in a gene can have distinct disease phenotypes depending on the positions of mutations [27,38].

The molecular functions that were overrepresented in the genes having nonsense SNPs included several molecular functions that were observed in human-specific pseudogenes[39], such as ATP binding, actin binding, calcium ion binding, extracellular matrix, nucleic acid binding and oxidoreductase. This is in accord with that nonsense mutations contribute to 'pseudogenization'. It is interesting that nonsense SNPs causing NMD were frequently found in genes that encode proteins involved in phosphorylation, cell-cell interaction, signal transduction and transport. This may be because changes in the length of polypeptides caused by nonsense mutations are under strong negative selection in the genes involved in signal transduction or transportation because abnormal translation products could cause dominant effects. Therefore, inactivation of translation by nonsense mutations in those genes could have milder effects than changes of the length of polypeptides.

Our results showed a low proportion of matches of nonsense SNPs with known pathological variants in OMIM, suggesting that the effects of most nonsense polymorphisms are unknown or not reported. Furthermore, the correspondence of the nonsense SNPs to the OMIM allelic variants (Table 6, Table 7) suggests that nonsense polymorphisms that are subject to NMD are more likely to be involved in phenotypic variations.

There is a possibility that the nonsense SNPs detected here have pathological effects, in particular, if non-dispensable genes have nonsense mutations. First, a defect in one gene by a nonsense mutation or a frame-shifting indels causing a premature termination codon could be a cause of genetic diseases including complex diseases[40]. Second, there is a possibility that nonsense mutations cause recessive lethal alleles that would not be detected as causative variant of diseases. Probably, focusing on nonsense polymorphisms observed in specific populations would be a good way of selection for finding variants with deleterious effects.

The effect of single nonsense SNPs can be compensated by the products of other genes having similar functions[41] and the other splicing isoforms of the gene [42]. Thus, single nonsense SNPs may not always cause severe phenotypic effects. In fact, some nonsense SNPs with high allele frequencies were found across populations[43]. There is a report of fixation of an inactive form of caspase 12 by a nonsense mutation (rs497116) in non-African populations[43], and this is an example supporting the 'less is more hypothesis'[44]. This example suggests that some of nonsense mutations are not disadvantageous and that the increase of frequency of a nonsense allele could be driven by positive selection.

Elongation of polypeptides by read-through mutations can affect protein folding and aggregation of proteins, which could affect phenotypic variations. Furthermore, a read-through mutation can cause more severe effects on translation when no additional stop codon follows. Such mutations are subject to 'non-stop decay' [32,33], and would result in no gene product. It has been suggested that non-stop decay and NMD serve to remove toxic, aberrant proteins [29]. It is unclear how frequently such mutations prevent mRNA from producing proteins. Therefore, it would be quite useful to be able to predict the effects of various types of genetic changes on mRNA.

Although the present results are based on representative transcripts (one transcript for one gene), the total number of

SNPs causing changes between amino acids and stop codons in all the splicing isoforms was much larger (2,234). These variations, which cause changes in the length of a polypeptide or which determine whether a protein is translated, may include pathological variants that have yet not been reported. Therefore, it is important to examine their presence in human populations.

Materials and Methods

Data of human genetic polymorphisms

As data of genetic polymorphisms of human genome, single nucleotide polymorphisms (SNPs) and insertions and deletions (indels) in dbSNP [1] were used in this study. The whole data of human SNPs and indels were downloaded from dbSNP (build 125). We used all SNPs and indels that were mapped on single position in the genome, except for 'large insertions' in dbSNP.

Data of human genes

The data of human gene structure were obtained from H-InvDB ver3.8 (http://www.h-invitational.jp/), created by the annotation project of human genes (H-Invitational project) [23,45]. Our analysis of all human genes that corresponds to H-InvDB (ver 3.8) predicted 36,712 protein coding loci. All protein-coding genes were annotated and classified based on similarity to known genes as follows; Category I, Identical to known human protein; Category II, Similar to known protein; Category III, IPR domain containing protein; Category IV, Conserved hypothetical protein; Category V, Hypothetical protein; Category VI, Hypothetical short protein; Category VII, pseudogene candidate. We used the following three kinds of data of the gene structure: 1) genomic location of exons to the human genome (build 35), 2) predicted ORF regions in transcripts, and 3) original and curated cDNA sequences.

Analysis

1. Analysis of polymorphism with exons and predicted ORFs. Selection of polymorphisms on exon regions. We selected polymorphisms in exon regions by comparing the genomic positions of polymorphisms and the start and end positions of exons that were obtained from mapping cDNA sequences to the human genome (Figure 1). Polymorphisms in introns were also selected in a same way.

Conversion of genomic position of polymorphism into nucleotide position in cDNA sequence. To analyze polymorphisms with a predicted ORF, nucleotide positions of polymorphisms in the human genome sequences were converted into the nucleotide positions in cDNA sequences. Because there could be gaps in the alignment of cDNA sequence and the human genome sequence, the nucleotide position was converted considering possible gaps in the alignment. When the cDNA sequence was corrected in ORF prediction because of frame-shifting and remaining intron, the nucleotide position of SNP was modified based on addition or deletion of nucleotides. For a quality control of polymorphism data used for classification, we conformed that one of the nucleotides in each pair of SNP alleles was the same nucleotide at the corresponding position in the cDNA sequence.

Classification of polymorphisms with predicted ORF. Polymorphisms within ORF were classified according to their effect on ORF. For SNPs with two alleles, alleles in nucleotide were converted into 'alleles in codon' by adding two other nucleotides in the codon from cDNA sequence. When a cDNA sequence was corrected in the annotation process by removing a remaining intron or by correcting a frameshift error, the corrected cDNA sequence was used. If these alleles in codon do not contain any stop codon, the alleles were classified into synonymous and nonsynonymous. In case a stop codon is included in the alleles in codon, they were classified into 1) premature termination (nonsense) codon, 2) read-through of original stop codon, and 3) synonymous at stop codon site, by assuming that the cDNA sequence has an ancestral allele. Indels were classified based on whether they are located in ORF. The indels within ORF were further classified by whether the insertion or deletion causes frame shifting in translation.

Inference of direction of nonsense and read-through mutations. Ancestral alleles were obtained from dbSNP (build 128) to check direction of mutations for SNPs causing changes between amino acids and stop codons. For nonsense SNPs in protein-coding regions, we checked whether the ancestral allele codes amino acids. In case that the ancestral allele codes stop codon, we do not regard this SNP as nonsense SNP, but is a read-through mutation assuming that there was a variant having a shorter ORF. For read-though SNPs at termination codon site, we checked whether the ancestral allele codes stop codon. In case that the ancestral allele codes amino acids, we regard this SNP not as a read-through mutation, but as a nonsense mutation in a variant having a longer ORF.

Number of sites for synonymous, nonsynonymous and nonsense mutations. To estimate densities of synonymous, nonsynonymous and nonsense SNPs, the numbers of potential synonymous, nonsynonymous and nonsense sites by single nucleotide changes were estimated for the ORF sequences. This is an extension of estimation of the numbers of synonymous and nonsynonymous sites[46]; the number of synonymous sites is calculated as the number of four-fold degenerate sites plus one-third of the number of two-fold degenerate sites. For 61 codons encoding amino acids, the numbers of nucleotide sites that would cause synonymous, nonsynonymous and nonsense mutations by a single nucleotide change were estimated with a model of nucleotide change. Here, the relative occurrence of a transitional mutation versus a transversional mutation (r) was set to be 4.0 (the expected ratio in the numbers of transitional and transversional mutations was 2.0). For example of the TTA codon for leucine, the number of nonsense sites was estimated to be 2.0/(r+2.0), because two types of transversional mutations at the second position cause nonsense mutations.

2. Correspondence to known pathological variants. To check whether the polymorphisms that alter proteins are known pathological variants with phenotypic effect, we examined correspondence of SNPs with data of known pathological variants. We used data of 'allelic variant' in the Online Mendelian Inheritance in Man (OMIM) database [18] as information of variants with phenotypic effect. For nonsynonymous and nonsense SNPs, their effects on translation and positions in ORF were compared with the 'list of alleles' in OMIM (e.g. described as "TRP324TER" or "ALA279THR" for the *NGAS* gene).

3. Prediction of nonsense SNPs causing NMD. Some of nonsense mutations cause nonsense-mediated decay (NMD), resulting in prevention of translation. It has been reported that mRNA would be destroyed if a stop codon occurs in the 5′ side of the boundary, which is 50–55 nucleotides upstream from the end of the second to last exon [30,31]. To predict whether a nonsense SNP causes NMD, we examined whether a nonsense SNP is located in the 3′ side of the boundary, which was set at 50 nucleotides upstream from the end of the second to last exon, in the exon-intron structure. This method is the same as the method in SNP2NMD [37] when 'NMD distance' is 50 nucleotides.

5. Functional bias of genes with nonsense SNPs. For each biological term from Gene Ontology (www.geneontology.org), a proportion of genes with the biological function in the genes having nonsense SNPs was compared with that in all human genes (representative transcripts in all human genes in H-InvDB ver 5.0), and the significance of over representation of a molecular function in the genes having nonsense SNPs was evaluated as the upper probability of the hypergeometric distribution.

Supporting Information

Results S1 Supplementary results and a table for analyses of nonsynonymous SNPs.
Found at: doi:10.1371/journal.pone.0003393.s001 (0.70 MB DOC)

Table S1 Nonsense SNPs and read-through SNPs on representative transcripts.
Found at: doi:10.1371/journal.pone.0003393.s002 (4.24 MB DOC)

Acknowledgments

We thank Drs. Craig Gough, Norikazu Yasuda, Shuhei Mano, Naoki Nagata, Yoshiyuki Suzuki and Hisakazu Iwama for helpful discussion. We also thank Ryuzou Matsumoto, Seigo Hosono, and all the members in the Integrated Database Team of BIRC, AIST for their technical assistance and providing data of gene structure and annotation.

Author Contributions

Conceived and designed the experiments: YYK SM RC TG TI. Analyzed the data: YYK MS YH. Wrote the paper: YYK MS TI.

References

1. Sherry ST, Ward M, Sirotkin K (1999) dbSNP-database for single nucleotide polymorphisms and other classes of minor genetic variation. Genome Res 9: 677–679.
2. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, et al. (2001) The sequence of the human genome. Science 291: 1304–1351.
3. Klein RJ, Zeiss C, Chew EY, Tsai JY, Sackler RS, et al. (2005) Complement factor H polymorphism in age-related macular degeneration. Science 308: 385–389.
4. The International HapMap Consortium (2005) A haplotype map of the human genome. Nature 437: 1299–1320.
5. Ozaki K, Ohnishi Y, Iida A, Sekine A, Yamada R, et al. (2002) Functional SNPs in the lymphotoxin-alpha gene that are associated with susceptibility to myocardial infarction. Nat Genet 32: 650–654.
6. The Wellcome Trust Case Control Consortium (2007) Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 447: 661–678.
7. Reumers J, Schymkowitz J, Ferkinghoff-Borg J, Stricher F, Serrano L, et al. (2005) SNPeffect: a database mapping molecular phenotypic effects of human non-synonymous coding SNPs. Nucleic Acids Res 33: D527–532.
8. Ng PC, Henikoff S (2002) Accounting for human polymorphisms predicted to affect protein function. Genome Res 12: 436–446.
9. Bao L, Zhou M, Cui Y (2005) nsSNPAnalyzer: identifying disease-associated nonsynonymous single nucleotide polymorphisms. Nucleic Acids Res 33: W480–482.
10. Sunyaev S, Ramensky V, Koch I, Lathe W 3rd, Kondrashov AS, et al. (2001) Prediction of deleterious human alleles. Hum Mol Genet 10: 591–597.
11. Ferrer-Costa C, Gelpi JL, Zamakola L, Parraga I, de la Cruz X, et al. (2005) PMUT: a web-based tool for the annotation of pathological mutations on proteins. Bioinformatics 21: 3176–3178.
12. Yue P, Melamud E, Moult J (2006) SNPs3D: candidate gene and SNP selection for association studies. BMC Bioinformatics 7: 166.
13. Stitziel NO, Binkowski TA, Tseng YY, Kasif S, Liang J (2004) topoSNP: a topographic database of non-synonymous single nucleotide polymorphisms with and without known disease association. Nucleic Acids Res 32: D520–522.
14. Karchin R, Diekhans M, Kelly L, Thomas DJ, Pieper U, et al. (2005) LS-SNP: large-scale annotation of coding non-synonymous SNPs based on multiple information sources. Bioinformatics 21: 2814–2820.
15. Burton PR, Clayton DG, Cardon LR, Craddock N, Deloukas P, et al. (2007) Association scan of 14,500 nonsynonymous SNPs in four diseases identifies autoimmunity variants. Nat Genet 39: 1329–1337.
16. Hampe J, Franke A, Rosenstiel P, Till A, Teuber M, et al. (2007) A genome-wide association scan of nonsynonymous SNPs identifies a susceptibility variant for Crohn disease in ATG16L1. Nat Genet 39: 207–211.
17. Minoshima S, Mitsuyama S, Ohtsubo M, Kawamura T, Ito S, et al. (2001) The KMDB/MutationView: a mutation database for human disease genes. Nucleic Acids Res 29: 327–328.
18. Hamosh A, Scott AF, Amberger J, Bocchini C, Valle D, et al. (2002) Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. Nucleic Acids Res 30: 52–55.
19. Morton NE, Crow JF, Muller HJ (1956) An Estimate of the Mutational Damage in Man from Data on Consanguineous Marriages. Proc Natl Acad Sci U S A 42: 855–863.
20. Cargill M, Altshuler D, Ireland J, Sklar P, Ardlie K, et al. (1999) Characterization of single-nucleotide polymorphisms in coding regions of human genes. Nat Genet 22: 231–238.
21. Fay JC, Wyckoff GJ, Wu CI (2001) Positive and negative selection on the human genome. Genetics 158: 1227–1234.
22. Halushka MK, Fan JB, Bentley K, Hsie L, Shen N, et al. (1999) Patterns of single-nucleotide polymorphisms in candidate genes for blood-pressure homeostasis. Nat Genet 22: 239–247.
23. Imanishi T, Itoh T, Suzuki Y, O'Donovan C, Fukuchi S, et al. (2004) Integrative annotation of 21,037 human genes validated by full-length cDNA clones. PLoS Biol 2: e162.
24. Yamasaki C, Murakami K, Fujii Y, Sato Y, Harada E, et al. (2008) The H-Invitational Database (H-InvDB), a comprehensive annotation resource for human genes and transcripts. Nucleic Acids Res 36: D793–799.
25. Chang JC, Kan YW (1979) beta 0 thalassemia, a nonsense mutation in man. Proc Natl Acad Sci U S A 76: 2886–2889.
26. Rosenfeld PJ, Cowley GS, McGee TL, Sandberg MA, Berson EL, et al. (1992) A null mutation in the rhodopsin gene causes rod photoreceptor dysfunction and autosomal recessive retinitis pigmentosa. Nat Genet 1: 209–213.
27. Inoue K, Khajavi M, Ohyama T, Hirabayashi S, Wilson J, et al. (2004) Molecular mechanism for distinct neurological phenotypes conveyed by allelic truncating mutations. Nat Genet 36: 361–369.
28. Mimori A, Hidaka Y, Wu VC, Tarle SA, Kamatani N, et al. (1991) A mutant allele common to the type I adenine phosphoribosyltransferase deficiency in Japanese subjects. Am J Hum Genet 48: 103–107.
29. Holbrook JA, Neu-Yilik G, Hentze MW, Kulozik AE (2004) Nonsense-mediated decay approaches the clinic. Nat Genet 36: 801–808.
30. Thermann R, Neu-Yilik G, Deters A, Frede U, Wehr K, et al. (1998) Binary specification of nonsense codons by splicing and cytoplasmic translation. Embo J 17: 3484–3494.
31. Zhang J, Sun X, Qian Y, LaDuca JP, Maquat LE (1998) At least one intron is required for the nonsense-mediated decay of triosephosphate isomerase mRNA: a possible link between nuclear splicing and cytoplasmic translation. Mol Cell Biol 18: 5272–5283.
32. Frischmeyer PA, van Hoof A, O'Donnell K, Guerrerio AL, Parker R, et al. (2002) An mRNA surveillance mechanism that eliminates transcripts lacking termination codons. Science 295: 2258–2261.
33. van Hoof A, Frischmeyer PA, Dietz HC, Parker R (2002) Exosome-mediated recognition and degradation of mRNAs lacking a termination codon. Science 295: 2262–2264.
34. Hughes AL, Packer B, Welch R, Bergen AW, Chanock SJ, et al. (2003) Widespread purifying selection at polymorphic sites in human protein-coding loci. Proc Natl Acad Sci U S A 100: 15754–15757.
35. Ehrlich M, Wang RY (1981) 5-Methylcytosine in eukaryotic DNA. Science 212: 1350–1357.
36. Kimura M (1968) Evolutionary rate at the molecular level. Nature 217: 624–626.
37. Han A, Kim WY, Park SM (2007) SNP2NMD: a database of human single nucleotide polymorphisms causing nonsense-mediated mRNA decay. Bioinformatics 23: 397–399.
38. Thein SL, Hesketh C, Taylor P, Temperley IJ, Hutchinson RM, et al. (1990) Molecular basis for dominantly inherited inclusion body beta-thalassemia. Proc Natl Acad Sci U S A 87: 3924–3928.
39. Wang X, Grus WE, Zhang J (2006) Gene losses during human origins. PLoS Biol 4: e52.
40. Senee V, Chelala C, Duchatelet S, Feng D, Blanc H, et al. (2006) Mutations in GLIS3 are responsible for a rare syndrome with neonatal diabetes mellitus and congenital hypothyroidism. Nat Genet 38: 682–687.
41. Gu Z, Steinmetz LM, Gu X, Scharfe C, Davis RW, et al. (2003) Role of duplicate genes in genetic robustness against null mutations. Nature 421: 63–66.
42. Takeda J, Suzuki Y, Nakao M, Barrero RA, Koyanagi KO, et al. (2006) Large-scale identification and characterization of alternative splicing variants of human gene transcripts using 56,419 completely sequenced and manually annotated full-length cDNAs. Nucleic Acids Res 34: 3917–3928.
43. Xue Y, Daly A, Yngvadottir B, Liu M, Coop G, et al. (2006) Spread of an inactive form of caspase-12 in humans is due to recent positive selection. Am J Hum Genet 78: 659–670.
44. Olson MV (1999) When less is more: gene loss as an engine of evolutionary change. Am J Hum Genet 64: 18–23.
45. Yamasaki C, Koyanagi KO, Fujii Y, Itoh T, Barrero R, et al. (2005) Investigation of protein functions through data-mining on integrated human transcriptome database, H-Invitational database (H-InvDB). Gene 364: 99–107.
46. Nei M, Gojobori T (1986) Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Mol Biol Evol 3: 418–426.

The future of biocuration

To thrive, the field that links biologists and their data urgently needs structure, recognition and support.

Doug Howe, Seung Yon Rhee *et al.*

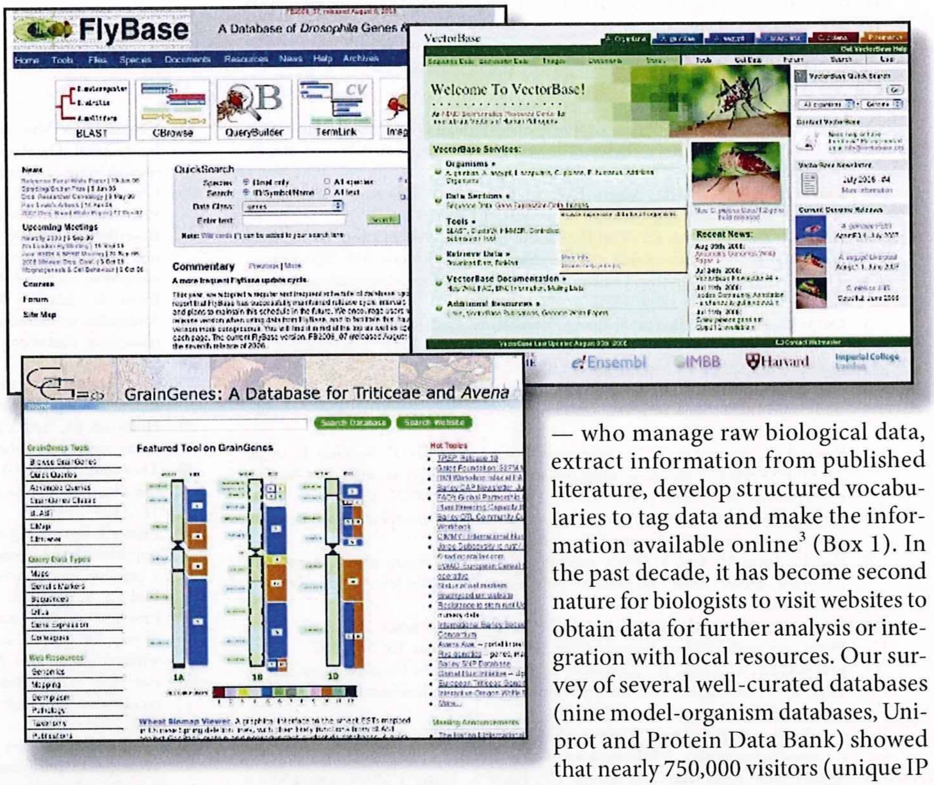
BIG DATA

The exponential growth in the amount of biological data means that revolutionary measures are needed for data management, analysis and accessibility. Online databases have become important avenues for publishing biological data. Biocuration, the activity of organizing, representing and making biological information accessible to both humans and computers, has become an essential part of biological discovery and biomedical research. But curation increasingly lags behind data generation in funding, development and recognition.

We propose three urgent actions to advance this key field. First, authors, journals and curators should immediately begin to work together to facilitate the exchange of data between journal publications and databases. Second, in the next five years, curators, researchers and university administrations should develop an accepted recognition structure to facilitate community-based curation efforts. Third, curators, researchers, academic institutions and funding agencies should, in the next ten years, increase the visibility and support of scientific curation as a professional career.

Failure to address these three issues will cause the available curated data to lag farther behind current biological knowledge. Researchers will observe an increasing occurrence of obvious gaps in knowledge. As these gaps expand, resources will become less effective for generating and testing hypotheses, and the usefulness of curated data will be seriously compromised.

When all the data produced or published are curated to a high standard and made accessible as soon as they become available, biological research will be conducted in a manner that is quite unlike the way it is done now. Researchers will be able to process massive amounts of complex data much more quickly. They will garner insight about the areas of their interest rapidly with the help of inference programs. Digesting information and generating hypotheses at the computer screen will be so much faster that researchers will get back to the bench quickly for more experiments. Experiments will be designed with more insight; this increased specificity will cause an exponential growth in knowledge, much as we are experiencing exponential growth in data today.



Data avalanche

Biology, like most scientific disciplines, is in an era of accelerated information accrual and scientists increasingly depend on the availability of each others' data. Large-scale sequencing centres, high-throughput analytical facilities and individual laboratories produce vast amounts of data such as nucleotide and protein sequences, protein crystal structures, gene-expression measurements, protein and genetic interactions and phenotype studies. By July 2008, more than 18 million articles had been indexed in PubMed and nucleotide sequences from more than 260,000 organisms had been submitted to GenBank[1,2]. The recently announced project to sequence 1,000 human genomes in three years to reveal DNA polymorphisms (www.1000genomes.org) is a tip of the data iceberg.

Such data, produced at great effort and expense, are only as useful as researchers' ability to locate, integrate and access them. In recent years, this challenge has been met by a growing cadre of biologists — 'biocurators' — who manage raw biological data, extract information from published literature, develop structured vocabularies to tag data and make the information available online[3] (Box 1). In the past decade, it has become second nature for biologists to visit websites to obtain data for further analysis or integration with local resources. Our survey of several well-curated databases (nine model-organism databases, Uniprot and Protein Data Bank) showed that nearly 750,000 visitors (unique IP addresses) viewed more than 20 million pages in just one month (March 2008, Eva Huala, Peter Rose, Rolf Apweiler, personal communications).

Despite the essential part that it plays in today's research, biocuration has been slow to develop. To provide a forum for the exchange of ideas and methods, and to facilitate collaborations and training, more than 150 biocurators met at two international conferences and created a mailing list and a website (www.biocurator.org). These meetings and discussions have honed in on the three actions, outlined above and elaborated on below, that must now be addressed to ensure scientists' continued access to the high-quality data on which their research depends.

Come together

Extracting, tagging with controlled vocabularies, and representing data from the literature, are some of the most important and time-consuming tasks in biocuration. Curated information from the literature serves as the gold-standard data set for computational analysis, quality assessment of high-throughput data and benchmarking of data-mining algorithms. Meanwhile, the boundaries of the biological domain that researchers study are widening rapidly, so researchers need faster and more reliable ways to understand unfamiliar domains. This too is facilitated by literature curation.

Typically, biocurators read the full text of articles and transfer the essence into a database. For a paper about the molecular biology of a particular gene, process or pathway, such information might include gene-expression patterns, mutant phenotypes, results of biochemical assays, protein-complex membership and the authors' inferences about the functions and roles of the gene products studied. As each paper uses different experimental and analysis methods, capturing this information in a consistent fashion requires intensive thought and effort. Limited resources and staff mean that most curation groups can't keep up with all the relevant literature.

How information is presented in the literature greatly affects how fast biocurators can identify and curate it. Papers still often report newly cloned genes without providing GenBank IDs or the species from which the genes were cloned. The entities discussed in a paper, including species, genes, proteins, genotypes and phenotypes must be unambiguously identified during curation. For example, using the HUGO Gene Nomenclature Committee resource (www.genenames.org), we find that the human gene *CDKN2A* has ten literature-based synonyms. One of those, *p14*, is also a synonym for five other genes: *CDK2AP2*, *CTNNBL1*, *RPP14*, *S100A9* and *SUB1*. To confirm the identity of the gene described, curators make inferences from synonyms, reported sequences, biological context and bibliographic citations. This time-consuming and error-prone step could be eliminated by compliance with data-reporting standards[4-9].

Most recent efforts in this direction have been developed by the communities that produce large-scale genomics data. The vast majority of the peer-reviewed literature does not yet have a reporting-structure standard. As publication has become a mainly digital endeavour, however, publications and biological databases are becoming increasingly similar. Properly cross-referenced and indexed, each could serve as an access point to the other[10]. Such collaboration between databases and journals would improve researchers' access to data and make their work more visible.

"To date, not much of the research community is rolling up its sleeves to annotate."

We recommend that all journals and reviewers require that a distinct section of the Methods (or a supplemental document) of all published articles includes approved gene symbols (which are inherently unstable) and model-organism database IDs (which do not change) for genes discussed; nucleotide or protein accession numbers (GenBank or UniProt ID) for isoforms of each gene or protein discussed; and descriptions of species, strains, cell types and genotypes used. Examples of sources for this information are listed in Table 1. This would accelerate literature curation, uphold information integrity, facilitate the proper linkage of data to other resources and support automated mining of data from papers. Another model is for authors to provide a 'structured digital abstract' — a machine-readable XML summary of pertinent facts in the article[11] — along with a manuscript. This approach is in an experimental phase at the journal *FEBS Letters*[12].

Box 1 | The role of biocurators

- To extract knowledge from published papers
- To connect information from different sources in a coherent and comprehensible way
- To inspect and correct automatically predicted gene structures and protein sequences to provide high-quality proteomes
- To develop and manage structured controlled vocabularies that are crucial for data relations and the logical retrieval of large data sets
- To integrate knowledge bases to represent complex systems such as metabolic pathways and protein-interaction networks.
- To correct inconsistencies and errors in data representation
- To help data users to render their research more productive in a timely manner
- To steer the design of web-based resources
- To interact with researchers to facilitate direct data submissions to databases

Journals should also mandate direct submission of data into appropriate databases as a part of publication. This has been implemented by the journal *Plant Physiology* and curators of The *Arabidopsis* Information Resource (TAIR) database[13]. On acceptance of a manuscript, the corresponding author must fill out a simple web-based form to provide appropriate genetic and molecular information about the *Arabidopsis* genes in the publication. The information is sent to TAIR for integration by biocurators, who work with the authors to ensure that the data reported are of high quality and accurate.

As this infrastructure develops, we would like to see authors routinely tagging all aspects of the data in their publication semantically using universally agreed tag standards. Examples of such tags include the National Center for Biotechnology Information (NCBI) Taxon IDs, the Gene Ontology (GO) IDs and Enzyme Commission (EC) numbers. This information should be embedded in the electronic versions of publications or provided in a supplemental file similar to the crystallographic information file (CIF) currently required for publication of a crystal structure. The CIF file is submitted to the Protein Data Bank (www.pdb.org), which offers software to assist in preparation and validation of such crystallographic data[14]. An analogous system to help authors identify, tag and validate the crucial basic information in their research reports before publication would accelerate the automated linkage of literature to key records in existing databases and improve the accuracy of the published data.

In short, authors and publishers must use the existing publication infrastructure to facilitate literature curation much more to the benefit of all parties.

Community curation

Curation of large-scale genomics and post-genomics data enjoys no such luxury of 'an existing publication infrastructure' to leverage, although emerging standards of data reporting are promising[4-9]. Sooner or later, the research community will need to be involved in the annotation effort to scale up to the rate of data generation. This transition will require annotation tools, standardized methods, oversight by expert curators and a combination of social infrastructure, tool development, training and feedback. Biocurators are especially important for establishing such an infrastructure and training to maintain consistency and accuracy.

To date, not much of the research community is rolling up its sleeves to annotate. What will be the tipping point? The main limitation in community annotation is the perceived lack of incentive. For example, several model-organism databases have requested that authors annotate the genes they publish. This has historically failed for one main reason: contributions by experts consist of information they already know, and do not increase the value of the resource to themselves. A mechanism tied to career or research advancement may be required before community curation can be established as a broadly accepted and productive scientific endeavour[15]. Incentives for researchers to curate data should include new information or insight for their research interests, improvement in academic reputation or impact, career advancement and better funding chances. Academic departments and funding agencies should consider community annotation as a productive contribution to the scientific research corpus and a natural extension of the publication process.

For example, in the *Daphnia* Genomics Consortium (http://daphnia.cgb.indiana.edu) collaboration wiki, a community of more than 300 contributors took ownership of annotation of the genome while it was being sequenced at the Joint Genome Institute in Walnut Creek, California, and shared publication authorship as a consortium. Similarly, the International *Glossina* Genomics Initiative (http://iggi.sanbi.ac.za) hosted an annotation jamboree for field workers, population geneticists and molecular biologists to annotate tsetse fly molecular data as the sequence information became available. This consortium-based publication mechanism is analogous to that used by other large-scale scientific projects such as the Sloan Digital Sky Survey (www.sdss.org). This is a viable course for communities that lack funding for dedicated curators, and offers a reward structure through consortium publication for participation and subsequent satellite papers.

The recently launched WikiProfessional Life Sciences (www.wikiprofessional.org) project links community curation with research and reputation gains. WikiProfessional indexed more than one million authors from PubMed and comparable numbers of biological concepts from authoritative databases and generated a simple way for researchers to update the information[16]. Because new potential 'facts' are mined from the network of associated concepts, the more accurate and comprehensive a particular concept is, the more chance it will have of being associated with other relevant ones, which in turn will lead to more potential new facts. All the updates researchers make are immediately publicly visible under their own name. Similarly, the Gene Wiki project generated thousands of wiki stubs in Wikipedia for human genes in an attempt to make it easier for the community to update the gene pages[17]. Although these wiki-based approaches provide an infrastructure for contributors to be recognized, there is not yet a standard practice for these contributions to be cited like a publication. It is imperative that the researchers, journal publishers and database curators start building a standard mechanism for citing annotation data sets.

Allowing anyone with a web browser, including the general public, to annotate entries would increase the number of potential annotators substantially, as pioneered in several astronomy projects. At Galaxy Zoo (www.galaxyzoo.org), 80,000 astronomers and members of the public manually classified the morphology of one million galaxies in less than three weeks. An analogous system to allow the public to contribute to biological annotation could be just as powerful if presented properly. For example, one could show a user an image of an *in situ* hybridization experiment and ask them to grade it as 'not expressed', 'restricted expression' or 'ubiquitous expression'. Even such basic information, if available for many thousands of genes, would be useful as first-pass annotation.

In sum, researchers (and even the general public) can be mobilized to provide the substantial resources needed to address the immense volume of data, if participation is appropriately rewarded. In the next five years, curators, funding agencies and academic institutions alike must find ways to consider substantial contributions to community curation efforts, much like a peer-reviewed publication, when it comes to issues of promotion, salary, hiring and funding.

Table 1 | Examples of knowledge-sharing databases

Species	Database	URL
Model organism databases		
Aedes aegypti	VectorBase	www.vectorbase.org
Anopheles gambiae	VectorBase	www.vectorbase.org
Arabidopsis thaliana	The *Arabidopsis* Information Resource	www.arabidopsis.org
Caenorhabditis elegans	WormBase	www.wormbase.org
Candida albicans	*Candida* Genome Database	www.candidagenome.org
Culex pipiens	VectorBase	www.vectorbase.org
Danio rerio	Zebrafish Information Network	http://zfin.org
Dictyostelium discoideum	dictyBase	http://dictybase.org
Drosophila sp.	FlyBase	http://flybase.org
Glycine max	SoyBase	www.soybase.org
Homo sapiens	HUGO Gene Nomenclature Committee	www.genenames.org
Hordeum vulgare	Barley Genetic Stocks Database	http://ace.untamo.net/bgs
Ixodes scapularis	VectorBase	www.vectorbase.org
Leishmania sp.	GeneDB	www.genedb.org
Mus musculus	Mouse Genome Informatics	www.informatics.jax.org
Oryza sp.	Gramene	http://gramene.org
Paramecium tetraurelia	ParameciumDB	http://paramecium.cgm.cnrs-gif.fr
Pediculus humanus	VectorBase	www.vectorbase.org
Rattus norvegicus	Rat Genome Database	http://rgd.mcw.edu
Saccharomyces cerevisiae	Saccharomyces Genome Database	www.yeastgenome.org
Schizosaccharomyces pombe	GeneDB	www.genedb.org
Solanaceae sp.	Sol Genomics Network	http://sgn.cornell.edu
Strongylocentrotus purpuratus	SpBase	http://sugp.caltech.edu/SpBase
Triticum sp.	GrainGenes	http://wheat.pw.usda.gov
Trypanosoma sp.	GeneDB	www.genedb.org
Xenopus laevis	Xenbase	www.xenbase.org
Xenopus tropicalis	Xenbase	www.xenbase.org
Zea mays	Maize Genetics and Genomics Database	www.maizegdb.org
Nucleotide, protein and structure databases		
All Species	GenBank	www.ncbi.nlm.nih.gov/Genbank
All Species	UniProt	www.pir.uniprot.org
All Species	Protein Data Bank	http://rcsb.org/pdb/home/home.do
Taxonomy		
All Species	NCBI Entrez Taxonomy	www.ncbi.nlm.nih.gov/sites/ entrez?db=taxonomy

Biological databases contain unique identifiers for the unambiguous identification of biological entities (scuh as genes, proteins, species and chemicals). These identifiers do not change as common biological names do. Authors should consult these databases for stable identifiers to cite in their publications.

Career path

How can biocuration mature faster as a career? Biocurators currently streamline submission to databases, automate curation, standardize data and facilitate contributions to annotation by research communities interested in the annotation process. To handle the increasing volume and types of data, journal publishers and researchers who generate data will need to be involved in the curation process and the roles of biocurators will expand to include editing and teaching. As biology moves towards more precise, quantitative science, biologists also need to adapt to thinking more quantitatively, systematically and objectively about their data; biocuration will need to become an inherent part of research and education in biology.

Biocuration requires a blend of skills and experience, including advanced scientific research and competence in database management systems, multiple operating systems and scripting languages. This type of background has typically been garnered through a combination of self-teaching and on-the-job experience, which can be narrow and spotty. Happily, formal education is becoming available. For example, the Graduate School of Library and Information Science at the University of Illinois at Urbana-Champaign offers a biological information specialist master's degree and a specialization in data curation[18]. Experienced biocurators must lead the way in establishing more and better formal training programmes. In the next 5–10 years, biology curricula should include courses in biocuration as this becomes an increasingly common activity for all biological researchers. And interdisciplinary programmes that include courses in biology, computer science and information science will be vital.

Attracting highly qualified individuals into this field has been challenging. The whole community must promote scientific curation as a professional career option. Funding agencies must assess the impact of curated data and support the development of innovative curation methods. To improve the profession, curators need a forum to share their experiences and publish their works. Oxford University Press plans to begin publishing a new journal in 2009 called *Database: The Journal of Biological Databases and Curation*. This may provide one such venue for publication of noteworthy advances in biocuration (www.database.oxfordjournals.org). Meanwhile, a committee of 20 biocurators and researchers is forming an International Society for Biocuration (www.biocurator.org/BiocuratorSociety.html) to make the discipline more visible and to promote it as an attractive career path. The official launch of the society is planned for the third International Biocuration Meeting next April in Berlin (http://projects.eml.org/Meeting2009).

Biology today needs more robust, expressive, computable, quantitative, accurate and precise ways to handle data. It is time to recognize that biocuration and biocurators are central to the future of the field. ■

1. Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J. & Wheeler, D. L. *Nucl. Acid. Res.* **36,** D25-D30 (2008).
2. Wheeler, D. L. *et al. Nucl. Acid. Res.* **36,** D13-D21 (2008).
3. Salimi, N. & Vita, R. *PLoS Comput. Biol.* **2,** e125 (2006).
4. Brazma, A. *et al. Nature Genet.* **29,** 365-371 (2001).
5. Deutsch, E. W. *et al. Nature Biotechnol.* **26,** 305-312 (2008).
6. Field, D. *et al. Nature Biotechnol.* **26,** 541-547 (2008).
7. Jenkins, H. *et al. Nature Biotechnol.* **22,** 1601-1606 (2004).
8. Orchard, S. *et al. Nature Biotechnol.* **25,** 894-898 (2007).
9. Taylor, C. F. *et al. Nature Biotechnol.* **25,** 887-893 (2007).
10. Bourne, P. *PLoS Comput. Biol.* **1,** 179-181 (2005).
11. Seringhaus, M. R. & Gerstein, M. B. *BMC Bioinformatics* **8,** 17 (2007).
12. Seringhaus, M. & Gerstein, M. *FEBS Lett.* **582,** 1170 (2008).
13. Ort, D. R. & Grennan, A. K. *Plant Physiol.* **146,** 1022-1023 (2008).
14. Burkhardt, K., Schneider, B. & Ory, J. *PLoS Comput. Biol.* **2,** e99 (2006).
15. Rhee, S. Y. *Plant Physiol.* **134,** 543-547 (2004).
16. Mons, B. *et al. Genome Biol.* **9,** R89 (2008).
17. Huss, J. W. *et al. PLoS Biol.* **6,** e175 (2008).
18. Palmer, C. L., Heidorn, P. B., Wright, D. & Cragin, M. H. *Int. J. Dig. Curation* **2,** 31-40 (2007).

Author information Correspondence and requests for materials should be addressed to D.H. (e-mail: dhowe@cs.uoregon.edu) and S.Y.R. (e-mail: rhee@acoma.stanford.edu).

See Editorial, page 1.

Authorship

Doug Howe[1], Maria Costanzo[2], Petra Fey[3], Takashi Gojobori[4], Linda Hannick[5], Winston Hide[6,7], David P. Hill[8], Renate Kania[9], Mary Schaeffer[10,11], Susan St Pierre[12], Simon Twigger[13], Owen White[14] and Seung Yon Rhee[15]

[1]The Zebrafish Information Network, 5291 University of Oregon, Eugene, Oregon 97403-5291, USA. [2]*Saccharomyces* and *Candida* Genome Databases, Stanford University, Stanford, California 94305-5120, USA. [3]dictyBase, Northwestern University Biomedical Informatics Center, 750 N. Lake Shore Drive, 11-175, Chicago, Illinois 60611, USA. [4]Centre for Information Biology and DNA Data Bank of Japan, National Institute of Genetics, Research Organization of Information and Systems, Yata, Mishima 411-8540, Japan. [5]J. Craig Venter Institute, Applied Bioinformatics, Rockville, Maryland 20850, USA. [6]South African National Bioinformatics Institute, University of the Western Cape, Private Bag X17, Bellville 7535, South Africa. [7]Department of Biostatistics, Harvard School of Public Health, 655 Huntington Avenue, Boston, Massachusetts 02115, USA. [8]Mouse Genome Informatics, The Jackson Laboratory, Bar Harbor, Maine 04609, USA. [9]Scientific Databases and Visualization, EML Research GmbH, Villa Bosch, Schloss-Wolfsbrunnenweg 33, D-69118 Heidelberg, Germany. [10]Division of Plant Sciences, University of Missouri, Columbia, Missouri, USA. [11]Plant Genetics Research Unit, Agricultural Research Service, United States Department of Agriculture, Columbia, Missouri 65211-7020, USA. [12]FlyBase, Harvard University, Cambridge, Massachusetts 02138, USA. [13]Rat Genome Database, Bioinformatics Research Center, Medical College of Wisconsin, 8701 Watertown Plank Rd, Milwaukee, Wisconsin 53226, USA. [14]Department of Epidemiology and Preventative Medicine, Institute for Genome Sciences, University of Maryland School of Medicine, Baltimore, Maryland 21201, USA. [15]The *Arabidopsis* Information Resource, Carnegie Institution for Science, Department of Plant Biology, 260 Panama Street, Stanford, California 94305, USA.