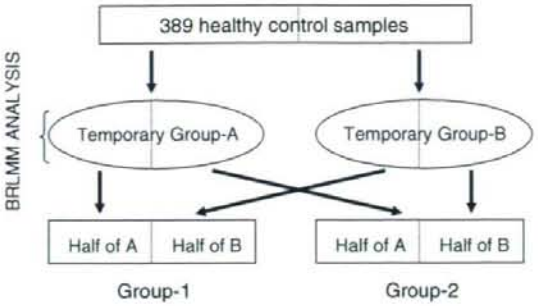

Fig. 1 The principle for making groups. The 389 healthy control samples were divided into two temporary groups (groups A and B). Each group was separately analyzed by the Bayesian Robust Linear Model with Mahalanobis (*BRLMM*) of GTYPE 4.1 software. Then, to remove the bias associated with the BRLMM analysis, the samples of groups A and B were equally subdivided into two new groups (groups 1 and 2); group 1 consisted half of group A' s and half of group B' s samples, and group 2 consisted of the remaining samples. Group 1 was compared with group 2 using the chi-square test for the difference between allele frequencies of each single nucleotide polymorphism. Next, for reshuffling analysis, 100 new combination sets were prepared using the same 389 healthy control samples. Each set was formed from two groups separately analyzed by the BRLMM algorithm, such as the set of groups A and B

The following four parameters were assessed in this surveillance:

1. SNP call rate: to remove an SNP for which genotyping was consistently problematic, SNPs with call rates ≥85, ≥90 and ≥95% in both groups were prepared.
2. Confidence score with the BRLMM: Confidence scores of 0.3, 0.4, 0.5, and 0.6 were applied. Each overall call rate was 93.3% (*Sty*I: 92.2%; *Nsp*I: 94.4%) for a confidence score of 0.3; 95.3% (*Sty*I: 94.4%; *Nsp*I: 96.1%) for 0.4; 96.6% (*Sty*I: 96.0%; *Nsp*I: 97.2%) for 0.5; and 97.6% (*Sty*I: 97.1%; *Nsp*I: 98.0%) for 0.6 before data cleaning. For a confidence score of 0.5, the distribution of call rates (per SNP and sample) is shown in Supplementary Figs. 1 and 2.
3. HWE: Deviations from HWE can occur by chance. However, they can also be due to genotyping errors, inbreeding, and population stratification. Testing for HWE can be helpful to check data (Balding 2006). We removed SNPs for which we observed genotype frequencies that significantly deviated from HWE (HWE $P < 0.001$ and $P < 0.01$). Evaluation of HWE was carried out using the chi-square test. In general, case-control studies, SNPs that deviate from HWE in a control group are removed. We considered one group in a set as controls. The possibility that a deviation from HWE is due to a deletion or duplication polymorphism, which could be important for disease susceptibility, should now be considered (Bailey and Eichler 2006; Conrad et al. 2006; Nielsen et al. 1998).
4. MAF: We removed SNPs in which the minor allele frequency was <1 or <5% in all samples.

SNPs with low MAF would produce inappropriately small P values in the chi-square test. However, in this study, the chi-square test could be used for the quasi-case-control studies and evaluation of HWE because SNPs with low MAF were removed for data cleaning.

First, the number of significant SNPs in a quasi-case-control study ($P < 0.0001$ and $P < 0.001$) was counted for each result after data cleaning and compared against the expected number calculated from each P value and the total number of SNPs. Next, the log quantile–quantile (QQ) P value (Balding 2006; Weir et al. 2004) was adopted for interpreting each result. The negative logarithm of P values was plotted against $-\log(i/(L+1))$, where L is the number of SNPs. Deviation from the expected number and the $y = x$ line corresponds to loci that deviate from the null hypothesis. The close adherence of P values to the expected number and the expected line, which corresponds to the null hypothesis over most of the range, is encouraging, as it implies that there are few systematic sources of spurious association because mutual healthy control groups were compared.

Results

Criteria for data cleaning

We compared two healthy groups (groups 1 and 2, Fig. 1) using the chi-square test as a quasi-case-control study. Then, we assessed the deviation of the results from the null hypothesis after each data cleaning (see "Materials and methods" for details). A small or no deviation implies that there are few systematic sources of spurious associations.

1. SNP call rate

As shown in Table 1, the number of SNPs with a call rate ≥95% was close to the expected number calculated from each P value and the total number of SNPs. For a confidence score of 0.5, the ratio of observed and expected number of SNPs with a call rate ≥95% was 1.05–1.50, whereas the ratio with call rates of ≥90% and ≥85% was 1.35–2.47 and 1.52–2.96, respectively (Table 1). The observed number of significant SNPs with call rates ≥90% and ≥85% was more inflated. For other confidence scores (0.3, 0.4, and 0.6), the ratio of observed and expected number of SNPs with a call rate ≥95% was 1.05–2.04, whereas the ratio with call rates ≥90% and ≥85% was 1.08–3.39 and 1.21–3.73, respectively (Supplementary Table 1). SNPs with call rates ≥90% and ≥85% also caused inflations. However, for a confidence score of 0.5, the ratio of observed and expected number of

Table 1 Comparison of group 1 and group 2 for a confidence score of 0.5

Confidence	MAF (%)	HWE	Call rate (%)	Total SNPs	Number of SNPs (obs. P values)		Number of SNPs (exp. P values)		Ratio of obs. number/ exp. number	
					$P < 0.0001$	$P < 0.001$	$P < 0.0001$	$P < 0.001$	$P < 0.0001$	$P < 0.001$
0.5	5	0.01	95	229,907	34	254	23	230	1.48	1.10
			90	305,001	71	415	31	305	2.33	1.36
			85	328,894	91	501	33	329	2.77	1.52
		0.001	95	233,023	35	258	23	233	1.50	1.11
			90	310,687	73	428	31	311	2.35	1.38
			85	336,275	93	520	34	336	2.77	1.55
	1	0.01	95	259,186	38	272	26	259	1.47	1.05
			90	349,900	86	471	35	350	2.46	1.35
			85	378,496	112	592	38	378	2.96	1.56
		0.001	95	262,854	39	277	26	263	1.48	1.05
			90	356,361	88	485	36	356	2.47	1.36
			85	386,702	114	612	39	387	2.95	1.58

Data cleaning was conducted using the following four parameters: single nucleotide polymorphism (*SNP*) call rate, confidence score in the Bayesian Robust Linear Model with Mahalanobis (BRLMM) genotype-calling algorithm, Hardy–Weinberg equilibrium (*HWE*), and minor allele frequency (*MAF*). When group 1 was compared with group 2 using the chi-square test for difference in allele frequencies of each SNP, the number of significant SNPs (observed P values) was counted for each result. The number of significant SNPs (expected P values) was logically calculated from the total number of SNPs

obs. observed, *exp.* expected

SNPs with a call rate ≥95% was up to 1.50. Nine significant SNPs with $P < 0.0001$ in the same region (within about 100 kb) were responsible for this random deviation. These SNPs, showing uniformly low P values, existed in a linkage disequilibrium block with a solid spine of $D' > 0.8$. It was shown that SNPs with a lower call rate are likely to contain genotyping errors, and SNP call rate is important for data cleaning.

Next, thresholds of SNP call rate ranging from 92% to 97% at increments of 1% were set, and the ratio of observed and expected number of SNPs in each threshold was calculated (Fig. 2). This ratio was saturated at SNPs with a call rate ≥95%, implying that they had few spurious associations and were considered to be the key threshold. We focused on SNPs with a call rate ≥95% in the following analyses using other parameters.

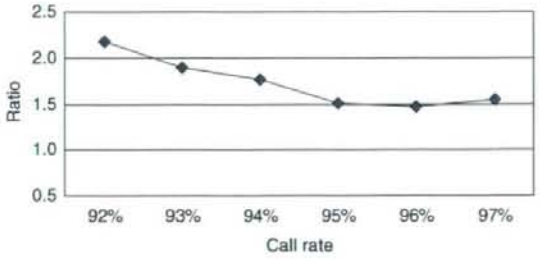

Fig. 2 Ratio of observed and expected number of single nucleotide polymorphisms (SNPs) with a call rate at 1% intervals between 92% and 97%. Ratio of observed and expected number of SNPs was calculated in each call rate, a confidence score of 0.5, Hardy–Weinberg equilibrium (HWE) $P \geq 0.001$ and minor allele frequency (MAF) ≥5%

2. Confidence score in BRLMM

In the BRLMM analysis, the ratio of observed and expected number of SNPs was 1.05–1.56 for confidence scores of 0.3, 0.4, and 0.5. However, this ranged from 1.13 to 2.04 for a confidence score of 0.6 (Table 2). Consequently, a confidence score of 0.6 would cause spurious associations. The total number of SNPs for a confidence score of 0.5 ranged from 229,907 to 262,854, whereas for confidence scores of 0.3 and 0.4, this was 146,469–217,818 (Table 2). On equivalent adequacy, the total number of SNPs for each confidence score (0.3, 0.4, and 0.5) should be taken into consideration. Then, the overall call rate for a confidence score was 93.3% for 0.3, 95.3% for 0.4, and 96.6% for 0.5. Therefore, it is suggested that a confidence score of 0.5 should be selected.

3. HWE

SNPs with an HWE of $P \geq 0.001$ or $P \geq 0.01$ did not result in unexpected inflations. For a confidence score of 0.5, the ratio of observed and expected number of SNPs with an HWE of $P \geq 0.001$ was 1.05–1.50, whereas for HWE with $P \geq 0.01$, it was 1.05–1.48 (Table 1). We can conclude that, as hundreds of thousands of SNPs were analyzed in this study, deviation from HWE might be caused by chance for SNPs with an HWE of $P \geq 0.001$.

Table 2 Results of confidence scores in Bayesian Robust Linear Model with Mahalanobis (BRLMM) for single nucleotide polymorphisms (*SNPs*) with a call rate ≥95%

Call rate (%)	MAF (%)	HWE	Confidence	Total SNPs	Number of SNPs (obs. *P* values)		Number of SNPs (exp. *P* values)		Ratio of obs. number/ exp. number	
					$P < 0.0001$	$P < 0.001$	$P < 0.0001$	$P < 0.001$	$P < 0.0001$	$P < 0.001$
95	5	0.01	0.3	146,469	21	161	15	146	1.43	1.10
			0.4	191,916	29	206	19	192	1.51	1.07
			0.5	229,907	34	254	23	230	1.48	1.10
			0.6	265,186	52	307	27	265	1.96	1.16
		0.001	0.3	148,065	21	165	15	148	1.42	1.11
			0.4	194,245	30	211	19	194	1.54	1.09
			0.5	233,023	35	258	23	233	1.50	1.11
			0.6	269,310	55	317	27	269	2.04	1.18
	1	0.01	0.3	163,484	25	178	16	163	1.53	1.09
			0.4	215,019	33	225	22	215	1.53	1.05
			0.5	259,186	38	272	26	259	1.47	1.05
			0.6	302,494	57	343	30	302	1.88	1.13
		0.001	0.3	165,407	25	182	17	165	1.51	1.10
			0.4	217,818	34	231	22	218	1.56	1.06
			0.5	262,854	39	277	26	263	1.48	1.05
			0.6	307,282	60	355	31	307	1.95	1.16

The results in comparison of group 1 and group 2 are arranged according to confidence scores in BRLMM

MAF minor allele frequency, *HWE* Hardy–Weinberg equilibrium, *obs.* observed, *exp.* expected

4. MAF

SNPs with MAF ≥5% or ≥1% did not exhibit unexpected inflations. For a confidence score of 0.5, the ratio of observed and expected number of SNPs for MAF ≥5% was 1.10–1.50, whereas for MAF ≥1%, this was 1.05–1.48 (Table 1). The decision regarding which SNPs with MAF ≥5% or ≥1% should be adopted depends on the sample size of each association study.

The log QQ *P* value plot was described after data cleaning using the criteria identified by the above analyses (SNP call rate ≥95%, confidence score 0.5, HWE $P \geq 0.001$, and MAF ≥5% or ≥1%; Fig. 3a, b). Plots of *P* values were close to the expected line ($y = x$). However, approximately 12 SNPs deviated from the expected line of low *P* values. The nine significant SNPs with $P < 0.0001$ in a linkage disequilibrium block with a solid spine of $D' > 0.8$ were also responsible for this random deviation.

Taken together, these results indicated that data cleaning could be appropriately conducted in our Japanese samples using SNP call rate ≥95%, a confidence score of 0.5, HWE $P \geq 0.001$, and MAF ≥5% or ≥1%.

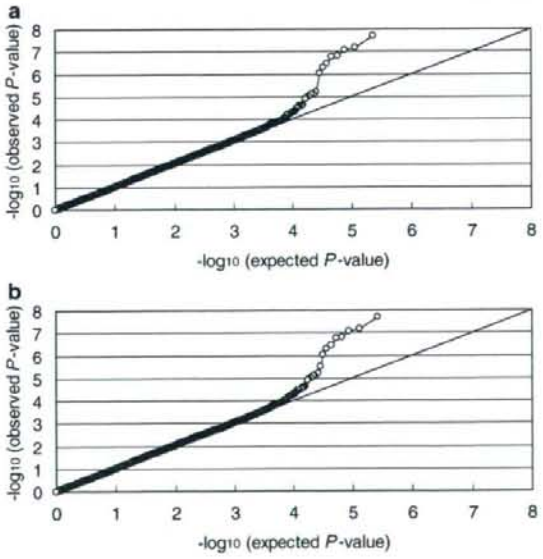

Fig. 3 Log quantile–quantile (QQ) *P* value plot for the results after data cleaning. **a** and **b** were analyzed using groups 1 and 2, respectively. Cleaning criteria were a single nucleotide polymorphism (SNP) call rate ≥95%, confidence score 0.5, Hardy–Weinberg equilibrium (HWE) $P \geq 0.001$, and minor allele frequency (MAF) ≥5%; **b** SNP call rate ≥95%, confidence score 0.5, HWE $P \geq 0.001$, and MAF ≥1%

Reshuffling analysis

Next, to confirm that the identified appropriate criteria were reproducible, 100 new combination sets were

Table 3 Reshuffling analysis

Confidence	MAF (%)	HWE	Call rate (%)	Total SNPs (SD)	Mean number of SNPs in obs. *P* values (SD)		Mean ratio of obs. number/ exp. number (SD)	
					$P < 0.0001$	$P < 0.001$	$P < 0.0001$	$P < 0.001$
0.5	5	0.001	95	235,227 (9,904)	22.38 (7.95)	230.64 (28.47)	0.95 (0.33)	0.98 (0.11)
			90	308,134 (13,769)	60.75 (12.67)	372.65 (39.63)	1.97 (0.40)	1.21 (0.12)
			85	334,375 (15,499)	375.30 (41.86)	916.84 (79.48)	11.20 (1.09)	2.74 (0.18)
	1		95	265,192 (11,265)	24.23 (8.10)	249.61 (29.46)	0.91 (0.30)	0.94 (0.10)
			90	352,067 (15,910)	127.45 (22.39)	628.48 (58.25)	3.62 (0.60)	1.79 (0.14)
			85	383,413 (17,935)	697.85 (72.62)	1814.64 (141.78)	18.18 (1.61)	4.73 (0.27)

To confirm whether the two criteria identified in the comparison of groups 1 and 2 are reproducible, 100 new combination sets were prepared, and each set was compared using the chi-square test for difference in allele frequencies of each SNP

MAF minor allele frequency, *HWE* Hardy–Weinberg equilibrium, *SNPs* single nucleotide polymorphisms, *obs.* observed, *exp.* expected, *SD* standard deviation

prepared. In addition, to consider that the bias associated with BRLMM analysis cannot be removed when other investigators make use of our frequency data, each set was formed from two groups separately analyzed by the BRLMM algorithm, such as the set of groups A and B (Fig. 1). Then, two groups in each set were compared as a quasi-case-control study. The mean ratio of observed and expected number of SNPs with a call rate of ≥95% was 0.91–0.98, whereas the mean ratio with call rates of ≥90% and ≥85% was 1.21–3.62 and 2.74–18.18, respectively (Table 3). Unexpected deviations were not observed in comparisons of these 100 sets when data cleaning was conducted using SNP call rate ≥95%, a confidence score of 0.5, HWE $P \geq 0.001$, and MAF ≥5 or ≥1%.

Discussion

GWAS have considerable potential for detecting susceptibility and/or resistant genes for various complex diseases. However, the considerable amount of data may occasionally create difficulties for investigators. One hundred thousand to 1 million SNPs are targeted for GWAS. It is important to note that the results obtained from genotyping so many SNPs are not always precise and that inaccurate data will unfavorably affect GWAS analyses. In this study, it was shown that spurious associations could be excluded using the criteria that we identified. However, it may not be possible to apply these criteria to data from the GeneChip Human Mapping 500K Array Set or other arrays used by other investigators, because the criteria may be affected by differences in overall call rates between studies. In such cases, the same analyses using four parameters (SNP call rate, confidence score in BRLMM, HWE, and MAF) would facilitate identification of appropriate criteria for each study. In the GeneChip Human Mapping 500K Array Set, these criteria can be applied to data that is of the same quality as our data, with an index for overall call rate (for a confidence score of 0.5, overall call rate was 96.6% before data cleaning in this study).

The tradeoff exists between overall call rate and accuracy. If a high accuracy is of greater importance than a high overall call rate, a higher-quality threshold for data cleaning should be selected. Alternatively, if a high overall call rate is of greater importance than a high accuracy, a lower quality threshold for data cleaning should be selected. In this study, we assessed each data-cleaning method by the deviation from the null hypothesis to obtain accurate data. Additionally, on equivalent adequacy, a higher overall call rate was considered. Therefore, the appropriate data cleaning methods we identified satisfy both overall call rate and accuracy.

In GWAS with the GeneChip Human Mapping 500K Array Set, the genotyping results of case–control samples were decided by the BRLMM algorithm. In the reshuffling analysis (100 sets), two groups in each set were separately analyzed by the BRLMM algorithm. The maximum mean ratio of observed and expected number in the reshuffling analysis was up to 18.18 (Table 2), whereas that in groups 1 and 2, equally subdivided from groups A and B, was 2.95 (Table 1). This result suggested that separate BRLMM analyses result in a bias of genotyping results. However, as shown in Table 3, unexpected deviations were not observed in the reshuffling analysis after the appropriate data cleaning (cleaning criteria: SNP call rate ≥95%, a confidence score of 0.5, HWE $P \geq 0.001$, and MAF ≥5 or ≥1%). We assume that the appropriate data cleaning could remove SNPs affected by the bias associated with BRLMM analysis.

Some studies required that samples pass a threshold of overall call rate (Buch et al. 2007; Rioux et al. 2007; The Wellcome Trust Case Control Consortium 2007), whereas we evaluated data cleaning methods for SNP selection. As

a result, the data cleaning methods we identified could exclude spurious associations, suggesting that it would not be necessary to require sample selection when the appropriate data cleaning for SNP selection is conducted.

The four parameters (SNP call rate, confidence score in BRLMM, HWE, and MAF) are mutually correlated. In these four parameters, SNP call rate was considered to be a key parameter because SNPs with a lower call rate, particularly, caused irrelevant inflations (Tables 1, 3; Fig. 2). In GWAS, it would be better to change the threshold of SNP call rate when the ratio of observed and expected number of SNPs with low *P* values is inflated. If the ratio is close to one according to the threshold of SNP call rate, it is suspected that SNPs with low *P* values include spurious associations due to errors as well as true associations with the target disease.

In the reshuffling analysis, we calculated deviations from expected *P* values of each chip (*Sty*I and *Nsp*I). Before data cleaning, the maximum mean ratio of observed and expected number of SNPs on *Sty*I chip was 47.08, whereas that of SNPs on *Nsp*I chip was 30.10 (Supplementary Table 2). Thus, the mean ratio for *Sty*I chip was more inflated. There might be a difference in the accuracy between the two chips before data cleaning. However, after the appropriate data cleaning, the maximum mean ratio of observed and expected number of SNPs on *Sty*I chip was 0.98, whereas that of SNPs on *Nsp*I chip was 0.93 (Supplementary Table 2). Accordingly, unexpected deviations were not observed in either chip after the appropriate data cleaning.

Even though data cleaning using appropriate criteria was conducted, sample size should be carefully considered when SNPs with an MAF $\geq$1% are used. The frequency of an SNP has a marked influence on statistical power. To identify SNPs with low MAF associated with complex disease, sample size must be large (Ohashi and Tokunaga 2001, 2002; Ohashi et al. 2001; Risch 2000). In instances where SNPs with low MAF are used, at the very least, desirable sample sizes should be calculated based on the frequencies of the targeted SNPs, and sufficient samples should be collected when planning the GWAS.

Generally, 300,000 SNPs might be required to capture most of the common genetic variation in a population (Balding 2006). The GeneChip Human Mapping 500K Array Set provides genotyping data for approximately 500,000 SNPs. However, only about 250,000 SNPs were extracted after data cleaning in our Japanese samples (Table 1). This reduction is caused by the presence of numerous SNPs with low MAF on the GeneChip Human Mapping 500K Array Set in Japanese. There are approximately 150,000 SNPs with an MAF <5% and 100,000 SNPs with an MAF <1% on this array set. Ideally, SNPs with low MAF are not likely to be suitable for GWAS (Ohashi and Tokunaga 2001, 2002; Ohashi et al. 2001; Risch 2000), and it is hoped that new arrays for different ethnic groups will be developed.

Acknowledgments We are deeply grateful to the people participating in this study. We thank JAMSAC (Japan Multiple System Atrophy Research Consortium) for a part of the samples. This study is supported by a grant-in-aid for Scientific Research on Priority Areas "Comprehensive Genomics" and "Applied Genomics" from the Ministry of Education, Culture, Sports, Science and Technology of Japan and a grant-in-aid for JSPS fellows.

References

Arking DE, Pfeufer A, Post W, Kao WH, Newton-Cheh C, Ikeda M, West K, Kashuk C, Akyol M, Perz S, Jalilzadeh S, Illig T, Gieger C, Guo CY, Larson MG, Wichmann HE, Marban E, O'Donnell CJ, Hirschhorn JN, Kaab S, Spooner PM, Meitinger T, Chakravarti A (2006) A common genetic variant in the NOS1 regulator NOS1AP modulates cardiac repolarization. Nat Genet 38:644–651

Bailey JA, Eichler EE (2006) Primate segmental duplications: crucibles of evolution, diversity and disease. Nat Rev Genet 7:552–564

Balding DJ (2006) A tutorial on statistical methods for population association studies. Nat Rev Genet 7:781–791

Buch S, Schafmayer C, Volzke H, Becker C, Franke A, von Eller-Eberstein H, Kluck C, Bassmann I, Brosch M, Lammert F, Miquel JF, Nervi F, Wittig M, Rosskopf D, Timm B, Holl C, Seeger M, ElSharawy A, Lu T, Egberts J, Fandrich F, Folsch UR, Krawczak M, Schreiber S, Nurnberg P, Tepel J, Hampe J (2007) A genome-wide association scan identifies the hepatic cholesterol transporter ABCG8 as a susceptibility factor for human gallstone disease. Nat Genet 39:995–999

Clayton DG, Walker NM, Smyth DJ, Pask R, Cooper JD, Maier LM, Smink LJ, Lam AC, Ovington NR, Stevens HE, Nutland S, Howson JM, Faham M, Moorhead M, Jones HB, Falkowski M, Hardenbol P, Willis TD, Todd JA (2005) Population structure, differential bias and genomic control in a large-scale, case-control association study. Nat Genet 37:1243–1246

Conrad DF, Andrews TD, Carter NP, Hurles ME, Pritchard JK (2006) A high-resolution survey of deletion polymorphism in the human genome. Nat Genet 38:75–81

Dewan A, Liu M, Hartman S, Zhang SS, Liu DT, Zhao C, Tam PO, Chan WM, Lam DS, Snyder M, Barnstable C, Pang CP, Hoh J (2006) HTRA1 promoter polymorphism in wet age-related macular degeneration. Science 314:989–992

Klein RJ, Zeiss C, Chew EY, Tsai JY, Sackler RS, Haynes C, Henning AK, SanGiovanni JP, Mane SM, Mayne ST, Bracken MB, Ferris FL, Ott J, Barnstable C, Hoh J (2005) Complement factor H polymorphism in age-related macular degeneration. Science 308:385–389

Kubo M, Hata J, Ninomiya T, Matsuda K, Yonemoto K, Nakano T, Matsushita T, Yamazaki K, Ohnishi Y, Saito S, Kitazono T, Ibayashi S, Sueishi K, Iida M, Nakamura Y, Kiyohara Y (2007) A nonsynonymous SNP in PRKCH (protein kinase C eta) increases the risk of cerebral infarction. Nat Genet 39:212–217

Matsuzaki H, Loi H, Dong S, Tsai YY, Fang J, Law J, Di X, Liu WM, Yang G, Liu G, Huang J, Kennedy GC, Ryder TB, Marcus GA, Walsh PS, Shriver MD, Puck JM, Jones KW, Mei R (2004) Parallel genotyping of over 10, 000 SNPs using a one-primer assay on a high-density oligonucleotide array. Genome Res 14:414–425

Nielsen DM, Ehm MG, Weir BS (1998) Detecting marker-disease association by testing for Hardy–Weinberg disequilibrium at a marker locus. Am J Hum Genet 63:1531–1540

Ohashi J, Tokunaga K (2001) The power of genome-wide association studies of complex disease genes: statistical limitations of indirect approaches using SNP markers. J Hum Genet 46:478–482

Ohashi J, Tokunaga K (2002) The expected power of genome-wide linkage disequilibrium testing using single nucleotide polymorphism markers for detecting a low-frequency disease variant. Ann Hum Genet 66:297–306

Ohashi J, Yamamoto S, Tsuchiya N, Hatta Y, Komata T, Matsushita M, Tokunaga K (2001) Comparison of statistical power between 2 × 2 allele frequency and allele positivity tables in case-control studies of complex disease genes. Ann Hum Genet 65:197–206

Oliphant A, Barker DL, Stuelpnagel JR, Chee MS (2002) BeadArray technology: enabling an accurate, cost-effective approach to high-throughput genotyping. Biotechniques Suppl:56–58, 60–61

Rabbee N, Speed TP (2006) A genotype calling algorithm for affymetrix SNP arrays. Bioinformatics 22:7–12

Rioux JD, Xavier RJ, Taylor KD, Silverberg MS, Goyette P, Huett A, Green T, Kuballa P, Barmada MM, Datta LW, Shugart YY, Griffiths AM, Targan SR, Ippoliti AF, Bernard EJ, Mei L, Nicolae DL, Regueiro M, Schumm LP, Steinhart AH, Rotter JI, Duerr RH, Cho JH, Daly MJ, Brant SR (2007) Genome-wide association study identifies new susceptibility loci for Crohn disease and implicates autophagy in disease pathogenesis. Nat Genet 39:596–604

Risch NJ (2000) Searching for genetic determinants in the new millennium. Nature 405:847–856

Sladek R, Rocheleau G, Rung J, Dina C, Shen L, Serre D, Boutin P, Vincent D, Belisle A, Hadjadj S, Balkau B, Heude B, Charpentier G, Hudson TJ, Montpetit A, Pshezhetsky AV, Prentki M, Posner BI, Balding DJ, Meyre D, Polychronakos C, Froguel P (2007) A genome-wide association study identifies novel risk loci for type 2 diabetes. Nature 445:881–885

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

Weir BS, Hill WG, Cardon LR (2004) Allelic association patterns for a dense SNP map. Genet Epidemiol 27:442–450

Zondervan KT, Cardon LR (2004) The complex interplay among factors that influence allelic association. Nat Rev Genet 5:89–100

ゲノムワイドSNPタイピング技術の現状と将来

Current and future state of genome-wide SNP typing technology

西田奈央(写真)　徳永勝士
Nao Nishida and Katsushi Tokunaga
東京大学大学院医学系研究科人類遺伝学分野

○ヒトゲノム計画をはじめとする遺伝情報解析の成果として，データベースに蓄積された1,100万種類を超える単一塩基多型(SNP)のうち，数十万～百万種類のSNPを同時にタイピングすることのできる手法が近年になって実用化された．著者らは最新のプラットフォームを用いてSNPタイピングを効率的に行うためのシステムを構築し，いくつかの多因子疾患を対象としてゲノムワイド関連分析を行っている．本稿ではゲノムワイドSNPタイピング技術の原理と概要について解説した後，日本人試料を用いたタイピングの結果を，いくつかの注意点や精度管理法を交えて紹介する．最後に，今後の課題および将来の展望についても触れたい．

Key word　単一塩基多型(SNP)，SNPタイピング，ケースコントロール関連分析，疾患感受性遺伝子

SNPタイピング技術の進展に伴って，ヒトのさまざまな多因子疾患にかかわる遺伝子を探索する戦略として，ゲノムワイド関連研究(genome-wide association study：GWAS)が近年大きな注目を浴びている．2007年5月には90万種を超えるSNP解析用プローブおよびCNV(copy number variation)解析用の94万種類を超えるプローブを搭載したキットが市販された(Affymetrix® Genome-Wide Human SNP Array 6.0：以下，SNP Array 6.0)．著者らの教室に設置したヒトSNPタイピングセンターでは，SNP Array 6.0によるSNPタイピングを効率的に行うためのシステムを構築し，いくつかの多因子疾患についてゲノムワイド関連分析を実施している．

本稿ではSNP Array 6.0の技術面を主としてとりあげ，日本人健常人200名のSNPタイピングデータを一例として紹介する．

ゲノムワイド関連研究(GWAS)の動向

ヒト多因子疾患の疾患感受性遺伝子を探索する統計遺伝学的手法には，大別して連鎖分析(linkage analysis)と関連分析(association analysis)がある．連鎖分析は，患者家族を対象として文字どおり疾患遺伝子と多型マーカーの連鎖を検出する手法であり，ゲノム全域に分布する1万から数万種類のSNPを用いれば十分である．一方，関連分析の代表であるケースコントロール関連分析法は，非血縁の患者群と健常対照群を対象として疾患遺伝子と多型マーカーの連鎖不平衡(linkage disequilibrium)を検出する手法であり，これをゲノム全域にわたって適用するGWASでは，数十万種類以上のSNPあるいは数万種以上のマイクロサテライトマーカーを解析することが必要となる[1](統計遺伝学手法の原理ならびに集団遺伝学的基礎については，別に解説した[2])．なお，このGWASは日本の研究者によって先駆的に行われ，これまでにいくつかのヒト多因子疾患の感受性遺伝子を特定することに成功している[3-6]．

しかし，SNP解析技術の著しい進展によって一挙に状況が変化し，2007年末，欧米を中心として大規模なゲノムワイド関連研究の成果がつぎつぎに報告されている．例をあげると，WTCCC(The Wellcome Trust Case Control Consortium)は7種類のcommon diseases(双極性感情障害：BD，冠動脈疾患：CAD，Crohn病：CD，高血圧：HT，関節リウマチ：RA，1型糖尿病：T1D，2型糖尿病：

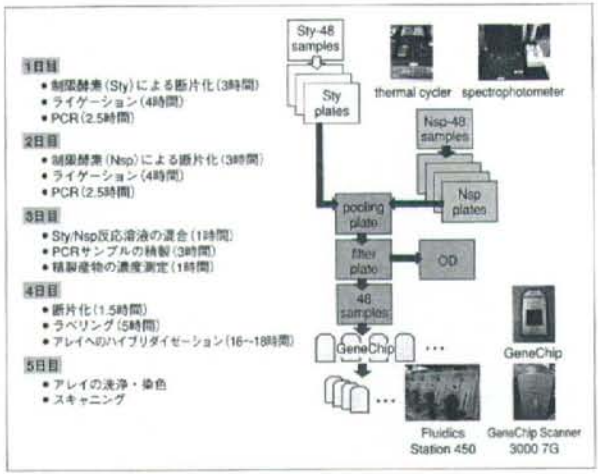

図 1　SNP Array 6.0によるSNPタイピングの流れ

制限酵素(StyI, NspI)による断片化反応からGeneChipのスキャニングまで，全5日の工程でSNPタイピングが行われる．1検体につき500 ngのゲノムDNAを用いて全909,622種類のSNPをタイピングすることができる．

T2D)について，それぞれ2,000人の患者と，コントロールとして健常人3,000人の計17,000人を対象とした大規模なゲノムワイド関連研究を行った[7]．また，アメリカNIHはGWAS計画を提案し，いくつかのcommon diseasesについて大規模な研究チームを公募した．さらに，大規模な疫学研究として知られるFramingham Heart Studyで収集された試料のうち9,000検体について心，肺，血液，睡眠疾患に関与する遺伝子変異を探索する計画が発表され，ゲノムワイド関連分析およびゲノムワイド連鎖解析などを行った結果が2007年にまとめて報告された[8]．

技術原理

SNP Array 6.0は，制限酵素によるゲノムDNAの断片化とマイクロアレイによるタイピングの手法に改良を加えることにより，大規模なタイピングを行える手法として確立された．解析対象となるSNPは，公共のSNPデータベースおよびPerlegen社に登録されている約220万種のSNPから遺伝学的情報量が最大化されるように，また連鎖不平衡やHapMapプロジェクトからの情報も考慮して選択された約44万種のSNPに，Tag SNP，X染色体およびY染色体に存在するSNPなどを加えた全909,622種類のSNPである．

SNP Array 6.0によるSNPタイピングは，ゲノムの複雑さを低減しマイクロアレイへのhybridization効率を上げるための酵素反応ステップと，洗浄・染色装置(Fluidics Station 450)およびマイクロアレイ用スキャナ(GeneChip Scanner 3000 7G)を用いた検出ステップに分けることができる(図1)．SNPタイピングは1検体について合計500 ngのゲノムDNAを使用し，2種類の制限酵素(StyI, NspI)を用いて実現される．制限酵素によるゲノムDNAの断片化を行った後，断片化されたゲノムDNAの両末端にアダプター配列をライゲーション反応により付加する．アダプター配列は続くPCRで使用されるプライマーと相同な配列をもち，また制限酵素認識配列を突出端としてもつ二本鎖DNAである．2種類の制限酵素(StyI, NspI)のそれぞれに対して用意されるアダプター配列は，制限酵素認識配列を除いて共通の配列をもっているので，共通のプライマーを使用してPCRを行うことができる．PCRでは目的の長さをもったゲノムDNA断片(250~1,100 bp)だけが選択的に増幅される．ここまでの酵素反応により，もともと30億塩基対のゲノムDNAが5億塩基対程度のPCR混合産物となる．

マイクロアレイへの効率的なhybridizationには，ゲノムの複雑さを低減することが大きな役割を果たすと考えられている．続いて，StyIおよびNspIそれぞれのPCR産物を混合した後，混合産物を精製し，DNase I制限酵素による断片化を行う．断片化されたPCR産物は平均長で180 bp以下となる．マイクロアレイへの効率的なhybridizationには，ゲノムの複雑さを低減することに加えてPCR産物の断片化が重要になる．最後に，terminal deoxynucleotidyl transferase酵素反応により，断片化されたPCR産物の末端にビオチンを導入する．

続いて，専用のマイクロアレイを用いてhybridizationを行う．マイクロアレイに固定されるプローブは25塩基長のオリゴDNAで，SNP部位を含む塩基配列をもっている．2種類のアリルを正確に識別するために，SNP部位を25塩基長のプローブの中心においたプローブを基本として，SNP部位を中心から4塩基上流(+4)にずらしたプローブから4塩基下流(−4)にずらしたプローブまで7種類のプローブ(−4, −2, −1.0, +1, +3, +4)を用意し，そのなかから最適な1種類のプローブを選択する．また，同一のプローブをマイクロアレイ上に4スポット用意することで，SNPタイピングデータの欠損を防ぐ工夫がなされている．

マイクロアレイへのhybridizationが終了した後，洗浄・染色装置を用いてマイクロアレイの洗浄および蛍光染色を行う．蛍光染色は，蛍光分子で標識されたストレプトアビジンが上述のビオチン導入されたPCR断片に結合することにより行われる．また，洗浄・染色装置内ではビオチン修飾された抗ストレプトアビジン抗体を用いてシグナルの増強が行われる．最後に，蛍光染色されたマイクロアレイを専用のスキャナーで画像データとして読み取り，続いて専用のソフトウェアを用いて各SNPの遺伝子型を決定する．

複数の施設で行われたSNP Array 6.0によるSNP解析の結果から，コール率(全909,622 SNPsのうち遺伝子型が決定されたSNPの割合)は平均99%以上となり，また，HapMapデータベースに登録されたSNPとの遺伝子型一致率は99.7%を超えることが，Affymetrix社から報告されている．また，タイピング結果が悪いことが明らかとなっている3,022 SNPsをクオリティーコントロール(QC)として用いて，QCコール率(3,022 SNPsのうち遺伝子型が決定されたSNPの割合)が86%を下まわる検体を除外したうえで全SNPの遺伝子型は決定される．

システム構築

1. ハードウェアの整備

SNP Array 6.0によるSNPタイピングを効率的に行うために，環境，装置を整備し，作業マニュアルを作成した．まず，ゲノムDNAへのPCR産物のコンタミネーションを防ぐために，試料調製室とSNP解析室を設けた．試料調製室にはゲノムDNAを保管し，PCR以前の酵素反応を行うのに必要な装置(サーマルサイクラーなど)を用意した．制限酵素による断片化からPCRの反応溶液の調製までは試料調製室で行い，PCR以降の酵素反応はSNP解析室で行った．また，4台の洗浄・染色装置を用意し，1回のランで計16枚のマイクロアレイを洗浄・染色することができるようにした．すべてのマイクロアレイはバーコードで管理され，洗浄・染色が終了したマイクロアレイはオートローダー付きのマイクロアレイ用スキャナに装填され，画像データが読み込まれる．オートローダー付きのマイクロアレイ用スキャナは，計64枚のマイクロアレイを装填することができ，バーコードを参照しながらすべてのマイクロアレイの画像データを自動的に読み込むことができる．

SNPタイピング作業のルーチン化にあたって，制限酵素による断片化からラベリングまでの5つの酵素反応ステップで使用するすべてのマイクロタイタープレートをバーコードで管理する．また，

1枚のマイクロタイタープレートで48検体分の酵素反応を行うこととし，マイクロタイタープレートのウェル位置をサンプルと対応させることでサンプルのID化を行った．マイクロタイタープレート上のレイアウトを変えずに酵素反応を進めることで，ウェル位置をサンプルIDとして解析結果を得ることができる．また，酵素反応の各工程を管理するためにチェックシートを作成し，反応工程の進行を随時チェックシートで確認しながら進める．PCRおよびDNase 1制限酵素による断片化の後にはアガロースゲル電気泳動を行い，PCR産物および断片化産物の平均長がそれぞれ250〜1,100 bp，180 bp以下となっていることを確認する．また，精製後のPCR産物の濃度が500〜600 ng/μlとなっていることを確認する．

2. ソフトウェアの開発

SNP Array 6.0によるSNPタイピングでは，メーカーが提供する2種類のソフトウェアを使用する．一方は洗浄・染色装置およびマイクロアレイ用スキャナを操作する際に使用し，また他方はマイクロアレイの蛍光強度データ(CELファイル)から遺伝子型を判定する際に使用する．決定された全909,622種類のSNPの遺伝子型は，検体ごとにテキストファイルとしてエクスポートすることができる．

著者らが開発したゲノムワイド関連分析用ソフトウェア(GeneChipAnalysis ver. 2.0.12)は，エクスポートしたファイルを直接入力ファイルとして用いることができる．また，ケースコントロール関連分析をするにあたって各検体から全SNPの遺伝子型データを抽出し，さらにすべての検体をケース群およびコントロール群に分けてあらたなテキストファイルとして作成する機能をソフトウェアに加えた．全909,622種類のSNPについてアリル頻度，ジェノタイプ頻度，優性・劣性モデルでのケースコントロール関連分析を行うことができ，関連分析の結果は専用のソフトウェア(GeneChipViewer ver. 2.1.1)を用いて視覚的に表示することができる．

日本人健常人200検体のSNPタイピングデータの解析結果

SNP Array 6.0による全909,622 SNPsのSNPタイピングでは，1検体につき500 ngのゲノムDNAを使用する．StyIおよびNspIによる断片化反応に用いるゲノムDNA量がそれぞれ250 ngとなるように調整することはSNPタイピングの精度に大きな影響を与えることが，これまでの実験結果から明らかとなっている．日本人健常人200検体のうち195検体のゲノムDNA濃度は規定濃度である50 ng/μlを満たしており，平均54.8 ng/μlであったが，5検体は規定濃度を下まわり平均41.1 ng/μlであった．そこで，規定濃度を下まわった5検体は，制限酵素断片化反応に6μlを持ち込み，ゲノムDNAの総量が約250 ngとなるように調整してタイピングを行うこととした．日本人健常人200検体のSNPタイピングを行った結果，QCコール率は平均97.37%となり，また，QCコール率が86%を下まわる検体は200検体のうち2検体であった．続いてQCコール率が86%を上まわった198検体を用いて全909,622 SNPsのコール率を決定したところ，平均99.71%となった(図2)．

日本人健常人200検体のSNPタイピングの結果から，SNP Array 6.0に搭載された全909,622 SNPsのうち，約19%に相当する170,921 SNPsにおいて多型性がみられなかった．また，遺伝子型が決定されたSNPのなかにはタイピング精度の悪いSNPが一部含まれており，それらのSNPは偽陽性関連の原因のひとつになると考えられる．これについては，マイナーアリル頻度(MAF)，ハーディー・ワインバーグ平衡およびSNPコール率(各SNPについてタイピングした全検体のうち遺伝子型を決定できた検体の割合)を指標として，タイピング精度の悪いSNPの大部分を排除することができる[9]．著者らの解析ではMAF>5%，HWE p 値>0.001，SNPコール率>95%を満たすSNPは588,932 SNPsとなり，また，MAF>1%，HWE p 値>0.001，SNPコール率>95%を満たすSNPは659,530 SNPsとなった．

将来の展望

ゲノムワイド関連研究により，さまざまな多因子疾患の疾患感受性遺伝子を特定したという報告がここ数年盛んになされている．著者らのヒトSNPタイピングセンターにおいても最新のプラットフォームであるSNP Array 6.0を導入し，大規模SNPタイピングを実施している．また，著者らはゲノムワイド連鎖分析あるいはゲノムワイド関連分析によって検出された候補領域において第一義的な疾患感受性遺伝子多型を特定するため(絞り込み)に適する技術として，DigiTag2法を確立した[10]．現在，いくつかの多因子疾患を対象とした多施設共同研究グループと協力してゲノムワイド関連分析を進めており，さまざまな集団に共通する遺伝因子だけでなく，日本人あるいはアジア人に特徴的な遺伝因子の特定をめざしている．

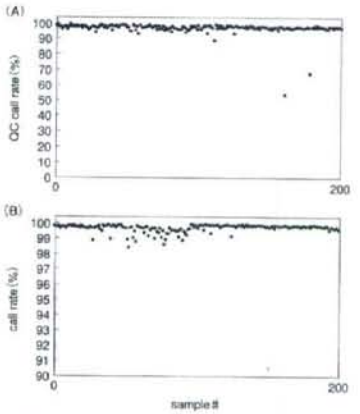

図2 SNP Array 6.0による日本人健常人200名のタイピング結果

A：クオリティーコントロール(QC)としてタイピングされた3,022 SNPsのコール率を示している．

B：QCコール率が86%を上まわった198検体を用いて決定された全909,622 SNPsのコール率を示している．

現在販売されているSNP Array 6.0には欧米人で頻度の高いSNPが優先的に搭載されているため，日本人試料では約19%のSNPについて多型性がみられなかった．疾患感受性候補領域を最大限に検出するためにも，今後アジア系集団に適したSNPセットを搭載したプラットフォームが作成されることを期待している．また，SNP Array 6.0はCNV(copy number variation)解析を行うためのプローブを搭載しているが，現段階ではCNVを対象とする大規模な関連解析を行うためのソフトウェアはできていない．また，100万種類のSNPといえども厳しい基準を採用した際のtagSNPおよびハプロタイプブロック外のSNPを網羅するものではないことから，さらに多数のSNPをタイピングできる技術の開発も予想される．いずれにせよ，ゲノムワイド多型解析情報は従来にない膨大なデータを産生することから，バイオインフォマティクスにかかわるさまざまな研究者にとって挑戦に値する多くの課題を提供してくれるとともに，その達成によって従来にない実り豊かな成果をわれわれにもたらしてくれるに違いない．

文献

1) Ohashi, J. and Tokunaga, K.: *J. Hum. Genet.*, **46**: 478-482, 2001.
2) 徳永勝士(編著): 人類遺伝学ノート. 南山堂, 2007.
3) Ohnishi, Y. et al.: *J. Hum. Genet.*, **46**: 471-477, 2001.
4) Ozaki, K. et al.: *Nat. Genet.*, **32**: 650-654, 2007.
5) Tamiya, G. et al.: *Hum. Mol. Genet.*, **14**: 2305-2321, 2005.
6) Kawashima, M. et al.: *Asu, J. Hum. Genet.*, **79**: 252-263, 2006.
7) The Wellcome Trust Case Control Consortium: *Nature*, **447**: 661-678, 2007.
8) Cupples, L. A. et al.: *BMC Med. Genet.*, **8**(Suppl. 1): s1, 2007.
9) Miyagawa, T. et al.: *J. Hum. Genet.*(submitted)
10) Nishida, N. et al.: *Anal. Biochem.*, **364**: 78-85, 2007.

ゲノムワイド関連解析データベースの開発

Development of genome-wide association database

小池麻子・西田奈央・徳永勝士

ゲノムの多型解析技術，とりわけSNPタイピング技術の飛躍的な向上により，ここ2年ほどで，検体あたり50～100万種のSNPについて高速で大容量なタイピングが可能となった．これをうけて，各種の複合疾患の感受性を規定する遺伝子の発見をめざした，大規模なゲノムワイド関連解析がそくそくと実施されている．これらの進行にともない，ゲノムワイド関連解析データの管理に関する問題も表出し，データの永続的な管理や研究者間での情報の共有による疾患遺伝子の探索研究の促進を目的として，ゲノムワイド関連解析データの集約的管理とその公開は世界的な潮流となりつつある．2007年度からはじまった文部科学省の統合データベースプロジェクトにおいて，筆者らは，ゲノムワイド関連解析データの国内における集約的管理をめざしたデータベースの構築を担当しており，このデータベースは本プロジェクトの中核機関によって公開される．本稿では，その概要について紹介する．

Key words ■ゲノムワイド関連解析 ■単一塩基多型（SNP） ■統合データベース

はじめに

ヒトゲノム解読の完了後，新規遺伝子やゲノム多型の発見とハプロタイプ地図の作成を目的としたHapMapプロジェクトが実行された．これにより，SNP（single nucleotide polymorphism，単一塩基多型）と表現型との関連をゲノム全体からみつけ出すために有用なタグSNPの数や位置情報が集積された．また，HapMapプロジェクトの進行と並行してさまざまなSNPタイピング技術が開発されたが，なかでも，DNAマイクロアレイによる高速で大容量なSNPタイピング技術の確立がこの分野の研究を劇的に加速させた．今日では，100万種近くのSNPを数日間で同時並列タイピングできるようになっており，疾患などの表現型と遺伝子型との関連性についての解析をゲノム全域にわたって行なうゲノムワイド関連解析（genome-wide association study；GWAS）により，疾患の感受性や抵抗性を規定する疾患関連SNPを発見する環境がほぼ整ったといえる[1]．

Asako Koike[1], Nao Nishida[2], Katsushi Tokunaga[2]
[1]日立製作所 中央研究所
[2]東京大学大学院医学系研究科 国際保健学専攻人類遺伝学分野
E-mail：asako.koike.ea@hitachi.com
nishida-75@umin.ac.jp
tokunaga@m.u-tokyo.ac.jp

このような状況のもと，英国Wellcome Trustや米国NIH（National Institutes of Health）がゲノムワイド関連解析へ積極的な資金援助を行ない，ひとつの疾患あたり数千検体ものSNPタイピングを行なう大規模なコホート研究やケースコントロール研究が企画・実行され，新たな疾患関連SNPがつぎつぎと報告されている[2,3]．

1991年にヒトゲノム計画がはじまってから15年あまりが経ち，疾患遺伝子の探索分野における収穫時期がようやく到来したといっても過言ではない．このことは，Science誌が2007年の科学界での最大のブレークスルーを"ヒトの遺伝的変異の研究"としたことにも現われている（報道で大きく取り上げられた，iPS細胞を含む細胞のリプログラミング研究は，第2位であった）．

このように，現在，多くの大規模なゲノムワイド関連解析が進展しており，これにともなって膨大な実験データや解析結果が産出されているが，これらのデータの共有・公開を促進する世界的な動きがある．すなわち，健常者集団と患者集団について数十万種のSNPを解析した結果，および，それらの統計遺伝解析の結果（おのおののSNPやハプロタイプについて，疾患との関連性の有無や強弱を統計的に解析した結果）などをデータベース化し，さらに，解析結果を可視化して，広く研究者間で情報を共有する枠組みが構築されつつある．たとえば，米国NCBI（National Center for Biotechnology Information）は，ゲノムワイド関連解析データを含む表現型と遺伝子型との関係に関するデータをdbGAP（database of genotype and phenotype）として2006年秋に公開し，着々と大型プロジェクトのデータが蓄積されている[4]．また，欧州のEBI（European Bioinformatics Institute）は，EGA（European Genotype Archive）を構築し，近日公開の予定である[5]．このような世界的潮流を踏まえ，文部科学省の統合データベースプロジェクトにおいて，筆者らの分担機関は，①ゲノムワイド関連解析データのデータベース化と解析手法の開発，②リシークエンシングによる疾患関連遺伝子の遺伝子情報および付随する臨床情報のデータベース構築とデータマイニングツールの開発，を行なっており，これらによって構築されたデータベースは，本プロジェクトの中核機関によって公開される．

①に関しては，文部科学省科学研究費特定領域"ゲノム"で実施しているゲノムワイド関連解析のデータをはじめ，厚生労働省科学研究費などのデータも蓄積しつつある．そのほか，国内のデータを広く受け入れるための準備を進めている．将来は国内にとどまらず，国外，とくに，アジア諸国からのデータも受け入れたいと考えている．②に関しては，神経疾患関連遺伝子に関するリシークエンス結果のデータベース化を行なうとともに，文献中の変異情報とそれに関する臨床情報なども広く収集・登録し，臨床にも役立つシステムの構築をめざしている．本稿では，①に関しての取り組みを概説する．

Ⅰ ゲノムワイド関連解析システム構築の意義

ゲノムワイド関連解析データの集約的管理とシステム公開は，永続的なデータの保管，データの再解釈による疾患研究の促進，実データを利用することによる統計遺伝学の計算手法研究の促進，という観点から重要な意義をもっている．

永続的なデータの保管については，まず，ゲノムワイド関連解析データが従来の疾患関連研究データとは比較にならない，大量のデータであることに注目してほしい．しかしながら，ほかのデータと同様，年月が経つにつれて研究責任者の異動などにより，実験のプロトコルやデータの質などの付随情報はもとより，データの保管場所すら不明になってしまう危険性は否めない．試料提供者への説明と同意の範囲，個人情報保護への対策に注意しつつも，永続的に研究者間でデータを共有できるよう集約的に管理することが望まれる．

データの再解釈による疾患研究の促進に関しては，ゲノムワイド関連解析を実施したグループ以外のグループによるデータの再解釈により，新たな疾患関連候補SNPの発見という可能性がある．なぜならば，多くのゲノムワイド関連解析研究においては，検体収集のむずかしさと予算上の問題からタイピングできる検体数が限られているうえ，疾患関連遺伝子の浸透率やオッズ比（疾患発症の危険率）がそれほど高くないことから，理論的にも実際にも，解析の第1段階のみでは一部の疾患関連SNPしか検出できない[6]，すなわち，第1段階でのゲノムワイド関連解析の結果のみでも従来の戦略では検出できなかった疾患関連候補SNPが明らかになることから，いくつかの論文が発表されているが，これらのデータのなかに，統計遺伝解析の結果だけからは疾患関連SNPとは断定できずにうもれている候補SNPが多数，存在するはずである．そこで，データベースに掲載されているビューアでさまざまな解析結果を可視化することにより，ゲノムワイド関連解析や疾患遺伝子探索の専門家のみならず，統計遺伝学の手法とは縁遠かった研究者にも広く結果の提供を行なうことができる．統計遺伝学的にはSNPと疾患との関連が不確かであっても，さまざまな遺伝子を個別に研究している研究者が独自にもっている機能情報と組み合わせて，新たな知見や仮説が浮き彫りになる可能性も少なくない．また，複数のデータ産出機関の結果を組み合わせてメタ解析などを行なうことも有効であろう[7]．

実データを利用することによる統計遺伝学の計算手法研究の促進に関しても，従来にはない大量の多型データあるいは変異解析データから医学・生物学的意義を抽出することが求められており，統計遺伝学およびそのほかの解析手法の発展をうながすことになろう．具体的には，多重検定の補正法，複数SNPの相互作用を検出する方法，異なるグループが産出したデータのメタ解析法，そのほかの遺伝学的・進化学的知見を得る手法などについて，実データを利用することにより開発研究が加速されることが期待される．

Ⅱ ゲノムワイド関連解析データベースの概要

筆者らは，品質管理を含む健常対照群データを情報する標準SNPデータベース，第1段階のゲノムワイドなSNP関連解析の結果を登録するデータベース，第2段階の再現性の検討・絞り込み結果を登録するデータベース，を構築している．いずれのデータベースも，筆者ら分担機関内に情

築する内部用データベースと，中核機関に設置される外部用のデータベースを用意し，一定の手続きを経たコミュニティー間で共有するデータは内部用データベースのみに登録され，幅広く公開するデータは外部用データベースにも格納される構成となっている．つぎに，これらを具体的に説明する．

1．標準データベース

ゲノムワイド関連解析の第1段階においては，数百人から数千人について，おのおの数十万種の大量SNPデータをあつかうため，統計遺伝学的な解析の以前に，データ解析に使用する検体とSNPタイピング結果についての品質管理が必須となる．そこで，品質管理に必要となる適切な基準を作成し，日本人対照群での標準化データベースとして公開する．品質管理は，SNPごとの品質，検体ごとの品質，集団としての品質，の3項目について行なっている．

SNPごとの品質については，SNPごとのcall rate（全検体のうち遺伝子型が判定できた割合）が十分に高いか，集団遺伝学における基本法則のひとつであるハーディ・ワインベルク平衡の検定量が閾値以上か，および，頻度が低いほうの対立遺伝子頻度が閾値以上か，などを検定している．検体ごとの品質については，検体としてのcall rateが閾値以下であるような質の悪い検体を除いている．集団としての品質については，同じ集団に属する非血縁の検体であるかを判定するため，構造化の解析や他集団の情報をくわえて主成分の分析を行なっている．このほかにも詳細な検討を経て品質管理の基準を通過したSNPについて，標準対立遺伝子頻度，遺伝子型頻度，ハーディ・ワインベルク平衡の検定量，ハプロタイプ頻度などを登録している．SNP標準データベースのトップページと検索結果の一部を，図1に示す．これらのデータは，各疾患のゲノムワイド関連解析における対照群データとしても使用している．現在，市販されているゲノムワイドなSNPタイピングのプラットフォームであるAffymetrix 500K，Affymetrix 6.0，Illumina 317Kの，おのおの約500，200，200の検体のデータが登録されており，今後も拡充していく予定である．

2．ケースコントロール関連解析データベース

第1段階の，ゲノムワイドなケースコントロール関連解析（患者群と健常者群のSNPタイピングの結果にもとづいて疾患と遺伝子・ゲノム多型の関連性の解析を行なう）の結果については，研究目的や疾患の診断基準，検体数，男女比などとともに，実験プロトコル，SNPタイピング結果，統計解析結果をデータベースに登録している．なお，解析結果の品質管理については，基本的に標準データと同じ手順で行なっている．また，解析結果を直感的に理解しやすいよう可視化するビューアも備えており，プロモーター領域や遺伝子コーディング領域であるかどうか，アミノ酸置換の有無などSNPの種類，染色体上の物理的位置などの遺伝子領域にあるかといった位置情報など，ゲノム情報も登録している．

図2に，解析結果の例を示す．図2(a)は，染色体2 Mbごとの領域における対立遺伝子頻度を比較したP値（有意確率）の最小値を色分けしたものである．赤色が濃いものほど

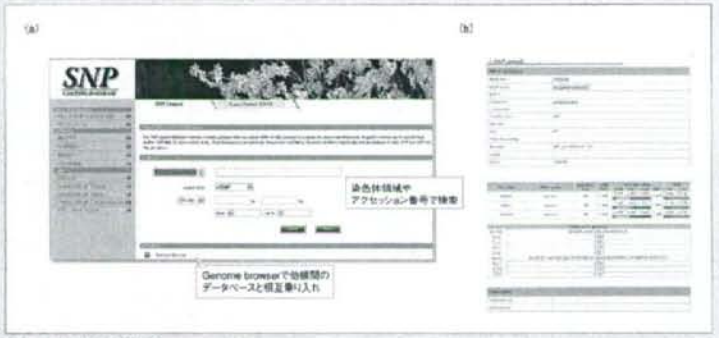

図1　SNP標準データベース
(a) トップページ
(b) 検索結果の例

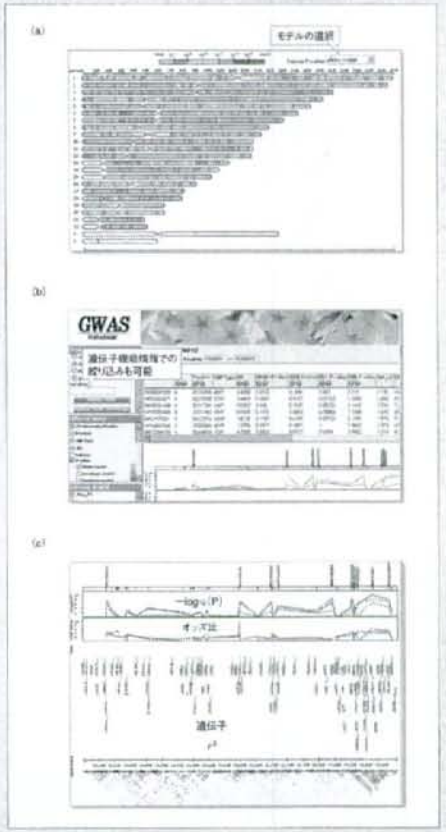

図2　ゲノムワイドなケースコントロール関連解析の結果
例として，網膜障害を起こす疾患であるナルコレプシーのデータ解析を示す．
(a) 色が濃い領域ほど，有意により強い関連を示す対立遺伝子が存在する．ここでは，ヒト第4染色体の30 Mb前後の領域が強い関連を示している．
(b) (a)において，ヒト第4染色体の30.0～30.2 Mbの領域をクリックした場合，当該領域での解析結果および複数の統計解析結果の表とグラフが表示される．全体データの一覧表示も可能．
(c) 各種の統計量の結果の−log10Pとオッズ比のグラフ，遺伝子の位置情報，および，連鎖不平衡の指標としてのrを表示した場合．−log10Pとオッズ比の折れ線グラフについては，高い値を示すほど疾患との関連性が強い．おのおのの計算モデルの結果を異なる色で表示している．

疾患との関連が強い領域となる．各領域をクリックするか領域を指定して検索すると，図2(b)に示すように，指定領域での実験結果および複数の統計解析結果の表とグラフが表示される．また，染色体表示ではなく，全体データの一覧も表示できる．図2(b)の右下部分が図2(c)となる．ここでは，各種の統計解析結果の $-\log_{10}P$ とオッズ比のグラフ，遺伝子の位置情報，および，連鎖不平衡(近隣の多型部位にみられる関連性)の指標である r^2 が表示されている．これ以外にも，エクソン-イントロン構造，OMIM (Online Mendelian Inheritance in Man) 登録遺伝子領域，コピー数変異(copy number variation；CNV)領域，マイクロサテライトの位置，HapMapデータベースに公開されている連鎖不平衡値，なども表示できる．また，染色体の位置情報だけでなく，遺伝子オントロジーによる遺伝子の機能情報からも絞り込みが可能である．すなわち，関連がみられたSNPのうち，指定の機能にかかわる遺伝子上に位置するもののみを選びだすこともできる．今後は，MEDLINEに記述されている遺伝子のアノテーション情報を利用することによって，ユーザが遺伝統計解析の結果を解釈できるシステムにつくりあげていく予定である．

第2段階の，再現性の検討・高密度関連マッピングのデータベースについては，第1段階において関連が検出されたSNPを再度タイピングするか(しばしば，第1段階とは独立なサンプルセットも解析する)，より高密度にSNPタイピングまたはリシークエンシングを行なった結果を，臨床情報および解析結果とともに登録する．このデータベースについては2009年度以降の構築となるが，第1段階のデータベースと同様，SNPタイピングの精度情報，ハーディ-ワインベルク平衡の検定量，対立遺伝子の頻度情報，遺伝子型の頻度情報にくわえて，各種の関連解析結果を表示する．さらに，これらにくわえて，いくつかの統計遺伝学的な手法やデータマイニング機能を搭載し，疾患関連SNPあるいはSNP組合せなどの情報を表示する予定である．結果を解釈する際にはほかのデータベースの情報も利用する場合があるため，DAS (distributed annotation server) としてほかのデータベースとのデータのやりとりをスムーズに行なえるよう対応していくとともに，SOAP (simple object access protocol) による解析結果の提供サービスも考えている．

Ⅲ データ公開のポリシー

米国NCBIのデータベースdbGAPのデータ公開の方針は，SNPごとの頻度データや各種統計値などは，個人を特定できるものではないためアクセス制限なしの公開であるが，遺伝子型の一覧表や臨床データなどについては，どのように公開するかはデータ提供元，および，データ提供元の研究費スポンサーに委ねられている．後者の多くについては，研究目的を明記したうえで利用を申請し承認された研究者にのみ公開をしている[1]．欧州EBIのデータベースEGAはまだシステムを公開していないが，dbGAPと同様の公開ポリシーを発表している(The European Genotype Archive, Background and Implementation, white paper, 2007)．

筆者らが構築しているゲノムワイド関連解析のデータ公開方針については，現在，倫理検討委員会で慎重に議論が進められているが，基本的にはdbGAPと同様，2つのデータレベルに分けたうえでの公開となるであろう．一般公開となるデータ部分については統合データベースプロジェクトの中核機関によって公開されるが，筆者ら分担機関に保管する内部用データベースには個人レベルの遺伝子型や遺伝子型判定前の生データなどが含まれるため，これらのデータは申請して承認を得た研究者にのみ，再配布禁止，所属機関の倫理審査委員会による承認など，いくつかの条件つきで提供される見込みである．

Ⅳ データのレポジトリー

統合データベースプロジェクトは，広くゲノムワイド関連解析データを受け入れ，研究者間でのデータの共有を図ることを目的のひとつとしている．Affymetrix社やIllumina社のアレイデータの生データのみでも受け入れることができ，また，データを提供した研究者自身が統計遺伝学的な解析を行なった結果についても，研究概要，品質基準，解析手法を明記したうえで受け入れて公開する方針である．また，データを登録した研究者が論文執筆の優先権を得るため，しかるべき期間(1年間)はデータ登録者もしくは指定グループのみに閲覧権限が与えられ，その期間ののち公開することも可能である．ゲノムワイド関連解析データの収集・登録に賛同していただける方は，ぜひ，筆者らにご連絡いただきたい．

おわりに：将来の方向性

現状の100万種のSNPタイピングといえども，ゲノム全域のカバー率においてはまだ不十分な点もあり[9]，今後，さらにSNP数が増加する可能性も考えられるので，データ量の急激な増加に解析手法もデータベース構築も対応していく必要がある．一方で，次世代型超並列シークエンサーの開発も急速に進展しており，米国NIHのかかげたロードマップにある1000ドルシークエンス(1000ドルで一個人の全ゲノムの塩基配列を決定する)をめざして，各社がしのぎを削っている．ここ数年間で，ゲノムの読み取り速度が急激に速くなるとともに，コストや労力は大きく下がることが予想される．ゲノムワイド関連解析などの疾患遺伝子の探索研究も，大量SNP解析とリシークエンスの併用となる時期がくるであろう．また，これらの高速・大容量の解析により，個別化医療への貢献のあり方も大きく変わってくるものと考えられる．統合データベースプロジェクトにおいては，データの保全管理と研究者間での情報の共有化だけでなく，ゲノム解析技術の発展をみすえたゲノムワイド関連解析システムの構築をめざしたい．

[文献]

1) 西田奈央・徳永勝士：実験医学，25，178-184 (2007)
2) The Wellcome Trust Case Control Consortium: Nature, 447, 661-678 (2007)
3) Easton, D. F. et al.: Nature, 447, 1087-1093 (2007)
4) Saxena, R. et al.: Science, 316, 1331-1336 (2007)
5) Vasan, R. S. et al.: BMC Med. Genet., 8, Suppl 1, S2 (2007)
6) Mailman, M. D. et al.: Nature Genet., 39, 1181-1186 (2007)
7) Williams, S. M. et al.: Science, 316, 1840-1842 (2007)
8) Zeggini, E. et al.: Science, 316, 1336-1341 (2007)
9) Barrett, J. C., Cardon, L. R.: Nature Genet., 38, 659-662 (2006)

小池麻子
略歴：京都大学理学系大学院修士課程 修了，現在，日立製作所中央研究所 主任研究員，2003～2006年 東京大学大学院情報理工学研究科 客員助教授 併任，2007年より東京大学大学院理学系研究科 非常勤講師 併任，理学博士．
研究テーマ：医学生物学系のテキストマイング，オントロジー蛋白質間相互作用のネットワーク解析．

▶シリーズ **掲載一覧**

掲載号	タイトル	執筆者
2007年 8月号	生命科学系データベース統合化の背景	大久保公策（国立遺伝研）
9月号	統合データベースセンターがめざすもの	高木利久（情報・システム研究機構）
10月号	医薬品の統合データベース	金久 實（京大化研）他
11月号	蛋白質構造情報の高度化と統合データベース	中村春木（阪大蛋白研）
12月号	ゲノム情報と統合データベース	五條堀 孝（国立遺伝研）
2008年 1月号	ライフサイエンスにおける データベース構築のための人材育成	瀬々 潤（お茶大院人間文化創成科学） 池村淑道（長浜バイオ大）
2月号	海外データベースとの連携	舘野義男（国立遺伝研）
3月号	統合データベースプロジェクトのサービスとその利用法	川本祥子・坊農秀雅（情報・システム研究機構）
4月号	利用者の立場からのコメント	
5月号	わが国における疾患データベースの統合化	田中 博（東京医科歯科大情報医科学セ）
6月号	ゲノムワイド関連解析データベースの開発	徳永勝士（東大院医）他
7月号	PNEバックナンバーの統合データベースにおける公開（予定）	

Published online 4 November 2008 *Nucleic Acids Research, 2009, Vol. 37, Database issue* **D333–D337**
doi:10.1093/nar/gkn855

The GTOP database in 2009: updated content and novel features to expand and deepen insights into protein structures and functions

Satoshi Fukuchi[1,*], Keiichi Homma[1], Shigetaka Sakamoto[2], Hideaki Sugawara[1], Yoshio Tateno[1], Takashi Gojobori[1] and Ken Nishikawa[3]

[1]Center for Information Biology and DNA Data Bank of Japan, National Institute of Genetics, Yata 1111, Mishima, Shizuoka 411-8540, [2]HOLONICS Corporation, Soeji 85, Numazu, Shizuoka 411-0803 and [3]Department of Bioinformatics, Maebashi Institute of Technology, Kamisadori 460-1, Maebashi, Gunma 371-0816, Japan

Received September 12, 2008; Revised October 15, 2008; Accepted October 16, 2008

ABSTRACT

The Genomes TO Protein Structures and Functions (GTOP) database (http://spock.genes.nig.ac.jp/~genome/gtop.html) freely provides an extensive collection of information on protein structures and functions obtained by application of various computational tools to the amino acid sequences of entirely sequenced genomes. GTOP contains annotations of 3D structures, protein families, functions, and other useful data of a protein of interest in user-friendly ways to give a deep insight into the protein structure. From the initial 1999 version, GTOP has been continually updated to reap the fruits of genome projects and augmented to supply novel information, in particular intrinsically disordered regions. As intrinsically disordered regions constitute a considerable fraction of proteins and often play crucial roles especially in eukaryotes, their assignments give important additional clues to the functionality of proteins. Additionally, we have incorporated the following features into GTOP: a platform independent structural viewer, results of HMM searches against SCOP and Pfam, secondary structure predictions, color display of exon boundaries in eukaryotic proteins, assignments of gene ontology terms, search tools, and master files.

INTRODUCTION

Proteins encoded by genomes generally function after adopting proper 3D structures. A rapid increase in the number of entirely sequenced genomes led to an unprecedented growth in the number of hypothetical proteins resulting from genome annotation. Protein structures and functions can be inferred from amino acid sequences by using advanced computer programs. There is no doubt in the importance of structural and functional annotations of hypothetical proteins. The GTOP project was started in 1999 as reported (1) and was taken over by the DNA Data Bank of Japan (2) in 2007, under which the database has been continuously updated. GTOP is a database that provides protein annotation of 3D structures and functions based on similarity searches against PDB (3), SCOP (4), and Swiss-Prot (5), 2D structure predictions, Pfam (6) protein families, PROSITE (7) functional motifs, prediction of trans-membrane regions, and others.

There are several databases of the 3D structures of all the genome-encoded proteins. For example, SUPERFAMILY (http://supfam.mrc-lmb.cam.ac.uk/SUPERFAMILY/) (8) provides SCOP domain assignments to proteins encoded by completely sequenced genomes. A collection of comparative protein 3D structure models is available at Modbase (http://modbase.compbio.ucsf.edu/modbase-cgi/index.cgi) (9) in some entirely sequenced genomes. Gene3D (http://gene3d.biochem.ucl.ac.uk/Gene3D/) (10) makes public CATH-based domain assignments and functional annotations to proteins in more than 500 genomes. Functional and domain assignments including intrinsically disordered (ID) regions can be found at PEDANT (http://pedant.gsf.de/) (11).

From the previous report, we have added a large body of data and tools to GTOP, for example ID region assignments, exon information on eukaryotic proteins, an efficient mechanism to search within a user-specified set of genomes, and tools for phylogenetic profile search. Since its inception, GTOP has employed a user-friendly interface to let the user grasp features of a query protein at a glance. The interface has been improved with the addition of new information. A GTOP user can readily obtain

*To whom correspondence should be addressed. Tel: +81 55 981 6837; Fax: +81 55 981 6889; Email: sfukuchi@genes.nig.ac.jp

comprehensive structural and functional data of all the proteins encoded by entirely sequenced genomes.

UPDATE IN GTOP THAT CONTRIBUTED TO IMPROVED STRUCTURAL ASSIGNMENTS

A list of the genomes stored in GTOP is available at http://spock.genes.nig.ac.jp/~genome/org.html, together with the abbreviations of organism names used in the database. In the 2002 paper, we reported that GTOP contained protein data of 41 genomes (1). The database has grown to cover a total of 797 genomes, with 41, 466, 114 and 176 genomes of archaea, eubacteria, eukaryota and bacteriophages, respectively. The following data are subject to regular renewal: (i) amino acid sequences encoded by genomes newly sequenced after the previous update, (ii) amino acid sequences that existed in the previous version but were subsequently modified and (iii) reference databases such as PDB, SCOP, Swiss-Prot, Prosite, and Pfam whose new versions were released. The sequences fallen in category (ii) were recalculated to keep annotations up-to-date. Update category (iii) is crucial to keep annotations up-to-date, because most annotations in GTOP are obtained by homology search programs or those based on homology search.

The main focus of GTOP is structural annotations made by homology searches against the PDB and SCOP databases. Although GTOP used PSI-BLAST (12) in the previous report, it now employs reverse-PSI-BLAST (13), as this method gives comparable results in drastically reduced computation time. HMM searches using the SUPERFAMILY profiles (8) of SCOP domains were additionally conducted, as they are particularly effective in identifying small domains such as DNA binding domains.

Figure 1 presents a time course of the number of the genomes stored and the average fractions of proteins with

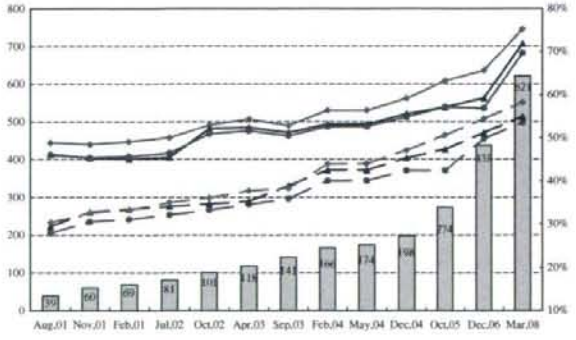

Figure 1. The time courses of the number of genomes included and the fraction of the sequences with homologs in the PDB. The line graphs represent the ratios of the sequences with homologs in the PDB, while the column graph stands for the number of genomes in GTOP. The scales for the fraction and the number of genomes are shown at the right and left ends, respectively. The blue, green, and red lines correspond to fruit fly, *E. coli*, and the overall average, respectively. The solid and dotted lines respectively show the ratios obtained using reverse PSI-BLAST, and those using BLAST. The exact numbers of genomes are displayed near the top of the rectangles.

3D annotations made by BLAST and reverse-PSI-BLAST. The fraction of sequences with alignments to PDB shows a steadily increasing trend, reflecting the growth of the PDB database. The fraction aligned by reverse-PSI-BLAST exceeds that by BLAST, reflecting the higher sensitivity of the former method. However, one should note that in this statistics a sequence is considered to be annotated if it has at least one PDB hit by BLAST or reverse PSI-BLAST and it may have large tracts of structurally undetermined regions. When statistics is evaluated residue-wise, the fractions of regions aligned to PDB sequences in the latest version in human and *Escherichia coli* proteins are 47% and 64%, respectively.

ID REGIONS

As most proteins do not entirely consist of structural domains, the fraction of residues with structural assignments will not reach unity; outside of globular domains there exist ID regions that assume no specific 3D structures by themselves, and tend to contain active regions in proteins involved in crucial biological processes such as signal transduction and transcriptional regulation (14–16). Recent research revealed that ID regions exist predominantly on the cytoplasmic side of eukaryotic proteins (17), play important roles in cell signaling, transcriptional control (18). We predicted ID regions in proteins stored in GTOP by the DISOPRED2 (19) program and presented them. Figure 2A shows a GTOP screen shot of human androgen receptor, a typical protein with long ID regions. As this example illustrates, GTOP graphically displays complex domain architectures of eukaryotic proteins composed of structural domains and ID regions.

EXON BOUNDARIES IN EUKARYOTIC PROTEINS

The existence of introns and exons is a unique feature of eukaryotic genes and the location of exon boundaries in the corresponding protein structure is of interest (20). We thus developed tools to display exon boundaries on amino acid sequences and 3D structures. Figure 2B shows an example of the exon boundary view. The exons are presented in 5 colors both in the 3D structure and the sequence displays, from which the boundaries can be clearly seen. We developed a 3D viewing system incorporating Jmol applet (http://www.jmol.org/) so that the user can view 3D structures in the browser without installing additional software. Alternatively Rasmol (21) or Chime (http://www.mdl.com/) can be used. Exon information is also presented in green and blue stripes (near the bottom of Figure 2A).

SEARCH TOOLS

GTOP strives to keep precomputed annotations of all the amino acid sequences of proteins derived from all the completely sequenced genomes. One clear benefit of having precomputed annotations beside the rapidity of supplying information is to make inter-genomic

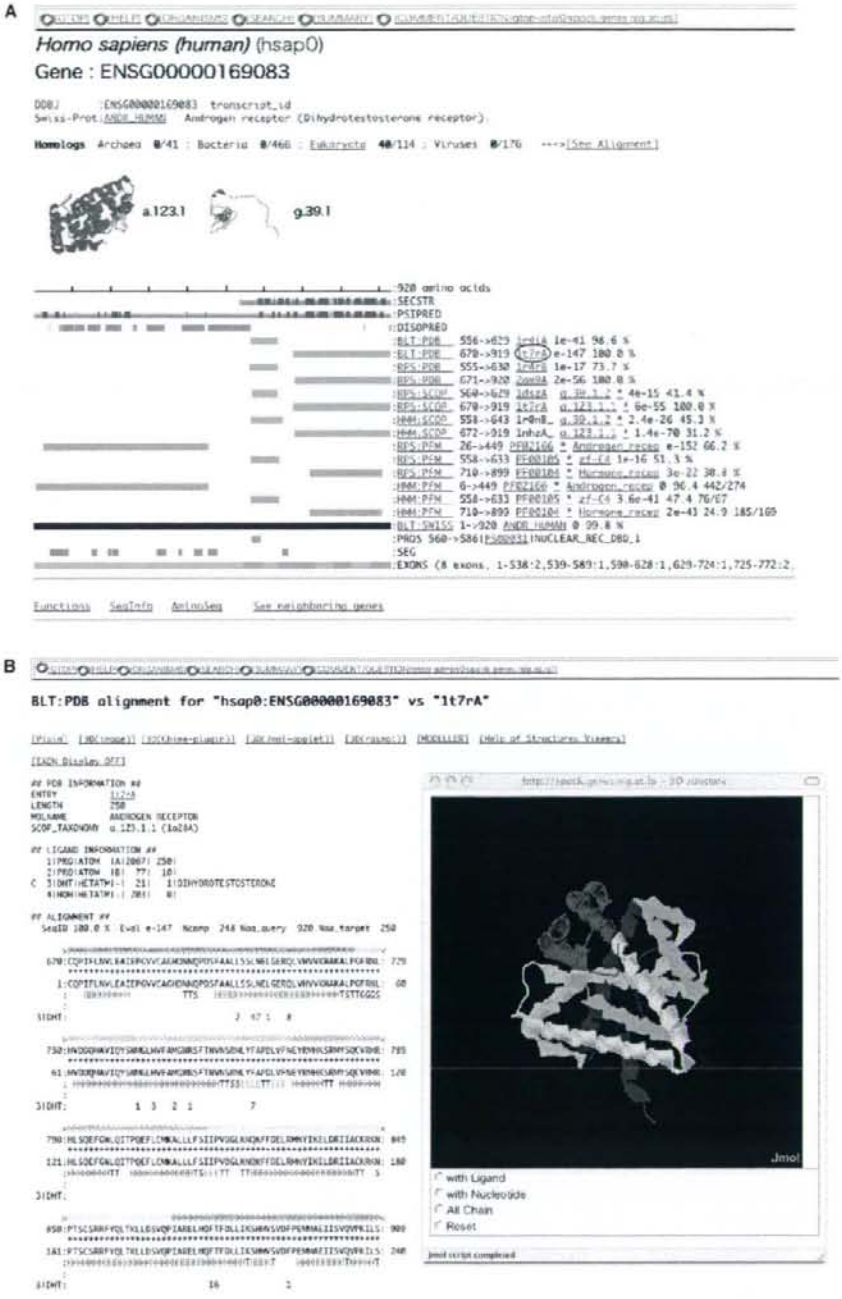

Figure 2. GTOP view examples. (**A**) The domain assignments of the human androgen receptor are presented in color bars to facilitate intuitive grasp of molecular architecture of the protein. This is a typical protein with long ID regions: the N-terminal half of the protein consists mainly of ID regions (18,22), consistent with the ID regions predicted by DISOPRED2 (gray bars on the line marked by DISOPRED). (**B**) A structurally aligned region of the same protein is shown in the exon view. This page can be obtained by clicking on the characters '1t7rA' circled in Figure 2A, and by clicking on the EXON Display and 3D (Jmol-applet) buttons in the top section of the pop-up screen. The 3D structure is shown in five colors. By the 3D viewer, the sequence alignment is displayed with the exons represented in the same colors.

comparative analyses possible. Phylogenetic profile search is one analytical tool that exploits this advantage: a user-specified search produces the presence and absence pattern of features such as SCOP folds, superfamilies, and families, Pfam domains, PROSITE motifs, and the number of trans-membrane helices. The user can conduct a search for a specific feature that are present in certain species and/or absent in others; for example, a search for a SCOP domain present in all the eubacterial species and absent in all the eukaryotic species in GTOP. The summary section of GTOP also offers comparative statistics, which has the ratio of 3D annotations in each genome, the frequencies of SCOP folds, superfamilies, and families, Pfam domains and PROSITE motifs.

Expansion of the database resulted in increased search time. The tools for keyword, homology, and text searches in GTOP were thus modified so that the user can reduce search time through selection of the genomes in which to conduct a search. The user can easily specify organisms with the use of check boxes placed next to organism names.

MASTER FILES

An annotation summary of each protein, consisting of abbreviated one-line descriptions, is saved in a master file. Master file information for each protein is displayed below a GTOP diagram of the type shown in Figure 2A. All the available data of each genome have been compiled in one file, freely downloadable from ftp://spock.genes.nig.ac.jp/pub/gtop/. Explanations of the meanings for each HEADER can be found at http://spock.genes.nig.ac.jp/~genome/mas-doc.html.

FUTURE DIRECTIONS

Despite the wealth of currently available structural data and use of sensitive programs, considerable fractions of most proteins have neither structural domains nor ID regions assigned. We are currently developing a system to accurately classify the fraction into structural domains and ID regions. Excitingly this will result in reliable identification of structural domains whose 3D structures remain undetermined. We expect that the installation of this system will provide further insights into the protein structure. We are also considering incorporation of protein–protein interaction data to enrich GTOP further.

FUNDING

The GTOP database is supported in part by the Target Protein Research Program from the Ministry of Education, Culture, Sports, Science and Technology of Japan, and in part by the Bioinformatics Research and Development Project from the Japan Science and Technology Agency. Funding for open access publication charge: the Ministry of Education, Culture, Sports, Science and Technology of Japan.

Conflict of Interest statement: None declared.

REFERENCES

1. Kawabata,T., Fukuchi,S., Homma,K., Ota,M., Araki,J., Ito,T., Ichiyoshi,N. and Nishikawa,K. (2002) GTOP: a database of protein structures predicted from genome sequences. *Nucleic Acids Res.*, **30**, 294–298.
2. Sugawara,H., Ogasawara,O., Okubo,K., Gojobori,T. and Tateno,Y. (2008) DDBJ with new system and face. *Nucleic Acids Res.*, **36**, D22–D24.
3. Henrick,K., Feng,Z., Bluhm,W.F., Dimitropoulos,D., Doreleijers,J.F., Dutta,S., Flippen-Anderson,J.L., Ionides,J., Kamada,C., Krissinel,E. *et al.* (2008) Remediation of the protein data bank archive. *Nucleic Acids Res.*, **36**, D426–D433.
4. Andreeva,A., Howorth,D., Chandonia,J.M., Brenner,S.E., Hubbard,T.J., Chothia,C. and Murzin,A.G. (2008) Data growth and its impact on the SCOP database: new developments. *Nucleic Acids Res.*, **36**, D419–D425.
5. UniProt_Consortium. (2008) The universal protein resource (UniProt). *Nucleic Acids Res.*, **36**, D190–D195.
6. Finn,R.D., Tate,J., Mistry,J., Coggill,P.C., Sammut,S.J., Hotz,H.R., Ceric,G., Forslund,K., Eddy,S.R., Sonnhammer,E.L. *et al.* (2008) The Pfam protein families database. *Nucleic Acids Res.*, **36**, D281–D288.
7. Hulo,N., Bairoch,A., Bulliard,V., Cerutti,L., Cuche,B.A., de Castro,E., Lachaize,C., Langendijk-Genevaux,P.S. and Sigrist,C.J. (2008) The 20 years of PROSITE. *Nucleic Acids Res.*, **36**, D245–D249.
8. Wilson,D., Madera,M., Vogel,C., Chothia,C. and Gough,J. (2007) The SUPERFAMILY database in 2007: families and functions. *Nucleic Acids Res.*, **35**, D308–D313.
9. Pieper,U., Eswar,N., Davis,F.P., Braberg,H., Madhusudhan,M.S., Rossi,A., Marti-Renom,M., Karchin,R., Webb,B.M., Eramian,D. *et al.* (2006) MODBASE: a database of annotated comparative protein structure models and associated resources. *Nucleic Acids Res.*, **34**, D291–D295.
10. Yeats,C., Lees,J., Reid,A., Kellam,P., Martin,N., Liu,X. and Orengo,C. (2008) Gene3D: comprehensive structural and functional annotation of genomes. *Nucleic Acids Res.*, **36**, D414–D418.
11. Riley,M.L., Schmidt,T., Artamonova,II, Wagner,C., Volz,A., Heumann,K., Mewes,H.W. and Frishman,D. (2007) PEDANT genome database: 10 years online. *Nucleic Acids Res.*, **35**, D354–D357.
12. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
13. Marchler-Bauer,A., Panchenko,A.R., Shoemaker,B.A., Thiessen,P.A., Geer,L.Y. and Bryant,S.H. (2002) CDD: a database of conserved domain alignments with links to domain three-dimensional structure. *Nucleic Acids Res.*, **30**, 281–283.
14. Dunker,A.K., Lawson,J.D., Brown,C.J., Williams,R.M., Romero,P., Oh,J.S., Oldfield,C.J., Campen,A.M., Ratliff,C.M., Hipps,K.W. *et al.* (2001) Intrinsically disordered protein. *J. Mol. Graph. Model.*, **19**, 26–59.
15. Tompa,P. (2005) The interplay between structure and function in intrinsically unstructured proteins. *FEBS Lett.*, **579**, 3346–3354.
16. Wright,P.E. and Dyson,H.J. (1999) Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm. *J. Mol. Biol.*, **293**, 321–331.
17. Minezaki,Y., Homma,K. and Nishikawa,K. (2007) Intrinsically disordered regions of human plasma membrane proteins preferentially occur in the cytoplasmic segment. *J. Mol. Biol.*, **368**, 902–913.
18. Minezaki,Y., Homma,K., Kinjo,A.R. and Nishikawa,K. (2006) Human transcription factors contain a high fraction of intrinsically disordered regions essential for transcriptional regulation. *J. Mol. Biol.*, **359**, 1137–1149.
19. Ward,J.J., Sodhi,J.S., McGuffin,L.J., Buxton,B.F. and Jones,D.T. (2004) Prediction and functional analysis of native disorder in

proteins from the three kingdoms of life. *J. Mol. Biol.*, **337**, 635–645.
20. Homma,K., Kikuno,R.F., Nagase,T., Ohara,O. and Nishikawa,K. (2004) Alternative splice variants encoding unstable protein domains exist in the human brain. *J. Mol. Biol.*, **343**, 1207–1220.
21. Sayle,R.A. and Milner-White,E.J. (1995) RASMOL: biomolecular graphics for all. *Trends Biochem. Sci.*, **20**, 374.
22. Kumar,R., Betney,R., Li,J., Thompson,E.B. and McEwan,I.J. (2004) Induced alpha-helix structure in AF1 of the androgen receptor upon binding transcription factor TFIIF. *Biochemistry*, **43**, 3008–3013.

D16–D18 *Nucleic Acids Research, 2009, Vol. 37, Database issue* *Published online 16 October 2008*
doi:10.1093/nar/gkn724

DDBJ dealing with mass data produced by the second generation sequencer

Hideaki Sugawara, Kazuho Ikeo, Satoshi Fukuchi, Takashi Gojobori and Yoshio Tateno*

Center for Information Biology and DNA Data Bank of Japan, National Institute of Genetics, Research Organization of Information and Systems, Yata, Mishima 411-8540, Japan

Received September 18, 2008; Accepted September 30, 2008

ABSTRACT

DNA Data Bank of Japan (DDBJ) (http://www.ddbj.nig.ac.jp) collected and released 2 368 110 entries or 1 415 106 598 bases in the period from July 2007 to June 2008. The releases in this period include genome scale data of *Bombyx mori*, *Oryzas latipes*, *Drosophila* and *Lotus japonicus*. In addition, from this year we collected and released trace archive data in collaboration with National Center for Biotechnology Information (NCBI). The first release contains those of *O. latipes* and bacterial meta genomes in human gut. To cope with the current progress of sequencing technology, we also accepted and released more than 100 million of short reads of parasitic protozoa and their hosts that were produced by using a Solexa sequencer.

INTRODUCTION

As a member of the International Nucleotide Sequence Database Collaboration (INSDC, http://www.insdc.org/), DDBJ has steadily collected, annotated, released and exchanged the original DNA sequence data, which, for example, is shown by a growth curve of the data submissions in the past years (visit http://www.ddbj.nig.ac.jp/images/breakdown_stats/percentage-e.gif). However, the current situation of data submissions is dramatically changing due to the emergence of ultra high speed or the 2nd generation sequencers (2GS), such as 454 (by 454 Life Sciences, Branford, USA), Solexa (by Illumina, Inc., San Diego, USA), SOLiD (by Applied Biosystems, Foster City, USA) and Helicos (by Helicos BioSciences Corporation, Cambridge, USA). With those machines the whole human genome could now be sequenced at one-thousandth or less speed of the first cases in 2001 (1,2). Recently, two reports announced that the whole genome was sequenced for two well-known persons (3,4), which was perhaps the beginning of personal genomics. Also known is the 1000 human genomes project that is underway in USA, Europe and China to obtain a complete and detailed catalogue of genetic variations of humans (http://www.1000genomes.org/page.php). Those activities warn us that the above growth curve will drastically be steepen. At present, INSDC release about 100 billion bases in total. This is the outcome of the collaboration among the three member banks for >20 years. However, this number will easily be surpassed when the 1000 human genomes project is completed and the result is submitted to INSDC in a few years, or even before that.

To cope with those activities INSDC collaborators discussed in 2008 the attitude towards handling mass submissions produced by 2GS. The common fear among the collaborators was limited computer storages that will sooner or later be filled with continuously coming mass submissions. Nevertheless, the collaborators agreed to collect, distribute and exchange mass data of transcriptomes, such as trace archives (TRA) and short reads (SR), upon the condition that the sequences are assembled. DDBJ has also started to accept and release such mass sequence data. In the following, DDBJ's activity is reported focusing mainly on mass data submissions from Japanese universities and institutes.

COLLECTION OF ORDINARY DATA IN THE PAST YEAR

In the period from July 2007 to June 2008, DDBJ collected, annotated and released the original data of 2 368 110 entries or 1 415 106 598 bases. More than 90% of the data came from Japanese researchers and Japan Patent Office (JPO), and the rest were mainly from researchers in China, Korea and Taiwan.

The released data newly include 282 117 entries of patent data from Korean Industrial Property Office (KIPO) that will continue to send their data to DDBJ for public release. The other portion of the released data contains WGS, GSS (fosmid ends and BAC ends) and HTG (BAC clones) of silkworm (*Bombyx mori*) submitted by National Institute of Agrobiological Sciences; EST entries of medaka (*Oryzas latipes*) submitted by National Institute of Basic Biology; EST entries of *Drosophila simulans*, *D. sechellia* and *D. auraria* submitted

*To whom correspondence should be addressed. Tel: +81 55 981 6857; Fax: +81 55 981 6858; Email: ytateno@genes.nig.ac.jp
The authors wish it to be known that, in their opinion, the all authors should be regarded as joint First Authors.

by Kyoto Institute of Technology and WGS and PLN of *Lotus japonicus* by Kazusa DNA Research Institute. Those data can be obtained at the DDBJ ftp site (http://www.ddbj.nig.ac.jp/ftp_soap-e.html).

It may be worthwhile to refer to the data on *L. japonicus* among them. This plant is widely used as a model organism to study symbiotic nitrogen fixation. This species experienced whole-genome duplication in evolution, and the genome is now composed of six linkage groups that together contain about 30 000 genes (5). The number of the genes is in agreement with that of *Arabidopsis thaliana* for which the number was estimated as 29 500 (6). These results may suggest that the number of genes for an angiosperm species is about 30 000, unless the species has experienced further genome duplication in evolution.

COLLECTION AND RELEASE OF TRA DATA

TRA is a repository of DNA sequence chromatograms (traces), base calls and quality estimates for a single-pass reads from a large-scale sequencing project. TRA data could be useful for confirming SNP sites in question, and, once assembled, provide information for finding new ORFs or genes. With the support by National Project of Integrating Life Science Databases in Japan (ILSD, http://dbcls.rois.ac.jp/en/), we are now able to collect and release TRA data at DDBJ. The released data are as follows.

(1) TRA data of *O. latipes* WGS sequences: The data were submitted by National Institute of Genetics and released at the DDBJ ftp site mentioned above. The data were also sent to National Center for Biotechnology Information (NCBI) TRA Repository (NTR, http://www.ncbi.nlm.nih.gov/Traces/home) and their TI numbers were given by NTR. The total number of entries is about 1.5 millions and the TI numbers without the first three digits (209) are 5 022 956–5 389 675, 5 396 176–6 435 759 and 6 858 496–6 933 759. The length of each entry is several thousand bases. Using any of these numbers one can retrieve at NTR and observe the chromatogram of the entry with the number. The data were also assembled to 24 entries with accession numbers, DG000001–DG00024, (see http://medaka.utgenome.org/ for more details).

(2) TRA data of meta bacterial-genomes in human gut: The data were submitted by University of Tokyo, RIKEN and other universities and institutes (7) and released at the DDBJ ftp site. The samples taken from 13 healthy individuals revealed 237 gene families in the adults and 136 gene families for the infants, though the names of the bacteria in the samples were not identified (7). Another interesting finding is the existence of a conjugative transposon family that could mediate gene transfer between bacteria in the samples (7). Similarly, TI numbers given by NTR without the first three digits (209) are 7 946 941–9 007 079.

COLLECTION OF DATA PRODUCED BY 2GS

2GS, Solexa for example, can produce more than 1 billion sequences per run with the accuracy of 99.9% in several days, though the length of each sequence is very short and thus called SR. However, SR could be valuable if the reference genome sequence to them is available, and assembled against it. In this sense, 2GS is quite powerful for the study of personal (or individual) genomics, population genetics and diagnostic medicine among others. SR data could also be useful for studying the gene expression patterns of a species. Therefore, INSDC set up an archive for SR data as Short Reads Archive (SRA). The participation of DDBJ in SRA is also supported by ILSD.

DDBJ received a tremendous amount of sequence data from Genome Sequence Center of Tokyo University. The submitters used a Solexa machine to sequence full-length cDNAs of eight species, *Plasmodium faliparum*, *P. vivax*, *P. yoelii*, *P. berghei*, *Toxoplasma gondii*, *Cryplosporidium* sp., *Anopheles stephensi* and *Glossina* sp. The first six are parasitic pathogens and the last two are host species. In particular, the first four and the seventh are known to be malarial pathogens and their host, respectively. The length of each entry is 36 or 48 bases due to the specification of Solexa, and the total number of entries is more than 100 millions in the present submission (Table 1). As long as the

Table 1. Species and amounts of submitted short reads

Species	Block	Read Length
Toxoplasma_v2	200	36
Toxoplasma_2nd	300	36
Toxoplasma_v1	300	36
Cryptosporidium_ref	300	36
Cryptosporidium_nref	300	36
Cryptosporidium_2nd	300	36
Plasmodium yoelii_ref	300	36
Plasmodium yoelii	300	36
Plasmodium yoelii_xzl_nref	300	36
Plasmodium yoelii_xzl_ref	300	36
Plasmodium yoelii_xzn_nref	300	36
Plasmodium yoelii_2nd1	300	36
Plasmodium yoelii_2nd2	300	36
P. falciparum_v1	300	36
P. falciparum_2nd1	300	36
P. falciparum_2nd2	300	36
P. falciparum_v1	300	36
P. falciparum_v2	300	36
P. vivax	200	36
P. vivax_ref1	100	36
P. vivax_ref2	100	36
P. vivax_nref	100	36
P. vivax_2nd2	100	36
P. vivax_2nd1	100	36
P. vivax_2nd3	100	36
Babesia bovis_2nd1	100	36
Babesia bovis_2nd2	100	36
P. berghei_2nd	300	36
P. berghei	200	36
Anopheles stephensi_tss	100	48
Anopheles stephensi2nd_1	100	48
Anopheles stephensi2nd_2	100	48
Anopheles stephensi2nd_3	100	48
Glossina_pup_tss	100	36
Glossina_pup_2nd_1	100	48
Glossina_pup_2nd_2	100	48
Glossina_lar_tss	100	36
Glossina_lar_2nd_1	100	48
Glossina_lar2nd_2	100	48

1 block contains 20 000–30 000 SR each of which is 38 or 48 bases in length.

number of entries is concerned, the present submission alone exceeds the total number of ordinary entries that INSDC together have collected and released since 1980. This implies something; the new sequencing technology will perhaps change biology considerably. Individualized biology could emerge in the near future. Namely, biologists would focus intensively on individual genomic characters and the difference between them to elucidate what life really is.

The SR data were released from DDBJ and the SRA repository at NCBI. We have been informed that more SR data will soon be submitted to DDBJ from Japanese universities and institutes. One problem with sending such a tremendous amount of data through Internet would be traffic congestion and an extremely slow rate, even if transmission is possible. We have learned that as long as the data amount is <50 GB the transmission can be done within a few hours. However, we have to resolve two problems to realize and promote individualized biology in the future, capacities of computer and Internet.

REMARKS

As personal genomes can be scrutinized now by the state-of-the-art sequencing technology, one problem emerges. One's genome is not only one's property but also one's ancestors' and descendants'. We are products of evolution. We will not be able to freely publicize the contents of our genomes. The genome of a person hides many recessive inferior genes that are shared with his parents and children (3). In general, children would oppose to sequencing the genome of their parents or *vice versa*. It is thus necessary to pay great care and attention in handling or dealing with person's genome contents.

ACKNOWLEDGEMENTS

We thank all staff of DDBJ for the data collection, annotation, release, management and software development. In particular, we are grateful to Tomohiro Koike and Makoto Yamamoto for their engagements in the collection and release of TRA and SR data.

FUNDING

DDBJ is funded by the Ministry of Education, Culture, Sports, Science and Technology (MEXT) with the management expenses grant for national university cooperation. DDBJ is also supported by a grant from National Project of Integrating Life Science Databases. Funding to pay the open access publication charges for this article was provided by the Japan Society for the Promotion of Science.

Conflict of interest statement. None declared.

REFERENCES

1. Lander,E.S., Linton,L.M., Birren,B., Nusbaum,C., Zody,M.C., Baldwin,J., Devon,K., Dewar,K., Doyle,M., FitzHugh,W. *et al.* (2001) Initial sequencing and analysis of the human genome. *Nature*, **409**, 860–921.
2. Venter,J.C., Adams,M.D., Myers,E.W., Li,P.W., Mural,R.J., Sutton,G.G., Smith,H.O., Yandell,M., Evans,C.A., Holt,R.A. *et al.* (2001) The sequence of the human genome. *Science*, **291**, 1304–1351.
3. Wheeler,D.A., Srinivasan,M., Egholm,M., Shen,Y., Chen,L., McGuire,A., He,W., Chen,Y.-J., Makhijani,V., Roth,G.T. *et al.* (2008) The complete genome of an individual by massively parallel DNA sequencing. *Nature*, **452**, 872–876.
4. Levy,S., Sutton,G., Ng,P.C., Feuk,L., Halpern,A.L., Walenz,B.P., Axelrod,N., Huang,J., Kirkness,E.F., Denisov,G. *et al.* (2008) The diploid genome sequence of an individual human. *PLoS Biol.*, **5**, 2113–2144.
5. Sato,S., Nakamura,Y., Kaneko,T., Asamizu,E., Kato,T., Nakao,M., Sasamoto,S., Watanabe,A., Ono,A., Kawashima,K. *et al.* (2008) Genome structure of the legume, *Lotus japonicus*, *DNA Res.* [Epub ahead of print; doi:10.1093/dnares/dsn008].
6. Alonso,J.M., Stepanova,A.N., Leisse,T.J., Kim,C.J., Chen,H., Shinn,P., Stevenson,D.K., Zimmerman,J., Barajas,P., Cheuk,R. *et al.* (2003) Genome-wide insertional mutagenesis of *Arabidopsis thaliana*. *Science*, **301**, 653–657.
7. Kurosawa,K., Itoh,T., Kuwahara,T., Oshima,K., Toh,H., Toyoda,A., Takami,H., Morita,H., Sharme,V.K., Srivastava,T.P. *et al.* (2007) Comparative metagenomics revealed commonly enriched gene sets in human gut microbiomes. *DNA Res.* [Epub ahead of print; doi: 10.1093/dnares/dsm018].