Table II. Relationship between 5 SNPs in the GHR and 6 linear measurements of body height and craniofacial morphology in 167 Japanese subjects

			Body height (cm)			*N-S (mm)*			*A'-PTM' (mm)*			*Gn-Co (mm)*			*Pog'-Go (mm)*			*Co-Go (mm)*		
		n	*Mean*	*SD*	P	*Mean*	*SD*	P	*Mean*	*SD*	P	*Mean*	*SD*	P	*Mean*	*SD*	P	*Mean*	*SD*	P
C422F	GG	135	161.6	7.9	0.16	69.7	3.4	0.66	50.0	4.8	0.95	122.9	9.3	0.63	79.5	5.6	0.78	61.6	6.5	0.02*
	GT	16	164.6	10.2		69.3	4.4		49.9	3.1		121.7	8.5		79.9	7.2		57.9	6.1	
S473S	CC	137	161.9	8.4	0.95	69.6	3.5	0.32	49.9	4.8	0.71	122.9	9.2	0.89	79.9	5.9	0.31	61.5	6.5	0.54
	CT	11	161.1	6.1		70.5	2.8		49.8	2.1		123.0	9.7		78.5	6.4		60.9	5.6	
P477T	CC	146	161.6	8.3	0.47	69.6	3.5	0.58	49.9	4.7	0.54	122.7	9.1	0.15	79.8	6.0	0.23	61.3	6.4	0.17
	CA	4	163.8	9.1		69.5	4.8		51.3	3.5		130.5	11.4		83.3	5.7		65.6	4.5	
I526L	AA	77	162.7	8.8	0.47	69.5	3.5	0.56	50.1	3.2	0.06	124.2	9.7	0.19	80.1	5.8	0.82	62.4	6.7	0.13
	AC	44	161.0	7.9		70.2	4.0		50.7	3.1		121.6	8.7		79.6	5.5		61.1	6.8	
	CC	32	161.4	6.9		69.4	2.6		48.1	7.9		121.4	7.7		79.5	6.8		59.7	5.4	
P561T	CC	135	161.6	7.9	0.16	69.7	3.4	0.66	50.0	4.8	0.95	122.9	9.3	0.63	79.5	5.6	0.78	61.6	6.5	0.02*
	CA	16	164.6	10.2		69.3	4.4		49.9	3.1		121.7	8.5		79.9	7.2		57.9	6.1	

N-S, Cranial base length; *A'-PTM'*, maxillary length; *Gn-Co*, overall mandibular length; *Pog'-Go*, mandibular corpus length; *Co-Go*, mandibular ramus height.
*$P < 0.05$.

parameter that measures the correlation between alleles showed that 1 SNP pair was tightly correlated (Table IV). Linkage disequilibrium analysis showed that 2 SNPs, C422F and P561T, were in complete linkage disequilibrium.

DISCUSSION

The quantitative genetic aspects of mandibular morphology were explored in reference to an association with the GHR polymorphisms in Japanese people, with an examination of the differences in the allelic frequencies between ethnicities. Mandibular growth greatly depends on cartilage growth and is a multifactorial phenomenon in which genetic disposition, nutrition, homeostasis, hormones, and growth factors interact.[25] The GHR plays an important role in cartilage growth, which directly affects other features of growth and development.[26] In this study, we showed a relationship between the P56IT and C422F variants at the GHR locus and mandibular ramus height in Japanese subjects.

We analyzed 5 SNPs of exon 10 of the GHR in Japanese subjects and found an association between exon 10 of the GHR polymorphisms, P561T and C422F, and mandibular ramus height. Even though a previous study of 95 Han Chinese reported a correlation between the I526L polymorphism of theGHR and mandibular ramus height, our study with 167 Japanese subjects did not replicate this observation. The reason for this discrepancy remains unclear, but we found widely discordant allele frequencies in exon 10 of the GHR between some ethnic groups. However, we could not obtain craniofacial measurements in other ethnic groups. The association of the GHR is different depending on ethnicity

Table III. Haplotypes constructed on the basis of genotypic data for 5 SNPs spanning the linkage disequilibrium block covering exon 10 of the GHR

Haplotype	*C422F*	*S473S*	*P477T*	*I526L*	*P561T*	*Estimated common haplotype frequency*
1	W	W	W	W	W	0.5016
2	W	W	W	v	W	0.3548
3	v	W	W	v	v	0.0693
4	W	v	W	v	W	0.0462

Haplotypes could be subdivided into 4 major haplotypes. Two major haplotypes were present in the Japanese population.
W, Major allele of each variation; *v*, minor allele of each variation.

in other cases such as Laron syndrome[11,12] and idiopathic short stature.[14-19] These differences might imply the need for independent studies of craniofacial morphology for the GHR in each ethnic group. The mandibles of Japanese subjects appear to be slightly smaller those of European Americans[27] or Caucasians.[28-30] Our work emphasizes the importance of close matching of ethnic groups, especially when craniofacial morphology is examined.

It has been hypothesized that an SNP outside exon 10 could also affect the receptor function to influence mandibular ramus height. Specifically, a polymorphism in the human GHR (d3/fl-GHR) resulting in genomic deletion of exon 3 was reported.[31,32] This common polymorphism of the GHR is associated with increased responsiveness to growth hormone: children carrying at least 1 d3-GHR allele show a 1.7 to 2 times greater response to growth hormone than children with the fl-GHR/fl-GHR homozygote allele.[33] Further epidemiologic studies in Japanese and other populations are

Table IV. Allelic correlations among exon 10 SNPs in the GHR

	C422F	*S473S*	*P477T*	*I526L*	*P561T*
D′					
C422F					
S473S	−1.000				
P477T	−1.000	−1.000			
I526L	1.000	0.672	−1.000		
P561T	1.000	−1.000	−1.000	1.000	
r2					
C422F					
S473S	0.004				
P477T	0.001	0.001			
I526L	0.082	0.028	0.009		
P561T	1.000	0.004	0.001	0.078	

Pairwise linkage disequilibrium between GHR SNPs was measured with 2 coefficients: D′ and r2 in Japanese people. The D′ parameter is close to or equal to 1, indicating that few historical recombinations have occurred in the locus. In contrast, the r2 coefficient, which measures the correlation between alleles, varies broadly. The linkage disequilibrium analysis showed that P561 T and C422F were completely linked with each other ($|D'| = 1.000$, $r2 = 1.000$).

required by genotyping other coding SNPs of the GHR including the exon 3 deletion.

Although various environmental factors have been found to contribute to the morphogenesis of the mandible, genetic factors play a substantial role.[34] However, few reports have examined the correlation between craniofacial morphology and genotype, and our results successfully reproduced the correlation between genotype and mandibular ramus height.[8,9,35,36] Sasaki et al[10] reported on a Japanese patient with ectodermal dysplasia and proposed that the P561T variant could be a genetic marker for mandibular growth. The clinical implication for growing patients is the potential advantage to predict mandibular ramus height growth by using DNA from a simple cheek swab. The sample can be collected without extensive training, potentially facilitating genetic studies in dentistry. It would then be possible to predict the orthopedic reaction before starting treatment, although ethnic differences should be considered. This genetic factor might be considered along with other factors associated with mandibular growth in treatment planning for influencing mandibular height, such as Herbst appliances,[37] functional appliances,[38] headgear,[39] and facemask therapy.[40]

We reported previously a genome-wide linkage analysis with 90 Asian sibling pairs with mandibular prognathism and mapped 3 chromosomal loci, including 1p36, 6q25, and 19p13.2.[36] The loci identified for mandibular prognathism are different from the GHR locus on chromosome 5. In our study, we found no SNPs with a relationship with mandibular corpus length or overall mandibular length, and neither relationship was found in the Chinese.[9] The gonial angle is strongly affected by the masseter muscle, and overall mandibular length is affected by the gonial angle.[41,42] The replication of these results with an independent data set should facilitate better understanding of the development of mandibular prognathism.

CONCLUSIONS

We confirmed an association between polymorphisms P561T and C422F that are in linkage disequilibrium, and mandibular ramus height using a larger sample size than our previous report.[8] Subjects with genotype CC of polymorphism P561T and genotype GG of polymorphism C422F had significantly greater mandibular height than those with genotypes CA and GT. Han Chinese, African Americans, European Americans, and Hispanics have different frequencies in exon 10 of the GHR compared with the Japanese. This knowledge provides insight into the molecular pathways associated with growth and development of the mandible, and might be useful for orthodontic diagnosis and orthopedic treatment of the mandible.

We thank Kozue Otaka for her expert technical assistance.

REFERENCES

1. Ramirez-Yanez GO, Smid JR, Young WG, Waters MJ. Influence of growth hormone on the craniofacial complex of transgenic mice. Eur J Orthod 2005;27:494-500.
2. Lewinson D, Bialik GM, Hochberg Z. Differential effects of hypothyroidism on the cartilage and the osteogenic process in the mandibular condyle: recovery by growth hormone and thyroxine. Endocrinology 1994;135:1504-10.
3. Godowski PJ, Leung DW, Meacham LR, Galgani JP, Hellmiss R, Keret R, et al. Characterization of the human growth hormone receptor gene and demonstration of a partial gene deletion in two patients with Laron-type dwarfism. Proc Nat Acad Sci 1989;86: 8083-7.
4. Kaji H, Nose O, Tajiri H, Takahashi Y, Iida K, Takahashi T, et al. Novel compound heterozygous mutations of growth hormone (GH) receptor gene in a patient with GH insensitivity syndrome. J Clin Endocrinol Metab 1997;82:3705-9.
5. Milward A, Metherell L, Maamra M, Barahona MJ, Wilkinson IR, Camacho-Hubner C, et al. Growth hormone (GH) insensitivity syndrome due to a GH receptor truncated after Box1, resulting in isolated failure of STAT 5 signal transduction. J Clin Endocrinol Metab 2004;89:1259-66.
6. Tiulpakov A, Rubstov P, Dedov I, Peterkova V, Bezlepkina O, Chrousos GP, et al. A novel C-terminal growth hormone receptor (GHR) mutation results in impaired GHR-STAT5 but normal STAT-3 signaling. J Clin Endocrinol Metab 2005;90:542-7.
7. Schaefer GB, Rosenbloom AL, Guevara-Aguirre J, Campbell EA, Ullrich F, Patil K, et al. Facial morphometry of Ecuadorian patients with growth hormone receptor deficiency/Laron syndrome. J Med Genet 1994;31:635-9.

8. Yamaguchi T, Maki K, Shibasaki Y. Growth hormone receptor gene variant and mandibular height in the normal Japanese population. Am J Orthod Dentofacial Orthop 2001;119:650-3.
9. Zhou J, Lu Y, Gao XH, Chen YC, Lu JJ, Bai YX, et al. The growth hormone receptor gene is associated with mandibular height in a Chinese population. J Dent Res 2005;84:1052-6.
10. Sasaki Y, Kaida C, Saitoh I, Fujiwara T, Nonaka K. Craniofacial growth and functional change in oligodontia with ectodermal dysplasia: a case report. J Oral Rehabil 2007;34:228-35.
11. Hopp M, Rosenbloom AL, Griffiths J, Kgwete S, Vaccarello MA. Growth hormone receptor deficiency (Laron syndrome) in black African siblings. S Afr Med J 1996;86:268-70.
12. Shevah O, Rubinstein M, Laron Z. Molecular defects of the growth hormone receptor gene, including a new mutation, in Laron syndrome patients in Israel: relationship between defects and ethnic groups. Isr Med Assoc J 2004;6:630-3.
13. Goddard AD, Covello R, Luoh SM, Clackson T, Attie KM, Gesundheit N, et al. Mutations of the growth hormone receptor in children with idiopathic short stature. The Growth Hormone Insensitivity Study Group. N Engl J Med 1995;333:1093-8.
14. Blum WF, Machinis K, Shavrikova EP, Keller A, Stobbe H, Pfaeffle RW, et al. The growth response to growth hormone (GH) treatment in children with isolated GH deficiency is independent of the presence of the exon 3-minus isoform of the GH receptor. J Clin Endocrinol Metab 2006;91:4171-4.
15. Hujeirat Y, Hess O, Shalev S, Tenenbaum-Rakover Y. Growth hormone receptor sequence changes do not play a role in determining height in children with idiopathic short stature. Horm Res 2006; 65:210-6.
16. Bonioli E, Taro M, Rosa CL, Citana A, Bertorelli R, Morcaldi G, et al. Heterozygous mutations of growth hormone receptor gene in children with idiopathic short stature. Growth Horm IGF Res 2005;15:405-10.
17. Sjoberg M, Salazar T, Espinosa C, Dagnino A, Avila A, Eggers M, et al. Study of GH sensitivity in Chilean patients with idiopathic short stature. J Clin Endocrinol Metab 2001;86:4375-81.
18. Sanchez JE, Perera E, Baumbach L, Cleveland WW. Growth hormone receptor mutations in children with idiopathic short stature. J Clin Endocrinol Metab 1998;83:4079-83.
19. Johnston LB, Pashankar F, Camacho-Hubner C, Savage MO, Clark AJ. Analysis of the intracellular signalling domain of the human growth hormone receptor in children with idiopathic short stature. Clin Endocrinol 2000;52:463-9.
20. Roeder K, Bacanu SA, Wasserman L, Devlin B. Using linkage genome scans to improve power of association in genome scans. Am J Hum Genet 2006;78:243-52.
21. Ambrosius WT, Lange EM, Langefeld CD. Power for genetic association studies with random allele frequencies and genotype distributions. Am J Hum Genet 2004;74:683-93.
22. Schork NJ. Power calculations for genetic association studies using estimated probability distributions. Am J Hum Genet 2002;70: 1480-9.
23. Longmate JA. Complexity and power in case-control association studies. Am J Hum Genet 2001;68:1229-37.
24. Schneider S, Kueffer JM, Roesslie D, Excoffier L. Arlequin: a software for population genetic data analysis. Geneva, Switzerland: University of Geneva; 2000.
25. Proffit WR, Fields HW Jr, editors. Contemporary orthodontics. 3rd ed. St Louis: Mosby; 2000.
26. Visnapuu V, Peltomaki T, Ronning O, Vahlberg T, Helenius HJ. Growth hormone and insulin-like growth factor I receptors in the temporomandibular joint of the rat. J Dent Res 2001;80: 1903-7.
27. Miyajima K, McNamara JA Jr, Kimura T, Murata S, Iizuka T. Craniofacial structure of Japanese and European-American adults with normal occlusions and well-balanced faces. Am J Orthod Dentofacial Orthop 1996;110:431-8.
28. Ishii N, Deguchi T, Hunt NP. Craniofacial morphology of Japanese girls with Class II division 1 malocclusion. J Orthod 2001; 28:211-5.
29. Ishii N, Deguchi T, Hunt NP. Morphological differences in the craniofacial structure between Japanese and Caucasian girls with Class II Division 1 malocclusions. Eur J Orthod 2002;24: 61-7.
30. Ishizuka K, Yamazaki T, Inoue K, Kouchi K, Ou B, Namura S. A morphological study of the cranial base and dentofacial structure of Japanese with Angle Class II, div. 1 malocclusion–as compared with American white with Angle Class II, div. 1 malocclusion. Nippon Kyosei Shika Gakkai Zasshi 1989;48:1-6.
31. Urbanek M, MacLeod JN, Cooke NE, Liebhaber SA. Expression of a human growth hormone (hGH) receptor isoform is predicted by tissue-specific alternative splicing of exon 3 of the hGH receptor gene transcript. Mol Endocrinol 1992;6:279-87.
32. Pantel J, Machinis K, Sobrier ML, Duquesnoy P, Goossens M, Amselem S. Species-specific alternative splice mimicry at the growth hormone receptor locus revealed by the lineage of retroelements during primate evolution. J Biol Chem 2000;275: 18664-9.
33. Dos-Santos C, Essioux L, Teinturier C, Tauber M, Goffin V, Bougneres P. A common polymorphism of the growth hormone receptor is associated with increased responsiveness to growth hormone. Nat Genet 2004;36:720-4.
34. Chang HP, Tseng YC, Chang HF. Treatment of mandibular prognathism. J Formos Med Assoc 2006;105:781-90.
35. Lee DG, Kim TW, Kang SC, Kim ST. Estrogen receptor gene polymorphism and craniofacial morphology in female TMJ osteoarthritis patients. Int J Oral Maxillofac Surg 2006;35:165-9.
36. Yamaguchi T, Park SB, Narita A, Maki K, Inoue I. Genome-wide linkage analysis of mandibular prognathism in Korean and Japanese patients. J Dent Res 2005;84:255-9.
37. Flores-Mir C, Ayeh A, Goswani A, Charkhandeh S. Skeletal and dental changes in Class II division 1 malocclusions treated with splint-type Herbst appliances. A systematic review. Angle Orthod 2007;77:376-81.
38. Faltin KJ, Faltin RM, Baccetti T, Franchi L, Ghiozzi B, McNamara JA Jr. Long-term effectiveness and treatment timing for bionator therapy. Angle Orthod 2003;73:221-30.
39. Ulger G, Arun T, Sayinsu K, Isik F. The role of cervical headgear and lower utility arch in the control of the vertical dimension. Am J Orthod Dentofacial Orthop 2006;130:492-501.
40. Baccetti T, Franchi L, McNamara JA Jr. Cephalometric variables predicting the long-term success or failure of combined rapid maxillary expansion and facial mask therapy. Am J Orthod Dentofacial Orthop 2004;126:16-22.
41. Benington PC, Gardener JE, Hunt NP. Masseter muscle volume measured using ultrasonography and its relationship with facial morphology. Eur J Orthod 1999;21:659-70.
42. Fogle LL, Glaros AG. Contributions of facial morphology, age, and gender to EMG activity under biting and resting conditions: a canonical correlation analysis. J Dent Res 1995;74: 1496-500.

Journal of Human Genetics (2009) 54, 543–546

www.nature.com/jhg

SHORT COMMUNICATION

Genome-wide association database developed in the Japanese Integrated Database Project

Asako Koike[1], Nao Nishida[2], Ituro Inoue[3], Shoji Tsuji[4] and Katsushi Tokunaga[2]

The establishment of high-throughput single-nucleotide polymorphism (SNP)-typing technologies has enabled astonishing progress to be made in genome-wide association studies (GWAS), and various novel genetic factors associated with complex diseases have been discovered. Our organization has created a public repository database (DB) to achieve a continuous and intensive management of GWAS data and to facilitate data sharing among researchers. In the GWAS DB, information on study design, quality control protocols, allele frequencies, genotype frequencies and statistical genetic analysis results are stored as publicly available data and can be accessed freely, whereas individual genotyping data and raw data are stored as restricted data and can only be accessed with authorization. All data are presented by a graphic viewer, which is designed to be user friendly for researchers who are not familiar with GWAS to accelerate disease-related studies. Furthermore, the DB allows users to compare various study results obtained by different institutions and on different platforms. The same data are also managed as a distributed annotation system to call up useful data from other DBs and to superimpose them on the GWAS data for help in interpretation. The DB is accessible at https://gwas.lifesciencedb.jp/.
Journal of Human Genetics (2009) **54**, 543–546; doi:10.1038/jhg.2009.68; published online 24 July 2009

Keywords: database; genome-wide association; SNP

INTRODUCTION

The accomplishment of sequencing of the entire human genome[1,2] and the HapMap project,[3] coupled with the development of cost-effective high-throughput dense single-nucleotide polymorphism (SNP)-typing techniques, have enabled a genome-wide exploration of various complex disease-associated variants. Currently, the high-throughput SNP-typing methods are expected to cover about 80% of the human genome in linkage disequilibrium.[4] A number of large-scale genome-wide cohort studies and case–control studies, such as seven common disease GWAS by the Wellcome Trust Case Control Consortium (WTCCC, 2007), have been planned, and some of them are underway. So far, more than 100 loci of disease-related/causing candidates for about 40 common diseases and traits have been identified,[5] and some loci have led to new insights into pathophysiology and etiological pathways. Because GWAS yields large amounts of raw data and analysis results, the management of GWAS data has become a matter of serious concern. Furthermore, more and more grant-funding agencies, journal editors and research communities are beginning to require the disclosure of GWAS data. Disclosure and data sharing of GWAS data will primarily lead to the following three possibilities: (1) meta-analysis using data sets produced in multiple studies to find novel disease-related SNP candidates; (2) re-use of GWAS data combined with other experimental data, including pathway data and expression data, to deepen the exploration of each disease; and (3) development of methods to analyze and compute genetic statistics. In the case of meta-analysis in particular, the use of raw data is indispensable for quality control and for consideration of population structures. Some studies have successfully found additional disease-related SNP candidates on the basis of meta-analysis.[6,7]

The National Center for Biotechnology Information launched the database (DB) of Genotype and Phenotype in the fall of 2006 as a centralized GWAS system to archive and distribute GWAS data. Currently, results funded by the Genetic Association Information Network and voluntarily submitted data have been accumulated. The European Genotype Archive was created in the spring of 2008 as a repository system for phenotype–genotype relationships, and results primarily from WTCCC have been accumulated and redistributed. To achieve a continuous and intensive management of GWAS data and data sharing among researchers, we established a new DB that is publicly available. This DB is expected to have an essential role in providing easily accessible GWAS data to researchers in various biomedical fields. Some disease-related SNPs are assumed to be buried because of their insufficient *P*-values caused by an insufficient number of case–control samples. It is possible that these SNPs will be revealed by combining the GWAS analysis results with other data possessed by users.

In this paper, we introduce the GWAS DB.

[1]Central Research Laboratory, Hitachi Ltd, Tokyo, Japan; [2]Department of Human Genetics, Graduate School of Medicine, University of Tokyo, Tokyo, Japan; [3]Department of Molecular Life Science and Molecular Medicine, Tokai University School of Medicine, Tokyo, Japan and [4]Department of Neurology, Graduate School of Medicine, University of Tokyo, Tokyo, Japan
Correspondence: Dr A Koike, Central Research Laboratory, Hitachi Ltd, 1-280 Higashi-koigakubo Kokubunji, Tokyo, Japan.
E-mail: asako.koike.ea@hitachi.com
Received 3 June 2009; accepted 27 June 2009; published online 24 July 2009

MATERIALS AND METHODS

Database structure

The DB system consists of an internal GWAS DB and a public GWAS DB. For a maximum of 1 year, or until the acceptance of publication, submitted data are stored in the internal GWAS DB and can be accessed only by the research team that submitted the data for greater convenience in data sharing among research team members living in various locations. Currently, the DB systems are implemented using mysql version 5.0 (http://dev.mysql.com/downloads/mysql/5.0.html), and some of the statistical analysis results are also accumulated in a distributed annotation system (DAS) server. A schematic drawing of the GWAS DB is shown in Figure 1.

In this DB, three types of data access, namely, (1) public access, (2) authorized access accompanied by a data use application, and (3) authorized access accompanied by a data use application and its review by a data access committee, are possible. Principally, frequency data of genotypes and alleles and statistical analysis results can be accessed freely. However, automatic access and frequent access are restricted to prevent the release of frequency data of genome-wide genotypes and alleles, as such a large volume of genotype/allele data leads to the specification of whether the given genome is contained in the case or in the control group, as reported previously.[8] These genome-wide frequency data can be obtained by submitting a data use application to the data access committee. For the use of genotype or raw data, an application that

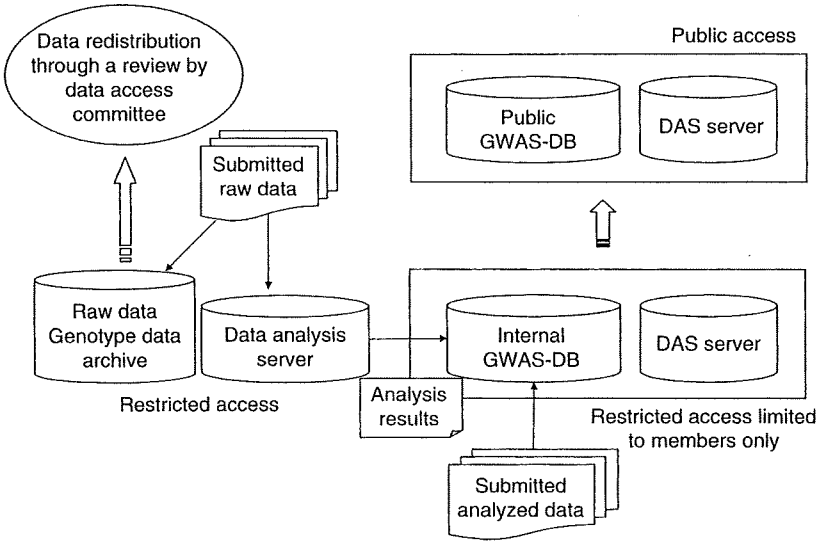

Figure 1 Schematic drawing of genome-wide association study (GWAS) database (DB) systems.

Table 1 Summary of database contents

Contents	*Data sources*
Statistics	
Frequencies of genotypes, alleles and haplotypes	
Statistical genetic analysis	
P-values and odds ratios on genotypic model and allelic model	
P-values and odds ratios on trend model, additive model and recessive model	
Permutation test results	
Bonferroni's corrections and false discovery rate for multiple testing using	
Akaike information criterion	
Hardy–Weinberg equilibrium test	
Haplotype-based χ^2-test	
Epistasis	
Linkage disequilibrium parameters (r^2, D', Lod)	
Other data	
mRNA, amino-acid sequence of each gene	NCBI (http://www.ncbi.nlm.nih.gov/)
mRNA, genome-mapped position	UCSC Hg. 18 (http://hgdownload.cse.ucsc.edu/)
SNP position and SNP kind (cSNP, sSNP, rSNP and so on)	NCBI (http://www.ncbi.nlm.nih.gov/)
OMIM	NCBI (http://www.ncbi.nlm.nih.gov/)
Copy number variation	DGV (http://projects.tcag.ca/variation/)
Gene function	Gene ontology (http://www.geneontology.org/)
Microsatellite polymorphism	UCSC (http://hgdownload.cse.ucsc.edu/)
Manually curated disease-related mutation information	

describes the research purpose and lists the research team members must be submitted to the data access committee. The data access committee deliberates on whether the applicant's research purpose meets the content of the consent form. Only applicants approved by the review committee can use individual genotype data and raw data in accordance with the data handling security rules required by the data access committee and following data use restrictions on the basis of informed consent.

Individual data and raw data are accumulated in the server in a secured computer environment that is different from the public DB server. Only authorized persons can access this server.

Data submission

In principal, both analysis results and unanalyzed data can be submitted. When data have already been analyzed, the analyzed data are accumulated in this DB, along with a detailed description of the analysis protocols. When data have not been analyzed yet, they are analyzed in our site, and the results are accumulated in this DB. When raw data are redistributable under certain conditions, they are also submitted with the contents of the consent form. All data must be submitted with documents explaining the design of the study, as well as ethical consideration.

Data cleaning for quality control

When data are submitted as individual data without analysis results, they are analyzed as follows: (1) SNPs with a call rate <95% and samples with a call rate <95% are removed. (2) SNPs, the Hardy–Weinberg equilibrium test result of which in a control group is less than 0.001 or the minor allele frequency of which is less than 0.05, are removed. (3) The principle component analysis (PCA) of these case–control data, along with HapMap data, is carried out using EIGENSTRAT[9] or other programs so that sample outliers and samples with a possible ethnic mixture or a different ethnicity are removed on the basis of the PCA result. Sample outliers in the plot of heterozygosity versus call rate are also removed. The quantile–quantile plot based on the allelic model is calculated and checked. When only genotype frequency data are submitted, PCA and heterozygosity checks are skipped,

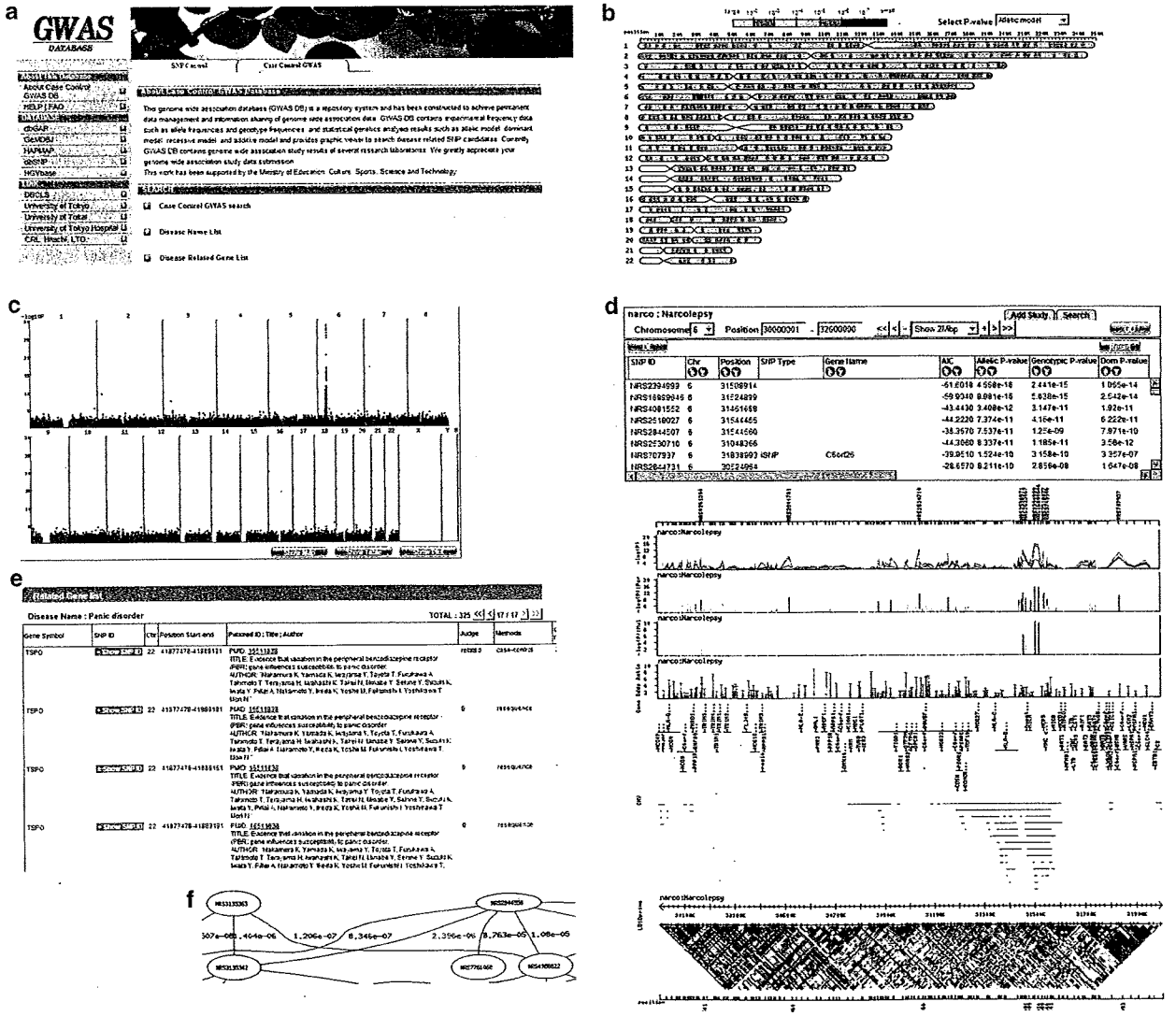

Figure 2 Snapshots of the genome-wide association study (GWAS) database. (**a**) Top page, (**b**) bird's-eye view, (**c**) Manhattan plot, (**d**) region table and graph, (**e**) disease-related gene/single-nucleotide polymorphism (SNP) lists (public data) and (**f**) SNP network based on epistasis.

as they require individual data. The cleaning results are linked from 'study details' on the web.

Data analysis

Standard statistical genetic analyses are performed by plink[10] and Haploview.[11] Additional analyses such as the Akaike information criterion, epistasis and more complicated ones (for example, genetic analysis considering potential case samples existing in the control samples, which sometimes becomes a concern for diseases that develop in old age) are calculated by internally developed programs. The major statistics include *P*-values based on an allelic model, genotypic model, trend model, dominant model, recessive model and permutation test results of these models; and Bonferroni's correction and false discovery rate for multiple testing. These methods are also shown in 'study details.' When submitted data consist of only genotype frequency data, the genome-wide permutation test is skipped.

Database contents and utility

The DB contents (as of April 2009) are summarized in Table 1.

User data other than GWAS data, such as expression data and epigenetic data, are also accumulated and can be displayed on the graph. Although clinical data are not currently accumulated in the DB, they can be added if submitted. Major tables are summarized in Supplementary Table 1.

A snapshot of the GWAS DB is shown in Figure 2. Figure 2a shows the top page of the GWAS DB. When the 'SNP control' tab is selected, the interface jumps to the SNP control DB, which is affiliated to the GWAS DB and contains allelic frequencies, genotypic frequencies, Hardy–Weinberg equilibrium tests and estimated haplotype frequencies of Japanese control samples. Bird's-eye view (Figure 2b) and Manhattan plot (Figure 2c) are provided to draw *P*-values of each model. A genome region can be selected from both (Figures 2b and c), and the results of statistical genetic analysis along with other information such as exon–intron information and copy number variations (CNVs) can be displayed in tables and graphs to facilitate the identification of disease-related SNPs, as shown in Figure 2d. Furthermore, comparisons among various study results obtained by different institutions and/or different platforms can be carried out easily by plotting their graphs on the web (using the 'add study' function in Figure 2d). When the published disease-related gene or SNP is registered as shown in Figure 2e, data are plotted as a known disease-related gene/SNP in the graph (Figure 2d). Epistasis data are also accumulated and drawn as a network graph using Graphviz (http://www.graphviz.org/), as shown in (Figure 2f). Data can be searched by SNP ID (dbSNP ID #rs, affymetrix SNP ID and so on), gene name, disease name and so on. The study design and analysis protocols can also be browsed.

Statistical results are also accumulated on a DAS server, and they can be browsed using the Gmod Gbrowse (http://gmod.org/wiki/Main_Page)-based browser (http://gwas.lifesciencedb.jp/cgi-bin/gbrowse/snpdb/). Furthermore, as a function of the DAS server, data on other DAS servers such as Ensemble can be called up. This function is useful to superimpose data from other DBs onto GWAS data. The GWAS DB is designed to be user friendly for researchers unfamiliar with GWAS to promote disease-related studies.

Further development

A recent topic of interest is genome-wide association analysis coupled with other data such as pathway data[12] to compensate for the low statistical power in disease-associated candidate SNPs. The function to browse or calculate SNP/SNP pair *P*-values on the basis of the GWAS result, along with other data, will be added to this DB to facilitate the generation and understanding of user hypotheses.

The relationships between CNVs and diseases have begun to emerge in recent studies.[13] Although concerns remain about the quality of detected CNVs, genomic locations and frequencies of CNV regions and their case–control association study results will be incorporated into this DB. Furthermore, in the near feature, new high-throughput techniques such as short-read sequencing will be applied for GWAS, and this DB will be improved to suit the new experimental techniques.

ACKNOWLEDGEMENTS

This work was supported by the contract research fund 'Integrated Database Project' from the Ministry of Education, Culture, Sports, Science, and Technology of Japan.

1 Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409,** 860–921 (2001).
2 Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G. *et al.* The sequence of the human genome. *Science* **291,** 1304–1351 (2001).
3 The International HapMap Consortium. A haplotype map of the human genome. *Nature* **437,** 1299–1320 (2005).
4 Barrett, J. C. & Cardon, L. R. Evaluating coverage of genome-wide association studies. *Nat. Genet.* **38,** 659–662 (2006).
5 Manolio, T. A., Brooks, L. D. & Collins, F. S. A HapMap harvest of insights into the genetics of common disease. *J. Clin. Invest.* **118,** 1590–1605 (2008).
6 Zeggini, E., Scott, L. J., Saxena, R., Voight, B. F., Marchini, J. L., Hu, T. *et al.* Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes. *Nat. Genet.* **40,** 638–645 (2008).
7 Houlston, R. S., Webb, E., Broderick, P., Pittman, A. M., Di Bernardo, M. C., Lubbe, S. *et al.* Meta-analysis of genome-wide association data identifies four new susceptibility loci for colorectal cancer. *Nat. Genet.* **40,** 1426–1435 (2008).
8 Homer, N., Szelinger, S., Redman, M., Duggan, D., Tembe, W., Muehling, J. *et al.* Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays. *PLoS Genet.* **4,** e000167 (2008).
9 Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A. & Reich, D. Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* **38,** 904–909 (2006).
10 Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81,** 559–575 (2007).
11 Barrett, J. C., Fry, B., Maller, J. & Daly, M. J. Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics* **21,** 263–265 (2005).
12 Baranzini, S. E., Galwey, N. W., Wang, J., Khankhanian, P., Lindberg, R., Pelletier, D. *et al.* Pathway and network-based analysis of genome-wide association studies in multiple sclerosis. *Hum. Mol. Genet.* **18,** 2078–2090 (2009).
13 McCarroll, S. A. Extending genome-wide association studies to copy-number variation. *Hum. Mol. Genet.* **17** (R2), R135–R142 (2008).

Supplementary Information accompanies the paper on Journal of Human Genetics website (http://www.nature.com/jhg)

INFORMATICS

Human Mutation

OFFICIAL JOURNAL

HGVS HUMAN GENOME VARIATION SOCIETY

www.hgvs.org

The Phenotype and Genotype Experiment Object Model (PaGE-OM): A Robust Data Structure for Information Related to DNA Variation

Anthony J. Brookes,[1]* Heikki Lehvaslaiho,[2] Juha Muilu,[3] Yasumasa Shigemoto,[4] Takashige Oroguchi,[5] Takeshi Tomiki,[6] Atsuhiro Mukaiyama,[7] Akihiko Konagaya,[8] Toshio Kojima,[9] Ituro Inoue,[10] Masako Kuroda,[11] Hiroshi Mizushima,[12] Gudmundur A. Thorisson,[1] Debasis Dash,[13] Haseena Rajeevan,[14] Matthew W. Darlison,[15] Mark Woon,[16] David Fredman,[17] Albert V. Smith,[18] Martin Senger,[19] Kimitoshi Naito,[5] and Hideaki Sugawara[20]

[1]University of Leicester, Department of Genetics, Leicester, United Kingdom; [2]South African National Bioinformatics Institute, University of Western Cape, Bellville, South Africa; [3]Institute for Molecular Medicine Finland (FIMM), Helsinki, Finland; [4]BioIT Business Development Unit, Fujitsu Limited, Tokyo, Japan; [5]Japan Biological Informatics Consortium, Strategic Planning Department, Tokyo, Japan; [6]NEC Soft, Ltd., VALWAY Technology Center, Tokyo, Japan; [7]AXIOHELIX Co. Ltd., Tokyo, Japan; [8]Department of Computer Science, Tokyo Institute of Technology, Tokyo, Japan; [9]Advanced Computational Sciences Department, RIKEN, Yokohama, Japan; [10]Department of Molecular Genetics, University of Tokai, Isehara, Japan; [11]Department of Advanced Databases, Japan Science and Technology Agency, Tokyo, Japan; [12]Information Center for Medical Sciences, Tokyo Medical and Dental University, Tokyo, Japan; [13]Institute of Genomics and Integrative Biology, Council of Scientific and Industrial Research (CSIR), Genomics Nanotechnology and Robotics (GNR) Knowledge Centre for Genome Informatics, Delhi, India; [14]Department of Genetics, Yale University, New Haven, Connecticut; [15]Centre for Health Informatics and Multiprofessional Education (CHIME) London, University College London (UCL), United Kingdom; [16]Department of Genetics, Stanford University, Stanford, California; [17]Bergen Center for Computational Science, University of Bergen, Bergen, Norway; [18]Icelandic Heart Association, Kopavogur, Iceland; [19]Crop Research Information Laboratory, International Rice Research Institute, Manila, Philippines; [20]Center for Information Biology and DNA Data Bank of Japan (DDBJ), National Institute of Genetics, Mishima, Japan

Communicated by Richard G. H. Cotton

Received 12 November 2008; accepted revised manuscript 19 December 2008.
Published online 18 March 2009 in Wiley InterScience (www.interscience.wiley.com). DOI 10.1002/humu.20973

ABSTRACT: Torrents of genotype–phenotype data are being generated, all of which must be captured, processed, integrated, and exploited. To do this optimally requires the use of standard and interoperable "object models," providing a description of how to partition the total spectrum of information being dealt with into elemental "objects" (such as "alleles," "genotypes," "phenotype values," "methods") with precisely stated logical inter-relationships (such as "A objects are made up from one or more B objects"). We herein propose the Phenotype and Genotype Experiment Object Model (PaGE-OM; www.pa-geom.org), which has been tested and implemented in conjunction with several major databases, and approved as a standard by the Object Management Group (OMG). PaGE-OM is open-source, ready for use by the wider community, and can be further developed as needs arise. It will help to improve information management, assist data integration, and simplify the task of informatics resource design and construction for genotype and phenotype data projects.
Hum Mutat 30, 968–977, 2009.

KEY WORDS: bioinformatics; data model; genotype–phenotype; database

Introduction

Individual genomes vary extensively, and much of this variation can impact disease and other phenotypes. Technological progress has made it possible to study such genotype to phenotype (G2P) relationships in a genome-wide manner, and deep whole-genome resequencing may soon be economically available as the ultimate experimental strategy [Mardis, 2008]. To complement this, clinical sample biobanks have been steadily growing in size and proficiency, providing large-scale resources to support the G2P field [Smith et al., 2005]. Consequently, new G2P correlations are being identified with increasing frequency, and the pressure is on to use this elemental information in the most optimal fashion—both for improved biomedical understanding and in the context of drug development and clinical practice. To enable this, databases and informatics resources must be developed to support the data-handling challenges posed by vast numbers of dispersed and multifarious G2P datasets. Those systems must be able to interoperate on many levels of data processing—such as security, validation, integration, exchange, interrogation, presentation, and analysis.

To achieve the desired widespread interoperability, G2P data systems must be based upon well-designed and robust standards. The role of standards and unified effort in modern biomedicine is

Heikki Lehvaslaiho and Juha Muilu contributed equally to this work.
David Fredman's current address: Department for Molecular Evolution and Development, University of Vienna, Vienna, Austria.
*Correspondence to: Anthony J. Brookes, University of Leicester, Department of Genetics, Leicester, UK. E-mail: ajb97@leicester.ac.uk

increasingly paramount, and reflected by coordination initiatives such as the Human Genome Epidemiology–Strengthening the Reporting of Genetic Association studies (HuGE/STREGA; www.cdc.gov/genomics/hugenet) and the National Cancer Institute–National Human Genome Research Institute (NCI-NHGRI) guidelines [Chanock et al., 2007] regarding genetic association studies, the Human Variome Project [Cotton et al., 2007], and the Public Population Project in Genomics (P3G) biobanking initiative [Knoppers et al., 2008]—all of which help to guide best practice in the creation of primary G2P datasets. But once created, these datasets need to be electronically disseminated and utilized. To standardize such operations, the way particular data components are named—the "semantics" of the data—must be carefully controlled. Precise and detailed ontologies, vocabularies, and nomenclatures are therefore being developed to support the G2P field. Finally, to enable informatics systems to work together in processing data content, the structure of the data—its "syntax"—must also be controlled so that it matches (or can be made to match) that of an agreed standard.

The structure of data is described by way of an "object model," which may also be called a "data model." This provides a way to compartmentalize the domain of interest into its principal elements, and define how these "objects" relate to one another. For example, a G2P object model could involve objects called *Genomic_variation* and *Variation_assay*, and associate these to indicate which *Variation_assay* can interrogate which *Genomic_variation*. This would suffice for singleplex assays, but some *Variation_assays* are multiplex in nature (i.e., able to score simultaneously more than one site of *Genomic_variation*). Therefore, one might wish to rename *Variation_assay* as *Multi_variation_assay* and include a third and distinct model component called *Variation_assay*—i.e., the concept of a subsection (e.g., oligonucleotides) of a *Multi_variation_assay* specifically involved in scoring one of the multiplex set of *Genomic_variations*. For users of the two above models to merge their lists of variations and assays, they must both be explicit regarding which model they are using, and rules must be available that dictate how to convert data from one structure to the other. Once this is done, and the specifications are published and made freely available, then future information technology (IT) developers can quickly and easily adopt optimal models without having to repeatedly tackle the same complex modeling challenges. The systems they develop will then be syntactically interoperable with other projects that use the same (or equivalent) object models, and tasks such as data submission to, or between, depositories will be greatly simplified. Furthermore, as the subject matter of the G2P field further evolves, new data features and modeling solutions can be fed back into the standard object model, thereby keeping G2P data resources current in design and fully interoperable.

Many object modeling projects are now underway across various biomedical domains, not least the MicroArray and Gene Expression (MAGE) object model [Spellman et al., 2002], the Proteomics Standard Initiative Model for Molecular Interaction (PSI-MI) data [Hermjakob et al., 2004], the Functional Genomics Experiment (FuGE) initiative [Jones et al., 2007], and the Health Level Seven Clinical Genomics Model (HL7-CGM; www.hl7.org). For G2P research, however, merely a few isolated projects have reported modeling initiatives; such as an Extensible Markup Language (XML)-specific model created by the Pharmacogenetics and Pharmacogenomics Knowledge Base (PharmGKB) database [Whirl-Carrillo et al., 2008], the Genomic Sequence Variation Markup Language (GSVML) (see entry for ISO/DIS 25720, Health Informatics–GSVML; www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber = 43182), and the Extensible Genotype and Phenotype Model (XGAP; www.xgap.org). Consequently, genetic investigations such as mutation detection, association analysis, linkage studies, gene knockouts, and (re-)sequencing presently lack a standard object model. To address this deficit, we assembled an international consortium of 20 groups engaged in genotype–phenotype projects, and formulated the Phenotype and Genotype Experiment Object Model (PaGE-OM), as presented here. Subsequent efforts will be needed to move towards full data interoperability between PaGE-OM and models from allied domains, such as those listed above, and cross-project collaborations would be helpful in bringing this about.

The current specification of PaGE-OM aims to strike a balance between being too generic (as would be required to support any and all G2P data management situations) and too specific (as would be required if it were to support just one experimental paradigm). Nevertheless, the goal is to enable the structured capture of at least the minimum amount of information required to properly report most genetic experiments involving genotype and/or phenotype information. The model's subcomponents could be tailored to suit particular applications—and any such further developments should be fed back into the PaGE-OM specification to increase its utility.

Materials and Methods

Technical Objective

The PaGE-OM project was instigated to create a specification for a platform-independent conceptual object model that is able to provide a common solution and framework for the management of DNA variation data, phenotype data, and G2P experimental findings. It is not intended to include a platform-specific implementation, such as a relational database or a World Wide Web Consortium (W3C) XML Schema—though the latter has been developed as part of the Object Management Group (OMG) validation process (XML schema v1.0b2 at the project website). The solution is not dependent upon, nor does it provide, any particular G2P domain ontology, though the names employed for its component objects are carefully chosen and precisely defined.

Technical Presentation

PaGE-OM was built around five core domains: GENOTYPE, PHENOTYPE, EXPERIMENT, SAMPLE, and COMMON. Within each domain, the range of information to be modeled was segmented into a number of logical, elemental, and precisely defined data objects. These components are joined by lines of "association" to indicate all the permitted, rational interrelationships between the various parts. These associations also specify possible cardinalities, for example to declare that "one" Genomic-variation can have "one or many" (but not "zero") component Alleles. In figures, open arrowheads signify subclass to superclass relationships, and open diamond arrowheads signify aggregation type relationships (wherein one class object represents the thing created by a collection of the other class).

The figures in this work are limited to those that present a high-level overview of the complete model, and these were generated directly from the most current development version (PaGE-OM v1.2), which itself is evolved from the formal OMG specification of December 2008 (PaGE-OM v1.0b2). For purposes of clarity and explanation, inherited attributes are not shown for subclasses, and singular and plural forms of class names are used interchangeably,

whereas only the singular form is valid in the formal PaGE-OM model. Each PaGE-OM object name is shown italicized when referred to in the text (i.e., as *Object_name*), and in use case examples in figures the object instances are shown capitalized (i.e., as OBJECT).

Development Procedure

PaGE-OM was developed by an international consortium of domain experts by way of a series of meetings and online collaboration. This consortium previously provided the Polymorphism Markup Language (PML) model, now registered by the OMG as the "Single Nucleotide Polymorphisms Specification" (www.omg.org/cgi-bin/apps/doc?dtc/05-06-02.pdf). PaGE-OM was developed from PML, and PaGE-OM v1.0 was accepted (March 2008) as an OMG standard, after which the model became a formal OMG specification after an implementation was demonstrated (December 2008). PaGE-OM is a fully-open standard, and community interaction and participation is strongly encouraged. Complete documentation, descriptions of emerging implementations, case examples (presented as "schemalets"), a first-version XML specification, and modes of communication are available online (www.pageom.org). When reviewing PaGE-OM at this website, it should be noted that class diagrams are reused from earlier versions of the model (modules "SNP" and "SNP2"), and so these should be considered as integral parts of PaGE-OM.

PaGE-OM development employed Enterprise Architect software (Sparx Systems, Creswick, Victoria, Australia; www.sparxsystems.com.au) and the Unified Modeling Language (UML). The UML model consists of classes that represent objects, and the associations between these objects. Most associations were made bidirectional, deferring directionality to specific implementations. This allows for flexible but consistent implementation of PaGE-OM to suit multiple purposes; e.g., to describe multiple assays per marker in a Laboratory Information Management System or multiple markers scored by a single assay in an association database entry.

Results

PaGE-OM is designed to support diverse activities involving data components related to the genome, the phenome, and data that correlate the two. The model is species-independent, and able to support both clinical and research undertakings. At the highest level, PaGE-OM separates genotype and phenotype information into two distinct domains (GENOTYPE and PHENOTYPE), with these being optionally connected via a third domain (EXPERIMENT). A SAMPLE domain is then provided to structure data pertaining to study subjects that may be investigated. Finally, there is a COMMON domain, which specifies various object concepts relevant throughout PaGE-OM. Below, we provide a simplified abstraction of PaGE-OM, to illustrate the main design features. Complete details of the model, case "schemalets," and an XML implementation, should be sought at the project website (www.pageom.org).

SAMPLE Domain

The SAMPLE domain specifies the PaGE-OM structure for information that characterizes study subjects and their derivative samples. It covers the various "classes" of biological resources that might be used to generate genotype, phenotype, or G2P data, namely; *Molecular_sample*, meaning biological samples such as blood DNA taken from a study subject; *Individual*, meaning a complete study subject; *Panel*, meaning a set of similar study subjects; and *Abstract_population*, meaning a broad collection or populace of one or more study subjects. Pedigrees are not formally modeled via a distinct class, but can be specified by simply listing all first degree relatives for each *Individual*. A family group could also, optionally, be listed as a *Panel*. Logical associations between the SAMPLE classes were then elaborated, as shown in Figure 1.

Panels are naturally comprised of *Individuals*, and the cardinality of this relationship is "zero or many to zero or many" (i.e., *Panels* can have no or up to many *Individuals* specified for them, and *Individuals* can be represented in no or up to many *Panels*). This aggregation type of relationship is indicated in the model by a line that joins these two entities, with an open diamond drawn where the line joins the *Panel* class along with "0..*" (asterisk meaning many) at each end. The *Panel* class additionally has a "zero or one to zero or many" association with itself, to allow for situations where one *Panel* may be split into many derivative *Panels*. This association is indicated by a line running from, and back to, this class. *Molecular_samples* are derived from *Individuals*, with one *Individual* potentially providing no or up to many *Molecular_samples*. In contrast, a *Molecular_sample* can only be stated to have originated from no or up to one *Individual*. Therefore, this association is represented by an adjoining line with "0..1" at the *Individual* end and "0..*" at the *Molecular_sample* end. The *Molecular_sample* class then has its own recursive association with itself, as *Molecular_samples* could be subdivided to give further *Molecular_samples*.

The *Abstract_population* class captures population specific information, such as ethnicity and language, that may apply to *Individuals* or *Panels*, but within PaGE-OM this class is not primarily intended to represent a population in the usual sense of the word (of any scale, either within or between studies). Instead, *Abstract_population* is being used as a modeling construct called a "superclass" to represent a generalization of other "subclasses"—in this case *Panel* and *Individual*. It can therefore be largely ignored by the casual reader. This kind of association is symbolized by adjoining lines that carry special open arrowheads, and no cardinality is specified for such relationships. In the modeling diagram, and in real-world implementations of PaGE-OM, the *Abstract_population* class is able to function as either of its subclasses while also allowing for additional data elements to be represented (e.g., ethnicity and language). Another way to state this is to say that *Panels* and *Individuals* are being handled in the model as specialized forms of *Abstract_population*. One important consequence of this is that any logical lines of associations drawn to *Abstract_population* from any other class would be equally valid if drawn directly to either of its subclasses.

Abstract_observation_target is the final class in the SAMPLE domain, and this provides a way to represent any biological entity upon which an investigation might be performed; i.e., a *Molecular_sample* or an *Abstract_population* (and therefore also its subclasses *Individual* and *Panel*). It is thus presented as a superclass to each of these subclasses. The *Abstract_observation_target* class provides a convenient means to represent the whole of the SAMPLE domain in high-level views of PaGE-OM.

GENOTYPE Domain

The GENOTYPE domain of PaGE-OM specifies a structure for data components that relate to the genome and its testing in the laboratory. It is based around modern genetic and genomic modes of experimentation. PaGE-OM should therefore support most

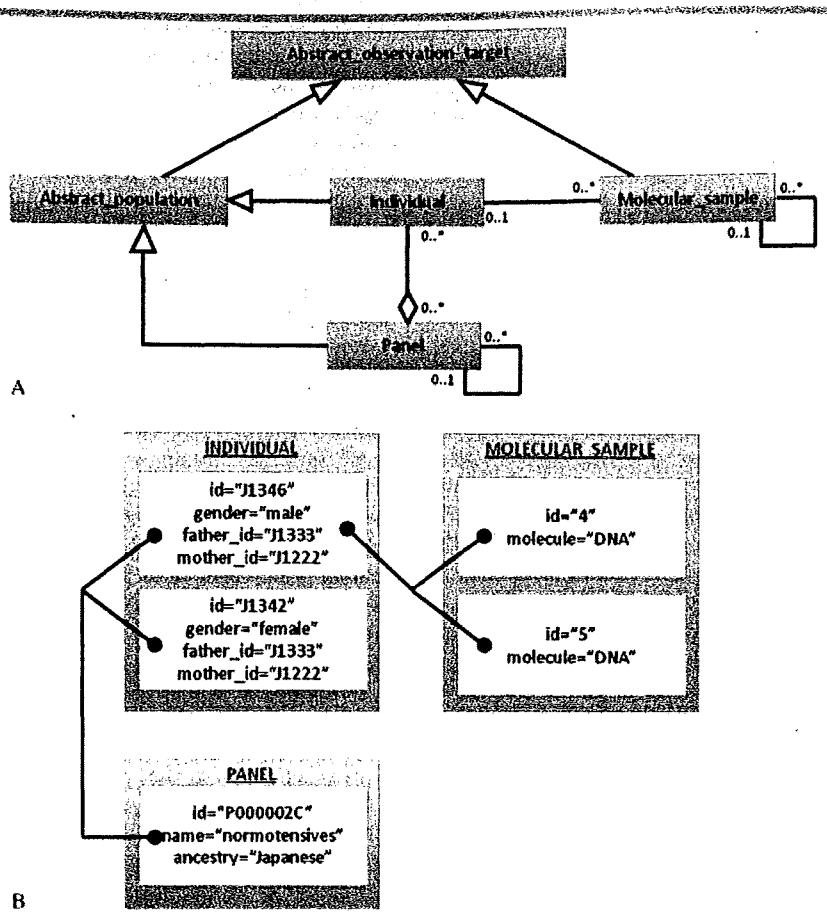

Figure 1. SAMPLE domain of PAGE-OM. **A:** The principal classes (colored blue) and class relationships from the SAMPLE domain, as described in the text. **B:** Shows how the model in (A) could be used to represent a cohort of normotensive Japanese, giving further details for a brother and sister from that cohort, and indicating two DNA samples taken from the male individual. The *Abstract_population* class is not used in this example use case, as its primary role is as a modeling superclass.

activities wherein singleplex or multiplex genotyping of predefined DNA sequences is performed to establish which of one or more possible alleles is present in one or more *Abstract_observation_targets*. Due to ongoing technical advances, this kind of data is growing rapidly in scale, implying an urgent need for a supporting object model. PaGE-OM should serve this purpose, at least for qualitative detection of "simple" sequences and sequence variations. The model has not yet been validated for use upon more complex challenges, such as quantitative genotyping of alleles, assessment of methylation, detection of DNA copy-number differences, or next-generation sequencing of extensive DNA stretches or genomes—though these activities should be possible to support via PaGE-OM, given small extensions to the model that would be allowed for by the system's flexible design. Such work is ongoing, driven by the consortium that has produced PaGE-OM to date, in partnership with the Genotype-to-Phenotype (GEN2-PHEN) project (www.gen2phen.org).

As shown in Figure 2, the GENOTYPE structure is built around the class called *Genomic_variation*, designed to represent what are commonly termed "markers"; i.e., short sequences of DNA from an organism's genome, within which a particular string of one or more bases may vary. The *Genomic_allele* class is used to list the one or more sequence alternatives for the variable segment (commonly termed "alleles"), and this is joined to the *Genomic_variation* class by an aggregation type of relationship. Each *Genomic_variation* may be genotyped by the deployment of zero or up to many *Variation_assays*, and additionally the model includes a *Multi_variation_assay* class that operates as elaborated in the Introduction (though for simplicity this is not shown in Fig. 2).

Upon using a *Variation_assay* to interrogate an *Abstract_observation_target* of type *Molecular_sample* or *Individual*, a single genotyping result is generated. This data is captured by the *Assayed_genomic_genotype* class, via its associations to *Abstract_observation_target* and *Variation_assay*, as well as by a direct relationship to the *Genomic_variation* class for scenarios in which no *Variation_assay* has been specified or recorded.

In genotyping studies, however, only certain *Assayed_genomic_genotypes* will be valid for any one *Genomic_variation*, based upon its constituent *Genomic_alleles* (e.g., testing a T/C human autosomal SNP could not generate a G:T heterozygote genotype), and so PaGE-OM includes a class called *Latent_genotype* to represent these valid alternatives. The *Latent_genotype* class is therefore associated via an aggregation type of relationship with the *Genomic_allele* class where its potential constituents would be listed, and it is also associated with the *Assayed_genomic_genotype* class to rationally constrain permitted values for each "measured genotype." But this is only the first of two possible ways the *Latent_genotype* concept can be used. It may also be employed to list the set of genotypes that a particular *Variation_assay* is actually able to detect—since some genotyping methods for some markers may fail to resolve all possible valid genotypes. This "detectable

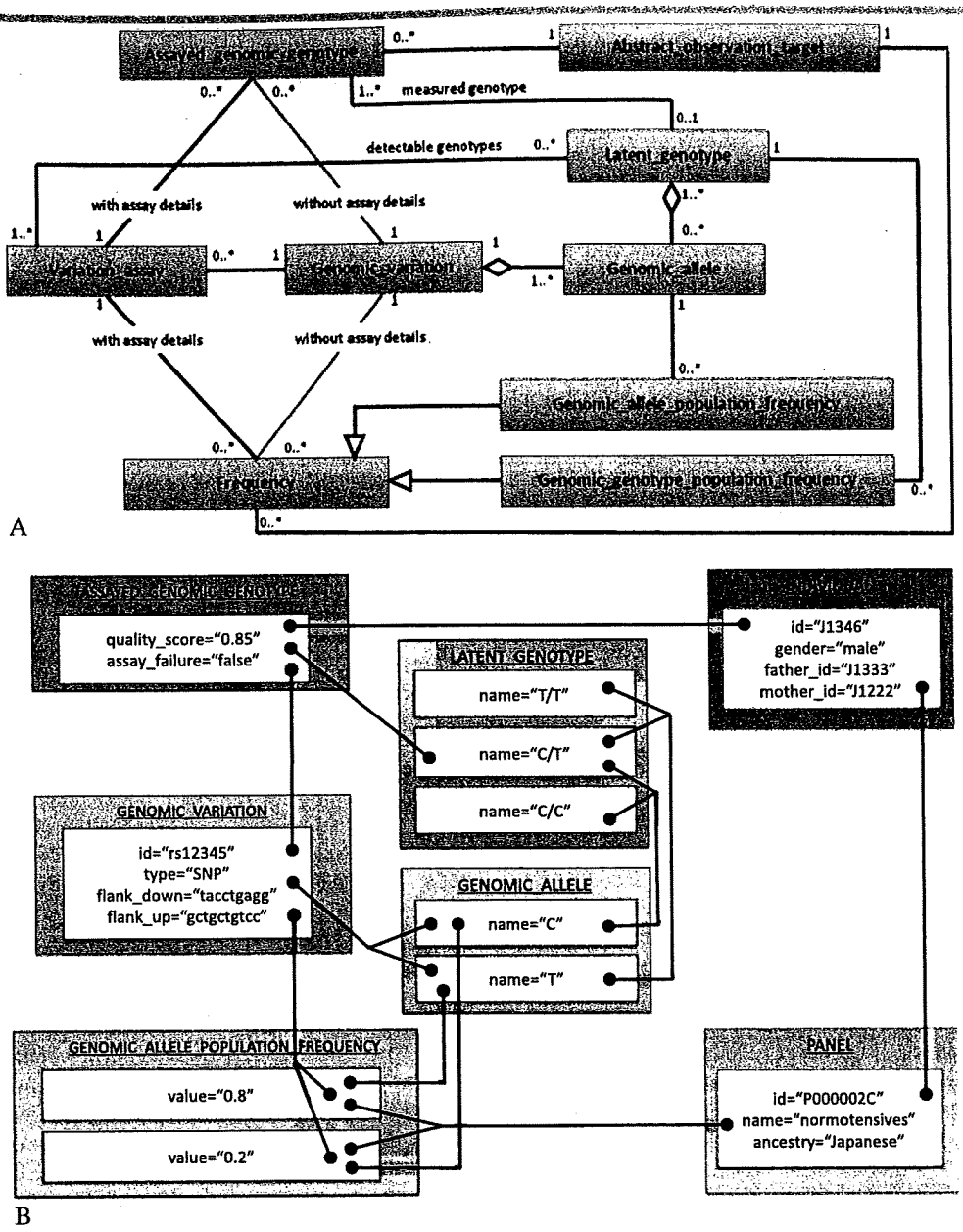

Figure 2. GENOTYPE domain of PAGE-OM. **A:** The principal classes (colored red) and class relationships from the GENOTYPE domain, as described in the text. One additional class (colored blue) is also included, taken from the SAMPLE domain. At the project website, sections of the model called Marker, Frequency, and Assay are provided to represent subsections of the GENOTYPE domain. As indicated, the model offers a choice between using interclass relationships "with assay details" and "without assay details," for scenarios in which assays details are or are not being considered, respectively. Similarly, the model makes a distinction between using the *Latent_genotype* class to process data on "detectable genotypes" (theoretical genotypes that an assay could produce) and "measured genotypes" (genotypes produced in a real sample). **B:** Shows how the model in (A) could be used to represent typical genotyping results, indicating the detection of a C/T genotype (1/3 possible genotypes) at marker rs12345 in one individual from a Japanese normotensive cohort, plus allele frequency data for this marker in that total cohort. Assay details are not being recorded in this example, but this would be possible via the *Variation_assay* class. Likewise, the cohort's genotype frequency data are not presented, but this would be possible via the *Genomic_genotype_population_frequency* class.

genotype" role is enabled via an association between *Latent_genotype* and *Variation_assay*, and it will become increasingly important as more complex forms of DNA variation become examined in the future.

In addition to single genotype results, marker frequency data also needs to be handled. This is achieved by including a *Frequency* class to carry actual frequency values, and connecting this to the *Abstract_observation_target* and *Variation_assay* classes. *Frequency* is also directly associated to the *Genomic_variation* class so that frequencies can be meaningfully presented in scenarios where no *Variation_assay* is identified. In reality of course, marker frequency data is made up of both allele frequency and genotype frequency data. Reflecting this, the *Frequency* concept represents a superclass that sits over two subclasses *Genomic_allele_population_frequency* and *Genomic_genotype_population_frequency*. The first of these is associated with the *Genomic_allele* class so that one can state which allele the frequency value refers to, and the second is associated with the *Latent_genotype* class to specify the valid genotype whose frequency is being stated. One further superclass of note is called *Genomic_observation*. This is not shown in Figure

2 for simplicity, but it sits over the subclasses *Assayed_genomic_genotype*, *Frequency*, and *Genomic_allele*, and it is intended to represent any of the above result types from a genetic analysis.

PHENOTYPE Domain

The PHENOTYPE domain of PaGE-OM specifies a structure for data that relates to any conceivable phenotype. The solution is designed to be equally applicable to human and model organism studies, to clinical and research phenotypes, to descriptions of molecules, cells, tissues, or whole organisms, and to quantitative as well as categorical traits. This implies extreme diversity and complexity for the phenotype realm that needs to be supported, and to solve this modeling problem we devised a simple and elegant way to partition the concept of "a phenotype" into its fundamental components.

In PaGE-OM the term "phenotype" is considered to have three fundamental elements. First, there is the "feature" of the phenotype, such as "blood pressure at rest"—meaning the concept that an individual at rest has a certain blood pressure that can be measured. Second, there is the "method" of the phenotype, such as "manual use of an upper arm pressure cuff plus stethoscope with subject seated and rested for 5 minutes"—meaning the precise way in which the phenotype was assessed. This component is important, because while some similar measurement regimes will be equivalent in what they assess, others will actually report on different phenotype features and/or have differing degrees of accuracy. For instance, the given example would not be equivalent to measuring blood pressure immediately after exercise, nor necessarily equivalent if the measurement were performed by an automated cuff and pulse detector. Third, there is the "value" of the phenotype, such as "high blood pressure of 160/90 mmHg"—meaning the actual finding generated by measuring the blood pressure. This example also nicely illustrates how there are two subconcepts in the value component: 1) any number of primary measurement values (in this case two values, 160 and 90 mmHg for systolic and diastolic pressures); and 2) the single value conclusion or inference (namely "high blood pressure"), which is typically derived from the primary measurements. Some phenotype value datasets will comprise information relating to both these subconcepts, whereas others may only need to use just one of them.

As shown in Figure 3, to reflect the feature+method+value conceptualization of a phenotype, PaGE-OM has classes named *Observable_feature*, *Observation_method*, and *Observed_value*. The root of these names is "Observation" rather than "phenotype," since as well as using these classes to support phenotype data we anticipate also using them to handle environmental data. Work is now underway to validate this utility, but until that is complete we do not formally sanction this extended use of the model. Nevertheless, to signal this intended dual usage, the *Observable_feature* class is here presented as a superclass over both *Phenotype_feature* and *Environment_feature* subclasses.

Sitting over *Observable_feature* is a class called *Observable_feature_category*, which provides a flexible means by which *Observable_features* can be variously classified. For example, one might implement a categorization based upon anatomic scale, and/or one based upon a disease classification, and/or one might use controlled keywords. These categorizations will sometimes derive their list of available options from formalized ontologies. Using ontologies here also means that the logical interrelationships between available categories is predefined, and such useful structures are then automatically propagated down to *Observable_features* connected to the various ontology terms (e.g., "Type II Diabetes Disease Status" might be defined in a disease ontology to have "subphenotypes" such as "Body Mass Index" and "Glucose Tolerance"). This organization of terms is managed in PaGE-OM via the recursive self-association indicated for *Observable_feature_category*.

A "one to zero or many" association connects the *Observable_feature* and the *Observation_method* classes, since each *Observable_feature* may be defined by no or up to many different phenotype methods (though preferably at least one). Similarly, a "one to zero or many" association is placed between the *Observation_method* and the *Observed_value* classes, since each *Observation_method* may be referencing no or up to many different sets of measurement values. The two level conceptual split of measurement values into measured and inferred types is conveniently allowed for by establishing a recursive self-association for the *Observed_value* class, with the manner of distinction between primary and inferred value types being discretionary and managed at the level of model implementation.

EXPERIMENT Domain

The EXPERIMENT domain of PaGE-OM specifies a structure that brings together data from the GENOTYPE and PHENOTYPE domains, along with experimental result information that elucidates how genetic variations influence phenotypic variation. It is based upon data elements traditionally employed for reporting experimental investigations in manuscripts and similar reports. In that respect, this part of PaGE-OM has a lot in common with the FuGE object model [Jones et al., 2007].

As shown in Figure 4, at the top of the EXPERIMENT domain lays the *Study* class, which acts to hold summary level information, such as the title, abstract, background, hypothesis, conclusion, and acknowledgement parts of a scientific manuscript. This class has an aggregation type of relationship to a class called *Genotype-phenotype_correlation_experiment*, representing the set of experiment subsections that would normally be listed in the results section of a G2P manuscript. As such, each *Genotype-phenotype_correlation_experiment* would typically be accompanied by statements regarding the experiment's objective and outcome. A class called *Experiment_result* is then provided to capture the distinct primary results that came out of an experiment (such as the allele-association p-value for a SNP tested in a case–control association study), and this is connected to *Genotype_phenotype_correlation_experiment* via a zero or many to zero or many relationship.

The *Experiment_result* class provides the natural location in the EXPERIMENT domain, where connections should to be made to components from the GENOTYPE and PHENOTYPE domains to substantiate the *Experiment_result* entry. To this end, associations are provided from *Experiment_result* to the following other classes: *Abstract_observation_target*, to state the utilized study subject materials; *Observable_feature*, to state the phenotype(s) being investigated; *Observed_value*, to state the phenotype measurement(s) being considered; *Genomic_variation*, to state the marker(s) examined; and *Genomic_observation*, to state the genotype measurements being considered.

COMMON Domain

The COMMON domain provides discrete classes of general utility, the need for which is common across PaGE-OM. Key examples include *Identifiable*, *Annotation*, and *Db_xref*, though there are several other such classes in the total model. *Identifiable* provides a standard way to provide an identifier value and a

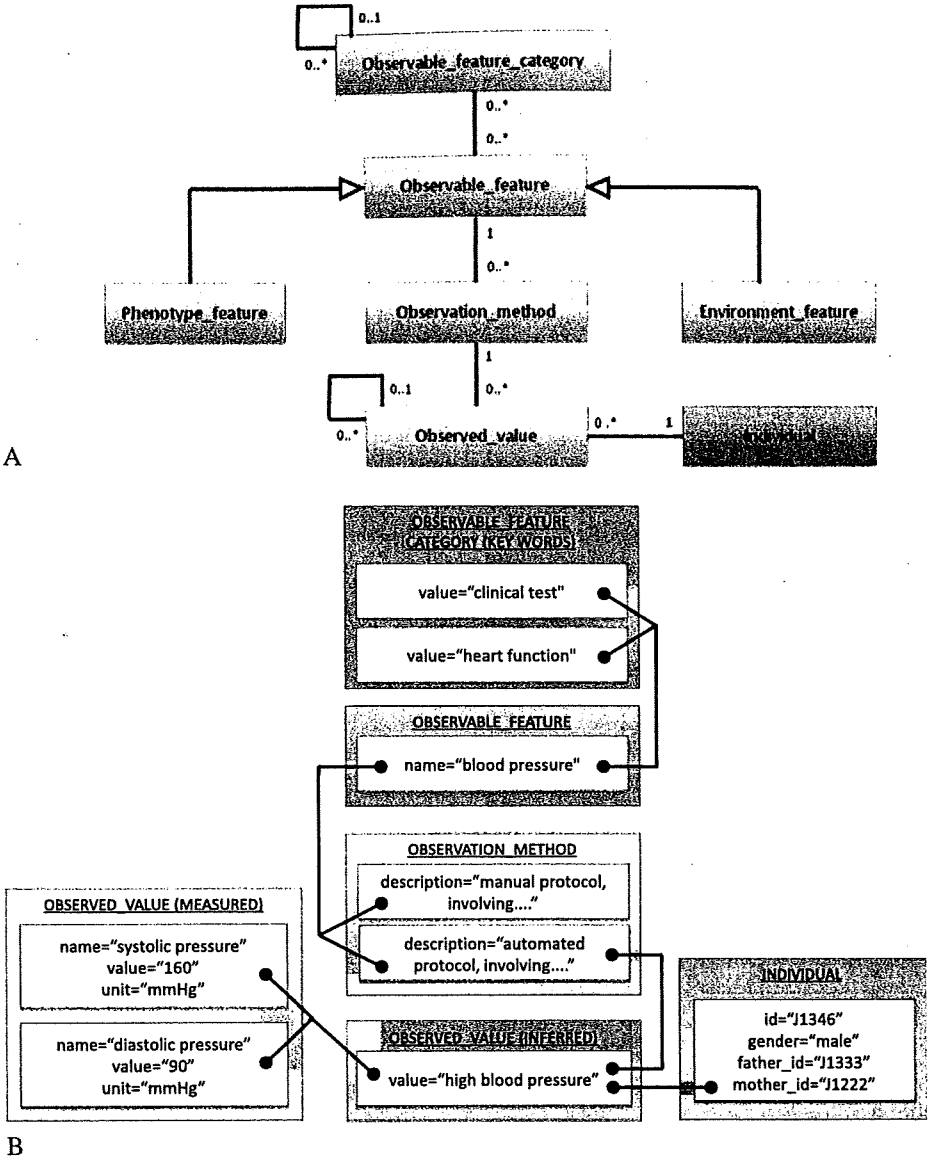

Figure 3. PHENOTYPE domain of PAGE-OM. **A:** The principal classes (colored purple) and class relationships from the PHENOTYPE domain, as described in the text. One additional class (colored blue) is also included, taken from the SAMPLE domain. **B:** Shows how this model could be used to represent a situation in which the blood pressure of an individual has been measured using a specific automated protocol (rather than an alternative manual protocol) and the systolic-diastolic blood pressure ratio is thereby found to be 160/90 mmHg, which is summarized as "high blood pressure." The "blood pressure" phenotype could be categorized in many different ways to aid in subsequent data analysis and integration, with this example showing the use of keywords, of which two are provided.

descriptive name for any other class in the model that can logically have such attributes. A special case of *Identifiable* would be *Ontology_term* (taken from FuGE [Jones et al., 2007]), which specifies a vocabulary system that must be used. *Annotation* likewise assists by providing a standard way to place annotations on entities, and *Db_xref* provides a universal means to assign cross-links to other websites or database entries on the web. Using these COMMON classes greatly simplifies data modeling and provides streamlined utility in implementations where all objects must be accessed on an equal footing. *Value* is another powerful support class in the COMMON domain, and it is used whenever the type of a value cannot be stated in advance. For example, the *Observed_value* for phenotypes might sometimes be a string or sometimes a numeric value, or even a set of values. The solution is, therefore, to simply reference the *Value* class, wherein the value type is stated and controlled as needed. Overall, the many different COMMON domain classes of PaGE-OM are very much aligned to those of equivalent domains in other data models.

Discussion

Current and future developments of PaGE-OM are occurring at a time of rapid change for the G2P data field. A recent review of this subject, which places into context both PaGE-OM and many of the resources and projects mentioned in this manuscript, has recently been published [Thorisson et al., 2009b]. It was against this backdrop that the PaGE-OM consortium became motivated by the urgent need for a robust G2P object model, given that no suitable generic solution yet existed.

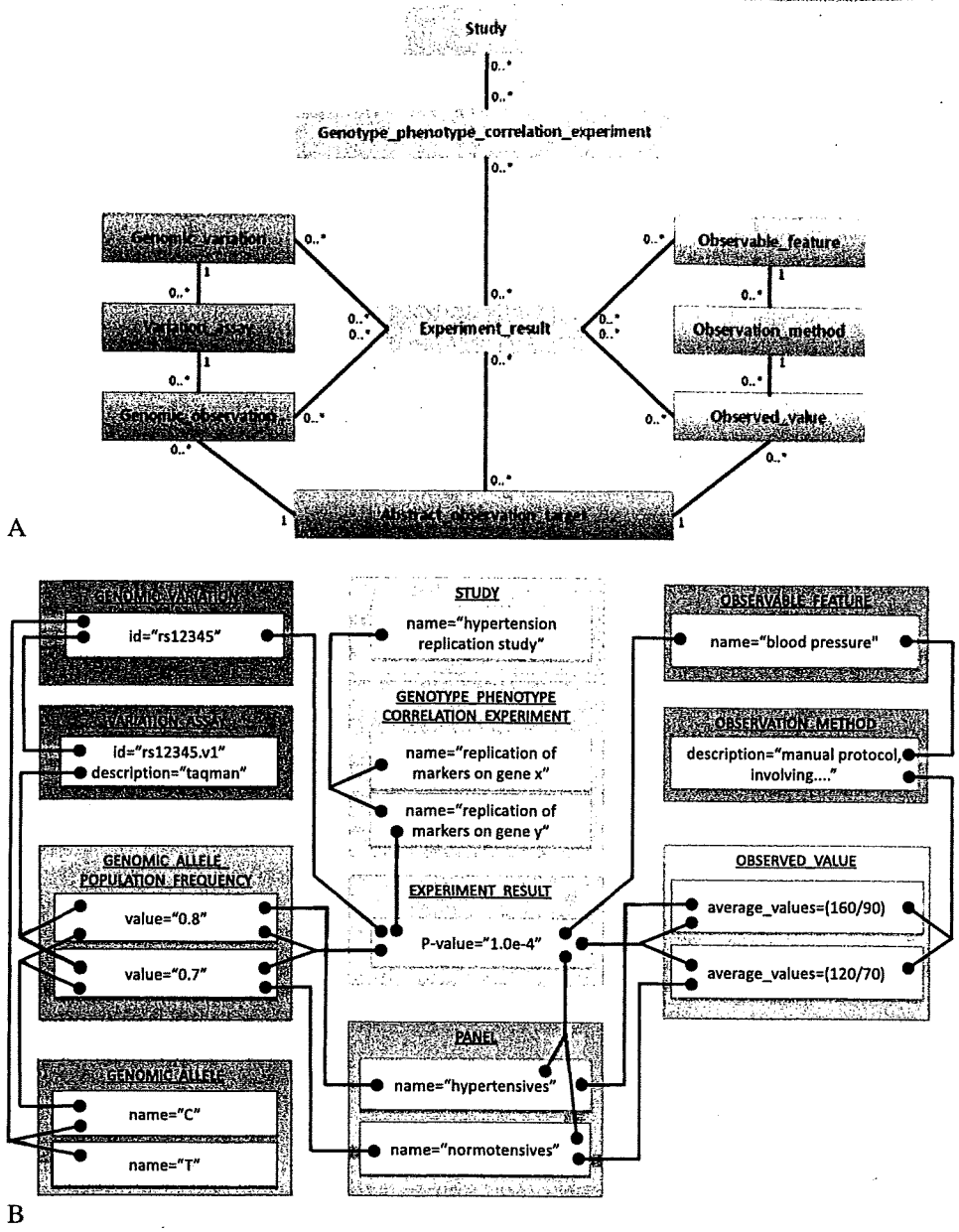

Figure 4. EXPERIMENT domain of PAGE-OM. **A:** Illustrates the principal classes (colored yellow) and class relationships from the EXPERIMENT domain, as described in the text. Additional classes are also included, taken from the SAMPLE (colored blue), GENOTYPE (colored red), and PHENOTYPE (colored purple) domains. **B:** Shows how this model could be used to represent data from a replication genetic association study into hypertension, composed of multiple experiments on different genes. Further details are given for the experiment on "gene y," specifically showing the outcome of a simple allele frequency association test on marker rs12345, which revealed the C allele to be a risk factor, given its increased frequency in hypertensives compared to normotensive controls. Generic and ancillary information about the study and its component experiments would be stored in those sections of the model. If there were redundancy regarding aspects of the Sample, Genotype, or Phenotype information underlying multiple results, then these data instances could be related directly to the experiment or study sections of the model, rather than to the individual results as presently shown.

Initial development efforts produced the PML, which was formally approved as a standard by the OMG in December 2005 (www.omg.org/technology/documents/formal/snp.htm). That basic model, which dealt with only DNA-related information, was further refined and extended to produce the complete PaGE-OM that itself has recently (March 2008) been accepted as an OMG standard, with formal approval being scheduled for mid-2009. PML comprised both a platform independent object model, as well as a platform-specific data exchange format based upon XML. Both the PML model and its exchange format were successfully tested with real datasets by the Human Genome Variation Database of Genotype-to-Phenotype Information (HGVbaseG2P; www.hgvbaseg2p.org) [Fredman et al., 2004], International Haplotype Mapping (HapMap) project database (www.hapmap.org) [Thorisson et al., 2005], dbSNP (www.ncbi.nlm.nih.gov/projects/SNP) [Sherry et al., 2001], PharmGKB (www.pharmgkb.org) [Altman, 2007], Indian Genome Variation database (IGVdb; http://igvdb.res.in) [Indian Genome Variation Consortium, 2005], Japanese SNP database (JSNP; http://snp.ims.u-tokyo.ac.jp) [Hirakawa et al., 2002], and Allele Frequency Database (ALFRED; http://alfred.med.yale.edu) [Rajeevan et al., 2003]. Small changes and several new classes were subsequently included to create the

PaGE-OM platform-independent object model, which has now been used effectively as the basis for a full database implementation to generate an XML exchange format specification, and the HGVbaseG2P database (www.hgvbaseg2p.org) [Thorisson et al., 2009a]. It has also been validated with respect to datasets from dbGaP (www.ncbi.nlm.nih.gov/gap), PharmGKB (www.pharmgkb.org) [Altman, 2007], and several locus specific databases. PaGE-OM continues to be improved, with the latest version available for inspection online (www.pageom.org).

Further work on PaGE-OM could proceed in a number of different directions. The field it supports continues to evolve rapidly (e.g., the emerging need to handle copy-number variation and resequencing data) and new use cases are arising all the time—implying the need to constantly evaluate and adapt the model to address these new challenges. Furthermore, the model could be increasingly aligned with other initiatives, such as MAGE and FUGE, to optimize data integration possibilities between fields. Such work is now underway, and will be reported elsewhere. Additionally, simpler versions of PaGE-OM could be extracted from the full model, tailored to the needs of particularly common use cases, and data exchange specifications for each could be created. Examples of this, called "schemalets," are available at the project website. Support tools could also be devised to aid groups in their uptake and further development of PaGE-OM. All these ideas for taking PaGE-OM forward are being considered, and several of them are being worked upon by the GEN2PHEN project (www.gen2phen.org). But it is important to emphasize that PaGE-OM is a fully-open-source project that is not "owned" by any team or institute, and any group that wishes to work further on the model are welcomed and encouraged to do so, either independently or in partnership with the authors of this work and/or the GEN2PHEN initiative.

In its current form, PaGE-OM will be of use in supporting many of the most common G2P data uses in the field, including data capture (from experiments and the published literature), data storage, and data exchange applications. For example; a company whose business involved DNA analysis kits might use only the *Genomic_variation* and *Variation_assay* parts of the model. In contrast, a genome variation database might employ multiple parts of the GENOTYPE and the SAMPLE domains. Projects involving clinical data would have a need for the PHENOTYPE and SAMPLE domains, and if their activities extended to DNA analysis then the GENOTYPE and the EXPERIMENT domains could also be deployed. These few examples illustrate the modularity and flexibility of PaGE-OM, as well as the general usability of the model in quite diverse scenarios.

Most domains of PaGE-OM encompass well-recognized data components for which the use of the model should be straightforward. The PHENOTYPE domain is, however, rather more open to interpretation and hence worthy of further explanation. First, the model's structure is such that an *Observable_feature* must always be accompanied by a sufficiently complete *Observation_method* if any *Observed_values* are to be given, as this method component is essential for meaningful interpretation of the phenotype data. Another benefit of recognizing the centrality of this method concept is that it enables one to clearly identify where one phenotype ends and another begins. The guiding principle would be that when one applies a single *Observation_method* then the results produced represent or demarcate the extent of one phenotype. In more complex situations, such as the use of questionnaires to gather phenotype data, each question should be entered as a distinct *Observable_feature* plus *Observation_method* pairing, so that the responses to identical questions can be integrated across results for different persons. The recursive association provided at the level of the *Observable_feature_category* can then be used, via a "list of questionnaires" categorization set, to group together the different questions within a questionnaire. Another complex use case would be the representation of quantitative phenotype data derived from a *Panel* of *Individuals*. In this situation, values that describe a distribution (e.g., maximum, minimum, median, standard deviation) would be entered as the primary *Observed_values*, and a summary statement for this distribution would be entered as the single *Observed_value* conclusion or inference.

In conclusion, PaGE-OM is now available as a useful object model to support G2P activities. However, it provides only one aspect of what is needed to move toward full data interoperability in this bioscience area. Infrastructure components, minimal dataset requirements, data exchange technologies, and ontologies must also be increasingly improved and harmonized. As a platform independent object model PaGE-OM in no way limits these options, and may even help guide some the choices that are made.

Acknowledgments

The research leading to these results has received funding from the University of Leicester, European Bioinformatics Institute, Karolinska Institute, University of Helsinki, National Center for Biotechnology Information, Cold Spring Harbor Laboratory, Stanford University, Yale University, Shanghai Center for Bioinformation Technology, Shanghai Information Center for Life Sciences, Tsinghua University, Indian Institute of Genomics & Integrative Biology, Japan National Institute of Genetics, Japan Science and Technology Agency, Japanese National Cancer Center Research Institute, Tokyo Institute of Technology, Japanese Ministry of Economy Trade and Industry, New Energy and Industrial Technology Development Organization, Functional Genomics Programme (FUGE) of the Research Council of Norway, YFF program of the Research Council of Norway and Bergen Forskningsstiftelse, GlaxoSmithKline, NIH grant U01GM61374 (PharmGKB project), NSF grant BCS0096588 (ALFRED Project), the European Community's Fifth Framework Programme under grant agreement QLG2-CT-2002-01254 (The GENOMEUTWIN project) and the European Community's Seventh Framework Programme under grant agreement 200754 (the GEN2PHEN project). We acknowledge the valuable intellectual contributions made by Masashi Tanaka (Tokyo Metropolitan Institute of Gerontology, Tokyo, Japan) and Tokio Kano (Japan Biological Informatics Consortium, Tokyo, Japan).

References

Altman RB. 2007. PharmGKB: a logical home for knowledge relating genotype to drug response phenotype. Nat Genet 39:426–426.

Chanock SJ, Manolio T, Boehnke M, Boerwinkle E, Hunter DJ, Thomas G, Hirschhorn JN, Abecasis G, Altshuler D, Bailey-Wilson JE, Brooks LD, Cardon LR, Daly M, Donnelly P, Fraumeni Jr JF, Freimer NB, Gerhard DS, Gunter C, Guttmacher AE, Guyer MS, Harris EL, Hoh J, Hoover R, Kong CA, Merikangas KR, Morton CC, Palmer LJ, Phimister EG, Rice JP, Roberts J, Rotimi C, Tucker MA, Vogan KJ, Wacholder S, Wijsman EM, Winn DM, Collins FS. 2007. Replicating genotype-phenotype associations. Nature 447:655–660.

Cotton RGH, Appelbe W, Auerbach AD, Becker K, Bodmer W, Boone DJ, Boulyjenkov V, Brahmachari S, Brody L, Brookes A, Brown AF, Byers P, Cantu JM, Cassiman JJ, Claustres M, Concannon P, Cotton RG, den Dunnen JT, Flicek P, Gibbs R, Hall J, Hasler J, Katz M, Kwok PY, Laradi S, Lindblom A, Maglott D, Marsh S, Masimirembwa CM, Minoshima S, de Ramirez AM, Pagon R, Ramesar R, Ravine D, Richards S, Rimoin D, Ring HZ, Scriver CR, Sherry S, Shimizu N, Stein L, Tadmouri GO, Taylor G, Watson M. 2007. Recommendations of the 2006 Human Variome Project meeting. Nat Genet 39:433–436.

Fredman D, Munns G, Rios D, Sjoholm F, Siegfried M, Lenhard B, Lehvaslaiho H, Brookes AJ. 2004. HGVbase: a curated resource describing human DNA variation and phenotype relationships. Nucleic Acids Res 32:D516–D519.

Hermjakob H, Montecchi-Palazzi L, Bader G, Wojcik J, Salwinski L, Ceol A, Moore S, Orchard S, Sarkans U, von Mering C, Roechert B, Poux S, Jung E, Mersch H, Kersey P, Lappe M, Li Y, Zeng R, Rana D, Nikolski M, Husi H, Brun C, Shanker K, Grant SG, Sander C, Bork P, Zhu W, Pandey A, Brazma A, Jacq B, Vidal M, Sherman D, Legrain P, Cesareni G, Xenarios I, Eisenberg D, Steipe B, Hogue C, Apweiler R. 2004. The HUPO PSI's molecular interaction format—a community standard for the representation of protein interaction data. Nat Biotechnol 22:177–183.

Hirakawa M, Tanaka T, Hashimoto Y, Kuroda M, Takagi T, Nakamura Y. 2002. JSNP: a database of common gene variations in the Japanese population. Nucleic Acids Res 30:158–162.

Indian Genome Variation Consortium. 2005. The Indian Genome Variation database (IGVdb): a project overview. Hum Genet 118:1–11.

Jones AR, Miller M, Aebersold R, Apweiler R, Ball CA, Brazma A, Degreef J, Hardy N, Hermjakob H, Hubbard SJ, Hussey P, Igra M, Jenkins H, Julian Jr RK, Laursen K, Oliver SG, Paton NW, Sansone SA, Sarkans U, Stoeckert Jr CJ, Taylor CF, Whetzel PL, White JA, Spellman P, Pizarro A. 2007. The Functional Genomics Experiment model (FuGE): an extensible framework for standards in functional genomics. Nat Biotechnol 25:1127–1133.

Knoppers B, Fortier I, Legault D, Burton P. 2008. Population genomics: the public population project in genomics (P(3)G): a proof of concept? Eur J Hum Genet 16:664–665.

Mardis ER. 2008. The impact of next-generation sequencing technology on genetics. Trends Genet 24:133–141.

Rajeevan H, Osier MV, Cheung KH, Deng H, Druskin L, Heinzen R, Kidd JR, Stein S, Pakstis AJ, Tosches NP, Yeh CC, Miller PL, Kidd KK. 2003. ALFRED: the ALelle FREquency Database. Update. Nucleic Acids Res 31:270–271.

Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, Sirotkin K. 2001. dbSNP: the NCBI database of genetic variation. Nucleic Acids Res 29: 308–311.

Smith GD, Ebrahim S, Lewis S, Hansell AL, Palmer LJ, Burton PR. 2005. Genetic epidemiology and public health: hope, hype, and future prospects. Lancet 366:1484–1498.

Spellman PT, Miller M, Stewart J, Troup C, Sarkans U, Chervitz S, Bernhart D, Sherlock G, Ball C, Lepage M, Swiatek M, Marks WL, Goncalves J, Markel S, Iordan D, Shojatalab M, Pizarro A, White J, Hubley R, Deutsch E, Senger M, Aronow BJ, Robinson A, Bassett D, Stoeckert CJ, Brazma A. 2002. Design and implementation of microarray gene expression markup language (MAGE-ML). Genome Biol 3:RESEARCH0046.

Thorisson GA, Smith AV, Krishnan L, Stein LD. 2005. The International HapMap Project Web site. Genome Res 15:1592–1593.

Thorisson GA, Lancaster O, Free RC, Hastings RK, Sarmah P, Dash D, Brahmachari SK, Brookes AJ. 2009a. HGVbaseG2P: a central genetic association database. Nucleic Acids Res 37(Database issue):D797–D802.

Thorisson GA, Muilu J, Brookes AJ. 2009b. Genotype-phenotype databases: challenges and solutions for the post-genomic era. Nat Reviews Genet 10:9–18.

Whirl-Carrillo M, Woon M, Thorn CF, Klein TE, Altman RB. 2008. An XML-based interchange format for genotype-phenotype data. Hum Mutat 29:212–219.

Journal of Human Genetics (2009) 54, 615–623

www.nature.com/jhg

REVIEW

Meta-analysis of genetic association studies: methodologies, between-study heterogeneity and winner's curse

Hirofumi Nakaoka[1,2] and Ituro Inoue[1]

Meta-analysis is a useful tool to increase the statistical power to detect gene–disease associations by combining results from the original and subsequent replication studies. Recently, consortium-based meta-analyses of several genome-wide association (GWA) data sets have discovered new susceptibility genes of common diseases. We reviewed the process and the methods of meta-analysis of genetic association studies. To conduct and report a transparent meta-analysis, the search strategy, the inclusion or exclusion criteria of studies and the statistical procedures should be fully described. Assessing consistency or heterogeneity of the associations across studies is an important aim of meta-analysis. Random effects model (REM) meta-analysis can incorporate between-study heterogeneity. We illustrated properties of test for and measures of between-study heterogeneity and the effect of between-study heterogeneity on conclusions of meta-analyses through simulations. Our simulation shows that the power of REM meta-analysis of GWA data sets (total case–control sample size: 5000–20 000) to detect a small genetic effect (odds ratio (OR)=1.4 under dominant model) decreases as between-study heterogeneity increases and then the mean of OR of the simulated meta-analyses passing the genome-wide significance threshold would be upwardly biased (*winner's curse* phenomenon). Addressing observed between-study heterogeneity may be challenging but give a new insight into the gene–disease association.
Journal of Human Genetics (2009) **54**, 615–623; doi:10.1038/jhg.2009.95; published online 23 October 2009

Keywords: genome-wide association study; heterogeneity; meta-analysis; winner's curse

INTRODUCTION

Population-based association studies provide a powerful approach to the identification of susceptibility genes underlying common diseases.[1,2] A very large amount of information about genetic variants in the human genome has been accumulated through the International Human Genome Sequencing Project and the International HapMap Project.[3–6] Combined with the establishment of high-throughput single-nucleotide polymorphism (SNP) typing systems, genome-wide association (GWA) studies have been widely applied.[7] Accordingly, gene–disease associations have been reported.

Replication studies were extensively implemented to establish the credibility of the initial positive findings. However, comprehensive reviews of the published literatures in the era of the candidate gene approach show that most of the initial positive associations were not reproduced in the subsequent replication studies.[8–13] These findings suggest that a large number of original findings were false-positive reports and another possibility is that most of the studies were underpowered to detect small genetic effect.[8,9] Furthermore, inconsistency or between-study heterogeneity of results of genetic associations can be observed regardless of whether the associations are true or not,[10,14] and it may be attributed to population stratification, genotyping errors, differences in the pattern of linkage disequilibrium (LD) structure and other factors.[15,16] In the era of GWA studies, this problem remains one of the most difficult issues of genetic association studies.[10,15,16] For example, the large-scale international study of Parkinson's disease failed to replicate 13 SNPs identified by the previous GWA study.[17]

In these circumstances, meta-analysis can be a useful tool to combine both statistically significant and nonsignificant results from individual studies on the same research question. In case–control study, the odds ratios (ORs) for individual studies are combined to calculate a summary OR. Meta-analysis improves the estimation of a summary OR and 95% confidence interval (CI) and increases the statistical power to detect gene–disease associations.[18] Therefore, conclusions from a meta-analysis are more robust than those from a single small study. In addition, meta-analysis is useful to investigate the consistency or heterogeneity of the associations across studies. Testing for and quantifying between-study heterogeneity is an

[1]Division of Molecular Life Science, School of Medicine, Tokai University, Isehara, Kanagawa, Japan and [2]The Japan Health Sciences Foundation, Chuo-ku, Tokyo, Japan
Correspondence: Professor I Inoue, Division of Molecular Life Science, Tokai University, School of Medicine, 143 Shimokasuya, Isehara, Kanagawa 259-1193, Japan.
E-mail: ituro@is.icc.u-tokai.ac.jp
Received 6 August 2009; revised 4 September 2009; accepted 15 September 2009; published online 23 October 2009

important aim of meta-analyses to determine whether there are differences underlying the results of the study.[19,20] Addressing the observed between-study heterogeneity could generate a new insight into the gene–disease association.[20]

In this review, we begin with describing the process of meta-analysis of genetic association studies. The statistical backgrounds, methodological issues and sources of between-study heterogeneity of meta-analysis of genetic association studies are briefly reviewed. Finally, we present the results of our simulation study to illustrate the effect of between-study heterogeneity on conclusions of meta-analyses.

LITERATURE-BASED META-ANALYSIS

In a basic meta-analysis, data are retrospectively collected from published literatures to assess whether a gene–disease association of interest is true or not.[18] When planning a meta-analysis, it is important to define precise search strategy beforehand.[21] If relevant studies are excluded or inadequate studies are included, conclusions of the meta-analysis may be biased.[22] The literature search is conducted in databases such as PubMed and EMBASE. The HuGe Published Literature database (http://www.cdc.gov/genomics/hugenet/) is also useful, as it includes published literatures on genetic associations and other human genome epidemiology.[23] It is important to collect the largest possible number of studies; therefore, we should use appropriate key words. Once the search has been completed, bibliographies of retrieved articles should be examined for further relevant publications.

These processes make up the essential part of the methods section of a meta-analysis, because literature-based meta-analysis is subjected to bias caused by difficulty to identify and include all conducted and relevant studies,[13,24] and small difference in selected literatures may alter conclusions of meta-analyses on the same genetic association.[25] However, the essential features of the search strategy have not fully reported in most meta-analyses of genetic association studies.[26] In order to avoid such biases, it may be recommended to have two or more different researchers conducting the same search.[21] When conducting and reporting a literature-based meta-analysis, flowchart detailing the exclusion and inclusion criteria and the number of studies excluded and included at each step of the literature search is useful (Figure 1).

Meta-analysis of genetic association studies may be subjected to publication bias.[18,26] Publication bias tends to occur when small studies showing negative or nonsignificant results remain unpublished and may result in the overestimation of the genetic effect. If the presence of publication bias is suspected by statistical tests,[27,28] conclusions from the meta-analysis should be cautiously reported and the potential impact of the publication bias should be mentioned.[18]

The results obtained from the meta-analysis would be assessed by the following: (i) the size of the summary OR; (ii) the extent and possible cause of between-study heterogeneity; and (iii) the sufficiency and stability of the meta-analysis by using the cumulative and recursive cumulative meta-analysis approaches.[29–31] In the cumulative meta-analysis, studies are sorted chronologically and a summary OR is calculated when a new study is added.[29] As a result, we can present how the summary OR has shifted over time. The recursive cumulative meta-analysis is an extension of the cumulative meta-analysis, where the relative change in the summary OR by adding a new study is evaluated.[30,31]

CONSORTIUM-BASED META-ANALYSIS

Consortium-based meta-analysis is the meta-analysis of individual patient data through the collaboration of consortium of investigators.

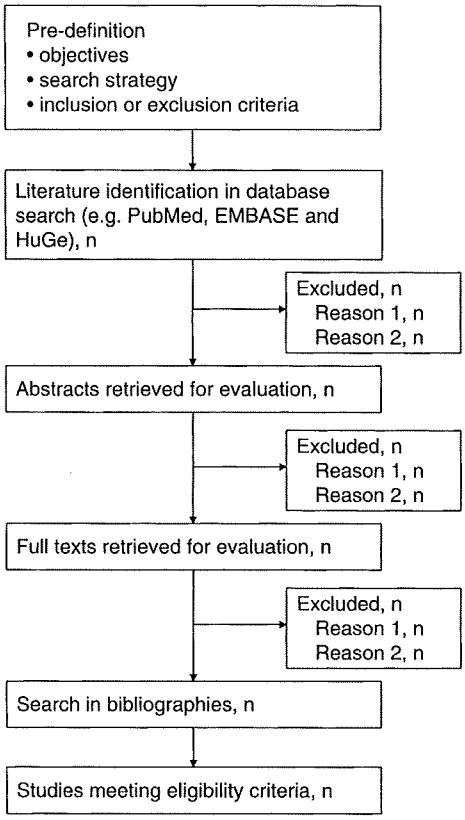

Figure 1 Flowchart detailing the exclusion and inclusion criteria and the number of studies excluded and included at each step of the literature search.

Consortium-based meta-analysis attains increased attention,[32–34] because integration of several GWA data sets has been designed and new susceptibility genes have been discovered.[35–39] Although meta-analysis of GWA studies can be implemented using reported ORs and 95% CIs or *P*-values from different GWA studies, it is preferable to reanalyze several GWA data sets with individual patient data.[35] In the latter case, one can use imputation techniques for missing data when SNPs have been genotyped in some platforms but not in others.[40] Barrett *et al.*[39] conducted a meta-analysis of three GWA data sets for Crohn's disease that used different genotyping platforms using imputation methods. The combined GWA data sets included 635 547 SNPs in 3230 cases and 4829 controls. They used the GWA data sets at the screening stage. The power of the meta-analysis was reported to be 0.74 to detect associations with per allele OR of 1.2 and with risk allele frequency of 0.2 at the significance level of $P=1.0\times10^{-5}$. The meta-analysis of the GWA data sets and additional replication data sets confirmed 11 previously reported loci and identified genome-wide significant signals for novel 21 loci.

GENETIC ASSOCIATION STUDY-SPECIFIC METHODOLOGICAL ISSUES

There are methodological issues relevant to meta-analysis of genetic association studies: (i) assessment of Hardy–Weinberg equilibrium (HWE) and (ii) definition of genetic models.

Deviation from HWE in control samples is the most commonly used test for genotyping error.[41] However, the test for HWE has relatively low statistical power to detect genotyping error.[42]

Furthermore, SNPs that are not in HWE can be used for inference about genetic model of disease susceptibility at the locus.[43] Although there is no consensus how meta-analyses should handle the studies that are not in HWE, three strategies have been applied: including all studies regardless of departure from HWE,[44] performing sensitivity analyses in order to evaluate whether the genetic effects are different between subgroups of studies classified according to test for HWE[26,45–47] and excluding studies showing statistically significant departure from HWE.[18] Reporting the extent of departure from HWE measured by such as α,[48] the inbreeding coefficient,[49] and the disequilibrium parameter[50] is also useful.[44]

In a genetic association study, subjects are classified into three exposure groups (*AA*, *Aa* and *aa*). Let *A* be the susceptibility allele, there are several methods of dichotomizing these exposure groups for conducting a meta-analysis:[26] by comparing allele frequency, by assuming a specific mode of inheritance (recessive, dominance, complete overdominant or codominant) and by performing multiple pairwise comparisons. All these methods, with exception of the method performing multiple pairwise comparisons, assume a particular genetic model. When performing multiple pairwise comparisons or testing multiple genetic models, results of all analyses undertaken should be reported. In order to choose most likely genetic model describing the genetic architecture underlying a disease of interest, Minelli *et al.*[51] presented a 'genetic model free' approach. Their procedure is based on the estimation of the ratio (λ) of the log OR of *Aa* versus *aa* compared with the log OR of *AA* versus *aa*. λ will be 0 under a recessive model, 0.5 under a codominant model and 1 under a dominant model.

ESTIMATION OF A SUMMARY OR AND TEST FOR AND MEASURE OF BETWEEN-STUDY HETEROGENEITY

The statistical methods of combining the results of different studies are described. We consider a meta-analysis of k separate genetic association studies to estimate the genetic effect (θ) for dichotomous disease outcome quantified by log OR. Let θ_i and $\hat{\theta}_i$ be the true and observed log OR for ith case–control study, respectively (i=1, ... ,k). Let v_i denote the variance of $\hat{\theta}_i$, the weight for ith study is given by w_i=1/v_i (that is, the inverse of the variance). OR for each study is given by $OR_i=a_id_i/b_ic_i$. $\hat{\theta}_i = \ln(OR_i)$. v_i is defined as $v_i=1/a_i+1/b_i+1/c_i+1/d_i$, where a_i and b_i correspond to numbers of affected individuals with and without the susceptible genotype, respectively, and c_i and d_i correspond to numbers of unaffected individuals with and without the susceptible genotype, respectively.

There are two commonly used procedures for combining $\hat{\theta}_i$s: 'fixed effects model' (FEM) and 'random effects model' (REM). FEM assumes that θ_is are homogeneous across studies (that is, $\theta_1=\theta_2=\ldots=\theta_k$) and all differences are due to chance. Inverse-variance, Mantel-Haenszel[52] and Peto's[53] methods are commonly used for FEM meta-analysis. Using the inverse-variance method for combining the results across studies, a summary log OR under FEM is calculated as a weighted average of the study estimates: $\hat{\theta}_{FEM} = (\sum_{i=1}^{k} w_i\hat{\theta}_i)/(\sum_{i=1}^{k} w_i)$. The variance of $\hat{\theta}_{FEM}$ is given by $v_{FEM} = 1/\sum_{i=1}^{k} w_i$.

The assumption underlying FEM should be examined with the test for heterogeneity, Cochran's Q test.[54] Test statistics of Cochran's Q test is

$$Q = \sum_{i=1}^{k} w_i\left(\hat{\theta}_i - \hat{\theta}_{FEM}\right)^2$$

Under the null hypothesis of homogeneity (that is, $\theta_1=\theta_2=\ldots=\theta_k$), this statistics approximately follows a χ^2 distribution with $k-1$ degrees of freedom. Cochran's Q test has relatively low statistical power to detect between-study heterogeneity, especially when the number of studies is small;[55] therefore, the test is usually preformed at the significance level of 0.1.[56]

REM assumes that the genetic effects may vary across studies because of genuine difference and/or differential biases. The estimate of the between-study variance (τ^2) is included into the weight as $w_i' = 1/(w_i^{-1}+\hat{\tau}^2)$. A summary log OR under REM are estimated as follows: $\hat{\theta}_{REM} = (\sum_{i=1}^{k} w_i'\hat{\theta}_i)/(\sum_{i=1}^{k} w_i')$. The variance of $\hat{\theta}_{REM}$ is approximated as $v_{REM} = 1/\sum_{i=1}^{k} w_i'$.

In DerSimonian and Laird[57] REM meta-analysis, the τ^2 is estimated as follows:

$$\hat{\tau}^2_{DL} = \frac{Q-(k-1)}{\sum_{i=1}^{k} w_i - \left(\sum_{i=1}^{k} w_i^2/\sum_{i=1}^{k} w_i\right)}$$

When $Q<k-1$, $\hat{\tau}^2_{DL}$ takes negative value. In practice, $\max\{0, \hat{\tau}^2_{DL}\}$ is used. Therefore, the precision of a summary log OR with REM ($1/v_{REM}$) can never exceed that with FEM ($1/v_{FEM}$).

The 95% CI for $\hat{\theta}$ is given by $\hat{\theta} \pm 1.96\times\sqrt{v}$. Test statistic of test for the genetic effect is given by $Z = \hat{\theta}/\sqrt{v}$. Under the null hypothesis, Z follows a standard normal distribution.

Higgins and Thompson[58] proposed three criteria (H, R and I^2) for measure of heterogeneity, which have following desired characteristics: (i) dependence on the extent of heterogeneity, (ii) scale invariance (that is, comparison can be made across meta-analyses with different scales and different outcomes) and (iii) size invariance (that is, independence on the number of studies included). $H = \sqrt{Q/(k-1)}$ is the relative excess of Q to its degrees of freedom. Mittlbock and Heinzl[59] proposed $H^2_M = \frac{Q-(k-1)}{k-1}$ as a modification of H. H^2_M is the proportion of between-study variance to within-study variance. In practice, $\max\{0, H^2_M\}$ is used. H^2_M values over 1.0 indicate considerable heterogeneity.[59] $R = \sqrt{v_{REM}/v_{FEM}}$ is the ratio of the standard error of a summary effect with REM to the standard error with FEM. R represents the inflation of the CI for REM compared with FEM. H and R coincide when all studies have equal weight.[58] $I^2 = 100\times\frac{Q-(k-1)}{Q}$. I^2 can take negative value, but $\max\{0, I^2\}$ is used in practice. I^2 represents the proportion of between-study variance to the total variation in study estimates and ranges from 0 to 100%. I^2 is most widely used for measure of heterogeneity. I^2 values over 50% indicate large heterogeneity.[58,60] Potential drawback of I^2 is that CIs are very large, especially when the number of studies is small.[61]

If heterogeneity is present or suspected by the statistical test or measures, there are several commonly used approaches: (i) performing sensitivity analysis by excluding one or more studies showing outlier effect size, (ii) stratifying the studies into homogeneous subgroups such as racial groups and applying FEM for each subgroup and (iii) implementing REM when observed heterogeneity could not be addressed. Some researchers recommend that the use of REM is preferable compared with FEM, because both models give similar summary effects when there is no between-study heterogeneity, FEM gives narrower CI for summary effect compared with REM when between-study heterogeneity exists and a negative result of test for heterogeneity does not always indicate homogeneity when the number of studies is small.[25]