Table 3. Univariate and multivariate Cox's proportional hazard model analysis of prognostic factors for progression-free survival.

	Univariate analysis		Multivariate analysis	
Prognostic factor	**Hazard ratio (95%CI*)**	***p*-value**	**Hazard ratio (95%CI)**	***p*-value**
A) Present study (n = 110)				
Age	0.99 (0.97–1.01)	0.41	1.00 (0.99–1.02)	0.68
Stage IV (vs Stage III)	1.40 (1.05–1.81)	0.022	0.93 (0.69–1.24)	0.65
Optimal Surgery (vs not optimal)	0.57 (0.45–0.72)	<0.0001	0.73 (0.56–0.94)	0.016
Grade				
Grade2 (vs Grade1)	1.21 (0.89–1.67)	0.23	1.08 (0.78–1.50)	0.66
Grade3 (vs Grade1)	1.44 (1.07–1.98)	0.016	1.34 (0.98–1.88)	0.065
Prognostic Index				
High (vs Low)	3.95 (2.85–5.74)	<0.0001	3.80 (2.68–5.61)	<0.0001
B) Tothill's dataset [20] (n = 87)				
Age	1.01 (0.98–1.03)	0.61	1.00 (0.98–1.03)	0.82
Stage IV (vs Stage III)	1.26 (0.51–2.28)	0.55	0.83 (0.33–1.55)	0.60
Optimal Surgery (vs not optimal)	0.78 (0.62–0.99)	0.049	0.76 (0.60–0.98)	0.035
Prognostic Index				
High (vs Low)	1.62 (1.26–2.09)	0.0001	1.64 (1.27–2.13)	0.0001

*CI denotes confidence interval.

doi:10.1371/journal.pone.0009615.t003

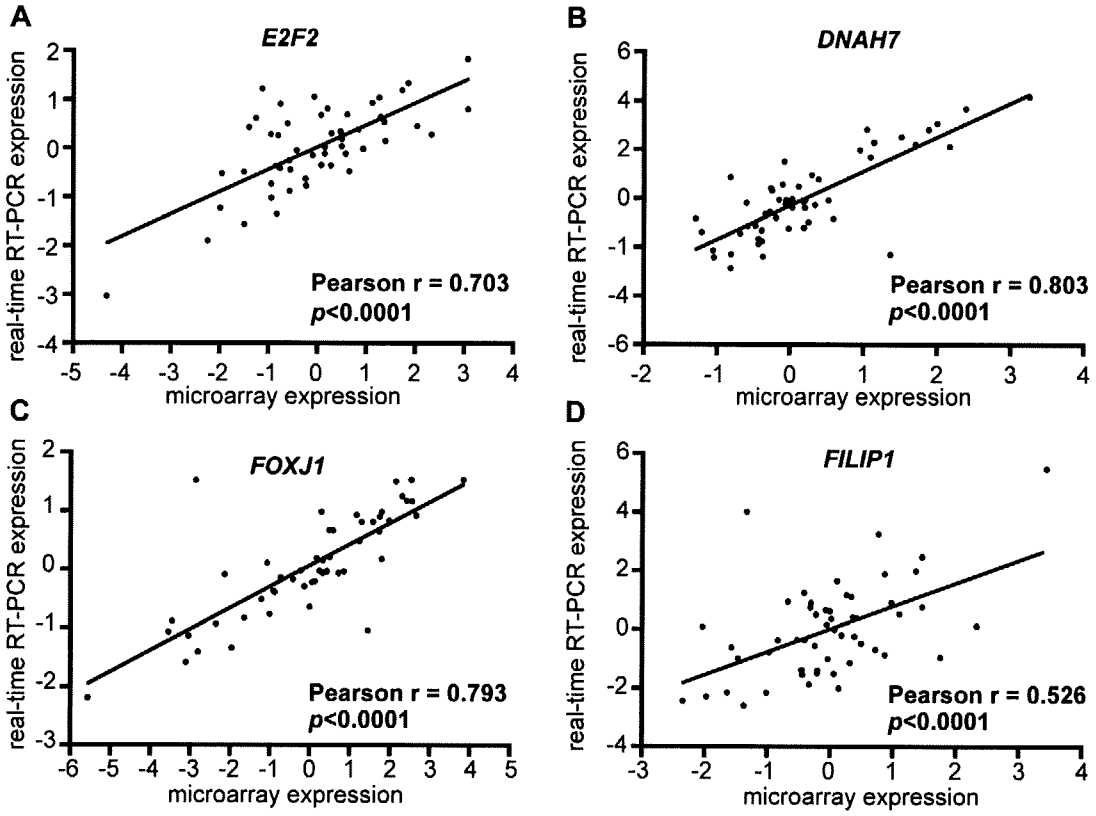

Figure 2. Validation of microarray expression data using quantitative real-time reverse transcript polymerase chain reaction (RT-PCR) analysis. There were significant correlations between microarray expression and real-time RT-PCR expression in (A) *E2F2*, (B) *DNAH7*, (C) *FOXJ1*, and (D) *FILIP1*.

doi:10.1371/journal.pone.0009615.g002

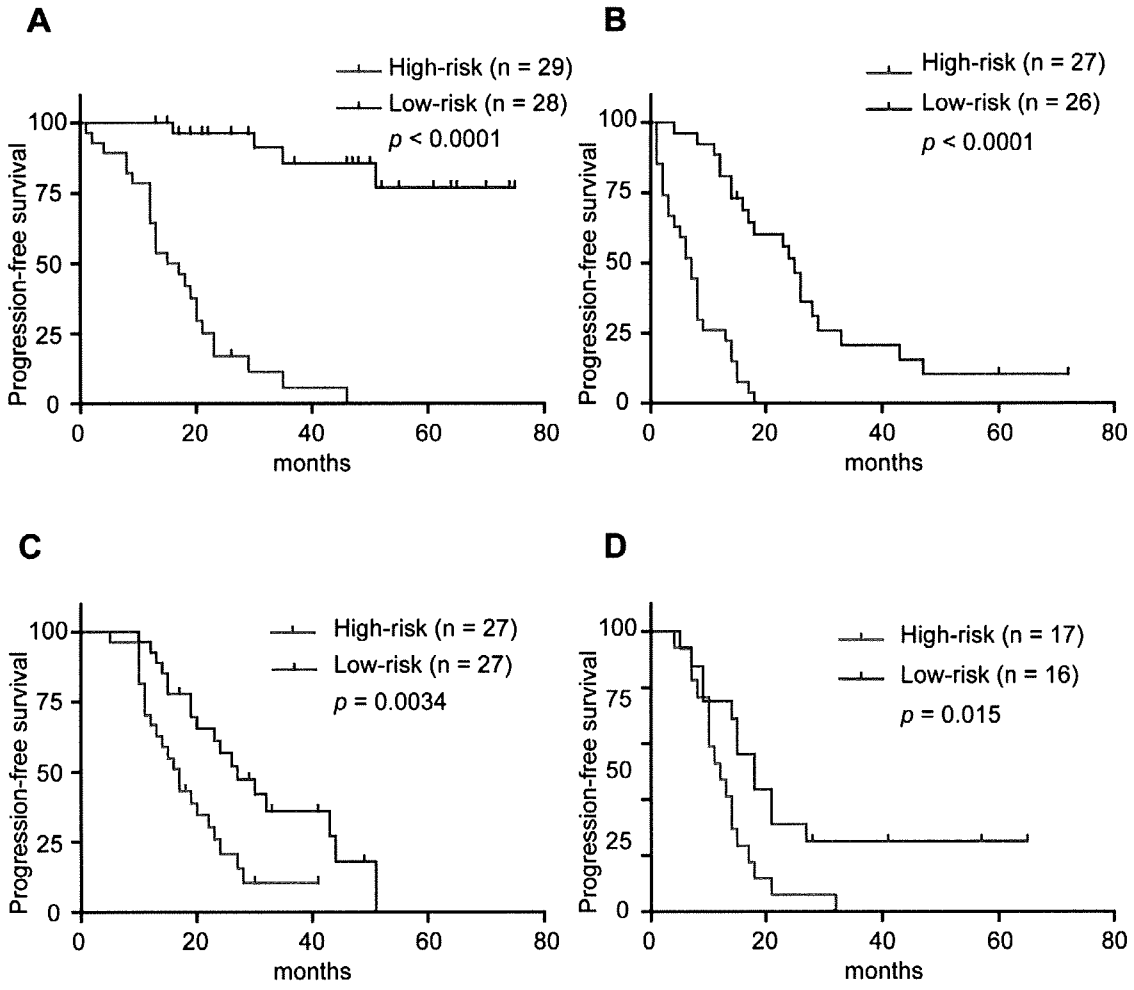

Figure 3. Prediction of prognosis in high-risk and low-risk patients based on the prognostic index after the stratification of patients according to the status of debulking surgery. High-risk patients had significantly short progression-free survival times compared to low-risk patients (A) in optimal (log rank test, $p<0.0001$) and (B) suboptimal group of discovery dataset (log rank test, $p<0.0001$). Similarly, high-risk patients had significantly shorter overall survival times compared to low-risk patients (C) in optimal (log rank test, $p=0.0034$) and (D) suboptimal group of the external dataset (log rank test, $p=0.015$).
doi:10.1371/journal.pone.0009615.g003

Cox model demonstrated the best performance in three datasets. Therefore, we used univariate Cox model only for selecting genes related to PFS time, and adjusted the regression coefficients by the ridge regression Cox model in order to increase the predictive performance of the prognostic index in our dataset.

The current study is intended to identify gene expression profile with a superior ability to predict prognosis than other clinicopathological factors. The stratification of patients with ovarian cancer according to clinicopathological prognostic factors is one of important analysis methods for the identification of highly accurate prognostic index [11]. After we stratified patients according to grade, FIGO stage, and status of debulking surgery, we investigated gene expression profile for predicting PFS time in stage III grade 2/3 serous ovarian cancer patients received optimal surgery or suboptimal surgery. However, we could find poorer predictive performance of the prognostic indices from the stratified analyses than that from the non-stratified analysis (Table S3). Besides the reduction of sample size in the discovery and external datasets after the stratification, a variety in clinical features and grading systems between the two datasets (Table S1) might influence the results from these stratified analyses. This is the main reason why we planned to identify prognostic index based on PFS-related genes in 110 advanced-stage serous ovarian cancers and then evaluate the significance of the prognostic index using multivariate analysis including grade, stage, and status of debulking surgery.

Although we enrolled ovarian cancer patients screened carefully by the following three categories: advanced-stage, histological serous-type, and platinum/taxane-based chemotherapy after primary surgery, we established no inclusion or exclusion criterion of histological grade for the enrollment as well as Crijns and colleagues did [12]. This is because a standard system for grading ovarian carcinomas is still under construction in the world, although several grading systems have been proposed for epithelial ovarian cancer [21–23,33,34]. According to the three criteria above, we recruited 110 Japanese ovarian cancer patients as a discovery set for the PFS analysis. The prognostic index for each patient was simply calculated by the ridge-regression-weighted sum of 88-gene expression values, and the prognostic power of our index was assessed using Tothill's dataset [20]. Further, subsequent stratified analysis according to debulking status, which was an independent prognostic factor in multivariate analysis of the discovery dataset, indicated that our prognostic index was associated with PFS time

Table 4. Univariate and multivariate Cox's proportional hazard model analysis of prognostic factors for overall survival.

Prognostic factor	Univariate analysis		Multivariate analysis	
	Hazard ratio (95%CI*)	*p*-value	Hazard ratio (95%CI)	*p*-value
A) Present study (n = 110)				
Age	1.01 (0.98–1.03)	0.56	-	-
Stage IV (vs Stage III)	1.14 (0.78–1.59)	0.49	0.75 (0.50–1.08)	0.12
Optimal Surgery (vs not optimal)	0.69 (0.50–0.92)	0.012	0.98 (0.70–1.35)	0.90
Grade				
Grade2 (vs Grade1)	1.30 (0.85–2.09)	0.23	1.23 (0.80–2.01)	0.35
Grade3 (vs Grade1)	1.68 (1.12–2.68)	0.012	1.83 (1.18–3.02)	0.0065
Prognostic Index				
High (vs Low)	2.72 (1.91–4.08)	<0.0001	2.99 (2.02–4.65)	<0.0001
B) Tothill's dataset [20] (n = 87)				
Age	1.01 (0.97–1.05)	0.73	1.00 (0.97–1.04)	0.88
Stage IV (vs Stage III)	2.13 (0.85–3.95)	0.093	1.60 (0.62–3.21)	0.28
Optimal Surgery (vs not optimal)	0.89 (0.62–1.23)	0.42	0.94 (0.66–1.37)	0.74
Prognostic Index				
High (vs Low)	1.76 (1.24–2.55)	0.0013	1.71 (1.20–2.49)	0.0029

*CI denotes confidence interval.

doi:10.1371/journal.pone.0009615.t004

independently of the debulking status. However, the sensitivity and specificity of the prognostic index for discriminating between early- and late-relapse patients were lower in Tothill's dataset than those in the discovery set. This might be caused by different backgrounds in respects of ethnicity or microarray platform. Although the differences in gene expression of cancer tissues among ethnicities have not been reported previously, several studies indicate that the proportions of clear cell and endometrioid histological types in

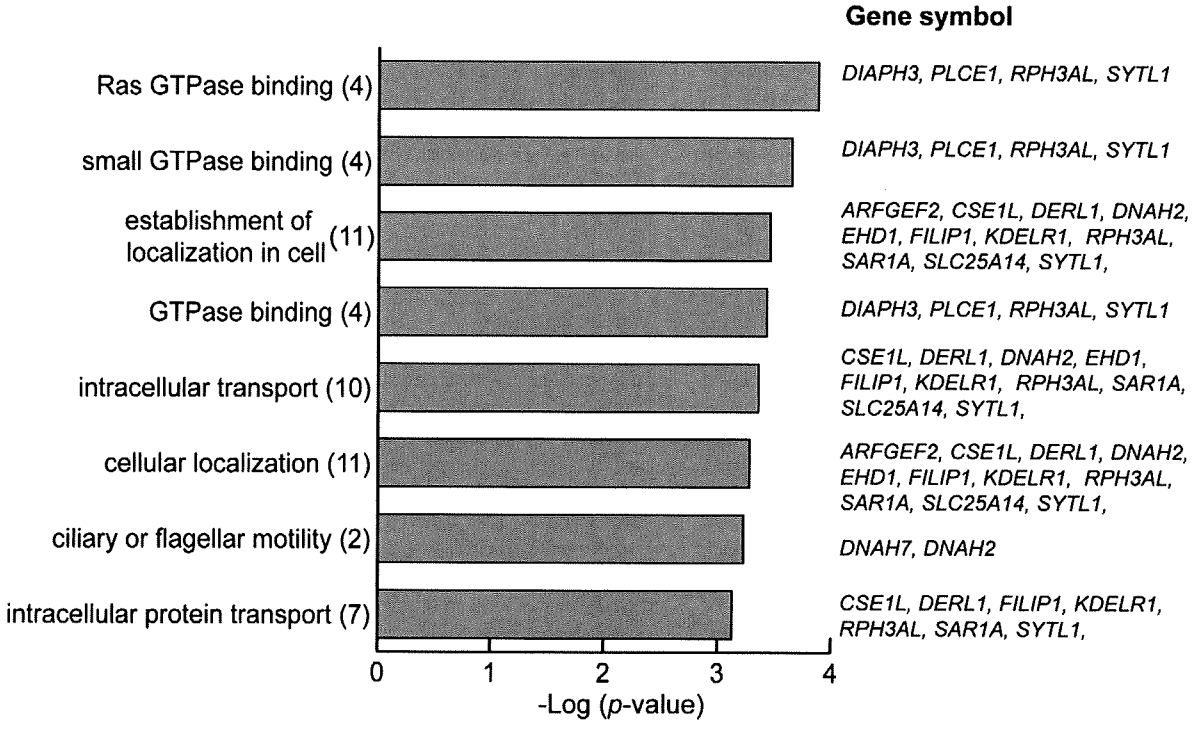

Figure 4. Biological characteristics of 88 progression-free survival-related genes. Significantly over-represented 8 gene ontology (GO) categories in GO-based profiling of 88 genes after multiple testing correction of the Benjamini–Hochberg false discovery rate method (FDR q-value<0.10). Over-represented GO categories were identified using all genes on Agilent platform as a background set of genes for the determining p-values. The actual number of the PFS-related genes involved in each category is given in parentheses.

doi:10.1371/journal.pone.0009615.g004

epithelial ovarian cancer in Asian population are higher than those in non-Asian populations [35,36]. Recent genome-wide association study has identified a single nucleotide polymorphism at 9p22 associated with ovarian cancer risk in subjects with European ancestry but not in non-European descendants [37]. This type of differences between studies could be also attributed to genetic as well as environmental factors. In addition, we cannot rule out the possibility that the present PFS-associated classifiers with ridge-regression-based weights still have insufficient generalization properties on the external dataset due to the problem of overfitting. Therefore, we will reconsider these important issues such as between-study differences in ethnicities and microarray platforms and the overfitting problem using a larger number of microarray data from advanced-stage serous ovarian cancer patients in order to obtain better classifiers for the prediction of prognosis. And to improve the accuracy of prognostic index, development of prognostic index after the stratification of patients will be a research agenda for further study.

Interestingly, the present 88-gene prognostic index for prediction of PFS time was also significantly associated with overall survival time in both our dataset and Tothill's dataset [20]. Moreover, we examined the predictive ability of our prognostic index in Dressman's dataset [25] since patients in their dataset received longer-term follow-up than those in the above two datasets. Although Dressman's dataset (n = 119) [25] included 34 patients treated with platinum/cyclophospamide chemotherapy and 3 with single-agent platinum, the significance of this prognostic index for overall survival was still statistically supported in the longer followed-up dataset. As treatments for recurrent ovarian cancer patients remain an open area of investigation aiming to lead to survival benefit [38], our prognostic index for patient with advanced-stage serous ovarian cancer displays a potential to predict not only PFS time but also overall survival time. In the future, we may apply the prognostic indices to estimation of risk of recurrence for serous ovarian cancer patients and select a novel treatment such as dose-dense chemotherapy [39] or molecular-targeted agent for the purpose of improving prognosis of high-risk patients.

There are small number of genes overlapped between our 88 PFS-related profile and previously reported expression-profiles that were related to prognosis or sensitivity of platinum/taxane-based chemotherapy [11–15,40,41]. Konstantinopoulos *et al.* [6] have discussed that these discrepancies might be related to the use of different microarray platforms with different normalization methods and different degree of contamination by noncancerous cells in a tumor sample, as well as differences in the patient populations under study. Nevertheless, several survival-associated genes such as *E2F2* and *HLA-DMB* [42,43] are included in 88 PFS-related genes. Reimer *et al.* [42] have reported that *E2F2* is associated with grade 3 ovarian tumors and residual disease (more than 2cm in diameter) after initial surgery, and that low *E2F2* expression is significantly associated with favorable disease-free and overall survival in epithelial ovarian cancer. Callahan *et al.* [43] have recently reported that the high expression of HLA-DMB in ovarian cancer cells is correlated with increased numbers of tumor-infiltrating CD8-positive T lymphocytes, and with good prognosis in advanced-stage high-grade serous ovarian cancer.

We performed GO analysis and IPA to assess biological characteristics of PFS-related genes. GO analysis revealed the significant associations of GTPase binding, intracellular transport, and ciliary or flagellar motility with PFS (Figure 4). *PLCE1* belongs to the GTPase binding category and activates MAP kinase or ERK as shown in IPA network 3 (Figure S7). In particular, previous report indicates that *PLCE1* activates the small G protein Ras/MAP kinase signaling [44], which is one of important pathways associated with cell growth and differentiation. Intriguingly, *CSE1L* included in the intracellular transport category is involved in the regulation of multiple cellular mechanisms, proliferation, and apoptosis [45]. Tanaka *et al.* [46] have reported that *CSE1L* is associated with regulated expression of p53 target genes, and that downregulation of *CSE1L* protects cancer cell from DNA damage-induced apoptosis. *DNAH2* and *DNAH7* are components of the inner dynein arm of cilialy axonemes, and axonemal dyneins are molecular motors that drive the beating of cilia and flagella. Plotnikova *et al.* [47] have reported that loss of cilia in cancer cells may contribute to the insensitivity of cancer cells to environmental repressive signals, partly owing to derangement of cell cycle checkpoints governed by cilia and centrosomes. On the other hand, IPA analysis showed several genes interacting with *SRC* or *MYC* (Figure S6), each of which was reported as a representative gene in oncogenic pathways of ovarian cancer [25,27]. Dressman *et al.* [25] have demonstrated that Src pathway activity is associated with chemotherapy response because of a significant correlation between the activation of Src pathway and poor prognosis in patients with platinum-resistant ovarian cancer. *MYC* is a multifunctional proto-oncogene and activated in about 30% of ovarian cancer by several mechanisms [48]. Iba *et al.* [49] report that MYC expression is associated with responsiveness to platinum-based chemotherapy and with prognosis in patients with epithelial ovarian cancer. Our PFS-related profile might have potentially functional relevance to altered activities of several oncogenic pathways. Although we identified several genes whose molecular function could be linked to prognosis in ovarian cancer patients, further functional study will be necessary to clarify the biological and pathological implications of the PFS-related profile.

These results suggest that the gene expression profile could be a useful tool to predict disease progression or recurrence of advanced-stage serous ovarian cancer. To apply the gene expression profile in clinical practice, we will need to improve the predictive ability of the profile and confirm the reliability of survival profile in a prospective multi-center study. Nevertheless, the survival-related profile could provide an optimization of the clinical management and development of new therapeutic strategies for the serous ovarian cancer patients.

Materials and Methods

Tissue Samples

One hundred ten Japanese patients who were diagnosed with advanced-stage serous ovarian cancer between July 1997 and June 2008 were included in this study. Fresh-frozen samples were obtained from primary tumor tissues during primary debulking surgery prior to chemotherapy. All patients with advanced-stage serous ovarian cancer were treated with platinum/taxane-based chemotherapy after surgery. In principle, patients were seen every 1 to 3 months for the first 2 years. Thereafter, follow-up visits had an interval of 3 to 6 months in the third to fifth year, and 6 to 12 months in the sixth to tenth year. At every follow-up visits, general physical and gynecologic examination were performed. CA125 serum levels were routinely determined. Staging of the disease was assessed according to the criteria of the International Federation of Gynecology and Obstetrics (FIGO) [19]. Optimal debulking surgery was defined as $\leq$1cm of gross residual disease. The histological characteristics of surgically resected specimens were assessed on formalin-fixed and paraffin-embedded hematoxylin and eosin sections by two or three gynecological pathologists belonging to the Japanese Society of Pathology at each institute,

and frozen tissues containing more than 80% of tumor cells upon histological evaluation were used for RNA extraction. In this study, the degree of histological differentiation is determined according to the increase in the proportion of solid growth within the adenocarcinoma as follows: grade 1, less than 5% solid growth; grade 2, 6-50% solid growth; grade 3, over 50% solid growth based on grading system proposed by Japan Society of Gynecologic Oncology.

PFS time was calculated as the interval from primary surgery to disease progression or recurrence. Based on standard Response Evaluation Criteria In Solid Tumors (RECIST) guidelines [50], disease progression was defined as at least 20% increase in the sum of the longest diameters of all target lesions or as the appearance of one or more new lesions and/or unequivocal progression existing non-target lesions. Overall survival time was calculated as the interval from primary surgery to the death due to ovarian cancer. This study was approved by the institutional ethics review board at Niigata University (No. 239, 282, 285, and 318), Niigata Cancer Center Hospital (No. 25), Jichi Medical University (G07-01), Kagoshima City Hospital (H19-21), Hiroshima University (Hi-11), Nagasaki University (080509), Kumamoto University (No. 309), and Tokai University (07I-29). All patients provided written informed consent for the collection of samples and subsequent analysis.

Microarray Experiments

Total RNA was extracted from tissue samples as previously described [17]. Five hundred nanograms of total RNA were converted into labeled cRNA with nucleotides coupled to a cyanine 3-CTP (Cy3) (PerkinElmer, Boston, MA, USA) using the Quick Amp Labeling Kit, one-color (Agilent Technologies). Cy3-labeled cRNA (1.65 μg) was hybridized for 17 hours at 65°C to an Agilent Whole Human Genome Oligo Microarray, which carries 60-mer probes to more than 40,000 human transcripts. The hybridized microarray was washed and then scanned in Cy3 channel with the Agilent DNA Microarray Scanner (model G2565AA). Signal intensity per spot was generated from the scanned image using Feature Extraction Software version 9.1 (Agilent Technologies) in the default settings. Spots that did not pass quality control procedures were flagged as "Absent". The MIAME-compliant microarray data were deposited into the Gene Expression Omnibus data repository (accession number GSE17260).

Microarray Data Analysis

We analyzed our dataset as a "discovery set" and the publicly available dataset as an "external dataset". Considering differences in microarray platforms, we selected common genes between the Agilent Whole Human Genome Oligo Microarray and Affymetrix Human Genome U133 Plus 2.0 Array, which was the platform in an external dataset (GSE9891) [20].

Data normalization was performed in GeneSpring GX 10 (Agilent Technologies) as follows: (i) Threshold raw signals were set to 1.0. (ii) 75th percentile normalization was chosen as normalized algorithm. (iii) Baseline was transformed to median of all samples. Furthermore, the expression level was normalized by Z-transformation (the mean expression was set to 0 and standard deviation to 1 for each gene in each dataset). In our dataset, 18,178 probes with expression levels marked as "Present" in all microarrays were used to remove missing and uncertain signals on gene expression.

The PFS-related genes from the 18,178 probes were identified by univariate Cox proportional hazard analysis, followed by a ridge regression, a penalized Cox regression analysis for survival prediction (Figure S2). We first identified 97 probes with expression levels correlating with the PFS time determined using the univariate Cox proportional hazard model ($p<0.01$). In case of multiple probes representing a given gene (so-called multiple tagged gene) in microarrays, only the probe with the largest magnitude (i.e., sum of the squares of per-individual expression values) was extracted as a representative probe for the gene [24]. To avoid the problem of overfitting, ridge regression extension of the multivariate Cox model was employed [18]. The ridge regression shrinks regression coefficients (β) of genes in multivariate Cox model by imposing a penalty on squared values of the coefficients, and is able to handle the problem of having larger number of expression values than individuals in an appropriate way [30]. We estimated regression coefficients of the prognostic genes by the ridge regression Cox model using M-files (available at http://www.med.uio.no/imb/stat/bmms/software/microsurv/) for MATLAB (Mathworks, Natick, MA, USA). Using 10-fold cross-validation, we obtained regression coefficients with optimal penalty parameter for the penalized Cox model, and calculated a prognostic index for each patient as defined by

$$\text{Prognostic index} = \sum_{i=1}^{88} \beta_i \times X_i \qquad (1)$$

where β_i is the estimated regression coefficient of each gene in discovery dataset under ridge regression multivariate Cox model and X_i is the Z-transformed expression value of each gene [18]. The estimated regression coefficient of each PFS-related gene given by ridge regression in the discovery set was also applied to calculate a prognostic index for each patient in external dataset using the equation above. We classified all patients into the two groups (high- and low-risk groups) by the median of the prognostic index in discovery set [9]. PFS between high- and low-risk groups was compared using Kaplan-Meier curves and the log rank test using GraphPad PRISM version 4.0 (GraphPad Software, San Diego, CA, USA). Furthermore, We then evaluated the prognostic index in the multivariate Cox proportional hazard model using JMP version 6 (SAS Institute, Cary, NC, USA). We also examined the discrimination performance of the prognostic index between early and late relapse in patients by plotting a receiver operating characteristic (ROC) curve for each dataset (JMP). Because 18 months is the median PFS time for advanced-stage ovarian cancer patients treated with cisplatin-paclitaxel [1], we used 18 months as the cut-off between early and late relapse. We performed ROC curve analysis for our prognostic index in only patients with follow-up for more than 18 months (Discovery set 103 samples; External dataset 84 samples).

To investigate the biological functions of PFS-related gene expression profiles, we used GO Ontology Browser, embedded in GeneSpring GX [17,51]. The GO Ontology Browser was used to analyze which categories of gene ontology were statistically overrepresented among the gene list obtained. Statistical significance was determined by Fisher's exact test, followed by multiple testing corrections by the Benjamini and Hochberg false discovery rate (FDR) method [26]. Furthermore, we tried to explore molecular interaction networks among the PFS-related genes using Ingenuity Pathway Analysis (IPA) [17].

Quantitative Real-Time Reverse Transcription Polymerase Chain Reaction (RT-PCR) Analysis

Real-time PCR was performed on *E2F2* (Hs00231667_m1, Applied Biosystems), *FOXJ1* (Hs00230964_m1, Applied Biosystems), *DNAH7* (Hs01022427_m1, Applied Biosystems), and *FILIP1* (Hs00325074_m1, Applied Biosystems) for a subset of serous

ovarian cancer (n = 53) as previously described [17]. The relative quantification method [52] was used to measure the amounts of the respective genes in serous ovarian cancer samples, normalized to *ACTB* (Hs99999903_m1, Applied Biosystems) and *TBP* (Hs99999910_m1, Applied Biosystems).

Evaluation of PFS-Related Genes in the External Dataset

To confirm whether our expression profile could predict prognosis of serous ovarian cancer patients in an independent data set, we selected to use publicly available microarray data (GSE9891) only because the data also disclosed individual clinical characteristics including PFS time. We examined clinical information of these dataset using supplementary data [20]. From this original dataset (n = 285), we selected 87 samples that were (i) diagnosed as advanced-stage serous adenocarcinoma, (ii) treated by platinum/taxane-based chemotherapy, (iii) obtained from primary lesion, and (iv) followed-up for more than 12 months (Table S1). Their samples are histologically graded by Silverberg classification [22] whose grading system is different from that in this study.

Supporting Information

Figure S1 Kaplan-Meier survival curves between 110 patients in this dataset and 87 in Tothill's dataset.
Found at: doi:10.1371/journal.pone.0009615.s001 (0.24 MB TIF)

Figure S2 Analytical process to develop a prognostic index for predicting survival.
Found at: doi:10.1371/journal.pone.0009615.s002 (0.48 MB TIF)

Figure S3 Assessment of the sensitivity and specificity of 88-gene prognostic index using receiver-operating characteristic (ROC) curves. When early relapse is positive in the analysis, the area under ROC curve to distinguish early-relapse patients with less than 18 months of progression-free survival times from late-relapse patients was 0.959 and 0.674 in (A) discovery set (early, n = 54; late, n = 49) and in (B) external set (early, n = 45; late, n = 39), respectively.
Found at: doi:10.1371/journal.pone.0009615.s003 (0.42 MB TIF)

Figure S4 Appling PFS-related gene expression profile to Dressman's dataset [25]. (A) Multivariate analysis showed a significant association of overall survival with the prognostic index estimated using the 88-gene linear combination model with the ridge regression coefficients from the present discovery set in Dresssman's dataset (HR, 1.51; 95% CI, 1.19–1.93, p = 0.0008) (B) Kaplan-Meier survival curves and the log rank test showed that high-risk patients had shorter overall survival compared to low-risk patients (median survival, 31 and 87 months for high- and low-risk patients, respectively; p = 0.0008).
Found at: doi:10.1371/journal.pone.0009615.s004 (0.23 MB TIF)

Figure S5 Molecular interaction networks of 88 progression-free survival-related genes using Ingenuity Pathway Analysis (IPA) software. The prognostic genes incorporated into the respective networks were marked as gray-colored.
Found at: doi:10.1371/journal.pone.0009615.s005 (2.42 MB TIF)

Figure S6 Molecular interaction networks of 88 progression-free survival-related genes using Ingenuity Pathway Analysis (IPA) software. The prognostic genes incorporated into the respective networks were marked as gray-colored.
Found at: doi:10.1371/journal.pone.0009615.s006 (1.68 MB TIF)

Figure S7 Molecular interaction networks of 88 progression-free survival-related genes using Ingenuity Pathway Analysis (IPA) software. The prognostic genes incorporated into the respective networks were marked as gray-colored.
Found at: doi:10.1371/journal.pone.0009615.s007 (1.82 MB TIF)

Table S1 Clinical characteristics of advanced-stage serous ovarian cancer patients in Tothill's dataset [20] (n = 87).
Found at: doi:10.1371/journal.pone.0009615.s008 (0.04 MB DOC)

Table S2 Univariate and multivariate Cox's proportional hazard model analysis of prognostic factors for progression-free survival.
Found at: doi:10.1371/journal.pone.0009615.s009 (0.04 MB DOC)

Table S3 Univariate Cox's proportional hazard model analysis of prognostic index for progression-free survival in the two datasets.
Found at: doi:10.1371/journal.pone.0009615.s010 (0.04 MB DOC)

Acknowledgments

We thank tissue donors and supporting medical staff for making this study possible. We are grateful to C. Seki and A. Yukawa for their technical assistance.

Author Contributions

Conceived and designed the experiments: KY AT TY II KT. Performed the experiments: KY AT. Analyzed the data: KY AT. Contributed reagents/materials/analysis tools: KY TY SK HF MS YO MH KS HF YK KK HM HT HK II KT. Wrote the paper: KY AT TY II KT.

References

1. McGuire WP, Hoskins WJ, Brady MF, Kucera PR, Partridge EE, et al. (1996) Cyclophosphamide and cisplatin compared with paclitaxel and cisplatin in patients with stage III and stage IV ovarian cancer. N Engl J Med 334: 1–6.
2. Piccart MJ, Bertelsen K, James K, Cassidy J, Mangioni C, et al. (2000) Randomized intergroup trial of cisplatin-paclitaxel versus cisplatin-cyclophosphamide in women with advanced epithelial ovarian cancer: three-year results. J Natl Cancer Inst 92: 699–708.
3. Cannistra SA (2004) Cancer of the ovary. N Engl J Med 351: 2519–29.
4. du Bois A, Reuss A, Pujade-Lauraine E, Harter P, Ray-Coquard I, et al. (2009) Role of surgical outcome as prognostic factor in advanced epithelial ovarian cancer: a combined exploratory analysis of 3 prospectively randomized phase 3 multicenter trials: by the Arbeitsgemeinschaft Gynaekologische Onkologie Studiengruppe Ovarialkarzinom (AGO-OVAR) and the Groupe d'Investigateurs Nationaux Pour les Etudes des Cancers de l'Ovaire (GINECO). Cancer 115: 1234–44.
5. Winter WE, 3rd, Maxwell GL, Tian C, Carlson JW, Ozols RF, et al. (2007) Prognostic factors for stage III epithelial ovarian cancer: a Gynecologic Oncology Group Study. J Clin Oncol 25: 3621–7.
6. Konstantinopoulos PA, Spentzos D, Cannistra SA (2008) Gene-expression profiling in epithelial ovarian cancer. Nat Clin Pract Oncol 5: 577–87.
7. van 't Veer LJ, Dai H, van de Vijver MJ, He YD, Hart AA, et al. (2002) Gene expression profiling predicts clinical outcome of breast cancer. Nature 415: 530–6.
8. Motoori M, Takemasa I, Yano M, Saito S, Miyata H, et al. (2005) Prediction of recurrence in advanced gastric cancer patients after curative resection by gene expression profiling. Int J Cancer 114: 963–8.
9. Chen HY, Yu SL, Chen CH, Chang GC, Chen CY, et al. (2007) A five-gene signature and clinical outcome in non-small-cell lung cancer. N Engl J Med 356: 11–20.
10. Schramm A, Schulte JH, Klein-Hitpass L, Havers W, Sieverts H, et al. (2005) Prediction of clinical outcome and biological characterization of neuroblastoma by expression profiling. Oncogene 24: 7902–12.
11. Bonome T, Levine DA, Shih J, Randonovich M, Pise-Masison CA, et al. (2008) A gene signature predicting for survival in suboptimally debulked patients with ovarian cancer. Cancer Res 68: 5478–86.
12. Crijns AP, Fehrmann RS, de Jong S, Gerbens F, Meersma GJ, et al. (2009) Survival-related profile, pathways, and transcription factors in ovarian cancer. PLoS Med 6: e24.
13. Denkert C, Budczies J, Darb-Esfahani S, Györffy B, Sehouli J, et al. (2009) A prognostic gene expression index in ovarian cancer - validation across different independent data sets. J Pathol 218: 273–80.

14. Hartmann LC, Lu KH, Linette GP, Cliby WA, Kalli KR, et al. (2005) Gene expression profiles predict early relapse in ovarian cancer after platinum-paclitaxel chemotherapy. Clin Cancer Res 11: 2149–55.
15. Spentzos D, Levine DA, Ramoni MF, Joseph M, Gu X, et al. (2004) Gene expression signature with independent prognostic significance in epithelial ovarian cancer. J Clin Oncol 22: 4700–10.
16. Agarwal R, Kaye SB (2006) Expression profiling and individualization of treatment for ovarian cancer. Curr Opin Pharmacol 6: 345–9.
17. Yoshihara K, Tajima A, Komata D, Yamamoto T, Kodama S, et al. (2009) Gene expression profiling of advanced-stage serous ovarian cancers distinguishes novel subclasses and implicates ZEB2 in tumor progression and prognosis. Cancer Sci 100: 1421–8.
18. Bovelstad HM, Nygård S, Storvold HL, Aldrin M, Borgan Ø, et al. (2007) Predicting survival from microarray data–a comparative study. Bioinformatics 23: 2080–7.
19. FIGO Cancer Committee. (1986) Staging Announcement: FIGO Cancer Committee. Gynecol Oncol 25: 383–5.
20. Tothill RW, Tinker AV, George J, Brown R, Fox SB, et al. (2008) Novel molecular subtypes of serous and endometrioid ovarian cancer linked to clinical outcome. Clin Cancer Res 14: 5198–208.
21. International Federation of Gynecology and Obstetrics (1971) Classification and staging of malignant tumours in the female pelvis. Acta Obstet Gynecol Scand 50: 1–7.
22. Silverberg SG (2000) Histopathologic grading of ovarian carcinoma: a review and proposal. Int J Gynecol Pathol 19: 7–15.
23. Kommoss S, Schmidt D, Kommoss F, Hedderich J, Harter P, et al. (2009) Histological grading in a large series of advanced stage ovarian carcinomas by three widely used grading systems: consistent lack of prognostic significance. A translational research subprotocol of a prospective randomized phase III study (AGO-OVAR 3 protocol). Virchows Arch 454: 249–56.
24. Woo HG, Park ES, Cheon JH, Kim JH, Lee JS, et al. (2008) Gene expression-based recurrence prediction of hepatitis B virus-related human hepatocellular carcinoma. Clin Cancer Res 14: 2056–64.
25. Dressman HK, Berchuck A, Chan G, Zhai J, Bild A, et al. (2007) An integrated genomic-based approach to individualized treatment of patients with advanced-stage ovarian cancer. J Clin Oncol 25: 517–25.
26. Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Statist Soc B 57: 289–300.
27. Bild AH, Yao G, Chang JT, Wang Q, Potti A, et al. (2006) Oncogenic pathway signatures in human cancers as a guide to targeted therapies. Nature 439: 353–7.
28. Dupuy A, Simon RM (2007) Critical review of published microarray studies for cancer outcome and guidelines on statistical analysis and reporting. J Natl Cancer Inst 99: 147–57.
29. Bair E, Tibshirani R (2004) Semi-supervised methods to predict patient survival from gene expression data. PLoS Biol 2: e108.
30. van Houwelingen HC, Bruinsma T, Hart AA, Van't Veer LJ, Wessels LF (2006) Cross-validated Cox regression on microarray gene expression data. Stat Med 25: 3201–16.
31. Rosenwald A, Wright G, Chan WC, Connors JM, Campo E, et al. (2002) The use of molecular profiling to predict survival after chemotherapy for diffuse large-B-cell lymphoma. N Engl J Med 346: 1937–47.
32. Sorlie T, Tibshirani R, Parker J, Hastie T, Marron JS, et al. (2003) Repeated observation of breast tumor subtypes in independent gene expression data sets. Proc Natl Acad Sci U S A 100: 8418–23.
33. Tawassoli FA, Devilee P (2003) Pathology and Genetics. Tumours of the Breast and Female Genital Organs. IARC Press, Lyon.
34. Malpica A, Deavers MT, Lu K, Bodurka DC, Atkinson EN, et al. (2004) Grading ovarian serous carcinoma using a two-tier system. Am J Surg Pathol 28: 496–504.
35. Goodman MT, Howe HL, Tung KH, Hotes J, Miller BA, et al. (2003) Incidence of ovarian cancer by race and ethnicity in the United States, 1992–1997. Cancer 97(10 Suppl): 2676–85.
36. McGuire V, Jesser CA, Whittemore AS (2002) Survival among U.S. women with invasive epithelial ovarian cancer. Gynecol Oncol 84: 399–403.
37. Song H, Ramus SJ, Tyrer J, Bolton KL, Gentry-Maharaj A, et al. (2009) A genome-wide association study identifies a new ovarian cancer susceptibility locus on 9p22.2. Nat Genet 41: 996–1000.
38. Ozols RF (2005) Treatment goals in ovarian cancer. Int J Gynecol Cancer 15 Suppl 1: 3–11.
39. Katsumata N, Yasuda M, Takahashi F, Isonishi S, Jobo T, et al. (2009) Dose-dense paclitaxel once a week in combination with carboplatin every 3 weeks for advanced ovarian cancer: a phase 3, open-label, randomised controlled trial. Lancet 374: 1331–8.
40. Berchuck A, Iversen ES, Luo J, Clarke JP, Horne H, et al. (2009) Microarray analysis of early stage serous ovarian cancers shows profiles predictive of favorable outcome. Clin Cancer Res 15: 2448–55.
41. Helleman J, Jansen MP, Span PN, van Staveren IL, Massuger LF, et al. (2006) Molecular profiling of platinum resistant ovarian cancer. Int J Cancer 118: 1963–71.
42. Reimer D, Sadr S, Wiedemair A, Stadlmann S, Concin N, et al. (2007) Clinical relevance of E2F family members in ovarian cancer–an evaluation in a training set of 77 patients. Clin Cancer Res 13: 144–51.
43. Callahan MJ, Nagymanyoki Z, Bonome T, Johnson ME, Litkouhi B, et al. (2008) Increased HLA-DMB expression in the tumor epithelium is associated with increased CTL infiltration and improved prognosis in advanced-stage serous ovarian cancer. Clin Cancer Res 14: 7667–73.
44. Lopez I, Mak EC, Ding J, Hamm HE, Lomasney JW (2001) A novel bifunctional phospholipase c that is regulated by Galpha 12 and stimulates the Ras/mitogen-activated protein kinase pathway. J Biol Chem 276: 2758–65.
45. Behrens P, Brinkmann U, Wellmann A (2003) CSE1L/CAS: its role in proliferation and apoptosis. Apoptosis 8: 39–44.
46. Tanaka T, Ohkubo S, Tatsuno I, Prives C (2007) hCAS/CSE1L associates with chromatin and regulates expression of select p53 target genes. Cell 130: 638–50.
47. Plotnikova OV, Golemis EA, Pugacheva EN (2008) Cell cycle-dependent ciliogenesis and cancer. Cancer Res 68: 2058–61.
48. Darcy KM, Brady WE, Blancato JK, Dickson RB, Hoskins WJ, et al. (2009) Prognostic relevance of c-MYC gene amplification and polysomy for chromosome 8 in suboptimally-resected, advanced stage epithelial ovarian cancers: a Gynecologic Oncology Group study. Gynecol Oncol 114: 472–9.
49. Iba T, Kigawa J, Kanamori Y, Itamochi H, Oishi T, et al. (2004) Expression of the c-myc gene as a predictor of chemotherapy response and a prognostic factor in patients with ovarian cancer. Cancer Sci 95: 418–23.
50. Therasse P, Arbuck SG, Eisenhauer EA, Wanders J, Kaplan RS, et al. (2000) New guidelines to evaluate the response to treatment in solid tumors. European Organization for Research and Treatment of Cancer, National Cancer Institute of the United States, National Cancer Institute of Canada. J Natl Cancer Inst 92: 205–16.
51. Okada H, Tajima A, Shichiri K, Tanaka A, Tanaka K, et al. (2008) Genome-wide expression of azoospermia testes demonstrates a specific profile and implicates ART3 in genetic susceptibility. PLoS Genet 4: e26.
52. Livak KJ, Schmittgen TD (2001) Analysis of Relative Gene Expression Data Using Real-Time Quantitative PCR and the $2^{-\Delta\Delta CT}$ Method. Methods 25: 402–8.

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

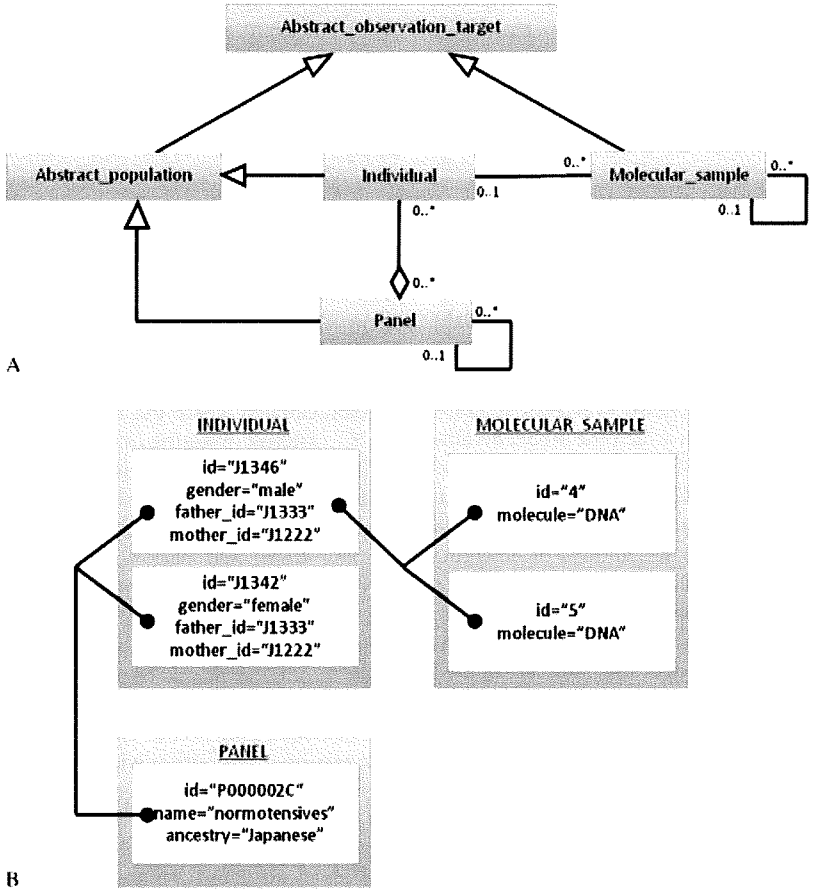

Figure 1. SAMPLE domain of PAGE-OM. **A:** The principal classes (colored blue) and class relationships from the SAMPLE domain, as described in the text. **B:** Shows how the model in (A) could be used to represent a cohort of normotensive Japanese, giving further details for a brother and sister from 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

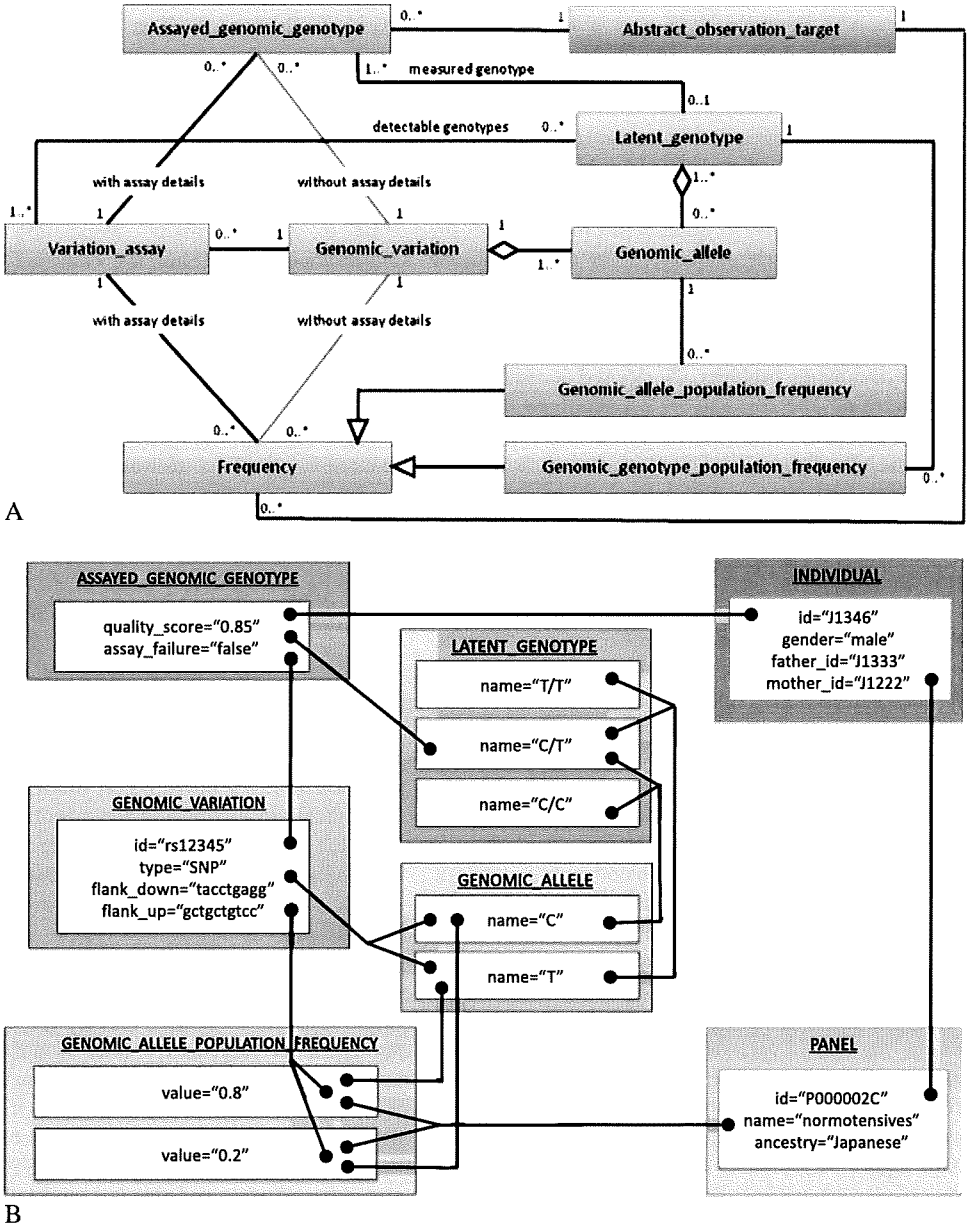

Figure 2. GENOTYPE domain of PAGE-OM. **A:** The principal classes (colored red) and class relationships from the GENOTYPE domain, as described in the text. One additional class (colored blue) is also included, taken from the SAMPLE domain. At the project website, sections of the model called Marker, 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

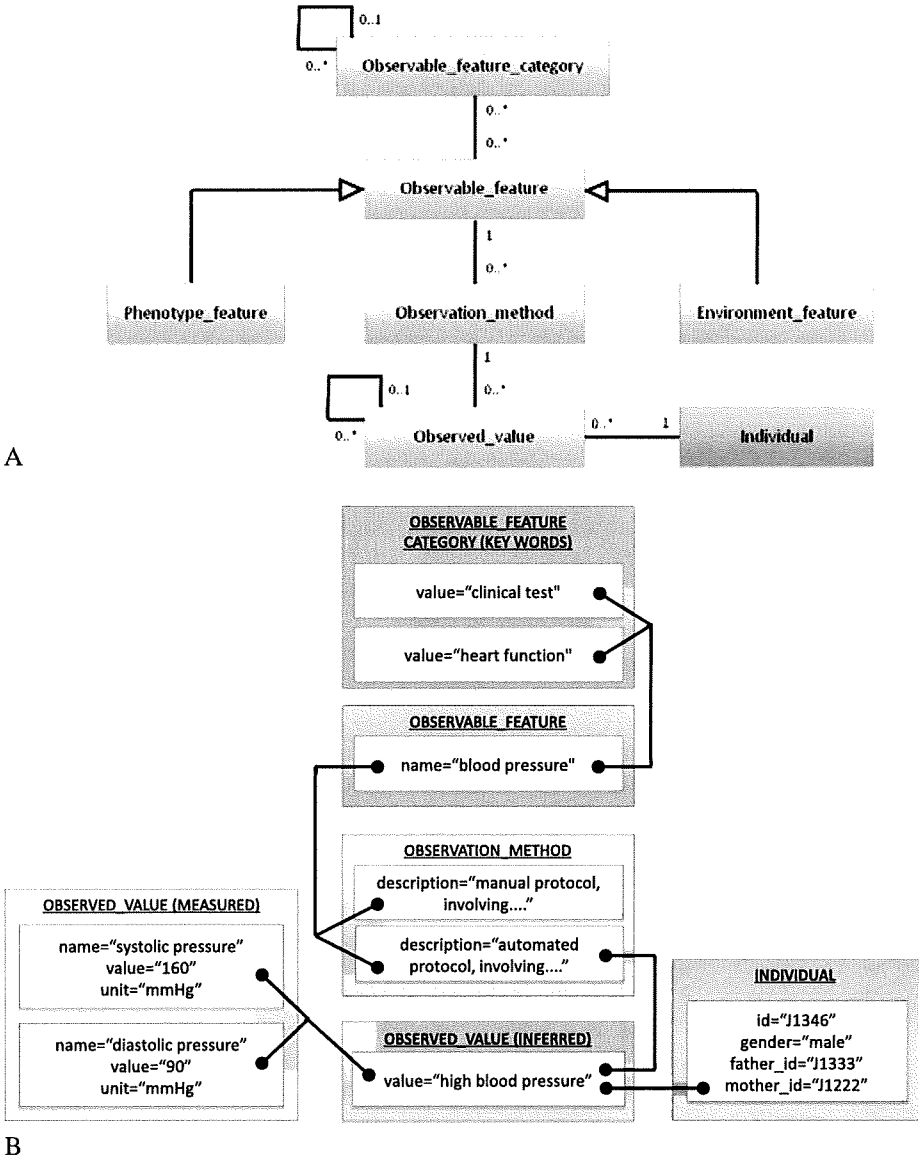

Figure 3. PHENOTYPE domain of PAGE-OM. **A:** The principal classes (colored purple) and class relationships from the PHENOTYPE domain, as described in the text. One additional class (colored blue) is also included, taken from the SAMPLE domain. **B:** Shows how this model could be used to represent 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.

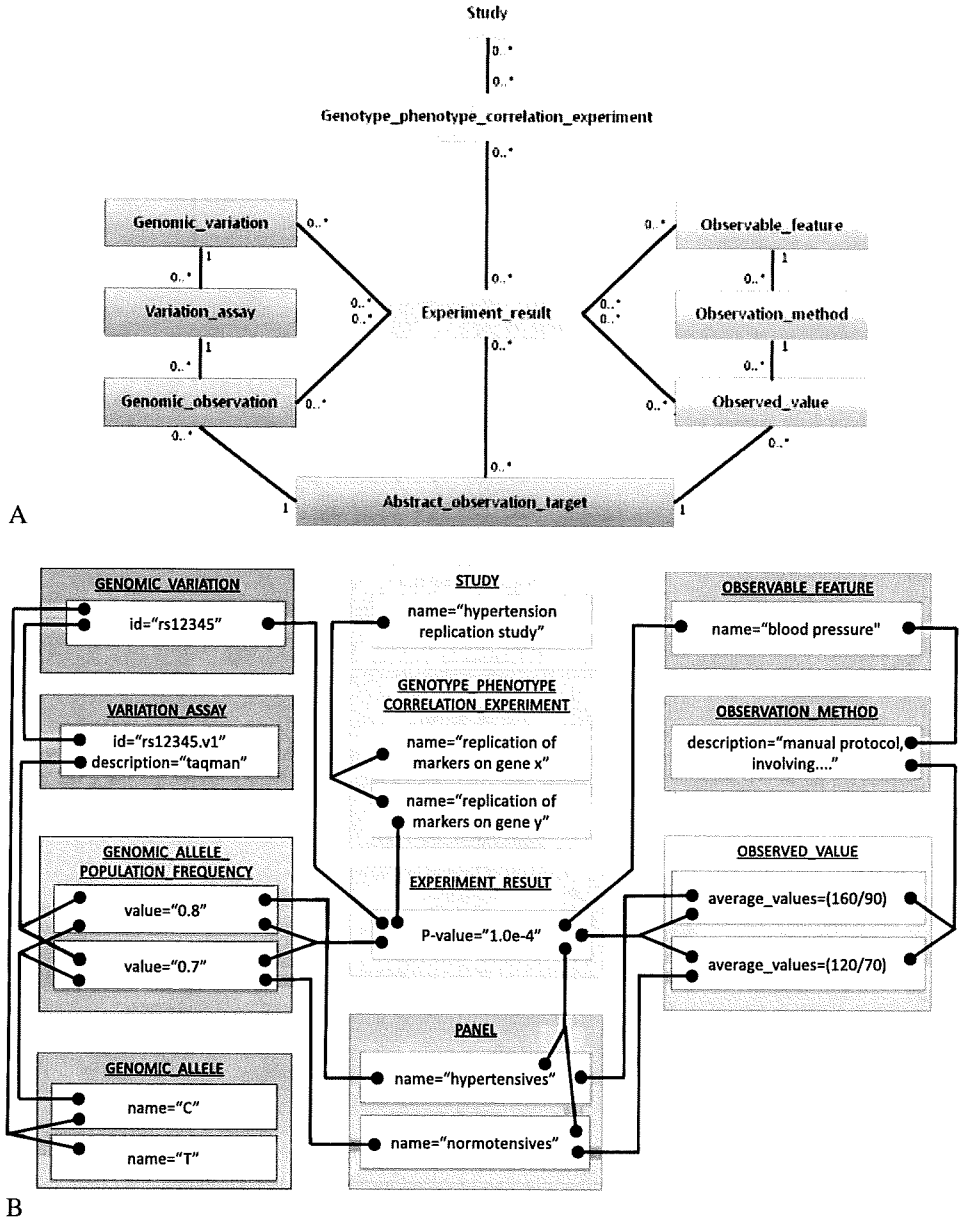

Figure 4. EXPERIMENT domain of PAGE-OM. **A:** Illustrates the principal classes (colored yellow) and class relationships from the EXPERIMENT domain, as described in the text. Additional classes are also included, taken from the SAMPLE (colored blue), GENOTYPE (colored red), and PHENOTYPE (colored purple) 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.

doi: 10.1111/j.1365-2222.2009.03229.x
Clinical & Experimental Allergy, 39, 972–981

ORIGINAL ARTICLE Asthma and Rhinitis

Gene-expression profiles in human nasal polyp tissues and identification of genetic susceptibility in aspirin-intolerant asthma

T. Sekigawa*,†, A. Tajima*, T. Hasegawa†, Y. Hasegawa‡, H. Inoue§, Y. Sano¶, S. Matsune‖, Y. Kurono‖ and I. Inoue*,**

*Department of Molecular Life Science, Tokai University School of Medicine, Isehara, Japan, †Division of Respiratory Medicine, Niigata University Graduate School of Medical and Dental Sciences, Niigata, Japan, ‡Department of Internal Medicine, Nagoya University School of Medicine, Nagoya, Japan, §Research Institute for Diseases of the Chest, Kyushu University Faculty of Medicine, Fukuoka, Japan, ¶Doai Memorial Hospital, Tokyo, Japan, ‖Department of Otolaryngology, Head and Neck Surgery, Kagoshima University Graduate School of Medical and Dental Sciences, Kagoshima, Japan and **Core Research for Evolutional Science and Technology, Japan Science and Technology Corporation, Kawaguchi, Japan

Clinical & Experimental Allergy

Summary

Background Aspirin-intolerant asthma (AIA) is a subtype of asthma induced by non-steroidal anti-inflammatory drugs and characterized by an aggressive mucosal inflammation of the lower airway (asthma) and the upper airways (rhinitis and nasal polyp). The lower airway lesion and the nasal polyp in AIA are postulated to have common pathogenic features involving aspirin sensitivity that would be reflected in the gene expression profile of AIA polyps.

Objective This study was conducted to clarify the pathogenesis of AIA using gene expression analysis in nasal polyps, and identify genetic susceptibilities underlying AIA in a case–control association study.

Methods Global gene expression of nasal polyps from nine AIA patients was examined using microarray technology in comparison with nasal polyps from five eosinophilic sinusitis (ES) patients, a related disease lacking aspirin sensitivity. Based on the AIA-specific gene expression profile of nasal polyp, candidate genes for AIA susceptibility were selected and screened by a case–control design of 219 AIA patients, 374 non-asthmatic control (CTR), and 282 aspirin-tolerant asthmatic (ATA) subjects.

Results One hundred and forty-three elevated and three decreased genes were identified as AIA-specific genes that were enriched in immune response according to Gene Ontology analysis. In addition, a *k*-means-based algorithm was applied to cluster the genes, and a subclass characteristic of AIA comprising 18 genes that were also enriched in immune response was identified. By examining the allelic associations of single nucleotide polymorphisms (SNPs) of AIA candidate genes relevant to an immune response with AIA, two SNPs, one each of *INDO* and *IL1R2*, showed significant associations with AIA ($P = 0.011$ and 0.026 after Bonferroni's correction, respectively, in AIA vs. CTR). In AIA–ATA association analysis, modest associations of the two SNPs with AIA were observed.

Conclusion These results indicate that *INDO* and *IL1R2*, which were identified from gene expression analyses of nasal polyps in AIA, represent susceptibility genes for AIA.

Keywords aspirin-intolerant asthma, candidate genes, genetic association, genome-wide gene expression, single nucleotide polymorphism
Submitted 5 March 2008; revised 11 January 2009; accepted 26 January 2009

Correspondence:
Atsushi Tajima, Department of Molecular Life Science, Tokai University School of Medicine, 143 Shimokasuya, Isehara, Kanagawa 259-1193, Japan.
E-mail: atajima@is.icc.u-tokai.ac.jp
Cite this as: T. Sekigawa, A. Tajima, T. Hasegawa, Y. Hasegawa, H. Inoue, Y. Sano, S. Matsune, Y. Kurono and I. Inoue, *Clinical & Experimental Allergy*, 2009 (39) 972–981.

Introduction

In some asthmatic patients, aspirin and several other non-steroidal anti-inflammatory drugs (NSAIDs) that inhibit cyclooxygenase enzymes (COXs) induce a severe asthmatic attack, a disease known as aspirin-intolerant asthma (AIA) [1, 2]. Several large surveys have concluded that the incidence of AIA in adult asthmatic patients is 5–15% based on patients' histories alone, but the frequency becomes two to three times higher when adult asthmatic

patients are challenged with aspirin. In women, AIA is overrepresented in a ratio of 2.3 : 1 and is more severe and has an earlier onset. AIA patients have typical clinical features including asthma, aspirin sensitivity, and bilateral nasal polyps, known as Samter's triad. Despite the well-defined pharmacological trigger, the molecular pathogenesis of AIA is still unclear. The usual hypothesis is a disturbance in the metabolism of arachidonic acid, because aspirin and NSAIDs target COXs, key enzymes of the prostaglandin biosynthetic pathway. However, the precise pathogenesis requires further investigation.

There is a moderate genetic background in AIA: the European Network on Aspirin-Induced Asthma found that 5.8% of 500 AIA patients had a family history of aspirin sensitivity [3]. First, a polymorphism in the promoter of leukotriene C_4 synthase, A-444C, was reported to be associated with AIA in Polish patients [4, 5]. A recent report showed that a haplotype of the 5-lipoxygenase gene was weakly associated with AIA in a Korean population [6]. With an extensive candidate gene analysis related to arachidonic acid metabolism, our group reported that single nucleotide polymorphisms (SNPs) in the prostaglandin E_2 receptor subtype 2 gene were significantly associated with AIA, and the functional impact of a promoter variant was further demonstrated [7]. Most recently, SNPs in prostaglandin E_2 receptor subtype 3 gene were associated in Korean population [8].

In the past few years, microarray techniques for gene expression profiling have been applied to a wide range of biological problems and have contributed to the discoveries of complex networks of biochemical processes underlying complex diseases. Microarray techniques have also helped to identify novel biomarkers, disease subtypes, and discrepancies of gene expression in human populations. Despite the advances in microarray techniques, application of the technology to identify susceptibility genes underlying complex diseases appears to be unsuccessful so far, with some exceptions [9, 10].

AIA is characterized by an aggressive mucosal inflammation of the lower airway (asthma) and the upper airways (rhinitis and nasal polyp). Rhinitis symptoms first occur in most AIA patients before the development of asthmatic intolerance to aspirin and other NSAIDs, whereas nasal polyps in AIA patients are first diagnosed at almost the same time aspirin intolerance appears [3]. We postulated that the lower airway lesion and the polyp in AIA have a common pathophysiology of aspirin intolerance, suggesting the nasal polyp as a pleiotropic genetic model of the bronchial inflammation of AIA. Global gene expression of the nasal polyps of AIA patients was examined using microarray technology for comparison with nasal polyps of eosinophilic sinusitis (ES) patients: ES is typically characterized by a nasal polyp with an inflammatory cell infiltration similar to that in an AIA polyp but without aspirin sensitivity, thus being an appropriate reference for the selection of AIA-specific genes.

Materials and methods

Nasal polyp tissues and Aspirin-Intolerant Asthma Subjects

Nasal polyp tissues for microarray analysis were obtained from nine Japanese patients (aged from 35 to 76 years, five males/four females) with AIA, five (aged from 34 to 73 years, three males/two females) with ES, and two (aged 61 and 71 years, both males) with only chronic sinusitis (CS) (Table 1). These patients had not been exposed to preoperative treatment with steroids for at least 1 year before surgery. According to the definition of rhinosinusitis, CS with nasal polyps with eosinophilic inflammatory features without fungal hyphae includes aspirin-sensitive and aspirin-tolerant types [11]. Thus, three groups of patients with nasal polyps were sequentially defined as follows: first, CS with nasal polyps was diagnosed based on clinical symptoms, such as nasal discharge, postnasal drip, headache, hyposmia, and nasal obstruction, and endonasal findings of muco-purulent secretion and nasal polyps with a paranasal shadow observed by CT examination [12]. Among CS patients with nasal polyps, ES patients were identified histologically by counting the number of eosinophils at ×200 magnification under light microscopy. Five fields were examined for each section,

Table 1. Clinical characteristics of patients with nasal polyps for microarray analysis

		Parameters in peripheral blood				
ID	Age/ gender	WBC (/mm^3)	Eosinophil (%)	Allergic rhinitis	Asthma	AIA episode
AIA#1	76/M	8000	3	–	+	+
AIA#2	48/M	5500	13	–	+	+
AIA#3	73/M	6500	3	–	+	+
AIA#4	59/F	9500	28	–	+	+
AIA#5	50/F	5720	14	–	+	+
AIA#6	40/M	9100	4	–	+	+
AIA#7	35/M	8800	6	–	+	+
AIA#8	50/F	6000	9	+	+	+
AIA#9	66/F	7000	8	–	+	+
ES#1	73/F	7200	2	–	+	–
ES#2	64/F	6400	23	–	+	–
ES#3	69/M	7700	4	+	–	–
ES#4	61/M	4900	5	–	+	–
ES#5	34/M	6300	3	+	+	–
CS#1	61/M	7400	10	–	+	–
CS#2	67/M	9700	10	–	–	–

M, male; F, female; WBC, white blood cell; –, no allergic rhinitis, no asthma, or no AIA episode; AIA, aspirin-intolerant asthma; CS, chronic sinusitis; ES, eosinophilic sinusitis.

and the average was considered to be the number of eosinophils infiltrating the sample. Nasal polyps having more than 100 eosinophils were classified as ES [12]. Among ES patients, those who had had apparent episodes of asthma attacks in response to aspirin and other NSAIDs were classified as AIA patients (AIA#1–9). The remaining five ES patients without AIA episodes (ES#1–5) had no troubles even after taking NSAIDs in postoperative courses during hospitalization. The oral provocation test for AIA patients was not performed in most of the patients due to potential risk, although severe reactions against the provocation were improbable [13], and only verbal history has yielded some false positives [14]. The ethics committees of Kagoshima University approved the study protocols, and each participant gave written informed consent.

DNA samples from 219 unrelated individuals with AIA (age: 55.7±13.5 years; 70 males/149 females) and 374 non-asthmatic controls (CTR) (age: 44.5±23.2 years; 181 males/193 females) were obtained as described previously [7]. For AIA-associated SNPs, 282 unrelated individuals with aspirin-tolerant asthma (ATA) (age: 56.0±16.1 years; 132 males/150 females) [7] were also genotyped, and used as asthmatic controls. The subjects were recruited at Niigata University Hospital, University of Tokyo Hospital, Nagoya University Hospital, Doai Memorial Hospital, and Kyushu University Hospital, with Institutional Review Board approvals. The diagnosis of AIA was based on a self-reported history due to the potential risk of a provocation test. ATA was defined as adult asthma diagnosed by expert physicians according to the American Thoracic Society criteria [15] and no history of aspirin or NSAID-induced asthmatic attack, and comprised of 154 atopic asthmatic (age: 48.0±15.6 years; 80 male/74 female) and 128 non-atopic asthmatic (age: 65.9±10.0 years; 52 male/76 female) subjects. CTR were outpatients with diseases (e.g., hypertension) other than respiratory diseases including asthma, and who self-reported no history of aspirin sensitivity. The patients and controls were all of Japanese ethnicity. Although the Japanese population is thought to be genetically homogenous, nearly identical numbers of patients and controls from the various locations were recruited to avoid geographical differences in allelic frequencies.

RNA extraction

The nasal polyp tissue was removed during endoscopic sinus surgery, submerged in RNA*later* regent (Ambion Inc., Austin, TX, USA) to avoid RNA degradation, and used for RNA extraction within 48 h after resection. Total RNA was extracted using TRIzol reagent (Invitrogen, Carlsbad, CA, USA) according to the manufacturer's instructions. The quality and quantity of the extracted RNA were analysed using the Agilent 2100 bioanalyzer (Agilent Technologies Inc., Palo Alto, CA, USA) with an RNA6000 Nano LabChip Kit (Agilent Technologies). RNAs from two CS patients were equally pooled, and used as a common reference in the two-colour microarray experiments, where a single microarray was used to compare each test sample from an AIA or an ES patient with the reference sample.

cRNA synthesis, labelling, hybridization, and expression profiling

For fluorescent cRNA synthesis, high-quality total RNA (150 ng) was labelled with the Low RNA Input Fluorescent Linear Amplification Kit (Agilent Technologies) according to the manufacturer's instructions. In this procedure, cyanine 5-CTP (Cy5) and cyanine 3-CTP (Cy3) (PerkinElmer, Boston, MA, USA) were used to generate labelled cRNA from the individual AIA or ES RNA and the pooled CS RNA as a reference, respectively. Labelled cRNAs (0.75 μg each) from the AIA, ES, or CS patients were fragmented in a hybridization mixture with the In Situ Hybridization Kit Plus (Agilent Technologies) according to the manufacturer's instructions. The mixture was hybridized for 17 h at 65 °C to an Agilent Human 1A(v2) Oligo Microarray. After hybridization, the microarray was washed with SSC buffer, and then scanned in Cy3 and Cy5 channels with the Agilent DNA Microarray Scanner, model G2565AA. Signal intensity per spot was generated from the scanned image with Feature Extraction Software ver7.5 (Agilent Technologies) in default settings. Spots that did not pass quality control procedures with the software were flagged and removed for further analysis.

GeneSpring software GX 7.3 (Agilent Technologies) was used for the Lowess (locally weighted linear regression curve fit) normalization of the ratio (Cy5/Cy3) of the signal intensities generated in each microarray and the subsequent data analysis. To determine the AIA-specific expression profile of nasal polyps, ES transcripts with ratios ranging from 0.5 to 2 were extracted, and the AIA transcripts with expression undergoing a twofold change or more were extracted as decreased or elevated genes. Of the transcripts overlapping the two groups, only those with statistically significant differences in expression between the AIA and CS nasal polyps (Benjamini and Hochberg false discovery rate (FDR)<0.01; [16]) were counted as AIA-specific genes. To identify novel expression patterns in nasal polyps from AIA patients, the *k*-means method [17], a well-known unsupervised partitional clustering approach, was applied to the AIA-specific genes. For functional subclassification of the AIA-specific genes, we applied the Gene Ontology (GO) classification for biological processes with DAVID 2.1 (http://david.abcc.ncifcrf.gov/), a web-accessible program [18]. A permutation test with 10 000 iterations was used for multiple test correction when nasal polyps from AIA

