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 interrelationships (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.pageom.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. © 2009 Wiley-Liss, Inc.

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

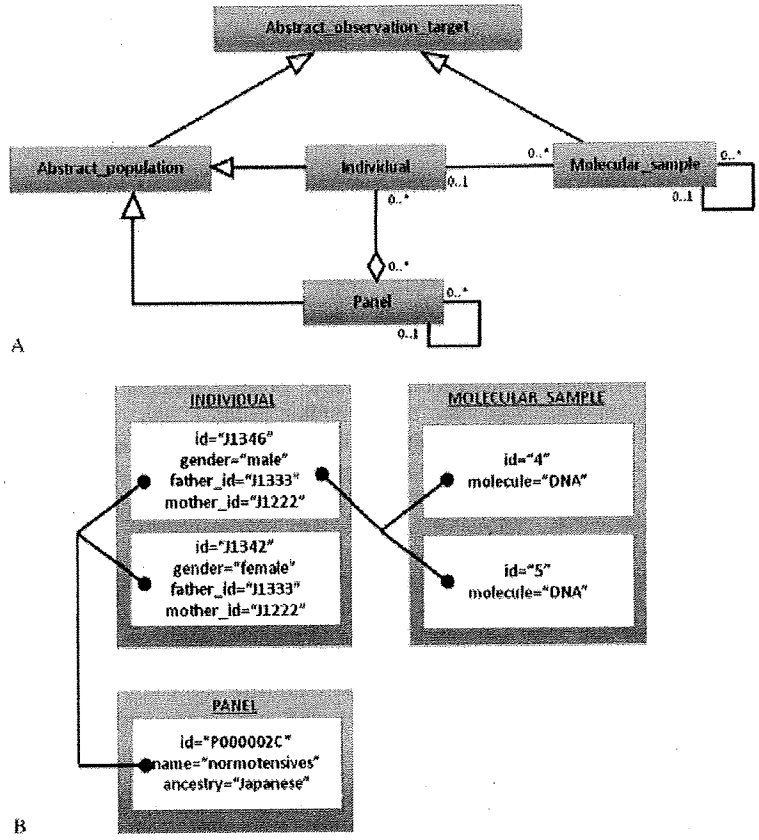

Figure 1. SAMPLE domain of PAGE-OM. **A:** The principal classes (colored blue) and class relationships from the SAMPLE domain, as described in the text. **B:** Shows how the model in (A) could be used to represent a cohort of normotensive Japanese, giving further details for a brother and sister from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

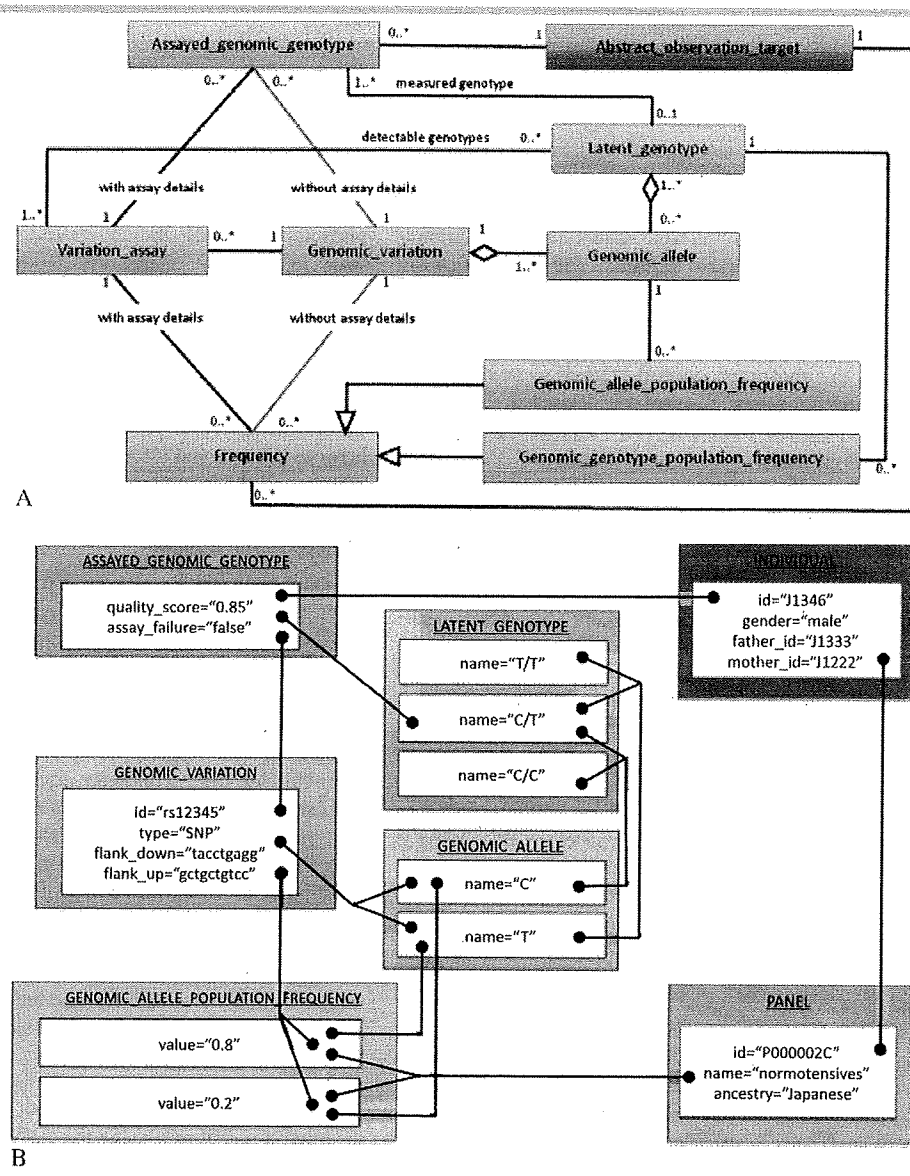

Figure 2. GENOTYPE domain of PAGE-OM. **A:** The principal classes (colored red) and class relationships from the GENOTYPE domain, as described in the text. One additional class (colored blue) is also included, taken from the SAMPLE domain. At the project website, sections of the model called Marker,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

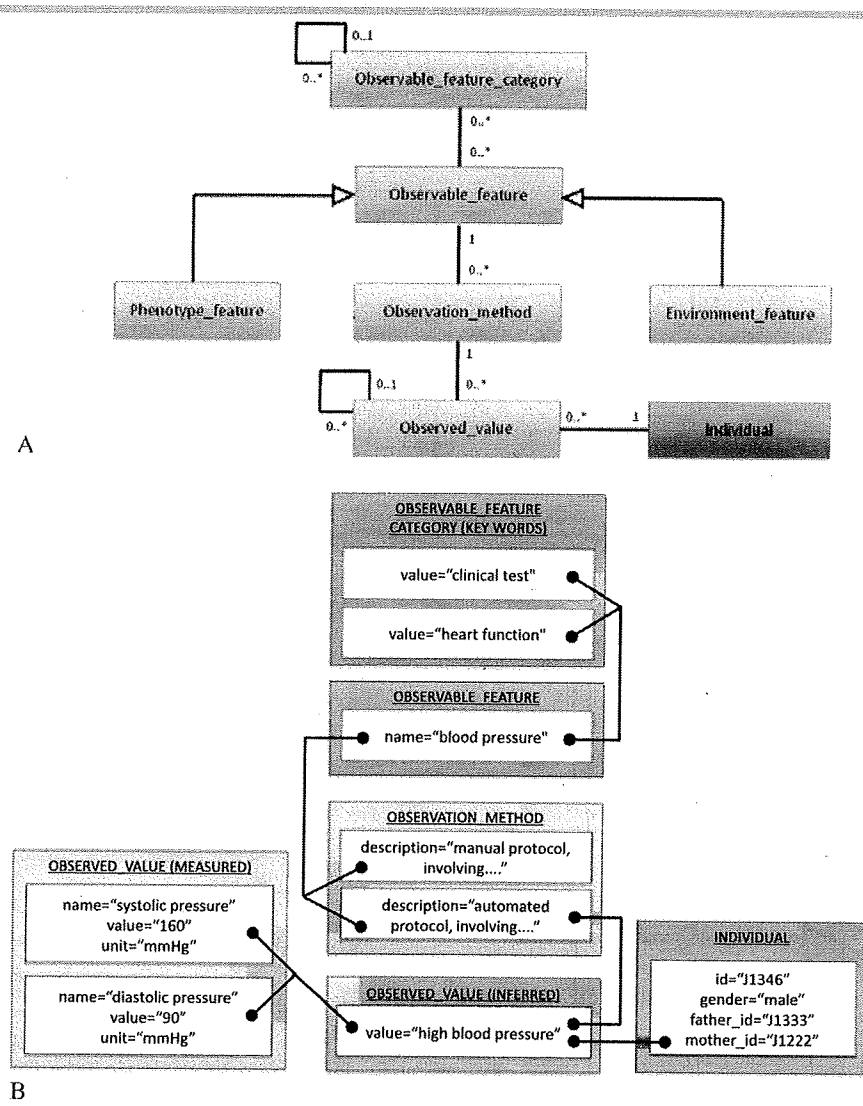

Figure 3. PHENOTYPE domain of PAGE-OM. **A:** The principal classes (colored purple) and class relationships from the PHENOTYPE domain, as described in the text. One additional class (colored blue) is also included, taken from the SAMPLE domain. **B:** Shows how this model could be used to represent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

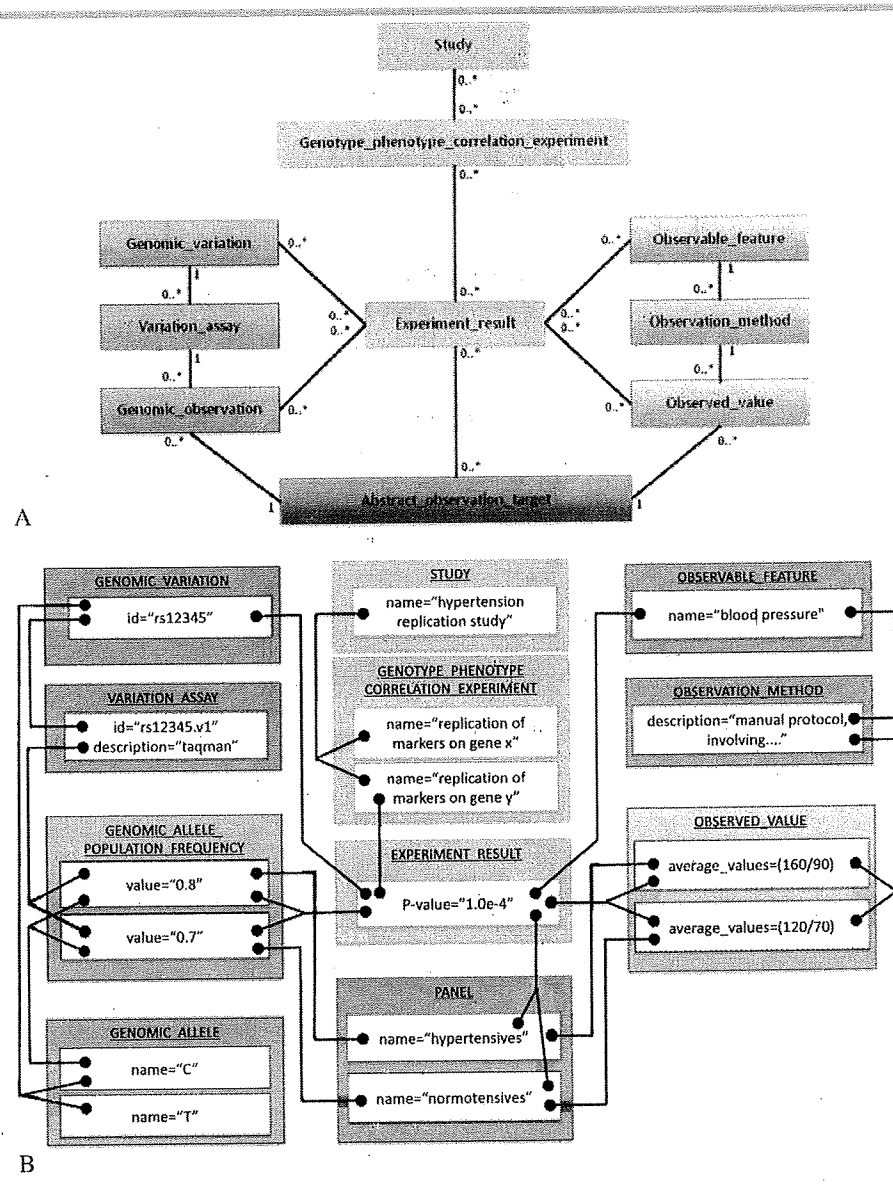

Figure 4. EXPERIMENT domain of PAGE-OM. **A:** Illustrates the principal classes (colored yellow) and class relationships from the EXPERIMENT domain, as described in the text. Additional classes are also included, taken from the SAMPLE (colored blue), GENOTYPE (colored red), and PHENOTYPE (colored purple)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

Gene expression profiling of advanced-stage serous ovarian cancers distinguishes novel subclasses and implicates *ZEB2* in tumor progression and prognosis

Kosuke Yoshihara,[1,10] Atsushi Tajima,[2,10] Dai Komata,[1] Tadashi Yamamoto,[3] Shoji Kodama,[4] Hiroyuki Fujiwara,[5] Mitsuaki Suzuki,[5] Yoshitaka Onishi,[6] Masayuki Hatae,[6] Kazunobu Sueyoshi,[7] Hisaya Fujiwara,[8] Yoshiki Kudo,[8] Ituro Inoue[2,9] and Kenichi Tanaka[1,9]

[1]Department of Obstetrics and Gynecology, Niigata University Graduate School of Medical and Dental Sciences, Niigata; [2]Department of Molecular Life Science, Tokai University School of Medicine, Isehara, Japan; [3]Department of Structural Pathology, Institute of Nephrology, Niigata University Graduate School of Medical and Dental Sciences; [4]Department of Gynecology, Niigata Cancer Center Hospital, Niigata; [5]Department of Obstetrics and Gynecology, Jichi Medical University, Shimotuke; [6]Department of Obstetrics and Gynecology, Kagoshima City Hospital; [7]Department of Pathology, Kagoshima City Hospital, Kagoshima; [8]Department of Obstetrics and Gynecology, Hiroshima University Graduate School of Biomedical Sciences, Hiroshima, Japan

(Received January 30, 2009/Revised April 16, 2009/Accepted April 24, 2009/Online publication May 26, 2009)

To elucidate the mechanisms of rapid progression of serous ovarian cancer, gene expression profiles from 43 ovarian cancer tissues comprising eight early stage and 35 advanced stage tissues were carried out using oligonucleotide microarrays of 18 716 genes. By non-negative matrix factorization analysis using 178 genes, which were extracted as stage-specific genes, 35 advanced stage cases were classified into two subclasses with superior ($n = 17$) and poor ($n = 18$) outcome evaluated by progression-free survival (log rank test, $P = 0.03$). Of the 178 stage-specific genes, 112 genes were identified as showing different expression between the two subclasses. Of the 48 genes selected for biological function by gene ontology analysis or Ingenuity Pathway Analysis, five genes (*ZEB2*, *CDH1*, *LTBP2*, *COL16A1*, and *ACTA2*) were extracted as candidates for prognostic factors associated with progression-free survival. The relationship between high *ZEB2* or low *CDH1* expression and shorter progression-free survival was validated by real-time RT-PCR experiments of 37 independent advanced stage cancer samples. *ZEB2* expression was negatively correlated with *CDH1* expression in advanced stage samples, whereas *ZEB2* knockdown in ovarian adenocarcinoma SKOV3 cells resulted in an increase in *CDH1* expression. Multivariate analysis showed that high *ZEB2* expression was independently associated with poor prognosis. Furthermore, the prognostic effect of E-cadherin encoded by *CDH1* was verified using immunohistochemical analysis of an independent advanced stage cancer samples set ($n = 74$). These findings suggest that the expression of epithelial–mesenchymal transition-related genes such as *ZEB2* and *CDH1* may play important roles in the invasion process of advanced stage serous ovarian cancer. (*Cancer Sci* 2009; 100: 1421–1428)

The serous type, comprising approximately 50% of ovarian cancers, is the most aggressive histology and has a tendency to be detected as advanced stage at the time of diagnosis.[1,2] Patients with advanced stage serous ovarian cancer are managed with surgical cytoreduction followed by platinum and taxane-based chemotherapy. Serous ovarian cancer is moderately chemosensitive and initially responds to postoperative chemotherapy, but the survival of patients with advanced stage remains poor. Because the majority of early stage ovarian cancers are asymptomatic and there is as yet no reliable screening test, it is difficult to diagnose early stage serous ovarian cancer. Therefore, the molecular mechanisms of progression in serous ovarian cancer should provide valuable clues for early detection and improved prognosis.

The development of microarray technology permits analysis of the expression levels of thousands of genes in cancer cells, and several studies have shown that microarrays can be used to identify gene expression profiles associated with surgery outcome, response to chemotherapy, grade, and survival in ovarian cancers.[3–17] However, there are limited reports of microarray analysis on tumor progression.[18–20] Serous ovarian cancer more rapidly progresses to advanced stage than other histological types.[21] In the present study, we used genome-wide expression microarray to distinguish between stage I (ovary confined) and stage III/IV serous ovarian cancers to focus on the molecular mechanisms of tumor progression and metastasis. Our microarray analysis identified 178 stage-specific genes, and also divided advanced stage (stage III/IV) ovarian cancers into two novel prognostic subclasses, by the NMF method. There were significant differences between the two subclasses in progression-free survival time. Furthermore, we extracted *CDH1* and its transcriptional repressor *ZEB2* from the 112 genes that were differentially expressed between the two novel subclasses, and found that the expression levels of these epithelial–mesenchymal transition-related genes[22,23] are associated with tumor progression and prognosis in advanced stage serous ovarian cancer patients.

Materials and Methods

Tissue samples. Eighty-nine patients (17 stage I; 72 stage III/IV) who were diagnosed with serous histological type ovarian cancer between July 1997 and October 2007 were recruited in this study. Fresh-frozen samples were obtained from primary tumor tissues at initial cytoreductive surgery. No patients received chemotherapy before surgery. All patients with advanced stage serous ovarian cancer ($n = 72$) were treated with platinum and taxane-based chemotherapy after surgery. The ethics committees of the participating institutions approved the study protocol, and each participant gave written, informed consent. Of the 89 samples, 43 were analyzed with microarray. The remaining 46 samples were used for subsequent validation analysis. There were no significant differences between the two samples sets regarding age of onset, stage, performance of optimal cytoreduction, histological grade, and follow-up period between the microarray set and validation set (Supplementary Table 1). Staging of the disease was assessed in accordance with the criteria of the International Federation of Gynecology and Obstetrics.[24] Optimal cytoreduction was defined as ≤1 cm of gross residual disease. The histological characteristics of surgically resected specimens

[9]To whom correspondence should be addressed.
E-mail: tanaken@med.niigata-u.ac.jp or ituro@is.icc.u-tokai.ac.jp
[10]These authors contributed equally to this work.

doi: 10.1111/j.1349-7006.2009.01204.x

were assessed on formalin-fixed and paraffin-embedded hematoxylin–eosin sections, and frozen tissues containing more than 80% tumor cells were used for RNA extraction. Normal peritoneum tissues were obtained from 10 patients having other procedures (such as hysterectomy for myoma uteri) at Niigata University. Tumors of 43 samples used for microarray analysis were screened for the presence of *TP53* somatic mutations using previously reported methods.[25] Four patients with family history of ovarian cancer in the microarray set were examined for germline mutations of *BRCA1* according to an in-house protocol,[26] and two patients showed mutations of *BRCA1*.

Microarray experiments. Total RNA, extracted from tissue samples using TRIzol reagent (Invitrogen, Carlsbad, CA, USA) was examined with a 2100 Bioanalyzer (Agilent Technologies, Palo Alto, CA, USA) using an RNA 6000 Nano LabChip (Agilent Technologies). Five hundred nanograms of total RNA was converted into labeled cRNA with nucleotides coupled to Cy3 (PerkinElmer, Boston, MA, USA) using the Low RNA Input Fluorescent Linear Amplification Kit (Agilent Technologies). Cy3-labeled cRNA (1.5 μg) was hybridized for 17 h at 65°C to an Agilent Human 1A (v2) Oligo Microarray, which carries 60-mer probes to 18 716 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 with Feature Extraction Software version 8.5 (Agilent Technologies) with the default settings. Spots that did not pass quality control procedures were flagged as 'absent'.

Microarray data analysis. Data normalization was carried out using GeneSpring GX 7.3 (Agilent Technologies) as follows: (i) values below 0.01 were set to 0.01, following background subtraction; and (ii) median percentile normalization was carried out using a per-chip 50th percentile of all measurements. Furthermore, genes with expression levels marked as 'absent' in more than 22 of 43 microarrays were excluded to analyze ovarian cancer-specific transcripts. When the gene expression patterns of two groups were compared, genes showing twofold or more mean expression differences between the groups were first determined by Welch's *t*-test in GeneSpring GX. For multiple testing corrections in this statistical analysis, the Benjamini–Hochberg procedure[27] of controlling the false discovery rate at the level of 0.05 was used.

To assess heterogeneity of the gene expression profile among serous ovarian cancer patients, we applied a NMF algorithm and hierarchical clustering using stage-specific gene expression profiles. NMF analysis was carried out according to Brunet *et al.*[28] as previously reported.[29]

To investigate the biological functions of the gene expression profiles, we used GO Ontology Browser, embedded in GeneSpring GX, and IPA (http://www.ingenuity.com). More detailed information about this analysis using the GO Ontology Browser and IPA is given in Supplementary methods.

Quantitative RT-PCR analysis. Total RNA (1 μg) from ovarian cancer was used as a template in first-strand cDNA synthesis with the SuperScript III First-Strand Synthesis System (Invitrogen). The cDNA was diluted one in ten for subsequent real-time PCR, which was carried out using TaqMan Gene Expression Assays (Applied Biosystems) with TaqMan Universal PCR Master Mix (Applied Biosystems) on a 7900HT Sequence Detection System (Applied Biosystems) according to the manufacturers' instructions. Detailed information on the 23 transcripts examined is summarized in Supplementary Table 2. The relative quantification method[30] was used to measure the amounts of the respective genes in serous ovarian cancer samples, normalized to *ACTB* and *TBP*.

Analysis of clinical and pathological parameters. All analyses except Cox's proportional hazard analysis were done using GraphPad PRISM version 4.0 (GraphPad Software, San Diego, CA, USA). Survival curves were investigated using the Kaplan–Meier method and log rank test (GraphPad PRISM). When clinicopathological parameters among ovarian cancer patients were compared, unpaired *t*-test, Fisher's exact test or χ^2-test was used depending on the purpose (GraphPad PRISM). Pearson's correlation coefficient was calculated for correlation between *ZEB2* expression and *CDH1* expression. Differences in gene expression levels between two subclasses were tested by Mann–Whitney test. Using a $\log_2$ transformation of expression data, Cox's proportional hazard model analysis was carried out using JMP version 6 (SAS Institute, Cary, NC, USA).

Results

Identification and characterization of molecular subclasses from advanced stage serous ovarian cancer cases. Using Agilent Human 1A(v2) Oligo microarray, we generated gene expression data for 43 serous ovarian cancers comprising eight stage I and 35 stage III/IV tumors, as well as 10 normal peritoneum tissues as a reference. First, 4275 ovarian cancer-specific genes that were differentially expressed between ovarian cancer and peritoneum tissues were isolated. Of these 4275 transcripts, 178 stage-specific genes showing significantly more than twofold upregulation or downregulation in stage III/IV samples compared to stage I samples; 107 transcripts were upregulated and 71 transcripts downregulated in stage III/IV serous ovarian cancers (Supplementary Fig. 1).

To clarify the heterogeneity of the samples at the transcriptome level, 43 serous ovarian cancer samples were analyzed by the NMF method[28,29,31] using the 178 transcriptomes that were differentially expressed between stage I samples and stage III/IV samples. Figure 1(A) shows reordered consensus matrices averaging 50 connective matrices generated for subclasses $K = 2$, 3, 4, and 5. The most distinct pattern of block partitioning was observed at the $K = 2$ model. Thus, the NMF method predicts the existence of robust subclasses of serous ovarian cancer samples for $K = 2$. This prediction was quantitatively supported by higher values of coph for NMF-clustered matrices. The NMF class assignment for $K = 2$ was the most robust with the highest coph value (coph = 0.999). Interestingly, one subclass in the $K = 2$ model was composed of eight stage I samples and 17 stage III/IV samples, whereas the other was composed of 18 stage III/IV samples. To verify the accuracy and robustness of the classification, a hierarchical clustering approach was also applied to log-transformed normalized data for stage-specific target genes. As depicted in Figure 1(B), 43 serous ovarian cancer samples were separated into two main branches showing similarity with the NMF-based subclassification. Thus, it was confirmed that the 35 advanced stage serous ovarian cancer samples were categorized into two distinct subclasses at the transcriptome level. A group composed of 17 stage III/IV samples with gene expression profiles similar to stage I samples was termed 'subclass 1', and the second group comprising 18 stage III/IV samples was termed 'subclass 2'. Two patients were identified as harboring *BRCA1* mutations: one patient belonged to stage I and the other to subclass 1 in the array analysis, but there was no particular gene expression pattern due to the mutations based on the expression levels of the 178 stage-specific genes.

We then investigated the possibility that the two subclasses of advanced stage serous ovarian cancers split by the NMF approach might represent clinically, pathologically, or genetically distinct characteristics. The distribution of several known prognostic factors is listed in Table 1. The two subclasses were similar in age of onset, stage, CA125 level before treatment, presence of tumor cells in ascites, histological grade, presence of lymph node metastasis, and frequency of *TP53* mutations, except that subclass 1 had a higher rate of optimal cytoreduction than subclass 2 (Fisher's exact test, $P = 0.09$). When the outcome of two

doi: 10.1111/j.1349-7006.2009.01204.x

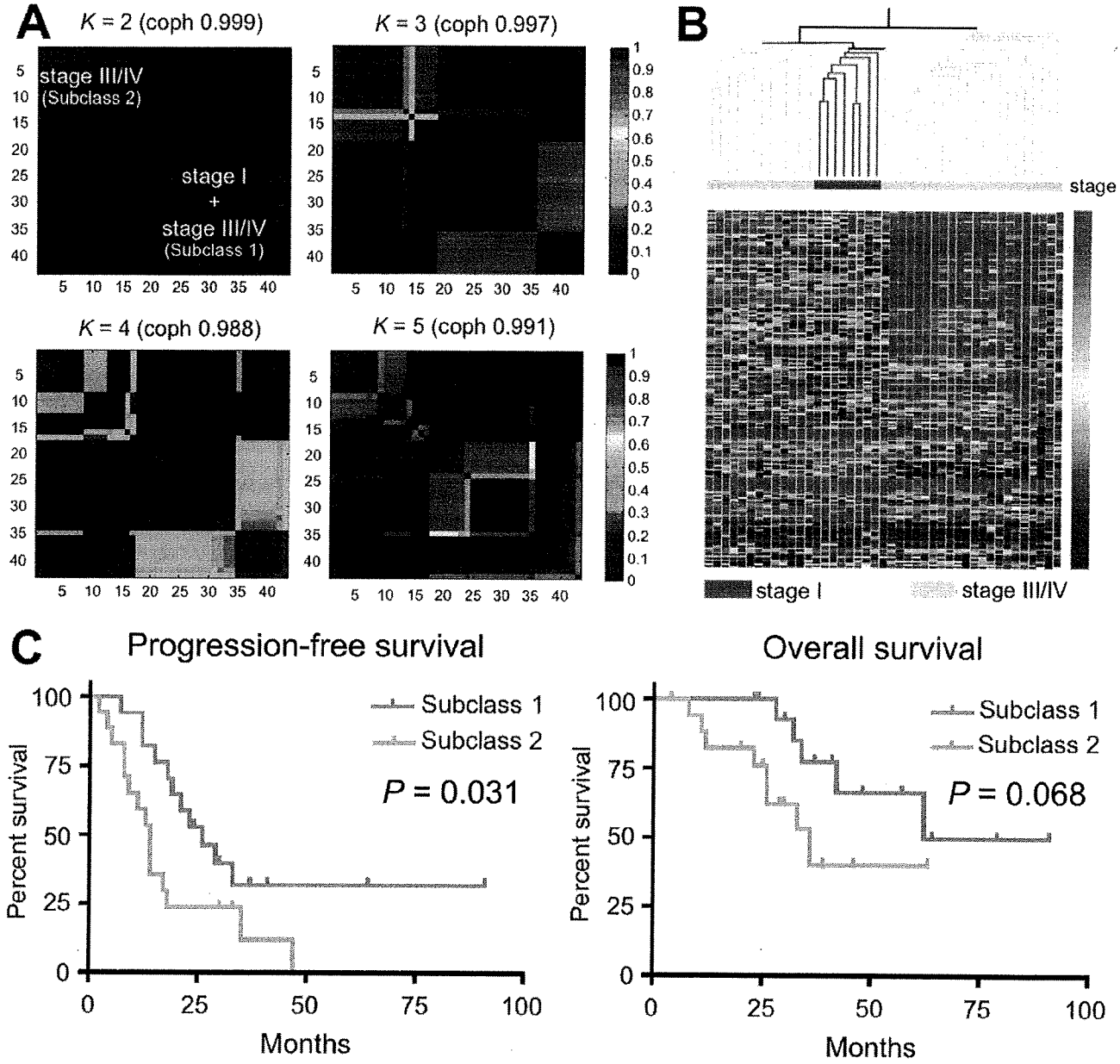

Fig. 1. Subclassification of 43 serous ovarian cancer samples and their prognosis. (A) By a non-negative matrix factorization (NMF) approach, NMF-consensus matrices averaging 50 connectivity matrices were computed at $K = 2$–5 (as the number of subclasses modeled) for the 43 serous ovarian cancer samples with 178 stage-specific genes. The NMF computation and model selection were carried out according to Brunet *et al.*[28] By accounting for the cophenetic correlation coefficients (coph) for NMF-clustered matrices, the NMF class assignment at $K = 2$ was the most robust. One subclass in the $K = 2$ model contained samples both from stage I ($n = 8$) and III/IV ($n = 17$), whereas the other contained only stage III/IV samples ($n = 18$). (B) A hierarchical clustering method was also used to classify serous ovarian cancer samples using 178 stage-specific genes. The 43 serous samples were largely separated into two clusters. The stage assignments for samples are: stage I, red; and stage III/IV, yellow. (C) Kaplan–Meier survival curves between two NMF-based subclasses of the 35 stage III/IV patients. Subclass 1, composed of 17 advanced stage patients with gene expression profiles similar to that of stage I, showed statistically prolonged progression-free survival (log rank test, $P = 0.031$), but no significant correlation with overall survival (log rank test, $P = 0.068$).

subclasses was compared for progression-free survival and overall survival, the Kaplan–Meier curves showed significantly better outcome in cases belonging to subclass 1 in progression-free survival (Fig. 1C, log rank test, $P = 0.031$) and fair outcome in overall survival (Fig. 1C, log rank test, $P = 0.068$).

Association of subclass-specific gene expression profile with prognosis of ovarian cancer patients. To characterize the gene expression differences associated with distinct prognoses between the two subclasses of advanced stage serous ovarian cancers, we identified 112 subclass-specific transcripts that were differentially expressed between the two subclasses; 25 transcripts were up-regulated in subclass 1 and 87 transcripts were upregulated in subclass 2 (Supplementary Table 3). We then examined the biological functions of the 112 subclass-specific genes using two analytic tools, GO analysis and IPA, to clarify the biological mechanism of tumor progression. The Gene Ontology Biological

Table 1. Clinical characteristics of two subclasses of advanced stage serous ovarian cancer samples

Characteristic	Subclass 1 ($n = 17$)	Subclass 2 ($n = 18$)	*P*-value
Age (years)	58.5 ± 8.6	61.1 ± 12.6	0.49†
Stage			
Stage III	16	15	1‡
Stage IV	1	3	
CA125 (IU)	1987 ± 2021	1178 ± 1057	0.14†
Cancer cell in abdominal fluid			
Positive	15	15	1‡
Negative	2	3	
Optimal cytoreduction			
Optimal (<1 cm)	12	7	0.09‡
Not optimal	5	11	
Lymph node metastasis			
Positive	6	4	1‡
Negative	9	6	
Unknown	2	8	
Grade			
Grade 1	6	4	0.18§
Grade 2	9	7	
Grade 3	2	7	
TP53 status			
Wild type	11	9	0.50‡
Mutated	6	9	

Differences in clinical characteristics between subclass 1 and subclass 2 were tested using the †unpaired *t*-test, ‡Fisher's exact test or §χ^2-test.

Table 2. Univariable Cox's proportional hazards model analysis of expression levels of five genes for progression-free survival and overall survival in patients with advanced stage serous ovarian cancers

Gene symbol	Hazard ratio (95% CI)	*P*-value
Microarray set (n = 35)		
Progression-free survival		
ZEB2	1.35 (1.06–1.77)	0.015*
CDH1	0.75 (0.62–0.94)	0.017*
LTBP2	1.63 (1.04–2.57)	0.032*
COL16A1	1.33 (1.02–1.74)	0.034*
ACTA2	1.21 (1.01–1.46)	0.036*
Overall survival		
ZEB2	1.56 (1.06–2.47)	0.023*
CDH1	0.81 (0.67–1.03)	0.081
LTBP2	2.53 (1.43–4.58)	0.0017*
COL16A1	1.66 (1.12–2.59)	0.012*
ACTA2	1.44 (1.10–1.95)	0.0087*
Validation set (n = 37)		
Progression-free survival		
ZEB2	1.74 (1.08–2.92)	0.023*
CDH1	0.20 (0.09–0.45)	0.00006*
LTBP2	1.16 (0.75–1.75)	0.49
COL16A1	1.18 (0.92–1.51)	0.20
ACTA2	1.22 (0.90–1.66)	0.19
Overall survival		
ZEB2	1.89 (1.06–3.64)	0.029*
CDH1	0.59 (0.26–1.30)	0.19
LTBP2	1.1 (0.70–1.66)	0.69
COL16A1	1.23 (0.93–1.66)	0.15
ACTA2	1.43 (0.99–2.13)	0.052

*$P < 0.05$.

Process categories over-represented among 112 subclass-specific genes are shown in Figure 2(A). After multiple testing corrections using the Benjamini–Hochberg FDR method, seven categories were significantly over-represented, and included 37 non-overlapping genes. Subclass-specific genes were involved in biological processes of transport (GO6817, GO15698, GO6820, and GO6811), development (GO48513 and GO1501), and cell adhesion (GO7155), and included a high proportion of extracellular matrix-related genes. In addition, when ID of Agilent probes of 112 subclass-specific transcripts were imported into the IPA software, a new pathway comprising 26 genes that were enriched in extracellular matrix genes was identified (Fig. 2B). Fifteen genes belonged to both the seven GO categories and the new network, and 48 non-redundant genes were biologically characterized.

To investigate whether the expression profile of the 48 genes extracted by GO analysis or IPA was implicated in the aggressive phenotype of ovarian cancer, we analyzed the association between the respective expression levels of the 48 genes and progression-free survival time using univariate Cox's proportional hazard model. The expression levels of *ZEB2*, *CDH1*, *LTBP2*, *COL16A1*, and *ACTA2* were significantly correlated with progression-free survival (Table 2). When overall survival also was evaluated by Cox's proportional hazard model, the expression of the above genes except *CDH1* was significantly correlated with overall survival.

Validation by quantitative real-time RT-PCR. To validate the microarray expression data, we measured expression levels of 23 randomly selected transcripts from the 112 subclass-specific transcripts by real-time RT-PCR analysis. In agreement with microarray results, there was a significant difference between the expression levels of the 23 transcripts measured by real-time RT-PCR of subclass 1 and subclass 2 (Supplementary Table 4).

To validate the previous findings that the expression levels of *ZEB2*, *CDH1*, *LTBP2*, *COL16A1*, and *ACTA2* are associated with progression-free survival, quantitative real-time RT-PCR was carried out on 46 samples comprising nine stage I samples and 37 stage III/IV samples recruited as an independent validation set. Cox's proportional hazard analysis showed that the expression levels of *ZEB2* and *CDH1* were again correlated with progression-free survival ($P = 0.023$ and 0.00006, respectively) (Table 2). Moreover, *ZEB2* expression was significantly associated with overall survival ($P = 0.029$). At the protein level, an association of the expression of E-cadherin (encoded by *CDH1*) with prognosis of advanced stage serous ovarian cancer patients was further verified by immunohistochemical analysis of independent samples ($n = 74$) (Supplementary Fig. 3) as previously reported.[32–35]

Interaction between *ZEB2* and *CDH1*. *ZEB2* directly interacted with *CDH1* in the IPA network, as shown in Figure 2(B). We also found a significantly negative correlation between *ZEB2* expression and *CDH1* expression (Pearson's correlation coefficient: –0.432, $P = 0.0002$) in advanced stage serous ovarian cancers using real-time RT-PCR data ($n = 72$). *ZEB2* acts on the promoter of *CDH1*, a well-known epithelial marker, and reduces its expression.[23,36] To confirm the interaction between *ZEB2* and *CDH1* in ovarian cancer cells, a siRNA approach was used. For this purpose, we selected the SKOV3 cell line expressing endogenously higher *ZEB2* and lower *CDH1* mRNA than other ovarian cancer cell lines (Supplementary Fig. 3A,B). In SKOV3 cells, siRNA-mediated transient silencing of *ZEB2* expression resulted in upregulation of *CDH1* expression and downregulation of *FN1* and *VIM* expression (Supplementary Fig. 3C–F).

For multivariate analysis, we selected *ZEB2* from the two genes as likely to be the more important prognostic factor owing to its functional significance as an upstream repressor of *CDH1*.[23] The prognostic capability of *ZEB2* was further compared with other prognosis-related variables such as clinicopathological factors including age, performance of optimal cytoreduction, and histological grade using multivariate Cox's proportional

doi: 10.1111/j.1349-7006.2009.01204.x

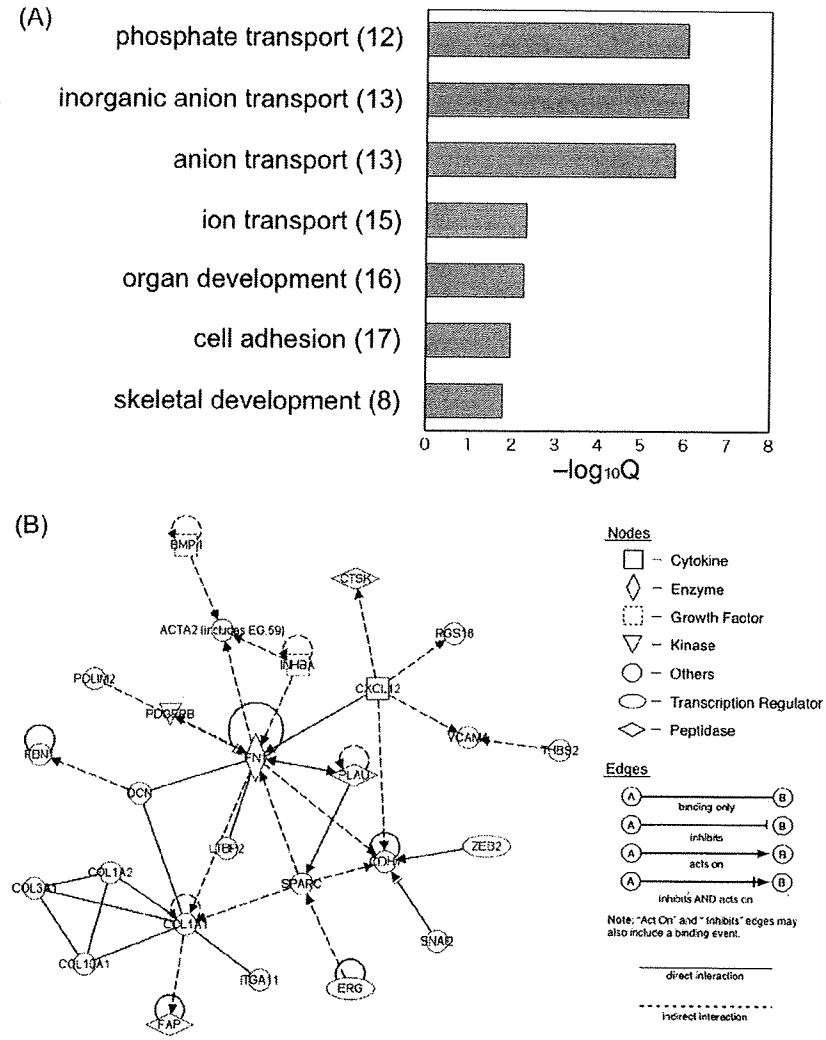

Fig. 2. Biological characterization of 112 subclass-specific genes using Gene Ontology analysis and Ingenuity Pathway Analysis (IPA). (A) Significant enrichments of gene ontology (GO) categories in GO-based profiling of 112 subclass-specific genes. Gray bars represent *q*-values (expressed as the negative logarithm [base 10]) after multiple testing correction of the Benjamini–Hochberg false discovery rate method for the significant ($q < 0.05$) GO categories over-represented in the 112 subclass-specific genes, using 4275 ovarian cancer-specific genes as a background set of genes for the determination of *q*-values. The actual number of the subclass-specific genes involved in each category is given in parentheses. (B) Twenty-six of 112 (24.1%) genes appeared in a new network based on the Ingenuity Pathway Knowledge base. Nodes represent genes, with their shapes showing IPA-defined functional classes of genes, and edges indicating biological relationships between nodes.

hazards analysis (Table 3). To increase the reliability of the multivariate analyses, all of the advanced stage serous ovarian cancer samples ($n = 72$) were analyzed by the real-time PCR technique. In Cox's proportional hazards model, *ZEB2* expression and the rate of optimal cytoreduction surgery were independent factors for progression-free survival time ($P = 0.014$ and 0.0011 respectively). The hazard ratio for relapse of *ZEB2* expression was 1.37 (95% confidence interval 1.07–1.78). Furthermore, when overall survival was evaluated by multivariate analysis, only the *ZEB2* expression level was independently associated with overall survival time ($P = 0.027$, hazard ratio = 1.53, 95% confidence interval 1.05–2.22).

***ZEB2* and *CDH1* expression and survival.** To clarify the details of the *ZEB2–CDH1* relationship, we analyzed the prognostic implications with regard to combinations of *ZEB2* and *CDH1* expression. For this purpose, we divided all of the samples into four groups, as shown in Table 4, when using a median expression level of each gene as a threshold for sample division. Multivariate Cox's proportional hazard model was used to compare survival among these four groups. The group showing high expression of *CDH1* and low expression of *ZEB2* served as a reference. Of the four groups, the only group with low expression of *CDH1* and high expression of *ZEB2* showed significantly poor prognosis in both progression-free survival and overall survival ($P = 0.0035$ and 0.013 respectively). When *ZEB2* expression was analyzed in combination with *CDH1* expression, the prognostic power of these genes became more significant.

Discussion

In this study, we evaluated the global gene expression profile to clarify the molecular etiology of the rapid progression specific to serous histological type cancers. We first attempted to subclassify our 43 serous type ovarian cancer tissues comprising eight stage I samples and 35 stage III/IV samples by a stepwise extraction of genes reflecting expression differences between samples (Supplementary Fig. 1). Although various classification methods have been proposed to characterize various cancer types at the molecular level using gene expression data, most of the methods tend to be unstable, producing different clusters with slightly different input or different choice of initial conditions.[37] Brunet *et al.*[28] showed that NMF is able to recover biologically significant phenotypes and appears superior to other methods especially when prior knowledge is lacking or undetermined. By applying the NMF algorithm, 35 patients with advanced stage serous ovarian cancer were grouped into two subclasses with 112 subclass-specific genes representing unique characteristics of tumor progression. Interestingly, one subclass, subclass 1 ($n = 17$), with a gene expression profile similar to that of stage I, showed a favorable outcome compared to the other subclass,

Table 3. Multivariable Cox's proportional hazards model analysis of prognostic factors for progression-free survival and overall survival in patients with advanced stage serous ovarian cancers ($n = 72$)

Variable	Hazard ratio (95% CI)	*P*-value
Progression-free survival		
ZEB2 expression	1.37 (1.07–1.78)	0.014*
Age	0.98 (0.96–1.00)	0.095
Optimal surgery (vs not optimal)	0.60 (0.44–0.82)	0.0011*
Grade 2 (vs Grade 1)	0.85 (0.58–1.24)	0.42
Grade 3 (vs Grade 1)	1.41 (0.98–2.06)	0.060
Overall survival		
ZEB2 expression	1.53 (1.05–2.22)	0.027*
Age	1.01 (0.96–1.04)	0.71
Optimal surgery (vs not optimal)	0.67 (0.41–1.05)	0.079
Grade 2 (vs Grade 1)	0.83 (0.47–1.50)	0.53
Grade 3 (vs Grade 1)	1.51 (0.93–2.62)	0.10

*Statistically significant ($P < 0.05$).

Table 4. Comparison of progression-free survival and overall survival in four groups with different expression profiles of *CDH1* and *ZEB2*

Serous ovarian cancer ($n = 72$)	Hazard ratio	95% CI	*P*-value
Progression-free survival			
CDH1 high/*ZEB2* low ($n = 23$)	1.00		
CDH1 high/*ZEB2* high ($n = 13$)	0.91	(0.53–1.43)	0.69
CDH1 low/*ZEB2* low ($n = 13$)	1.29	(0.83–1.94)	0.25
CDH1 low/*ZEB2* high ($n = 23$)	1.65	(1.18–2.35)	0.0035*
Overall survival			
CDH1 high/*ZEB2* low ($n = 23$)	1.00		
CDH1 high/*ZEB2* high ($n = 13$)	0.96	(0.37–1.96)	0.91
CDH1 low/*ZEB2* low ($n = 13$)	1.12	(0.63–1.95)	0.70
CDH1 low/*ZEB2* high ($n = 23$)	1.77	(1.12–2.92)	0.013*

subclass 2 ($n = 18$). This result was compatible with findings by Berchuck *et al.*[7] demonstrating similarities in gene expression between early stage serous ovarian cancers and a subset of advanced stage serous ovarian cancers that had favorable prognosis. Regarding the sample size in the current microarray analysis, one can realize that this may be first-stage evidence on ovarian expression profile associated with tumor progression. However, we successfully provided valuable insights that clarify the molecular mechanism of tumor progression using NMF algorithm.

Kurman *et al.* divide epithelial ovarian cancers into two groups designated type I and type II based on clinical, pathological, and molecular genetic studies.[21] Type I tumors are low grade and slow growing (including endometrioid, mucinous, and low-grade serous). Type II tumors (including high grade serous and undifferentiated) are rapidly growing, more aggressive, and are frequently associated with *TP53* mutation. In our experiments, the frequency of *TP53* mutation was higher in cases belonging to subclass 2 (9/18, 50%) compared to those belonging to subclass 1 + stage I (8/25, 32%). Although the frequency difference was not statistically significant, our novel subclassification based on gene expression profile might have a potential relationship with that of the two-type classification model of ovarian cancer proposed by Kurman *et al.*[21] Further study will be necessary to elucidate other biological and pathological implications except tumor progression in our subclassification.

After screening genes associated with tumor progression and subsequent validation of the association, we identified the expression of *ZEB2* and *CDH1* as prognostic factors for serous ovarian cancers. Although other genome-wide expression analyses[7–10] have identified gene expression profiles with prognosis values in patients with ovarian cancer, *ZEB2* and *CDH1* are not listed in their profiles. Previous studies using the expression microarrays investigate directly the association between gene expression level and survival time in patients with ovarian cancer, whereas we first extracted gene expression profiles reflecting tumor progression by a stepwise approach (Supplementary Fig. 1), and selected survival-associated genes with biological function from these genes. Furthermore, differences in microarray platforms, normalization methods, degrees of contamination by non-cancer cells in a given tumor specimen, and the patient populations under study[38] were observed between previous reports and ours. These points might contribute to the development of inconsistencies in lists of survival-associated genes from the microarray studies.

Our data also suggest that reduced *CDH1* expression is a key to subclassify advanced stage serous ovarian cancers. Recently Tothill *et al.* reported that six molecular subtypes of ovarian cancers, including serous and endometrioid histological types, were identified by a *k*-means clustering method according to genome-wide expression data from 285 ovarian cancer samples.[39] Of the six molecular subtypes, one subtype (C5 in the paper), comprising mainly high grade serous ovarian cancer samples, is characterized by reduced E-cadherin. Despite the difference in experimental design of the two studies, our data are compatible with their finding that a molecular subtype of ovarian cancers can be tagged by E-cadherin expression. E-cadherin is a hallmark of epithelial–mesenchymal transition, and a reduction of E-cadherin is thought to result in dysfunction of the cell–cell junction system, triggering cancer invasion in various human malignancies. In our experiment, E-cadherin expression was significantly associated with prognosis in patients with advanced stage serous ovarian cancer at both the mRNA and protein levels. Therefore, it is important to clarify the regulatory mechanisms of *CDH1* expression[40] in serous ovarian cancer in terms of tumor progression and prognosis, as well as subclassification.

Recent study shows that the interaction of Snail, ZEB, and bHLH factors regulates *CDH1* repression and epithelial–mesenchymal transition.[23] Besides *ZEB2*, other transcriptional repressors may reduce *CDH1* expression and lead to epithelial–mesenchymal transition.[41] Indeed, *SNAI2* was included in the 112 subclass-specific genes, and was found to directly interact with *CDH1* in the newly obtained IPA network (Fig. 2B). Previous reports show that other transcriptional repressors such as Snail 1 and Twist are related to prognosis in ovarian cancer, using immunohistochemical analysis.[35,42] Hosono *et al.*[42] have reported that expression of Twist is a significant prognostic factor in non-serous type but not in serous type tumors. Our results demonstrate that expression of *ZEB2* is negatively correlated with *CDH1* expression, and that the expression signature of increased *ZEB2* and reduced *CDH1* in ovarian tumor tissues is related to poor prognosis in serous ovarian cancer patients (Table 4). Furthermore, siRNA-mediated suppression of *ZEB2* in the serous type of ovarian cancer SKOV3 cells leads to an increase in *CDH1* expression (Supplementary Fig. 3), suggesting that *ZEB2* regulates *CDH1* expression in serous histological type tumors. To validate that *ZEB2* expression at the protein level is a significant prognostic factor, we would like to analyze *ZEB2* expression in a larger number of patients stratified according to individual histological types using immunohistochemical staining.

Park *et al.* have recently reported that microRNA-200 directly targets the mRNA of *ZEB2* as well as that of *ZEB1*, and indirectly controls the expression level of *CDH1* in cancer cell lines.[43] Further investigation is required to elucidate the more detailed mechanisms by which the *ZEB2–CDH1* axis in epithelial–mesenchymal transition is regulated in the process of ovarian cancer progression. Clarification of the mechanisms for the regulation of *ZEB2–CDH1* expression may provide plausible targets for the development of therapeutic strategies in the clinical management of serous ovarian cancers.

doi: 10.1111/j.1349-7006.2009.01204.x

Acknowledgments

This work was supported in part by a Grant-in-Aid for the Third-term Cancer Control Strategy Program from the Ministry of Health, Labor and Welfare, Japan. We are grateful to Hiroshi Kamiguchi and Tadayuki Satoh (Teaching and Research Support Center, Tokai University School of Medicine) for their technical support in the microarray experiment, and also thank Yoshiko Sakamoto, Eriko Tokubo, Hiromi Kamura, and Kozue Otaka for their technical assistance.

Abbreviations

ACTA2	actin, alpha 2, smooth muscle, aorta
ACTB	actin, beta
bHLH	basic helix-loop-helix
BRCA1	breast cancer 1, early onset
CA125	carbohydrate antigen 125
CDH1	cadherin 1
COL16A1	collagen, type XVI, alpha 1
coph	cophenetic correlation coefficient
Cy3	cyamine 3-CTP
FN1	fibronectin 1
GO	gene ontology
IPA	Ingenuity Pathway Analysis
LTBP2	latent transforming growth factor beta binding protein 2
NMF	non-negative matrix factorization
SNAI1	snail homolog 1
TBP	TATA box binding protein
TP53	Tumor Protein p53
VIM	vimentin
ZEB2	zinc finger E-box binding homeobox 2

References

1 Disaia PJ, Creasman WT. Epithelial ovarian cancer. In: Disaia PJ, Creasman WT, eds. *Clinical Gynecologic Oncology*, 6th edn. St Louis: Mosby, 2002; 289–350.
2 Cannistra SA. Cancer of the ovary. *N Engl J Med* 2004; **351**: 2519–29.
3 Agarwal R, Kaye SB. Expression profiling and individualization of treatment for ovarian cancer. *Curr Opin Pharmacol* 2006; **6**: 345–9.
4 Olivier RI, van Beurden M, van't Veer LJ. The role of gene expression profiling in the clinical management of ovarian cancer. *Eur J Cancer* 2006; **42**: 2930–8.
5 Fehrmann RS, Li XY, van der Zee AG *et al.* Profiling studies in ovarian cancer: a review. *Oncologist* 2007; **12**: 960–6.
6 Spentzos D, Levine DA, Ramoni MF *et al.* Gene expression signature with independent prognostic significance in epithelial ovarian cancer. *J Clin Oncol* 2004; **22**: 4700–10.
7 Berchuck A, Iversen ES, Lancaster JM *et al.* Patterns of gene expression that characterize long-term survival in advanced stage serous ovarian cancers. *Clin Cancer Res* 2005; **11**: 3686–96.
8 Hartmann LC, Lu KH, Linette GP *et al.* Gene expression profiles predict early relapse in ovarian cancer after platinum-paclitaxel chemotherapy. *Clin Cancer Res* 2005; **11**: 2149–55.
9 Bonome T, Levine DA, Shih J *et al.* A gene signature predicting for survival in suboptimally debulked patients with ovarian cancer. *Cancer Res* 2008; **68**: 5478–86.
10 Le Page C, Ouellet V, Quinn MC, Tonin PN, Provencher DM, Mes-Masson AM. BTF4/BTNA3.2 and GCS as candidate mRNA prognostic markers in epithelial ovarian cancer. *Cancer Epidemiol Biomarkers Prev* 2008; **17**: 913–20.
11 Dressman HK, Berchuck A, Chan G *et al.* An integrated genomic-based approach to individualized treatment of patients with advanced-stage ovarian cancer. *J Clin Oncol* 2007; **25**: 517–25.
12 Newton TR, Parsons PG, Lincoln DJ *et al.* Expression profiling correlates with treatment response in women with advanced serous epithelial ovarian cancer. *Int J Cancer* 2006; **119**: 875–83.
13 Partheen K, Levan K, Osterberg L, Horvath G. Expression analysis of stage III serous ovarian adenocarcinoma distinguishes a sub-group of survivors. *Eur J Cancer* 2006; **42**: 2846–54.
14 Okamoto A, Nikaido T, Ochiai K *et al.* Indoleamine 2,3-dioxygenase serves as a marker of poor prognosis in gene expression profiles of serous ovarian cancer cells. *Clin Cancer Res* 2005; **11**: 6030–9.
15 Donninger H, Bonome T, Radonovich M *et al.* Whole genome expression profiling of advance stage papillary serous ovarian cancer reveals activated pathways. *Oncogene* 2004; **23**: 8065–77.
16 Meinhold-Heerlein I, Bauerschlag D, Hilpert F *et al.* Molecular and prognostic distinction between serous ovarian carcinomas of varying grade and malignant potential. *Oncogene* 2005; **24**: 1053–65.
17 Bonome T, Lee JY, Park DC *et al.* Expression profiling of serous low malignant potential, low-grade, and high-grade tumors of the ovary. *Cancer Res* 2005; **65**: 10602–12.
18 Shridhar V, Lee J, Pandita A *et al.* Genetic analysis of early-versus late-stage ovarian tumors. *Cancer Res* 2001; **61**: 5895–904.
19 De Cecco L, Marchionni L, Gariboldi M *et al.* Gene expression profiling of advanced ovarian cancer: characterization of a molecular signature involving fibroblast growth factor 2. *Oncogene* 2004; **23**: 8171–83.
20 Lancaster JM, Dressman HK, Clarke JP *et al.* Identification of genes associated with ovarian cancer metastasis using microarray expression analysis. *Int J Gynecol Cancer* 2006; **16**: 1733–45.
21 Kurman RJ, Visvanathan K, Roden R, Wu TC, Shin IeM. Early detection and treatment of ovarian cancer: shifting from early stage to minimal volume disease based on a new model of carcinogenesis. *Am J Obstet Gynecol* 2008; **198**: 351–6.
22 Thiery JP. Epithelial mesenchymal transitions in tumour progression. *Nat Rev Cancer* 2002; **2**: 442–54.
23 Peinado H, Olmeda D, Cano A. Snail, Zeb and bHLH factors in tumour progression: an alliance against the epithelial phenotype? *Nat Rev Cancer* 2007; **7**: 415–28.
24 FIGO Cancer Committee. Staging Announcement: FIGO Cancer Committee. *Gynecol Oncol* 1986; **25**: 383–5.
25 Amikura T, Sekine M, Hirai Y *et al.* Mutational analysis of TP53 and p21 in familial and sporadic ovarian cancer in Japan. *Gynecol Oncol* 2006; **100**: 365–71.
26 Sekine M, Nagata H, Tsuji S *et al.* Mutational analysis of BRCA1 and BRCA2 and clinicopathologic analysis of ovarian cancer in 82 ovarian cancer families: two common founder mutations of BRCA1 in Japanese population. *Clin Cancer Res* 2001; **7**: 3144–50.
27 Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Statist Soc B* 1995; **57**: 289–300.
28 Brunet JP, Tamayo P, Golub TR, Mesirov JP. Metagenes and molecular pattern discovery using matrix factorization. *Proc Natl Acad Sci USA* 2004; **101**: 4164–9.
29 Okada H, Tajima A, Shichiri K, Tanaka A, Tanaka K, Inoue I. Genome-wide expression of azoospermia testes demonstrates a specific profile and implicates ART3 in genetic susceptibility. *PLoS Genet* 2008; **4**: e26.
30 Livak KJ, Schmittgen TD. Analysis of relative gene expression data using real-time quantitative PCR and the $2^{-\Delta\Delta CT}$ method. *Methods* 2001; **25**: 402–8.
31 Inamura K, Fujiwara T, Hoshida Y *et al.* Two subclasses of lung squamous cell carcinoma with different gene expression profiles and prognosis identified by hierarchical clustering and non-negative matrix factorization. *Oncogene* 2005; **24**: 7105–13.
32 Daraï E, Scoazec JY, Walker-Combrouze F *et al.* Expression of cadherins in benign, borderline, and malignant ovarian epithelial tumors: a clinicopathologic study of 60 cases. *Hum Pathol* 1997; **28**: 922–8.
33 Faleiro-Rodrigues C, Macedo-Pinto I, Pereira D, Lopes CS. Prognostic value of E-cadherin immunoexpression in patients with primary ovarian carcinomas. *Ann Oncol* 2004; **15**: 1535–42.
34 Voutilainen KA, Anttila MA, Sillanpää SM *et al.* Prognostic significance of E-cadherin–catenin complex in epithelial ovarian cancer. *J Clin Pathol* 2006; **59**: 460–7.
35 Blechschmidt K, Sassen S, Schmalfeldt B, Schuster T, Höfler H, Becker KF. The E-cadherin repressor Snail is associated with lower overall survival of ovarian cancer patients. *Br J Cancer* 2007; **98**: 489–95.
36 Imamichi Y, König A, Gress T, Menke A. Collagen type I-induced Smad-interacting protein 1 expression downregulates E-cadherin in pancreatic cancer. *Oncogene* 2007; **26**: 2381–5.
37 Gao Y, Church G. Improving molecular cancer class discovery through sparse non-negative matrix factorization. *Bioinformatics* 2005; **21**: 3970–5.
38 Konstantinopoulos PA, Spentzos D, Cannistra SA. Gene-expression profiling in epithelial ovarian cancer. *Nat Clin Pract Oncol* 2008; **5**: 577–87.
39 Tothill RW, Tinker AV, George J *et al.* Novel molecular subtypes of serous and endometrioid ovarian cancer linked to clinical outcome. *Clin Cancer Res* 2008; **14**: 5198–208.
40 Liu YN, Lee WW, Wang CY, Chao TH, Chen Y, Chen JH. Regulatory mechanisms controlling human E-cadherin gene expression. *Oncogene* 2005; **24**: 8277–90.
41 Imai T, Horiuchi A, Wang C *et al.* Hypoxia attenuates the expression of E-cadherin via up-regulation of SNAIL in ovarian carcinoma cells. *Am J Pathol* 2003; **163**: 1437–47.

42 Hosono S, Kajiyama H, Terauchi M *et al*. Expression of Twist increases the risk for recurrence and for poor survival in epithelial ovarian carcinoma patients. *Br J Cancer* 2007; **96**: 314–20.

43 Park SM, Gaur AB, Lengyel E, Peter ME. The miR-200 family determines the epithelial phenotype of cancer cells by targeting the E-cadherin repressors ZEB1 and ZEB2. *Genes Dev* 2008; **22**: 894–907.

Supporting Information

Additional Supporting Information may be found in the online version of this article:

Fig. S1. Analytical process to extract 'subclass-specific genes'.

Fig. S2. Association between E-cadherin expression and prognosis of advanced stage serous ovarian cancers validated by immunohistochemical analyses.

Fig. S3. Interaction between *ZEB2* and *CDH1*.

Table S1. Comparison of clinicopathological characteristics between microarray set and validation set

Table S2. List of 23 transcripts analyzed by quantitative real-time RT-PCR in this study

Table S3. One hundred and twelve transcripts representing statistically significant expression differences between two subclasses of advanced stage serous ovarian cancers

Table S4. Expression levels of 23 genes by quantitative real-time RT-PCR were significantly different between subclass 1 (S1) and subclass 2 (S2)

Supplementary Methods Methods about GO analysis, Pathway analysis, siRNA experiments, and immunohistochemical analysis

Please note: Wiley-Blackwell are not responsible for the content or functionality of any supporting materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.

doi: 10.1111/j.1349-7006.2009.01204.x

OPEN ACCESS Freely available online

PLoS one

Gene Expression Profile for Predicting Survival in Advanced-Stage Serous Ovarian Cancer Across Two Independent Datasets

Kosuke Yoshihara[1], Atsushi Tajima[2], Tetsuro Yahata[1], Shoji Kodama[3], Hiroyuki Fujiwara[4], Mitsuaki Suzuki[4], Yoshitaka Onishi[5], Masayuki Hatae[5], Kazunobu Sueyoshi[6], Hisaya Fujiwara[7], Yoshiki Kudo[7], Kohei Kotera[8], Hideaki Masuzaki[9], Hironori Tashiro[10], Hidetaka Katabuchi[10], Ituro Inoue[2], Kenichi Tanaka[1]*

1 Department of Obstetrics and Gynecology, Niigata University Graduate School of Medical and Dental Sciences, Niigata, Japan, **2** Department of Molecular Life Science, Tokai University School of Medicine, Isehara, Japan, **3** Department of Gynecology, Niigata Cancer Center Hospital, Niigata, Japan, **4** Department of Obstetrics and Gynecology, Jichi Medical University, Shimotsuke, Japan, **5** Department of Obstetrics and Gynecology, Kagoshima City Hospital, Kagoshima, Japan, **6** Department of Pathology, Kagoshima City Hospital, Kagoshima, Japan, **7** Department of Obstetrics and Gynecology, Hiroshima University Graduate School of Biomedical Sciences, Hiroshima, Japan, **8** Department of Obstetrics and Gynecology, Nagasaki Municipal Hospital, Nagasaki, Japan, **9** Department of Obstetrics and Gynecology, Nagasaki University Graduate School of Biomedical Sciences, Nagasaki, Japan, **10** Department of Gynecology, Faculty of Medical and Pharmaceutical Sciences, Kumamoto University, Kumamoto, Japan

Abstract

Background: Advanced-stage ovarian cancer patients are generally treated with platinum/taxane-based chemotherapy after primary debulking surgery. However, there is a wide range of outcomes for individual patients. Therefore, the clinicopathological factors alone are insufficient for predicting prognosis. Our aim is to identify a progression-free survival (PFS)-related molecular profile for predicting survival of patients with advanced-stage serous ovarian cancer.

Methodology/Principal Findings: Advanced-stage serous ovarian cancer tissues from 110 Japanese patients who underwent primary surgery and platinum/taxane-based chemotherapy were profiled using oligonucleotide microarrays. We selected 88 PFS-related genes by a univariate Cox model ($p<0.01$) and generated the prognostic index based on 88 PFS-related genes after adjustment of regression coefficients of the respective genes by ridge regression Cox model using 10-fold cross-validation. The prognostic index was independently associated with PFS time compared to other clinical factors in multivariate analysis [hazard ratio (HR), 3.72; 95% confidence interval (CI), 2.66–5.43; $p<0.0001$]. In an external dataset, multivariate analysis revealed that this prognostic index was significantly correlated with PFS time (HR, 1.54; 95% CI, 1.20–1.98; $p=0.0008$). Furthermore, the correlation between the prognostic index and overall survival time was confirmed in the two independent external datasets (log rank test, $p=0.0010$ and 0.0008).

Conclusions/Significance: The prognostic ability of our index based on the 88-gene expression profile in ridge regression Cox hazard model was shown to be independent of other clinical factors in predicting cancer prognosis across two distinct datasets. Further study will be necessary to improve predictive accuracy of the prognostic index toward clinical application for evaluation of the risk of recurrence in patients with advanced-stage serous ovarian cancer.

Citation: Yoshihara K, Tajima A, Yahata T, Kodama S, Fujiwara H, et al. (2010) Gene Expression Profile for Predicting Survival in Advanced-Stage Serous Ovarian Cancer Across Two Independent Datasets. PLoS ONE 5(3): e9615. doi:10.1371/journal.pone.0009615

Editor: Zoltán Bochdanovits, VU University Medical Center and Center for Neurogenomics and Cognitive Research, The Netherlands

Received November 3, 2009; **Accepted** February 16, 2010; **Published** March 12, 2010

Funding: This work was supported in part by a Grant-in-Aid for the Third-term Cancer Control Strategy Program from the Ministry of Health, Labor and Welfare, Japan (KT), and 2009 Research and Study Program of Tokai University Educational System General Research Organization (AT). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

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

* E-mail: tanaken@med.niigata-u.ac.jp

Introduction

Patients with advanced-stage ovarian cancer generally undergo primary debulking surgery followed by platinum/taxane-based chemotherapy. Although postoperative introduction of taxane drug has improved the 5-year survival rate for advanced-stage ovarian cancer, patients with this cancer have a 5-year survival rate of only 30% [1–3]. Clinicopathological characteristics, such as debulking status after primary surgery, are clinically considered important indicators of prognosis [4,5]. However, recurrence after optimal debulking surgery occurs in some patients, while disease-free status after incomplete surgery is maintained in others. In fact, it has been reported that 34% of patients treated with optimal surgery and platinum-taxane combination chemotherapy for advanced-stage ovarian cancer recur within 12 months [4]. Therefore, these clinicopathological factors alone are insufficient for predicting prognosis and elucidating the pathological mechanisms of disease progression or recurrence. Molecular biology

approaches can be used to identify new prognosis-related profiles leading to elucidation of pathological issues of advanced-stage serous ovarian cancer.

Microarray technology has been developing very rapidly, and it has become relatively easy to analyze the expression levels of thousands of genes within cancer cells. Although many studies have reported the associations of gene expression profiles with prognoses in cancer patients [6–10], a limited number of such profiles are used in clinical settings. Microarray technology is clinically applied for predicting prognosis in breast cancer patients. MammaPrint ™ (Agendia BV, Amsterdam, the Netherlands) has been already put to practical use for the purpose. Meanwhile, there are no microarray kits for clinical diagnosis and management in patients with ovarian cancer yet.

Three studies have recently reported gene expression profiles that predict overall survival (OS) in ovarian cancer patients using microarray techniques [11–13]. These studies use a relative large sample size (n>80) for establishing a survival-related profile in a discovery phase of the experiment and an external independent dataset as the validation set to solve the problem that the number of the genomic variables examined is much larger than that of subjects. Thus, research on the overall survival-related profiles in ovarian cancer patients has progressed, whereas there are no extensive studies based on multicenter validation of gene expression profiles for prediction of disease progression or recurrence in patients with ovarian cancer [14–15]. Prediction of the risk of recurrence in patients with advanced-stage ovarian cancer receiving standard treatments (primary surgery+platinum/taxane-based chemotherapy) is more important with respect to optimization of clinical management [16].

We have recently reported that there are high similarities in gene expression between early-stage and a subset of advanced-stage serous ovarian cancer patients that have favorable prognoses, and two molecular subgroups among patients with advanced-stage serous ovarian cancer according to gene expression profiles reflecting tumor progression and prognosis [17]. In this study, we focused on progression-free survival (PFS) time in a larger number of patients only with advanced-stage serous ovarian cancer treated with platinum/taxane-based chemotherapy, and tried to identify PFS-related gene expression profile using a new survival analysis method: ridge regression Cox model [18]. We then assessed the correlation between our PFS-related genes expression profile and survival time in an external independent dataset of advanced-stage serous ovarian cancer.

Results

Clinical Characteristics

The clinical characteristics of 110 Japanese patients with advanced-stage serous ovarian cancer are summarized in Table 1. In the discovery set, 93 patients (84.5%) were diagnosed as the International Federation of Gynecology and Obstetrics (FIGO) stage III, and 17 patients (15.5%) as FIGO stage IV [19]. All patients received platinum/taxane-based chemotherapy after primary surgery. The median progression-free and overall survival times were 17 and 31 months, respectively.

On the other hand, we used a part of publicly available microarray data (GSE9891) as an external independent dataset (See Materials and Methods) [20]. The clinical characteristics of 87 patients with advanced-stage serous ovarian cancer in the external dataset are listed in Table S1 [20]. Kaplan-Meier survival analysis showed that there were no significant differences in PFS and OS time between patients of the discovery dataset and those of the external dataset (Figure S1). When we compared clinicopathological characteristics between the discovery set and the external dataset, there were significant differences in frequencies of stage (Table S1). Because grading system adopted in the external dataset was distinct from that in the discovery set [21–23], we could not make a simple comparison of malignant grade between the two datasets. Then we examined the association between clinicopathological features and PFS time in patients with advanced-stage serous ovarian cancer of each dataset. Multivariate analysis revealed that only optimal surgery was an independent prognostic factor for PFS in the discovery dataset (Table S2) and that there was marginally significant correlation between debulking status of primary surgery and PFS time in the external dataset (Table S2). Therefore, we planned first to develop a prognostic index based on PFS-related genes in the discovery dataset, secondarily to evaluate the prognostic ability of our index in the external dataset using multivariate analysis, and then thirdly to assess predictive performance of the prognostic index again after the stratification of patients according to the debulking status of primary surgery.

Table 1. Clinical characteristics of advanced-stage serous ovarian cancer patients.

	Present Dataset (n = 110)	Percentage
Median age, years (range)	58 (23–85)	
Stage		
Stage III	93	84.5
Stage IV	17	15.5
CA125 (IU) (n = 99)	1960±3519	
Optimal Cytoreduction		
Optimal (<1cm)	57	51.8
Not optimal	53	48.2
Grade		
Grade 1	26	23.6
Grade 2	41	37.3
Grade 3	43	39.1
Median survival time, months (range)	31 (1–81)	

doi:10.1371/journal.pone.0009615.t001

Identification of PFS-Related Profile

Using Agilent Whole Human Genome Oligo microarray, we generated gene expression data for 110 advanced-stage serous ovarian cancer patients. Then this dataset was used as a discovery set for the identification of PFS-related profile in patients with advanced-stage serous ovarian cancer. To further evaluate the PFS-related profile, we prepared a part of the GSE9891 dataset as an external independent dataset using Affymetrix Human Genome U133 Plus 2.0 Array (See Materials and Methods) [20]. To deal with cross-platform microarray data appropriately, we analyzed only common genes (28304 probes in Agilent platform; 38497 probes in Affymetrix platform) between the two platforms in this study. Of 28304 Agilent probes, 18178 probes with expression levels marked as "Present" in all of the 110 microarray data from the discovery set was further extracted to remove missing and uncertain signals on gene expression, and then the data were per-gene normalized in each dataset by transforming the expression of each gene to a mean of 0 and standard deviation of 1 (Figure S2).