Minimum Information Specification For In Situ Hybridization and Immunohistochemistry Experiments (MISFISHIE)

Eric W Deutsch[1], Catherine A Ball[2], Jules J Berman[3], G Steven Bova[4], Alvis Brazma[5], Roger E Bumgarner[6], David Campbell[1], Helen C Causton[7], Jeffrey H Christiansen[8], Fabrice Daian[9], Delphine Dauga[9], Duncan R Davidson[8], Gregory Gimenez[9], Young Ah Goo[1,10], Sean Grimmond[11], Thorsten Henrich[12], Bernhard G Herrmann[13], Michael H Johnson[1], Martin Korb[1], Jason C Mills[14], Asa J Oudes[1,10], Helen E Parkinson[5], Laura E Pascal[1,10], Nicolas Pollet[15], John Quackenbush[16], Mirana Ramialison[12], Martin Ringwald[17], David Salgado[9], Susanna-Assunta Sansone[5], Gavin Sherlock[18], Christian J Stoeckert Jr.[19], Jason Swedlow[20], Ronald C Taylor[21], Laura Walashek[1,10], Anthony Warford[22], David G Wilkinson[23], Yi Zhou[24], Leonard I Zon[24], Alvin Y Liu[1,10], and Lawrence D True[25]

[1]Institute for Systems Biology, 1441 N 34th Street, Seattle, WA 98103, USA [2]Department of Biochemistry, Stanford University School of Medicine, 279 Campus Drive West, Stanford, CA 94305, USA [3]Association for Pathology Informatics, 9650 Rockville Pike, Bethesda, MD 20814, USA [4]Johns Hopkins University School of Medicine, PELICAN Laboratory, Departments of Pathology , Health Information Sciences, Genetic Medicine, Oncology, and Urology, Baltimore, MD, USA [5]EMBL-EBI, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK [6]Department of Microbiology, University of Washington, Seattle, WA 98195, USA [7]MRC CSC Microarray Centre, Imperial College School of Medicine, London, W12 ONN, UK [8]MRC Human Genetics Unit, Western General Hospital, Crewe Road, Edinburgh, UK EH4 2XU [9]Institut de Biologie du Développement de Marseille, UMR 6216, CNRS/Université de la Méditerranée, Parc Scientifique de Luminy, Case 907, F-13288 Marseille Cedex 9, France [10]Department of Urology, University of Washington, Seattle, WA 98195, USA [11]Institute for Molecular Bioscience, University of Queensland, St. Lucia, Qld, Australia [12]EMBL Heidelberg, Meyerhofstrasse 1, D-69117 Heidelberg, Germany [13]Institute for medical Genetics, Charité-CBF, and Dept. Developmental Genetics, Max-Planck-Institute for molecular Genetics, Ihnestrasse 73, 14195 Berlin, Germany [14]Department of Pathology and Immunology, Washington University School of Medicine, St. Louis, MO 63110, USA [15]CNRS UMR 8080, Université Paris-Sud, 91405 Orsay, France [16]Dana-Farber Cancer Institute, 44 Binney Street, M232, Boston, MA 02115, USA [17]The Jackson Laboratory, 600 Main Street, Bar Harbor, ME 04609, USA [18]Department of Genetics, Stanford University School of Medicine, 300 Pasteur Drive, Stanford, CA 94305-5120, USA [19]Center for Bioinformatics and Department of Genetics, University of Pennsylvania, Philadelphia, PA 19104, USA [20]Division of Gene Regulation and Expression, Wellcome Trust Biocentre, University of Dundee, Dow Street, Dundee DD1 5EH, Scotland [21]Computational Biology & Bioinformatics Group, Pacific Northwest National Laboratory, PO Box 999, MS K7-90, Richland, WA 99352, USA [22]Wellcome Trust Sanger Institute, Hinxton, Cambridge CB10 IHH, UK [23]Division of Developmental Neurobiology, National Institute for Medical Research, The Ridgeway, Mill Hill, London NW7 1AA, UK [24]Stem Cell Program, Division of Hematology/ Oncology, Children's Hospital Boston and Harvard Medical School, 300 Longwood Ave, Boston,

MA 02115, USA [25]Department of Pathology, University of Washington, Seattle, WA 98195-6100, USA

Abstract

One purpose of the biomedical literature is to report results in sufficient detail so that the methods of data collection and analysis can be independently replicated and verified. Here we present for consideration a minimum information specification for gene expression localization experiments, called the "Minimum Information Specification For In Situ Hybridization and Immunohistochemistry Experiments (MISFISHIE)". It is modelled after the MIAME (Minimum Information About a Microarray Experiment) specification for microarray experiments. Data specifications like MIAME and MISFISHIE specify the information *content* without dictating a *format* for encoding that information. The MISFISHIE specification describes six types of information that should be provided for each experiment: Experimental Design, Biomaterials and Treatments, Reporters, Staining, Imaging Data, and Image Characterizations. This specification has benefited the consortium within which it was initially developed and is expected to benefit the wider research community. We welcome feedback from the scientific community to help improve our proposal.

Background

High-throughput analyses of gene expression in biological samples (*e.g.*, transcript abundance using microarrays or protein abundance using proteomics) often do not provide information about the cell types or spatial domains within tissues that express the genes of interest, and may not reveal dynamic or transient gene expression. Consequently, such analyses are often followed by experiments to confirm the location and degree of gene expression by specific cell types within the tissue by probing with specific reporters for the genes of interest. In addition, the wealth of clinical information associated with tissue samples in large collections all over the world provide a powerful tool to validate or expand the conclusions made using such high-throughput analysis of fresh samples.

However, it is often the case that studies that make use of *in situ* hybridization (ISH) and immunohistochemistry (IHC) staining, and/or their resulting images are presented without the information needed to interpret the images or the methodology that produced them. Furthermore, neither the reagents and methods used in the experiments, nor the results are easily searchable through current biomedical literature databases like PubMed. Since the interpretation of ISH and IHC stains could differ between observers, between different image analysis platforms and programs, and even between different sessions using the same image analysis platform and program[1], communicating the methods and criteria used are critically important for teaching others and to permit critical evaluation of a published work.

Data annotation specifications that have been developed by the wider microarray community[2-4] have begun to show benefits for the biomedical research community. First and foremost, the debate initiated by the proposal for specifications engaged many researchers, and the current specifications included the contributions of many different

interests within the microarray data generating community. Common exchange formats and the willingness of researchers to put their data into the public domain upon publication have significantly increased the accessibility of data to all researchers. The open-source software and ontologies developed in conjunction with the data specifications resulted from the efforts of many different groups in the community. General discussion forums facilitated interaction between manufacturers and experimenters working towards development of the specifications for better experiments and better publications. Similar specifications are currently under development for other high-throughput technologies[5-10].

Others have proposed data formats to better enable exchange of microscopy image data. For example, an XML data format specifically for tissue microarrays has been proposed[11]. However, no minimum amount of information is specified, and users are free to include only as much information as they wish. Also available is Open Microscopy Environment (OME), which provides a flexible XML data format for storing and transmitting metadata for microscopy image datasets (http://www.openmicroscopy.org/). However, there is no comprehensive specification for facilitating the exchange of data from visual interpretation-based tissue protein and transcript abundance/localization experiments (hereafter referred to as 'gene expression localization experiments'), such as ISH and IHC.

Results and Discussion

To maximize the benefit of new gene expression localization experiments to the biomedical research community, we propose a minimum information specification, the "Minimum Information Specification For In Situ Hybridization and Immunohistochemistry Experiments (MISFISHIE)". This specification provides guidelines for the minimum information that should be provided when publishing, making public, or exchanging results from visual interpretation-based tissue gene expression localization experiments such as ISH, IHC, lectin affinity histochemistry, and reporter gene constructs (e.g., green fluorescent protein [GFP], β-galactosidase). Compliance with this specification is expected to provide researchers at different laboratories with enough information to fully evaluate the data and to reproduce the experiment. Although MISFISHIE facilitates the identification of specific sources of variability, it cannot, and does not aim to, reduce this variability. However, if complete information, including raw image data, is always provided, the original interpretations may be re-evaluated by other researchers.

Modelled after the widely accepted MIAME specification for microarray experiments[2], MISFISHIE only prescribes the kind of information that should be provided. It does not include every parameter that could be specified about an experiment, but rather broad categories of detail that should be addressed, relying on the data producers and reviewers to ensure that each section contains enough information for readers to be able to fully assess the validity of, and accurately reproduce the experiment described. Just as MIAME has been a guide to help authors provide enough information about a microarray experiment such that its interpretation could be verified or refuted[12], we hope that MISFISHIE will be used in the same way for gene expression localization experiments.

This specification does not dictate a specific format for reporting the information. We expect to develop a data model based on the concepts of MAGE-OM (MicroArray Gene Expression Object Model) and software based on the MAGEstk (MicroArray Gene Expression software tool kit)[3]. It is this model and the associated XML-based mark-up language that will provide a data format for archiving or transferring data. Since a major revision of MAGE-OM, named the FuGE-OM (Functional Genomics Experiment Object Model)[13], is currently being developed to accommodate data from other functional genomics experiments, it is likely that the MISFISHIE-derived object model will be an extension of FuGE-OM and not a separate construct. A simpler, non-XML format following the concepts of MAGE-TAB[14] may also facilitate data sharing in cases where simplicity is most important[15]. It is intended that MISFISHIE should function together with other technology-related specifications such as MIAME and MIAPE (Minimum Information About a Proteomics Experiment)[16] to support functional genomics investigations. We anticipate that MISFISHIE will be integrated with other MGED (Microarray and Gene Expression Data) Society standards[17] through the Reporting Structure for Biological Investigation (RSBI) working group[18] and the Minimum Information for Biological and Biomedical Investigations (MIBBI) project[19], in particular. This is especially important since the goal of integrating different data types will most easily be realized when a common reporting structure is used. Separation of the minimum information specification and the data format is important because there should be scope for the provision of unlimited additional information beyond the minimum specification and encoding of incomplete information for optimal flexibility. Furthermore, broad acceptance of the minimum information required would greatly aid the design of a data model.

To facilitate data transfer between some existing expression databases, a MISFISHIE-compliant XML data format has been developed. A Document Type Definition (DTD) was developed for three expression databases, ANISEED[20], COMPARE[21] and 4DXpress[22]. It defines a format that follows the MISFISHIE specification. This DTD and an associated example are available at http://crfb.univmrs.fr/aniseed/exchange_format.php and at http://compare.ibdml.univ-mrs.fr/exchange_format.php.

It has long been appreciated that improved standards for IHC are needed. However, standardization discussions have largely been focused on the development of standardized technical protocols that might lead to more uniform staining[23], or efforts towards reducing the subjectivity in interpretation of histological sections[24]. Here we do not attempt to endorse standardized methodologies or data interpretation, but rather seek to promote complete disclosure of the methodologies used so that experiments may be replicated by others employing the same procedures as the original investigators.

A set of guidelines specifically for tumor marker prognostic studies called REMARK[25] has recently been established. REMARK encompasses the domain of outcome studies based on tumor markers of any kind, not just those of IHC. MISFISHIE encompasses the domain of studies employing IHC or ISH techniques; it may be a tumor marker study or a zebrafish embryo study. We believe that MISFISHIE presents a subset of guidelines applicable to nearly any IHC or ISH study regardless of context. We fully expect that specialized

subdomains (such as clinical prognostic studies) will want to add applicable requirements for that subdomain.

While no accepted minimum specification for this type of data yet exists, there have been several efforts at organizing gene expression localization data in databases. Such database designs provide a useful framework from which to build a specification. Two databases for the mouse research community, the Mouse Gene Expression Database (GXD)[26] and the Edinburgh Mouse Atlas Gene Expression (EMAGE) database[27], have influenced the design of MISFISHIE. Mouse-specific fields in these databases were removed in favor of more organism-neutral ones. Several fields in these databases were deemed useful but are not part of a minimum requirement and, consequently, were not included. Also, in these databases many experiments that were entered by curators using the information provided in journal articles have empty fields because they had not been described in sufficient detail in the papers. Achieving MISFISHIE compliance in a publication will result in more complete reporting of experiments and, therefore, more reproducible experiments in these and other databases in the future. Although MISFISHIE is primarily designed as a specification for peer-reviewed journal articles, it will guide database development as well. For example, the release of ANISEED V3.0 is based on MISFISHIE rules and the new schema of the database is MISFISHIE-compliant. The inclusion of specific experimental details, such as tissue type, reagents and methods, will allow investigators to find precedent for experiments they are considering more efficiently. For example, an investigator might be able to rapidly search all publications that reported immunoperoxidase localization of CD10/MME in the human prostate using the database and retrieve information on how the gene localization experiments were conducted.

This specification describes the type of information that should be provided for publication of gene expression localization experiments in six sections (Figure 1):

1. Experimental Design

2. Biomaterials (specimens) and Treatments (section or whole-mount preparation)

3. Reporters (probes or antibodies)

4. Staining

5. Imaging Data

6. Image Characterizations

The following description provides guidelines for ensuring that data are compliant with the specification. It is intended to be useful to researchers preparing to publish data as well as to manuscript reviewers and editors checking for MISFISHIE compliance. The use of ontologies, such as the MGED Ontology (MO)[4] or Ontology for Biomedical Investigations (OBI; formerly named FuGO)[28], facilitates computational searches of data and are therefore extremely advantageous as a source of descriptors. For terms outside the scope of OBI, such as those in anatomy, another appropriate ontology may be used. A good list of ontologies is maintained at the Ontologies for Biology Organization (OBO) web site http://obo.sourceforge.net/. Use of OBI and other ontologies will be especially important as

MISFISHIE-supporting applications and databases are developed. Many of the terms used in this specification are already defined in OBI.

Experimental Design

This section should contain information about the gene expression localization experiment as a whole including a brief description of the project, experimental factors, and the methods. For example, this would include the variables between the assays in the experiment, and how and where to get more information about the experiment (web sites and contact persons). We propose that the following types of information be used to describe the overall design of an experiment:

A. Experiment description: a short summary of the aims of the experiment.

B. Assay type(s): e.g., IHC, ISH, lectin affinity histochemistry, cell-lineage- or tissue-specific reporter expression.

C. Experiment design type: e.g., is it a comparison of normal *vs*. diseased tissue, of multiple tissue/embryo specimens of similar type, of multiple probes/antibodies applied to the same tissue, or a localization screen, etc.? The MGED Ontology ExperimentDesignType has many entries categorizing design type.

D. Experimental factors: the parameters or conditions that are tested, such as probe/antibody, disease state, genetic variation, structural unit, age, etc. Again, the MGED Ontology is a rich source of terms that can be used to describe the factors being tested.

E. Total number of assays performed in the experiment: an assay is defined as one instance of a hybridization/stain of a single specimen with a single reporter. Thus, the result of a tissue microarray consisting of a 10×10 array of tissues would be counted as 100 assays. If replicates or reruns are a component of the experimental design, provide details that should include number of replicates per tissue, per reporter, etc.

F. URL of any websites or database accession numbers (if available) pertinent to the experiment.

G. Contact information for communicating with the experimenters.

Biomaterials (specimens) and Treatments (section or whole-mount preparation)

Describing specimens comprehensively is challenging, since they may have dozens or even hundreds of characteristics, especially for patient material when clinical information is available. The guiding principle in sample description is to supply enough information for an independent researcher to carry out a similar experiment. Characteristics that are known to differ among specimens should be provided with each specimen; while common attributes of all the specimens may be provided only once. The MISFISHE proposal lists characteristics of a biological sample that should be described:

A. Origin of the biological specimens. Information required includes:

i. Attributes of the individual(s). The organism species must be named, preferably using the NCBI taxonomy, and for non-human organisms the strain and mutant alleles should be named according to the accepted standards for that organism. Additional attributes may include, but are not limited to, sex, age, developmental stage, genotype, phenotype.

ii. Physiologic state of the individual(s) (normal *vs*. diseased).

iii. Relevant exogenous factors (e.g., treatment, special diet).

iv. Anatomic source of the tissue or cell sample.

v. Provider of the specimens.

All information critical for other researchers to reproduce the biomaterials as closely as possible should be provided. The information is not limited to the above examples. Referencing an established ontology or controlled vocabulary for the terms used is *highly* encouraged. Ontologies and controlled vocabularies are available from many sources in a variety of formats, including on-line references and reference textbooks. Since we are still at an early stage in the development and widespread use of databases to store sample information, a standardized set of terms and a single, widely accepted ontology is not yet available. The rationale for providing specific structural detail is that the location of an object, such as a cell type that is being studied may correlate with expression of a specific gene by that cell type. Structural detail may be important not only for cases where gene expression is dependent on tissue handling (e.g., there is stronger labelling at the specimen edges) but also in cases where, even within a single microanatomical unit, there is heterogeneity (e.g., in lung tumors, cell cycle regulatory genes are highest at the periphery)[29].

B. Manner of preparation of the specimens for the study. Information required includes:

i. Nature of the specimens (e.g., whole tissue, whole mounts of tissue, tissue sections, thickness of sections, whole cells, or sections of cells).

ii. Manner in which the specimens were prepared for the experiments (e.g., fixation with type of fixative and duration of fixation *vs*. fresh, non-fixed, non-frozen specimens or frozen specimens, sections mounted on slides *vs*. sections floating in reagents).

iii. Protocols used. Referencing previously published protocols is permissible if the protocols are appropriately detailed and were strictly followed.

Sensitivity of the immunoreaction of some gene products to fixation is exemplified by the observation that p27 was least frequent and least intense in prostate cancer cells that were farthest from the cut surface of a fixed tissue. These were the cells that are least rapidly fixed[30].

Reporters (probes or antibodies)

It is critical to provide full information about the reporters (probes, lectins, or antibodies) used, since these can differ in reactivity from lot to lot and manufacturer to manufacturer. A manufacturer's literature usually provides most of the needed information; however, the manufacturer's literature may not be permanent. For privately produced reporters, enough information needs to be provided so that another lab could produce the same compounds. MISFISHIE specifies several requirements necessary to best describe the molecules used to label a tissue sample. It was noted in the review of this manuscript that thorough validation of reporters is very often poorly done in current literature. This specification does not at present require that researchers validate each reporter used in a particular way, but such validation is encouraged and should be reported when performed.

A. Unambiguous genomic identification of each reporter:

  i. For *in situ* hybridizations, provide the corresponding GenBank/EMBL/ DDBJ accession number and, if applicable, the start and end nucleotide positions of the probe within that sequence. Also, provide the accession number version or database release version.

  ii. For antibodies, provide the protein identifier, including specific version information for the accession number or database release.

B. Full sequence of each probe, or clone number of each antibody. For fluorescent protein experiments, the promoter sequence should be specified. In each case, provide the method by which the reporter was characterized.

  i. If the sequence or clone number is not known, then the template or clone must be made publicly available. Provide specific details on how the template or clone may be obtained.

  ii. Some tissue localization experiments are based on the principle that the gene being localized is detected when the gene promoter activates a fluorescent protein reporter, such as GFP. In such experiments, the sequence of the reporter (i.e., GFP) is not important. Rather, the sequence of the promoter is critical and confers cell and tissue specificity to the reporter since the promoter is specific to that cell.

C. Protocol(s) for how the reporters were designed and produced or the source from which they were obtained.

  i. For reporters purchased from a company, the company name, address, catalogue number, and lot number should be provided.

  ii. For a custom-made antibody, the putative antigen and references to studies that characterize the sensitivity and specificity of the antibody in tissue immunostains should be given.

D. Additional attributes of the reporter:

> **i.** For antibodies, the type of primary antibody (monoclonal or polyclonal), the immunoglobulin isotype, and the organism in which the antibody was generated.
>
> **ii.** For lectins, the full name (e.g., *Dolichos biflorus*), the source of the lectin (e.g., which company produced it), how it was detected (e.g., whether it was fluorescently labelled or biotinylated, with follow-up histochemical analysis), and how it was labelled (e.g., if the investigators labelled the lectin themselves, the source of the reagents, the method and/or the labeling kit should be provided).

Staining

The protocols used for staining vary considerably among experimenters. The merits of standardizing these protocols have been discussed extensively in the literature. This specification merely requires that the protocol used is provided and is sufficiently detailed that another researcher may follow it. The following types of information should be provided to adequately describe the staining protocols and parameters:

A. Number of detectable reporters in the hybridization or stain (e.g., more than one for multiple-dye fluorescence microscopy) plus specific details about the detection method:

> **i.** Detection reagent used (e.g., fluors used, enzyme-substrates, gold particles).
>
> **ii.** Source of the detection system plus sufficient detail to reproduce the reaction.

B. Protocol used to produce the hybridization or immunostain. This should include a description of how the tissue (organism, organ, or section) was mounted onto the slide/substrate and treatments of the section, e.g., IHC protocol inclusive of parameters such as buffer, temperature, post-wash conditions, etc. Referencing previously published protocols is permissible if the protocols are appropriately detailed and were strictly followed. Also include:

> **i.** What steps, if any, were taken to decrease non-specific reaction product. For example, in immunoperoxidase experiments there might be pre-incubation of the specimen preparation with (a) albumin solution to block non-specific binding, (b) peroxide solution to block signal due to endogenous peroxidase.
>
> **ii.** Use of an antigen or gene product retrieval method.

C. Information about assay controls: the nature of both positive and negative tissue and reporter controls (or state if controls were not performed). The same level of detail of the tissue controls should be reported as for the cells or tissues that are being studied. Optionally provide specificity reporter controls, such as competitive inhibition with either purified protein or peptide in IHC.

Imaging Data

Although the MIAME specification stops short of requiring microarray image data, we propose that MISFISHIE require that representative IHC or ISH images be provided since the interpretation of these images varies with the experience and training of the observer. While the images are not needed to facilitate reproducibility of an experiment, they greatly aid in the interpretation and analysis, and in determining reasons for discordant results. Both positive and negative results should be reported; this information is potentially useful for other work outside the scope of the reported experiment.

For several model organisms, there are already repositories for gene expression localization experiment images, including GXD[26] and EMAGE[27] for mouse, ZFIN[31] for zebrafish, and others. However, for many organisms including human, there may not be such a dedicated database. It would be of tremendous value to the research community to have a general, organism-independent database for archiving gene expression localization experiment images. Such an archive could provide examples of tissue localization studies, and could be a reference site for investigators who want to verify the tissue localizations of reporter reagents they are considering using. More importantly, a general-purpose repository to which researchers could submit their images for permanent storage with accession numbers for publications would be very valuable for facilitating MISFISHIE compliance and in realizing the full value of these data for future research. MorphBank (http://www.morphbank.net) is an available general purpose image repository for biological research. BioImage is an image repository under construction at http://www.bioimage.org/ [32].

The MISFISHIE specification suggests that the following information should be provided:

A. Digital images for each assay included in the study should be digitally available for download without additional charge. The images should be of sufficient resolution to allow independent characterization, and provided in a standard file format (e.g., JPEG, PNG, GIF, TIFF). The images should be named or tagged with the reporter and specimen that they represent.

B. Detection method by which hybridization or staining is observed (e.g., for each channel a fluorescent wavelength if multiple reporters are used). If the detection method is the same for all images, it need only be mentioned once.

C. Images for the controls are not required, although may optionally be provided.

Image Characterizations

The results as interpreted by the original researchers should be reported in a clearly articulated, concise and consistent manner. This permits reviewers to ensure that the characterizations are consistent with and representative of the data, and that the conclusions are reasonable. The characterizations should also be provided in such a way that they can be easily stored in a database, queried, and compared with other expression data.

The types of characterization recorded can vary depending on the experimental design. The following guidelines specify a minimum set of characterization features. Additional

characterization of the images as required by the experimental design could also be provided.

A. Ontology entries, including reference to the ontology (e.g., refs. [33-36], note that some ontologies, such as SNOMED CT and NIH/NLM's Unified Medical Language System (UMLS), may contain licensing restrictions that make them unavailable to some or limit the use of the terms; a MISFISHIE-compliant document that contains SNOMED CT entries or some UMLS entries may not be legally redistributable[37]), terms, accession numbers, or terms and definitions if sufficient detail cannot be found in an existing ontology for individual structural units used for classification. Structural units could be an organ, tissue, cell, subcellular component, etc. Note that only the structural units relevant to the experiment need to be listed and characterized. It is not necessary to list (and characterize) structural units visible in the assays or slides but not relevant to the experiment or report.

B. Intensity scale, ideally choosing one from the MGED Ontology. For example, a three-level scale of present, absent, or equivocal might be appropriate for evaluating IHC stains. However, any scale that the investigators feel is appropriate may be used as long as each gradation of intensity in the scale is defined in a manner that enables an independent investigator to understand or apply the same criteria.

C. Per each structural unit (relevant to the experiment) in each assay (or in each image), provide:

   i. Staining intensity or the fraction of the structural unit's population exhibiting each intensity (see example below).

   ii. Other optional annotations/characterizations of the structural unit, e.g., feature density, qualitative characteristics or spatial distribution of the structural unit or staining. The use of referenced ontology terms is encouraged. Both positive and negative calls of staining relevant to the experiment should be reported. It is quite useful to provide negative expression results; it is understood that a negative result is actually an upper limit to the expression level, where the limit is usually not well known. If some structural units cannot be characterized for some reporters, corresponding calls may be null. For example:

   Luminal epithelial cell: present

   Basal epithelial cell: absent

   etc.

   is sufficient; or, when appropriate for the type of analysis being done, more detail:

   Luminal epithelial cell: 90% present, 10% equivocal, 0% absent
   Basal epithelial cell: 10% present, 10% equivocal, 80% absent etc.

Unless only a few expression calls are presented, it is clearest if the calls are presented in tabular form, either within the main text or as supplemental material as appropriate.

D. Optionally as a best practice, the protocol for the characterization and information about the basic technique for characterizing the assays. For example, this information may include how many observers performed the characterizations, whether the characterizations were performed from the images themselves or visually through the instrument, any exceptions or assumptions made in characterizing the data, etc. We refer to one example of a well-described characterization protocol[38]. We also note that it has been reported that performing the characterization from digital images has advantages in terms of replication, decreased intraobserver and interobserver variability[39].

Some examples of real experimental data annotated according to MISFISHIE are posted at the MISFISHIE web site, available as a link from the MGED workgroup web page http://www.mged.org/Workgroups/ . We also provide an abbreviated checklist (Figure 2) to aid in assessing MISFISHIE compliance. It should be used in conjunction with the full description, not in place of it. A printable version is supplied as Supplementary Table 1.

Survey of the recent literature

To assess how the MISFISHIE specification compares with what appears to be standard practice for publication today, a selection of articles reporting on IHC or ISH from the last seven years were assessed for compliance with the six sections of the MISFISHIE specification. Three articles[40-42] were assessed and discussed by all ten *ad hoc* reviewers so that inter-reviewer variability could be minimized. Another 29 articles[43-70] were assigned to individual reviewers for assessment. Each reviewer assessed each of his or her assigned articles in the context of a scenario of a journal referee reviewing a submitted manuscript. As part of the review, the MISFISHIE compliance checklist (see Supplementary Table 1) was completed by the reviewer as if it were the journal's policy to require MISFISHIE compliance.

Compliance for each MISFISHIE subsection was rated by the reviewers on the scale of 0 to 10, where a 10 indicates that the authors provided all information that the reviewer needed to understand or reproduce the experiment without needing to make any assumptions. Scores lower than 10 correspond to how incomplete the information was that the reviewer thought necessary to understand or reproduce the work. Scores of 8 and 9 were considered a low pass; the reviewer could reproduce the experiment although with a few assumptions. It was therefore possible for a paper to leave out a few details that the reviewers deemed ought to have been provided, but still pass. Compliance with each section was somewhat subjective as the strictness of each reviewer was not uniform, as would presumably be the case for *bona fide* journal reviewers. Therefore, the MISFISHIE specification itself is subject to individual interpretation. Since this cannot be avoided, we hope that the checklist will minimize subjectivity.

This exercise not only proved useful in testing the proposed MISFISHIE specification, but also allowed us to determine if any section seemed too onerous a requirement. Of the 32 papers assessed, only four (13%) were deemed MISFISHIE compliant in all six sections. An additional 28% were out of compliance with only one section, and 31% did not comply in two sections. The review considered that more than 90% of the papers were compliant with MISFISHIE sections 1 and 2 (Experimental Design; Biomaterials and Treatments). Compliance for sections 3 and 4 (Reporters and Staining) was about 75%. Section 5 (Imaging Data) proved to be the most troublesome, with only 16% of the articles compliant. Finally, about 47% complied with section 6 (Image Characterizations). These results are summarized in Table 1.

Although few of the surveyed articles complied fully, the reviewers felt that the majority of non-compliant papers would require only modest additions to become compliant, with the possible exception of section 5. This section requires that at least one representative image of each assay be made electronically available. This may be within a model organism database, a generic image database, a journal's supplemental data web site, or even the author's web site, although the last is the least preferable. It is not necessary for all images to be reproduced within the manuscript itself. One might feel that making all images accessible to others is unduly burdensome. However, we feel that since image interpretation is variable, it is necessary that the original images be made available in a digital format for subsequent review, ideally in a centrally-managed public repository. Some model organism databases already provide such a facility. MorphBank provides an example of a general-purpose image repository for any organism, although it does not appear to be well suited to store the accompanying characterizations in an easily queryable format.

We provide as one example of a paper that was deemed MISFISHIE compliant the work of Santagata et al.[65] Our review of this article concluded that it provides sufficient detail for all MISFISHIE sections; all images used for the study are available at their own web site.

Conclusions

This specification was jointly developed by members of the NIH/NIDDK Stem Cell Genome Anatomy Projects consortium to facilitate data sharing within the consortium. After use and refinement within the consortium, and based on discussions with additional members of the larger research community, we offer this specification, published here as MISFISHIE version 1.0 as a proposal to the whole research community. The history of the creation of MISFISHIE and the lessons learned from it[71] may be helpful for others aiming to create a similar specification for other data types.

We expect that MISFISHIE will undergo updates, leading to future editions, as other localization methods, such as DNA *in situ* hybridization experiments to chromosomes, are implemented and the need for a specification is expressed. The eventual accepted specification cannot be dictated, but rather must be achieved through discussion and consensus. Suggestions from the community are actively encouraged and will be collected and folded into an eventual second release, published at the MISFISHIE domain of the MGED web site: http://www.mged.org/Workgroups/MISFISHIE/. Comments may be

addressed to the email distribution list dedicated to discussion about MISFISHIE: mgedmisfishie@lists.sourceforge.net. We note that there is still considerable room for researching the scientific best practice for performing and reporting these types of studies. We have attempted here to define a minimum set of information and have provided a few optional better practices that were deemed not quite appropriate as a requirement for all publications.

After a suitable period of dialogue and revision by the community, and should the community accept the final proposal, we would encourage reviewers, journal editors and funding agencies to promote compliance with MISFISHIE for all studies that report gene expression localization data so that all published data and resulting conclusions may be correctly interpreted, and that independent investigators would have the necessary information that would enable them to validate the experiment. Our survey of recent articles indicated that only about 15% of published works are fully compliant with this specification, and most fail by not making images of assays used in the study digitally accessible to the research community. Most of the surveyed papers could be brought into compliance by uploading the images into a repository and adding fewer than a dozen additional sentences of description. If article length constraint hinders full MISFISHIE compliance, it would be encouraged that the information be provided in supplemental material.

Several of the model organism databases are already able to accept and archive the results from a publication that provides all information that MISFISHIE specifies. We highly encourage authors to submit their data to these databases via the provided database submission process upon submission of the manuscript.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

List of Abbreviations

ANISEED	Ascidian Network for In Situ Expression and Embryological Data
EMAGE	Edinburgh Mouse Atlas Gene Expression database (http://genex.hgu.mrc.ac.uk/)
FuGE-OM	Functional Genomics Experiment Object Model
FuGO	Functional Genomics Ontology (renamed OBI in Oct 2006)
GFP	green fluorescent protein

GXD	Gene Expression Database (http://www.informatics.jax.org/)
IHC	immunohistochemistry
ISH	*in situ* hybridization
MAGE-OM/ML	MicroArray Gene Expression Object Model/ Markup Language
MGED	Microarray and Gene Expression Data Society (http://www.mged.org/)
MIAME	Minimum Information About a Microarray Experiment
MIAPE	Minimum Information About a Proteomics Experiment
MISFISHIE	Minimum Information Specification For In Situ Hybridization and Immunohistochemistry Experiments
MO	MGED Ontology
NIDDK	National Institute of Diabetes & Digestive & Kidney Diseases
NIH	National Institutes of Health
NLM	National Library of Medicine
OBI	Ontology for Biomedical Investigations (formerly FuGO)
OME	Open Microscopy Environment (http://www.openmicroscopy.org/)
PEDRo	Proteomics Experiment Data Repository (http://pedro.man.ac.uk/)
RSBI	Reporting Structure for Biological Investigation
UMLS	Unified Medical Language System
XML	Extensible Markup Language
ZFIN	Zebrafish Information Network (http://www.zfin.org)

References

1. True LD. Quantitative immunohistochemistry: a new tool for surgical pathology? Am J Clin Pathol. 1988; 90:324–325. [PubMed: 3046322]

2. Brazma A, et al. Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. Nat Genet. 2001; 29:365–371. [PubMed: 11726920]

3. Spellman PT, et al. Design and Implementation of Microarray Gene Expression Markup Language (MAGE-ML). Genome Biol. 2002; 3 RESEARCH0046.

4. Stoeckert CJ, Parkinson H. The MGED ontology: A framework for describing functional genomics experiments. Comparat and Funct Genomics. 2003; 4:127–132. http://mged.sourceforge.net/ontologies/MGEDontology.php.

5. Taylor CF, et al. A Systematic Approach to Modeling Capturing and Disseminating Proteomics Experimental Data. Nat Biotechnol. 2003; 21:247. [PubMed: 12610571]

6. Garwood K, et al. PEDRo: A database for storing, searching and disseminating experimental proteomics data. BMC Genomics. 2004; 5:68. [PubMed: 15377392]

7. Jones A, Hunt E, Wastling JM, Pizarro A, Stoeckert CJ Jr. An object model and database for functional genomics. Bioinformatics. 2004; 20:1583–1590. [PubMed: 15145818]

8. Xirasagar S, et al. CEBS object model for systems biology data, SysBio-OM. Bioinformatics. 2004; 20:2004–2015. [PubMed: 15044233]

9. Jenkins H, et al. A proposed framework for the description of plant metabolomics experiments and their results. Nat Biotechnol. 2004; 22:1601–1606. [PubMed: 15583675]

10. Lindon JC, et al. Summary recommendations for standardization and reporting of metabolic analyses. Nat Biotechnol. 2005; 23:833–838. [PubMed: 16003371]

11. Berman JJ, Edgerton ME, Friedman BA. The tissue microarray data exchange specification: a community-based, open source tool for sharing tissue microarray data. BMC Med Inform Decis Mak. 2003; 3:5. [PubMed: 12769826]

12. Stoeckert CJ, Quackenbush J, Brazma A, Ball CA. Minimum information about a functional genomics experiment: the state of microarray standards and their extension to other technologies. Drug Discovery Today: TARGETS. 2004; 3:159–164.

13. Jones AR, et al. The Functional Genomics Experiment model (FuGE): an extensible framework for standards in functional genomics. Nat Biotechnol. 2007; 25:1127–1133. [PubMed: 17921998]

14. Rayner TF, et al. A simple spreadsheet-based, MIAME-supportive format for microarray data: MAGE-TAB. BMC Bioinformatics. 2006; 7:489. [PubMed: 17087822]

15. Brazma A, Krestyaninova M, Sarkans U. Standards for systems biology. Nat Rev Genet. 2006; 7:593–605. [PubMed: 16847461]

16. Taylor CF, et al. HUPO - Proteomics Standards Initiative (PSI). IOMICS: A Journal of Integrative Biology. 2006 in press.

17. Ball CA, Brazma A. MGED Standards. OMICS: A Journal of Integrative Biology. 2006; 10:138–144. [PubMed: 16901218]

18. Sansone S-A, et al. A Strategy Capitalizing on Synergies: The Reporting Structure for Biological Investigation (RSBI) Working Group. OMICS: A Journal of Integrative Biology. 2006; 10:164–171. [PubMed: 16901222]

19. Taylor CF, et al. Promoting Coherent Minimum Reporting Requirements for Biological and Biomedical Investigations: The MIBBI Project. Nat Biotechnol. (submitted).

20. Tassy, O., et al. Digital representation of embryonic development: the ANISEED (Ascidian Network for In Situ Expression and Embryological Data) system. http://crfb.univ-mrs.fr/aniseed/(in preparation)

21. Salgado D, Gimenez G, Coulier F, Marcelle C. COMPARE, a multi-organism system for cross-species data comparison and transfer of information. Bioinformatics. (submitted).

22. Haudry Y, et al. 4DXpress: a database for cross-species expression pattern comparisons. Nuc Acids Res. 2007 gkm797.

23. Swanson PE. Methodologic Standardization in Immunohistochemistry: A Doorway Opens. Appl Immunohist. 1993; 1:229–231.

24. Taylor CR. An exaltation of experts: concerted efforts in the standardization of immunohistochemistry. Hum Pathol. 1994; 25:2–11. [PubMed: 7508882]

25. McShane LM, et al. Reporting recommendations for tumor marker prognostic studies (REMARK). J Natl Cancer Inst. 2005; 97:1180–1184. [PubMed: 16106022]

26. Smith CM, et al. The mouse Gene Expression Database (GXD): 2007 update. Nucleic Acids Res. 2007; 35:D618–623. [PubMed: 17130151]

27. Baldock RA, et al. EMAP and EMAGE: a framework for understanding spatially organized data. Neuroinformatics. 2003; 1:309–325. [PubMed: 15043218]

28. Whetzel PL, et al. Development of FuGO – an Ontology for Functional Genomics Experiments. OMICS: A Journal of Integrative Biology. 2006; 10:199–204. [PubMed: 16901226]

29. Dobashi Y, et al. Active cyclin A-CDK2 complex, a possible critical factor for cell proliferation in human primary lung carcinomas. Am J Pathol. 1998; 153:963–972. [PubMed: 9736045]

30. De Marzo AM, Fedor HH, Gage WR, Rubin MA. Inadequate formalin fixation decreases reliability of p27 immunohistochemical staining: probing optimal fixation time using high-density tissue microarrays. Hum Pathol. 2002; 33:756–760. [PubMed: 12196928]

31. Sprague J, et al. The Zebrafish Information Network (ZFIN): the zebrafish model organism database. Nucleic Acids Res. 2003; 31:241–243. [PubMed: 12519991]

32. Carazo JM, Stelzer EH. The BioImage Database Project: organizing multidimensional biological images in an object-relational database. J Struct Biol. 1999; 125:97–102. [PubMed: 10222266]

33. Rosse C, Mejino JL Jr. A reference ontology for biomedical informatics: the Foundational Model of Anatomy. J Biomed Inform. 2003; 36:478–500. [PubMed: 14759820]

34. Bard J, Rhee SY, Ashburner M. An ontology for cell types. Genome Biol. 2005; 6:R21. [PubMed: 15693950]

35. Bard JL, et al. An internet-accessible database of mouse developmental anatomy based on a systematic nomenclature. Mech Dev. 1998; 74:111–120. [PubMed: 9651497]

36. Hayamizu TF, Mangan M, Corradi JP, Kadin JA, Ringwald M. The Adult Mouse Anatomical Dictionary: a tool for annotating and integrating data. Genome Biol. 2005; 6:R29. [PubMed: 15774030]

37. Berman JJ. A tool for sharing annotated research data: the "Category 0" UMLS (Unified Medical Language System) vocabularies. BMC Med Inform Decis Mak. 2003; 3:6. [PubMed: 12809560]

38. Abd El-Rehim DM, et al. Expression of luminal and basal cytokeratins in human breast carcinoma. J Pathol. 2004; 203:661–671. [PubMed: 15141381]

39. Bova GS, et al. Web-based tissue microarray image data analysis: initial validation testing through prostate cancer Gleason grading. Hum Pathol. 2001; 32:417–427. [PubMed: 11331959]

40. Liu AY, True LD. Characterization of prostate cell types by CD cell surface molecules. Am J Pathol. 2002; 160:37–43. [PubMed: 11786396]

41. Kernek KM, et al. Fluorescence in situ hybridization analysis of chromosome 12p in paraffin-embedded tissue is useful for establishing germ cell origin of metastatic tumors. Mod Pathol. 2004; 17:1309–1313. [PubMed: 15167939]

42. McKenney JK, et al. Basal cell proliferations of the prostate other than usual basal cell hyperplasia: a clinicopathologic study of 23 cases, including four carcinomas, with a proposed classification. Am J Surg Pathol. 2004; 28:1289–1298. [PubMed: 15371944]

43. Amara N, et al. Prostate stem cell antigen is overexpressed in human transitional cell carcinoma. Cancer Res. 2001; 61:4660–4665. [PubMed: 11406532]

44. Ayala G, et al. High levels of phosphorylated form of Akt-1 in prostate cancer and non-neoplastic prostate tissues are strong predictors of biochemical recurrence. Clin Cancer Res. 2004; 10:6572–6578. [PubMed: 15475446]

45. Bart J, et al. The distribution of drug-efflux pumps, P-gp, BCRP, MRP1 and MRP2, in the normal blood-testis barrier and in primary testicular tumours. Eur J Cancer. 2004; 40:2064–2070. [PubMed: 15341980]

46. Browne TJ, et al. Prospective evaluation of AMACR (P504S) and basal cell markers in the assessment of routine prostate needle biopsy specimens. Hum Pathol. 2004; 35:1462–1468. [PubMed: 15619204]

47. Chen D, et al. Syndecan-1 expression in locally invasive and metastatic prostate cancer. Urology. 2004; 63:402–407. [PubMed: 14972511]

48. Clayton H, Titley I, Vivanco M. Growth and differentiation of progenitor/stem cells derived from the human mammary gland. Exp Cell Res. 2004; 297:444–460. [PubMed: 15212947]

49. Cooray HC, Blackmore CG, Maskell L, Barrand MA. Localisation of breast cancer resistance protein in microvessel endothelium of human brain. Neuroreport. 2002; 13:2059–2063. [PubMed: 12438926]

50. Giangreco A, Shen H, Reynolds SD, Stripp BR. Molecular phenotype of airway side population cells. Am J Physiol Lung Cell Mol Physiol. 2004; 286:L624–630. [PubMed: 12909587]

51. Gmyrek GA, et al. Normal and malignant prostate epithelial cells differ in their response to hepatocyte growth factor/scatter factor. Am J Pathol. 2001; 159:579–590. [PubMed: 11485916]

52. Hwang JH, et al. Isolation of muscle derived stem cells from rat and its smooth muscle differentiation [corrected]. Mol Cells. 2004; 17:57–61. [PubMed: 15055528]

53. Jonker JW, et al. The breast cancer resistance protein BCRP (ABCG2) concentrates drugs and carcinogenic xenotoxins into milk. Nat Med. 2005; 11:127–129. [PubMed: 15685169]

54. Knudsen BS, et al. High expression of the Met receptor in prostate cancer metastasis to bone. Urology. 2002; 60:1113–1117. [PubMed: 12475693]

55. Larkin A, et al. Investigation of MRP-1 protein and MDR-1 P-glycoprotein expression in invasive breast cancer: a prognostic study. Int J Cancer. 2004; 112:286–294. [PubMed: 15352042]

56. Lee K, Klein-Szanto AJ, Kruh GD. Analysis of the MRP4 drug resistance profile in transfected NIH3T3 cells. J Natl Cancer Inst. 2000; 92:1934–1940. [PubMed: 11106685]

57. Li R, et al. High level of androgen receptor is associated with aggressive clinicopathologic features and decreased biochemical recurrence-free survival in prostate: cancer patients treated with radical prostatectomy. Am J Surg Pathol. 2004; 28:928–934. [PubMed: 15223964]

58. Martin CM, et al. Persistent expression of the ATP-binding cassette transporter, Abcg2, identifies cardiac SP cells in the developing and adult heart. Dev Biol. 2004; 265:262–275. [PubMed: 14697368]

59. Martin MJ, Muotri A, Gage F, Varki A. Human embryonic stem cells express an immunogenic nonhuman sialic acid. Nat Med. 2005; 11:228–232. [PubMed: 15685172]

60. Master VA, Wei G, Liu W, Baskin LS. Urothlelium facilitates the recruitment and trans-differentiation of fibroblasts into smooth muscle in acellular matrix. J Urol. 2003; 170:1628–1632. [PubMed: 14501678]

61. Piotrowska AP, et al. Alterations in smooth muscle contractile and cytoskeleton proteins and interstitial cells of Cajal in megacystis microcolon intestinal hypoperistalsis syndrome. J Pediatr Surg. 2003; 38:749–755. [PubMed: 12720186]

62. Ricciardelli C, et al. Androgen receptor levels in prostate cancer epithelial and peritumoral stromal cells identify non-organ confined disease. Prostate. 2005; 63:19–28. [PubMed: 15378523]

63. Roudier MP, et al. Phenotypic heterogeneity of end-stage prostate carcinoma metastatic to bone. Hum Pathol. 2003; 34:646–653. [PubMed: 12874759]

64. Rubin MA, et al. Quantitative determination of expression of the prostate cancer protein alpha-methylacyl-CoA racemase using automated quantitative analysis (AQUA): a novel paradigm for automated and continuous biomarker measurements. Am J Pathol. 2004; 164:831–840. [PubMed: 14982837]

65. Santagata S, et al. JAGGED1 expression is associated with prostate cancer metastasis and recurrence. Cancer Res. 2004; 64:6854–6857. [PubMed: 15466172]

66. Scotlandi K, et al. C-kit receptor expression in Ewing's sarcoma: lack of prognostic value but therapeutic targeting opportunities in appropriate conditions. J Clin Oncol. 2003; 21:1952–1960. [PubMed: 12743148]

67. Shah RB, et al. Androgen-independent prostate cancer is a heterogeneous group of diseases: lessons from a rapid autopsy program. Cancer Res. 2004; 64:9209–9216. [PubMed: 15604294]

68. St Croix B, et al. Genes expressed in human tumor endothelium. Science. 2000; 289:1197–1202. [PubMed: 10947988]

69. Wang Z, et al. Expression of the human cachexia-associated protein (HCAP) in prostate cancer and in a prostate cancer animal model of cachexia. Int J Cancer. 2003; 105:123–129. [PubMed: 12672042]

70. Zhigang Z, Wenlv S. Prostate stem cell antigen (PSCA) expression in human prostate cancer tissues: implications for prostate carcinogenesis and progression of prostate cancer. Jpn J Clin Oncol. 2004; 34:414–419. [PubMed: 15342669]

71. Deutsch EW, et al. Minimum Information Specification For In Situ Hybridization and Immunohistochemistry Experiments (MISFISHIE). OMICS. 2006; 10:205–208. [PubMed: 16901227]

Figure 1.
The six sections of the MISFISHIE specification.

Figure 2.
An abbreviated checklist for the full MISFISHIE specification. This checklist should be
used in conjunction with the full specification, not instead of it.

Table 1

Summary of statistics from the MISFISHIE assessment survey of a cohort of selected current literature.

N	Percent	Statistic
32	100%	Number of articles assessed for compliance
4	13%	Number of articles considered to be fully MISFISHIE compliant
9	28%	Number of articles for which MISFISHIE information is missing for one section
10	31%	Number of articles for which MISFISHIE information is missing for two sections
6	19%	Number of articles for which MISFISHIE information is missing for more than two sections
31	97%	Number of articles that meet the data content requirements for section 1 (Experimental Design)
29	91%	Number of articles that meet the data content requirements for section 2 (Biomaterials and Treatments)
24	75%	Number of articles that meet the data content requirements for section 3 (Reporters)
24	75%	Number of articles that meet the data content requirements for section 4 (Staining)
5	16%	Number of articles that meet the data content requirements for section 5 (Imaging Data)
15	47%	Number of articles that meet the data content requirements for section 6 (Image Characterizations)