

# MS1, MS2, and SQT—three unified, compact, and easily parsed file formats for the storage of shotgun proteomic spectra and identifications

W. Hayes McDonald<sup>1†</sup>, David L. Tabb<sup>1,3†</sup>, Rovshan G. Sadygov<sup>1</sup>, Michael J. MacCoss<sup>1</sup>, John Venable<sup>1</sup>, Johannes Graumann<sup>2</sup>, Jeff R. Johnson<sup>1</sup>, Daniel Cociorva<sup>1</sup> and John R. Yates III<sup>1\*</sup>

<sup>1</sup>Department of Cell Biology, The Scripps Research Institute, La Jolla, CA, USA

<sup>2</sup>Department of Biology, California Institute of Technology, Pasadena, CA, USA

<sup>3</sup>Department of Genome Sciences, The University of Washington, Seattle, WA, USA

Received 26 April 2004; Revised 20 July 2004; Accepted 20 July 2004

As the speed with which proteomic labs generate data increases along with the scale of projects they are undertaking, the resulting data storage and data processing problems will continue to challenge computational resources. This is especially true for shotgun proteomic techniques that can generate tens of thousands of spectra per instrument each day. One design factor leading to many of these problems is caused by storing spectra and the database identifications for a given spectrum as individual files. While these problems can be addressed by storing all of the spectra and search results in large relational databases, the infrastructure to implement such a strategy can be beyond the means of academic labs. We report here a series of unified text file formats for storing spectral data (MS1 and MS2) and search results (SQT) that are compact, easily parsed by both machine and humans, and yet flexible enough to be coupled with new algorithms and data-mining strategies. Copyright © 2004 John Wiley & Sons, Ltd.

Proteomic technologies are helping to change the scale at which biological experiments can be performed. Unfortunately, they also generate such voluminous data that they can result in a computational quagmire. Shotgun proteomic strategies, in which tandem mass spectra (MS/MS) are collected on mixtures of thousands of peptides, require the collection of tens of thousands of spectra.<sup>1</sup> Incorporating multidimensional separation strategies such as MudPIT (Multidimensional Protein Identification Technology) can easily balloon this into hundreds of thousands of spectra.<sup>2–5</sup> Spectra must be stored for identification via database search software such as SEQUEST. In many cases, multiple identification strategies and even algorithms are applied in the course of a complete analysis.<sup>6,7</sup>

Since its inception as the file format recognized by the SEQUEST database search algorithm,<sup>8</sup> the DTA file has seen widespread use as a format to store individual MS/MS spectra. It and the SEQUEST-generated OUT file were sufficient for experiments producing merely hundreds of spectra. However, as the scope and complexity of these experiments has expanded, the limitations of these files have become increasingly evident; the number of files produced

for individual experiments makes directory management problematic, and the storage space wasted in these formats is problematic as well. To deal with some of these limitations we have developed and implemented a new set of unified file formats that are simple, compact, and yet retain their flexibility.

There were several major goals associated with moving towards unified formats. First, unified formats dramatically reduced the number of files required to represent a proteomic data set. This was important because the huge number of files (hundreds of thousands) added to the file servers each week were taxing their stability and exceeded file system limitations. Second, the new formats reduced the amount of storage space used. Many of the individual files were small enough to be below the minimum block size limit for a file (typically 4K or 8K). Thus, simply concatenating the files together reduced the total amount of disk space required to store this information. Third, switching to unified file formats enabled greater efficiency in data storage by removing fields that had been repeated in each file and by grouping data that had been distributed to multiple files. Finally, the unified formats were formatted for automated parsing and designed for extensibility. In keeping with this final goal, the formats had to be adaptable to existing programs and able to accommodate future code developments.

In order to accomplish these goals, we adopted three unified file formats—MS1, MS2, and SQT. All of these store their particular information type for an entire experimental

\*Correspondence to: J. R. Yates III, Department of Cell Biology, The Scripps Research Institute, 10550 North Torrey Pines Road, La Jolla, CA, USA.

E-mail: jyates@scripps.edu

†Present address: Oak Ridge National Laboratory, Oak Ridge, TN, USA.

step, e.g. an entire LC/MS/MS experiment or a single salt step from a MudPIT run. The naming convention is simple; the base filename of the instrument-generated file is used with a new extension. For instance, the ThermoFinnigan LCQ file *salt\_step.RAW* would generate *salt\_step.ms1* and *salt\_step.ms2* files and, after database searching, would yield a *salt\_step.sqt* file.

The MS1 file contains full-scan data and is used for analyses that require this type of data such as quantitation or measurement of chromatographic efficiency. The MS2 file stores MS/MS data and replaces a folder of thousands of DTA files; it contains all the spectral information necessary for database searching algorithms. Finally, the SQT file unifies the database search results. While initially designed to replace the SEQUEST OUT file, it has proven flexible enough to work quite well with other algorithms used in the lab, e.g. PEP\_PROBE<sup>9</sup> and GutenTag.<sup>10</sup>

### Format descriptions

These formats were intended to store all necessary information in as compact and accessible a format as possible while retaining human legibility. In general, they contain information generic to all records in a header at the start of the file while data for specific spectra are stored individually through the body of the file. Unless specifically noted, all fields are tab-delimited within an individual line. A comprehensive description of these formats with examples is available on our website.<sup>11</sup>

The MS1 format is the simplest of the three. It contains four types of lines: *H*, *S*, *I*, and [m/z intensity]. The header is defined by a series of *H* lines. Each line includes a field label and its corresponding value (string, integer, or floating point). The following fields are required in the header: *CreationDate*, *Extractor*, *ExtractorVersion*, and *ExtractorOptions*. The values for these are the date the file was created, the program used, its version, and any specific options used in the program (Fig. 1). Some optional field labels include: *InstrumentType* (ion trap, q-tof, tof-tof, etc.), *InstrumentSN* (serial number), and *Comment* (other information general to the file).

In the body of the file, each full scan in the experiment begins with an *S* line which contains the scan number. This can be followed by the optional *I* line which can be used to store any ancillary information such as retention time (*IRTime* 33.2). Next comes a series of [m/z intensity pairs] (space separated) representing the spectral data for that entire full mass scan. This pattern (*S*, [m/z intensity]<sup>n</sup> or *S*, *I*, [m/z intensity]<sup>n</sup>) is repeated for each full mass scan in the experiment. If necessary, multiple *I* lines can be used for a given spectrum (e.g. LC retention time).

The format for the MS2 file is similar to the MS1 file except that MS/MS spectra are stored. It shares the *H*, *S*, *I*, and [m/z intensity] lines with the MS1 file but adds two additional lines, *Z* and *D*, to store charge-state-dependent information. The header itself has additional field labels such as *IAnalyzer*, to denote a program which does not consider the charge state of the precursor ion (e.g. spectral quality filtering), and *DAnalyzer*, to denote a program that analyzes charge-state-specific features (e.g. charge-state discrimination or neutral losses off the precursor). Any specific features noted by the

*DAnalyzer* programs are annotated in the *D* line following their specific *Z* line (see below). One advantage of the MS2 file format is that the file format logs which algorithms have been applied serially to the file.

The MS2 file body is structured similarly to the MS1 except for the addition of the *Z* and *D* lines. The description of each spectrum begins with the *S* line which has fields for the [start scan], [end scan], and [precursor m/z]. This can be followed by the optional *I* line which contains a datum or analytical result that is independent of the charge-state prediction of the precursor, such as a spectral quality score or instructions to the search program not to query this particular spectrum. Next comes the *Z* line with the [charge state] and [predicted [M+H]<sup>+</sup>] fields. The optional *D* line may follow and can be used to store information specific to the charge state of the preceding *Z* line. This can include annotations of a particular structural feature that might necessitate the use of a different search algorithm, e.g. neutral loss of phosphoric acid off of the precursor as an indication of a phosphorylated peptide. There can be multiple *Z* and optional *D* lines for a given spectrum depending on how well the precursor charge state is able to be discriminated. The spectral data are stored in the same manner as the MS1 file with and [m/z intensity] stored for every peak in the experimental spectrum. The minimum pattern to represent a spectrum is *S*, *Z*, [m/z intensity]<sup>n</sup>, but, as previously mentioned, can also contain multiple *Z* lines and the optional *I* and *D* lines to encompass additional information. One obvious advantage of this format over the DTA format is that to have the search algorithm to consider an additional charge state, all one has to do is add another *Z* line rather than producing a separate DTA file (currently each charge state to be considered has a corresponding DTA file), with all of the [m/z intensity]<sup>n</sup> information repeated.

The SQT file format is a greater departure from the SEQUEST OUT files it replaces. The design aims were to provide the same information as reported in the OUT file but to do so in a more compact, more easily parsed format while retaining a degree of human legibility. It is comprised of *H*, *S*, *M*, and *L* lines. The header lines are similar in format to both the MS1 and MS2 files, except that it has its own distinct set of field labels and values. Figure 2 shows the required fields and an example of how they are employed. Invariant information usually stored in each OUT file is stored just once in the header of the SQT file. The data characterizing a particular identification are stored in a block of lines lower in the file.

Each search result for a spectral entry is denoted by the following generic line pattern: *S*(*M*(*L*)<sup>k</sup>)<sup>n</sup>. (Fig. 3). The *S* line contains information specific for that spectrum and search. It is followed by an *M* line which describes a particular matching sequence along with its characteristic scores. Next comes at least one *L* line that notes which protein in the database contains this particular peptide sequence; there can be multiple *L* lines depending on how many proteins within the database contain the matched peptide sequence. The *M* line for the second highest scoring peptide match is followed by its respective *L* line(s). This pattern of *M* and *L* lines continues for as many search results as were set to be stored in the search parameters (typically 5–10). The SQT file also allows the inclusion of a column which stores 'manual evaluation' information. The state can be either the default of

## A

symbol	meaning	generic form		required
H	header	H	[field label] [corresponding value]	yes
S	scan	S	[first scan] [second scan]	yes
I	charge independent analysis	I	[field label] [corresponding value]	
[data]	mass intensity pairs	[m/z]	[intensity]	yes
<b>"H" line required field labels</b>				
<b>field label</b>	<b>description</b>			
CreationDate	The date and time on which this file was created.			
Extractor	The name of the software used to create the MS2 file.			
ExtractorVersion	The version number of the Extractor software.			
ExtractorOptions	The options used in running the Extractor software.			
<b>"H" line optional field labels</b>				
<b>field label</b>	<b>description</b>			
IAnalyzer	If software to conduct charge state-independent analysis of the spectra is used to analyze the MS2 file (such as removing sparse spectra), the name of the program should be recorded in the header.			
IAnalyzerVersion	The version number of the IAnalyzer			
IAnalyzerOptions	The options used for the IAnalyzer			
InstrumentType	The type of mass analyzer used.			
Comment	Remarks, ownership, and copyright information may be included. Multiple comment lines may be included.			
InstrumentSN	The serial number of the mass spectrometer used.			

## B

```

H      CreationDate      07/10/2003
H      Extractor         RAWXtract
H      ExtractorVersion      1.0
H      ExtractorOptions
S      0001      0001
I      RTime      0.01
400.3  828908
400.9  695425
401.9  683908
403.1  556803
405.8  906291
407.2  293691
408.2  2475720
409.4  305360
412.4  324123
413.6  1974477
414.8  2339087
415.6  242656
416.5  1402914
417.4  1635225
418.1  354
419.8  681890
420.5  63
[data continued ...]

```

**Figure 1.** MS1 file format description. (A) General description of required fields and format used in the MS1 file. Both required and optional lines and field descriptions are noted along with a generic pattern for data storage. (B) Example MS1 file and a partial spectrum. Following the *H* lines of the header each full-scan spectrum begins with an *S* line denoting its scan number. Next, optional *I* lines give additional information about that scan such as retention times. Finally, the spectral data are stored in as series of *m/z* intensity pairs. This pattern of *S(I)[m/z intensity]<sup>n</sup>* continues until each full-scan spectrum has been represented.

U—unevaluated, Y—yes, N—no, or M—maybe. The inclusion of this field allows the SQT file format to store manual validation information that cannot be stored in OUT files.

To institute such file format changes we made modifications to several pre-search (extraction and filtering), searching (*SEQUEST*), organization and summary (*DTASelect*<sup>12</sup>), and visualization (results display) programs. For spectral extraction we have the Linux-compiled, *makems2* to extract MS/MS from a *.dat* (ICIS) file into an MS2 format and perform limited charge-state selection and filtering. It is basically a unified format version of the *extractms* program. For spectral extraction directly from *.RAW* files, we have *MSMaker* which extracts both MS and MS/MS spectra, but does only limited charge-state selection. Charge-state selection is done primarily using *2to3u*.<sup>13</sup> *DTASelect* has been designed to accommodate a variety of file formats, including these new ones.

For visualization purposes we developed a new CGI, *show*,<sup>14</sup> which gathers information from both the SQT and MS2 files and passes them to an applet version of the *DTASelect* ion display graphical interface. It is also back-compatible with *DTA*, *OUT*, and a variety of intermediate file formats. Finally, we developed a PERL script, *Unitemare*, which transcodes previously searched DTAs and OUTs into the new unified formats. With the exception of *SEQUEST*, these programs are freely available to academic and other non-profit groups; see the group website for details.<sup>15</sup>

For performance comparisons a single MudPIT cycle was chosen from a six-cycle analysis of a previously described protein mixture.<sup>16</sup> These data were extracted either into *DTA* or *MS2* format without filtering and only rudimentary charge-state discrimination. The 2886 non-blank MS/MS spectra generated a total of 5642 *DTA* files to be searched (as a

A

symbol	meaning	generic form	required
H	header	H [field label] [corresponding value]	yes
S	scan	S [first scan] [second scan] [pre m/z]	yes
I	charge-independent analysis	I [field label] [corresponding value]	
Z	charge	Z [charge] [mass]	yes
D	charge-dependent analysis	D [field label] [corresponding value]	
[data]	mass intensity pairs	[m/z] [intensity]	yes
<b>"H" line required field labels</b>			
field label	description		
CreationDate	The date and time on which this file was created.		
Extractor	The name of the software used to create the MS2 file.		
ExtractorVersion	The version number of the Extractor software.		
ExtractorOptions	The options used in running the Extractor software.		
<b>"H" line optional field labels</b>			
field label	description		
IAnalyzer	If software to conduct charge state-independent analysis of the spectra is used to analyze the MS2 file (such as removing sparse spectra), the name of the program should be recorded in the header.		
IAnalyzerVersion	The version number of the IAnalyzer		
IAnalyzerOptions	The options used for the IAnalyzer		
DAnalyzer	If software to conduct charge state-dependent analysis of the spectra is used to analyze the MS2 file (such as removing possible precursor charge states), the name of the program should be recorded in the header.		
DAnalyzerVersion	The version number of the DAnalyzer		
DAnalyzerOptions	The options used for the DAnalyzer		
SortedBy	If a program is used to sort the spectra, which field are they sorted by?		
InstrumentType	The type of mass analyzer used.		
Comment	Remarks, ownership, and copyright information may be included. Multiple comment lines may be included.		
InstrumentSN	The serial number of the mass spectrometer used.		

B

```

H      CreationDate      unknown
H      Extractor          RAWXtract
H      ExtractorVersion   1.0
H      ExtractorOptions
S      0002      0002      534.04
I      RTime      0.02
Z      2          1067.08
Z      3          1600.12
151.0  22869
153.0  14453
155.1  35721
157.1  17102
158.1  12948
166.3  24733
167.0  103456
168.0  11485
171.0  92158
172.1  62988
173.1  136408
[data continued ...]

```

**Figure 2.** MS2 file format description. (A) General description of required fields and format used in the MS2 file. The format follows the general conventions of the MS1 file format except that MS/MS information is stored and with the addition of the required Z lines and the optional D lines. (B) Example MS2 file and partial spectrum. As with the MS1 file, each spectral description in the MS2 file begins with an S line. Z lines denote which charge states are to be considered for the spectrum. Like the MS1 file, the spectrum itself is represented as a series of [m/z intensity] pairs. Optional I and D lines can be used to store charge-state-independent and charge-state-dependent information, respectively. The general pattern  $S(I)[Z(D)]^k[m/z \text{ intensity}]^n$  continues until all MS/MS spectra are represented.

result of the need to store a separate DTA file if multiple charge states were to be considered). Both formats were searched using SEQUEST against the same database using identical settings to generate OUT and SQT files. Disk usage measures were performed either using the Linux *du-b* command or *folder properties* in WindowsXP.

## RESULTS AND DISCUSSION

One of the primary design goals for these files was to reduce the number of files which must be stored on the hard drive. Clearly, this was accomplished since the MS/MS spectra and

search results of a typical MudPIT cycle were able to be stored in two files rather than the 5642 DTAs and their corresponding 5642 OUTs present in the example file. While it is difficult to quantify the impact that this has had on the stability of our file servers, anecdotally we noted a dramatic increase in uptime with a concomitant decrease in errors (primarily network file system, NFS, errors). We experienced this increased stability in spite of going from ~1/3 terabytes of data to >1.5 terabytes of data stored on our two Linux file servers.

A potential problem with aggregating results into a single file is that accessing a specific spectrum or search result can be much slower than having them split out as single files. We



**Table 1.** Savings per file type

		Bytes	% Saved
MS/MS	DTAs	35 115 008	
Data	MS2	10 207 232	70.93%
Search	OUTs	23 355 392	
Results	SQT	3 858 432	83.48%

result files would be predicted to help since many of these files were smaller than the block size or minimum file size. This could be seen by measuring the total disk usage for a folder of DTA and OUT files, 58 908 672 bytes total, versus the total usage if all the DTA files and OUT files were put together into two files, 35 729 408 bytes. However, the redundancy of the headers in the OUT files and the file repetitions needed to represent multiple charge states allowed a total saving of about 75% for the SQT and MS2 file formats (Table 2). The largest reduction in disk space requirements was seen in going from the OUT files to the SQT file, an 83% saving (Table 1). Similar differences were seen when the files were stored on a WindowsXP machine (NTFS formatted partitions) (Table 2). These savings scaled proportionally to the number of LC/MS/MS runs (data not shown). Since it is now trivial to collect 100 s of gigabytes of data in a relatively short period of time, being able to store data in one-quarter of the space is significant, especially for those academic labs that are unable to afford or maintain multi-terabyte storage arrays. As newer and faster instruments emerge these considerations will become even more significant; for instance, the next generation ThermoFinnigan linear ion trap (LTQ) collects about four and a half times as many spectra as the LCQ in the same amount of time.

Another advantage we have noted is that programs which have to read through every result in a given experiment, such as *DTASelect*, are able to parse through these files more quickly. For those same 5642 OUT files, *DTASelect* required 4.2 s to read through and gather all the necessary information. The corresponding SQT files were parsed in a less than a second. Again, while not substantial for a single file, this parsing can take a great deal of time when the dataset consists of >50 MudPIT cycles which in turn were searched against multiple databases.<sup>4</sup> In comparison to their predecessors, the design of the unified file formats makes it relatively easy to develop software to read them, especially when compared to the difficulty of dealing with all of the subtle differences in OUT files produced by different SEQUEST revisions.

These file formats also allowed for a more streamlined, logical, and flexible workflow. First, the MS2 file logs the serial application of multiple programs during the data analysis workflow. After initial extraction of the various files, these could include spectral quality filtering/scoring, charge-state selection, and feature annotation. Which programs have

been run is stored as header information and the *I* and *D* lines provide the flexibility to annotate specific spectra and/or specific charge states of those spectra. This expansion room without compromising file size is an important feature that was missing from the original DTA file format.

Another workflow advantage can be seen when one wishes to analyze the same data versus multiple databases or for a variety of post-translational modifications (e.g. see MacCoss *et al.*<sup>6</sup>). Instead of having to search copies of the DTA files it becomes quite practical to use symbolic links back to the original MS2 files. Even more dramatic space savings can be realized under such a scenario. Using symbolic links to the MS2 files, six different searches could be performed on our example LC/MS/MS run with a total savings of 94% over DTAs and OUTs (~352 megabytes vs. ~22 megabytes). In fact, after all of these searches have been performed, it is possible to collate the various answers back into a single aggregate SQT file. Extensive analysis of complex datasets requires flexible formats to bring these results together into an easily digested final output. The MS1 file format, for instance, could be used to extract full-scan chromatograms of individual peptides for purposes of quantitation or characterizing chromatographic efficiency.

Ongoing discussions seek to standardize file formats for proteomic data, with the ultimate goal of moving towards a common database schema that can be employed globally.<sup>18,19</sup> However, there is an evident need for an intermediate step moving from either the single spectrum or proprietary instrument manufacturer formats to this ultimate goal. Several groups are proposing moving towards common XML (extensible markup language) formats in order to store all the data concerning a particular experiment. There are many advantages to this idea in terms of tools available to deal with XML data, and of course, a common language spoken by all proteomics labs. However, XML files typically spend many bytes on formatting information, potentially increasing rather than decreasing storage capacity requirements. The extensibility to new instruments and experimental strategies possible with XML formatting may not prove an adequate gain for the cost in file size.

The MS1, MS2, and SQT file formats do not meet all of the goals of these upcoming database and XML standards, and are not intended to substitute for them. However, since they are compact, flexible, easily parsed, and mature in their implementation, we propose that they can serve a very useful role in the proteomic mass spectrometry community. They should be particularly appealing to small-scale labs that are still able to generate large volumes of data, but have necessarily limited computational and storage resources. The ease with which these files can be parsed and the existing suite of tools under continuing development in our group and others make them an appealing platform. The use of tab-delimited text files makes the creation of translation software to produce other formats of data trivial, allowing export to whatever industry standards are ultimately adopted. We propose that they are a viable alternative to an XML-based single file format because of advantages in disk space required, developmental flexibility, and ease in later translation to an industry-standard XML or database format for dissemination and sharing of data.

**Table 2.** Total savings

	Linux	% Saved	WinXP	% Saved
DTAs and OUTs	58 462 208		59 469 923	
Concatenated	35 274 752	39.66%	36 630 528	38.40%
MS2 and SQT	14 061 566	75.95%	14 987 264	74.80%

**REFERENCES**

1. McCormack AL, Schieltz DM, Goode B, Yang S, Barnes G, Drubin D, Yates JR III. *Anal. Chem.* 1997; **69**: 767.
2. Link AJ, Eng J, Schieltz DM, Carmack E, Mize GJ, Morris DR, Garvik BM, Yates JR III. *Nat. Biotechnol.* 1999; **17**: 676.
3. Washburn MP, Wolters D, Yates JR III. *Nat. Biotechnol.* 2001; **19**: 242.
4. Florens L, Washburn MP, Raine JD, Anthony RM, Grainger M, Haynes JD, Moch JK, Muster N, Sacci JB, Tabb DL, Witney AA, Wolters D, Wu Y, Gardner MJ, Holder AA, Sinden RE, Yates JR, Carucci DJ. *Nature* 2002; **419**: 520.
5. Peng J, Elias JE, Thoreen CC, Licklider LJ, Gygi SP. *J. Proteome Res.* 2003; **2**: 43.
6. MacCoss MJ, McDonald WH, Saraf A, Sadygov R, Clark JM, Tasto JJ, Gould KL, Wolters D, Washburn M, Weiss A, Clark JL, Yates JR III. *Proc. Natl. Acad. Sci. USA* 2002; **99**: 7900.
7. Gatlin CL, Eng JK, Cross ST, Detter JC, Yates JR III. *Anal. Chem.* 2000; **72**: 757.
8. Eng JK, McCormack AL, Yates JR III. *J. Am. Soc. Mass Spectrom.* 1994; **5**: 976.
9. Sadygov RG, Yates JR III. *Anal. Chem.* 2003; **75**: 3792.
10. Tabb DL, Saraf A, Yates JR III. *Anal. Chem.* 2003; **75**: 6415.
11. Available: <http://fields.scripps.edu/sequest/unified>.
12. Tabb DL, McDonald WH, Yates JR III. *J. Proteome Res.* 2002; **1**: 21.
13. Sadygov RG, Eng J, Durr E, Saraf A, McDonald H, MacCoss MJ, Yates JR III. *J. Proteome Res.* 2002; **1**: 211.
14. Available: <http://fields.scripps.edu/sequest/show/index.html>.
15. Available: <http://fields.scripps.edu>.
16. McDonald WH, Ohi R, Miyamoto D, Mitchison TJ, Yates JR III. *Int. J. Mass Spectrom.* 2002; **219**: 245.
17. Available: <http://fields.scripps.edu/sequest/SQTSort.html>.
18. Orchard S, Hermjakob H, Apweiler R. *Proteomics* 2003; **3**: 1374.
19. Taylor CF, Paton NW, Garwood KL, Kirby PD, Stead DA, Yin Z, Deutsch EW, Selway L, Walker J, Riba-Garcia I, Mohammed S, Deery MJ, Howard JA, Dunkley T, Aebersold R, Kell DB, Lilley KS, Roepstorff P, Yates JR, Brass A, Brown AJ, Cash P, Gaskell SJ, Hubbard SJ, Oliver SG. *Nat. Biotechnol.* 2003; **21**: 247.