Appendix A. Description of the Boolean Tool

A.1. The database

The algorithms use CGAP's relational database, which is composed of a library list (Figure 1S), a transcript list (Figure 2S) and an expression table (Figure 3S). All are available from CGAP's website (http://cgap.nci.nih.gov/Info/CGAPDownload). For searching the library database is converted into a table with the field headings as the columns and each library occupying its own row. As of 25 February 2012, this table had 8,335 rows (including field headings) and 72 columns. The transcript database is similarly treated (124,132 rows including field headings, and 19 columns), while the expression database is converted to a two dimensional array with each library represented by each column and each transcript represented by each row, in the same order of appearance as in their respective tables (this therefore had 8,335 columns and 124,132 rows). The number of ESTs mapping onto a particular transcript in a specific library appears where the relevant row and column intersect.

Two tools were created using Visual Basic for Applications and Microsoft Excel Excel, one based on CGAP's cDNA xProfiler (and therefore reporting presence of absence in a Boolean manner) and the other based on CGAP'scDNA DGED (and therefore presenting quantitative expression levels). For both new tools, the overall aim is to make a greater range of investigations possible within the same time frame. The Boolean tool is described here.

A.2. How the Boolean tool works

This tool reports the presence or absence of a transcript in the libraries studied for each of the user-selected tissues.

Procedure 1: Selecting libraries to use

This procedure reports a list of EST libraries from the required tissues. Steps 1.1 – 1.10 below apply to each pool individually. Steps 1.11 – 1.13 are applied to all pools together. This procedure requires the CGAP library database (Figure 1S). As with the CGAP algorithms, the distribution of libraries between pools can be subsequently altered by the user.

Step 1.1. Selection of the pools to use

The number of library pools is set by the user. The default is 8, this accounts for most foreseeable cases. Any number of pools can be omitted (as indicated by the user). A larger number of pools is possible in principle (not allowed for simplicity in the current version)

This functionality enables the finding of transcripts in a specific tissue but which are not expressed in unrelated, connected or proximal tissues. For example, to find transcripts expressed only in thyroid and not in a range of related or proximal tissues, one could request:

1. Pool 1: thyroid.
2. Pool 2: parathyroid.
3. Pool 3: vascular.
4. Pool 4: peripheral nervous system.
5. Pool 5: oesophagus.
6. Pool 6: muscle.

Once the list of transcripts was produced using Procedure 2, the results would be filtered to display only those transcripts present in Pool 1 (thyroid) and not in any of the tissues presented in the other pools.

This is different from the CGAP cDNA xProfiler tool, which has a built-in limit of 2 pools (no less no more). For the above example, individual study of each of tissues 2 – 6 would not be possible because they would have to be grouped together in the same pool. Individual study of those tissues would require multiple searches and the merging of the results using other software.

Step 1.2. Selection of the tissue type for each pool

The user may optionally specify a tissue type for each pool. The required tissue type is found by finding libraries whose "UNIQUE_TISSUE" field entry (see Figure 1S) is identical to the required phrase for the tissue selected.

For example, if "prostate" is requested, libraries whose "UNIQUE_TISSUE" annotation is "prostate" (as in the example shown in Figure 1S) will be reported in that pool, while libraries with any other term as the "UNIQUE_TISSUE" entry (for example, "muscle" or "kidney") will not be presented in the pool. This means that if "ear" is requested, the only libraries reported are those with "ear" as their "UNIQUE_TISSUE" annotations, for example the library named "Morton Fetal Cochlea", with no irrelevant libraries being presented. Irrelevant libraries that are not presented include "Atrium cDNA library Human heart", "Atrium cDNA library Human heart" and "Human heart cDNA (YNakamura)" (all from heart).

This is unlike the CGAP tools, which search the keywords field entry "KEYWORDS" (see Figure 1S) for the required phrase within a longer string (partial match), for example "prostate" within the "KEYWORDS" entry in Figure 1S. This has previously resulted in libraries containing the phrase "ear" as part of any other string. The libraries whose "KEYWORDS" fields are shown below were all reported when "ear" is requested, in addition to those which only contain the phrase "ear".

Soares_fetal_heart_NbHH19W:
"fetus, heart, normal, bulk, normalized, CGAP, EST, Soares normalized, phagemid, oligo-dT primed",
from heart

NCI_CGAP_Kid5:
"kidney, bulk, normalized, CGAP, EST, Soares normalized, unknown developmental stage, phagemid, oligo-dT primed, clear cell renal carcinoma",
from kidney.

NCI_CGAP_Ov23:
"metastasis, ovary, non-normalized, bulk, CGAP, EST, LTI non-normalized, unknown developmental stage, size fractionated, directionally cloned, phagemid, oligo-dT primed, papillary ovarian carcinoma, serous ovarian tumor, clear cell ovarian tumor, spindle cell ovarian tumor, mixed tumors, mixed mullerian tumor of the ovary",
from ovary.

In summary, the main difference of our algorithm to those of CGAP is the reporting of libraries from exactly the requested tissue, rather than including those which contain it within a longer phrase. Although many differences in the annotations can be picked up by selecting a different field, the key is in using correct querying approaches, such as searching only for the exact match, instead of accepting partial matches within a longer string.

Step 1.3. Selection of dependent tissues

The new algorithms allow the user to specify whether or not dependent tissues are reported. This allows one to study conditions in, for example, parts of the eye except for the retina, instead of being limited to studying conditions which also affect the retina.

For example, if "eye" is requested and dependent tissues are not required, the only libraries presented will be those whose "UNIQUE_TISSUE" annotation is "eye". However, if dependent tissues are required, libraries will also be reported whose "UNIQUE_TISSUE" annotation is "retina". The association of dependent tissues to parent tissues is illustrated in Table 1S.

This is unlike the CGAP tools which does not provide control over whether or not to include dependencies and would arbitrarily report them in nearly all cases. The exception was bone libraries with the phrase "bone marrow" in their "KEYWORDS" entry are not presented when "bone" was requested, while libraries containing the phrase "retina" in their "KEYWORDS" field were included when "eye" was requested.

In summary, the main difference of our algorithm to those of CGAP is the provision of control over the inclusion of dependencies.

Step 1.4. Selection of the requested tissue preparation

Both CGAP's tools and the new algorithms provide the user with the option to specify a tissue preparation for each pool. Using the new tools, the specified tissue preparation is found by finding libraries whose "UNIQUE_PREPARATION" field entry (see Figure 1S) is identical to the tissue preparation selected.

For example, if "microdissected" is requested, libraries whose "UNIQUE_PREPARATION" annotation is "microdissected" (as in the example shown in Figure 1S) will be reported in that pool, while libraries with any other term as the "UNIQUE_PREPARATION" annotation will not be presented in the pool. This means that if "multiple preparation" is requested, the only libraries reported are those with "multiple preparation" as their "UNIQUE_PREPARATION" annotation, with no other libraries shown.

This is unlike the CGAP tools, which search the "KEYWORDS" field for the requested phrase within a longer string, for example "microdissected" within the "KEYWORDS" annotation in Figure 1S. This results in libraries with multiple tissue preparation annotations being reported when one of the tissue preparations concerned is requested, and not being available as a separate option. For example, the library named "Stratagene colon HT29 (#937221)" would be presented when either bulk or cell line is selected because its "KEYWORDS" entry is "non-normalized, colon, bulk, cell line, adult, EST, female, size fractionated, directionally cloned, phagemid, oligo-dT primed, adenocarcinoma of the colon".

In summary, the main difference of our algorithm to those of CGAP is the reporting of libraries created using the requested tissue preparation only, instead of also showing those which contain the annotation within a longer phrase.

Step 1.5. Selection of the appropriate library protocol

Both CGAP's tools and the new algorithms allow the user the option to specify a library protocol for each pool. With the new algorithms, the required library protocol is selected by finding libraries whose "UNIQUE_PROTOCOL" field entry (see figure 1S) is identical to the library protocol chosen.

For example, if "non-normalized" is requested, libraries whose "UNIQUE_PROTOCOL" annotation is "non-normalized" (as in the example shown in Figure 1S) will be reported in that pool, while libraries with any other terms as the "UNIQUE_PROTOCOL" annotation will not be presented. This means that if "multiple treatment" is requested, the only libraries reported are those with "multiple treatment" as their "UNIQUE_PREPARATION" annotation.

This is unlike the CGAP algorithms, which search the "KEYWORDS" field for the requested phrase as part of a longer string, for example "non-normalized" within the "KEYWORDS" annotation in Figure 1S. The result is libraries with multiple library protocol annotations being reported when one of library protocols concerned is required, and not being available as a separate option. For example, "NIH_MGC_20" would be reported when either non-normalised or normalised is selected because its "KEYWORDS" entry is "skin, melanoma, non-normalized, cell line, EST, unknown developmental stage, MGC, Rubin normalized, size fractionated, plasmid vector, directionally cloned, oligo-dT primed".

In summary, the main difference of our algorithm to those of CGAP is the presentation of libraries prepared using the required library protocol only, instead of also showing those which contain the requested annotation as part of a longer phrase.

Step 1.6. Selection of the correct tissue histology

Both the new algorithms and CGAP's original tools give the user the option to specify a tissue histology for each pool. In the new algorithms, the required tissue histology is found by finding libraries whose "UNIQUE_HISTOLOGY" field entry (see Figure 1S) is the exact tissue histology requested.

For example, if "neoplasia" is requested, libraries whose "UNIQUE_HISTOLOGY" entry is "neoplasia" (as in the example shown in Figure 1S) will be reported in that pool, while libraries with any other terms as the "UNIQUE_HISTOLOGY" annotation will not be presented in that pool. This means that if "multiple histology" is required, the only libraries reported are those whose "UNIQUE_HISTOLOGY" annotation is "multiple histology".

This is unlike the CGAP tools, which searched the "KEYWORDS" field for the requested phrase as part of a longer string, for example "prostatic intraepithelial neoplasia" within the "KEYWORDS" annotation in Figure 1S. This results in libraries with multiple histology annotations being reported when one of the tissue histology annotations concerned is requested, and not being available as a separate option. For example, the library entitled "NIH_MGC_56" would be presented when either non-normalised or normalised is selected because its "KEYWORDS" entry is "brain, normal, non-normalized, cell line, full-length enriched, EST, unknown developmental stage, MGC, Clontech non-normalized, primitive

neuroectodermal tumor of the CNS" (where "primitive neuroectodermal tumor of the CNS" is used to assign "neoplasia" as one of the histology annotations of the library.

In summary, the main difference of our algorithm to those of CGAP is the reporting of libraries with the required histology annotation, instead of also presenting those which contain that annotation as part of a longer phrase.

Step 1.7. Selection of the chosen developmental stage

The new algorithms provide the user with the option to only present libraries in a pool which match a specified developmental stage. The specified developmental stage is found by finding libraries whose "KEYWORDS" field entry (see Figure 1S) contains the developmental stage requested as part of a longer string.

For example, if "adult" is requested, libraries whose "KEYWORDS" entry contains "adult" (as in the example shown in Figure 1S) will be reported in that pool, while libraries with any other developmental stage term in their "KEYWORDS" annotation will not be presented in that pool. As with the other settings except for library name (Step 1.10) and library size (Steps 1.12 and 1.13), the user is required to choose the requested term from a list.

This is unlike the CGAP tools, which do not provide options for selecting a developmental stage. This would be needed if, for example, the aim is to study adult tissue alone.

Step 1.8. Selection of the specified gender

The developed algorithms enable the user to optionally specify a gender for each pool. The required gender is found by finding libraries whose "KEYWORDS" field entry (see Figure 1S) contains the gender requested as part of a longer string.

For example if "male" is requested, libraries whose "KEYWORDS" entry contains "male" (as in the example shown in Figure 1S) will be presented in that pool, while libraries with any other gender term in their "KEYWORDS" annotation will not be reported in that pool.

This is unlike the CGAP tools, which do not provide the facility for choosing a gender. This is useful because gene expression levels in each gender may be different and the aim of an investigation may be to only study cancer in males or females.

Step 1.9. Selection of the requested pregnancy state

The new tools enable the user to optionally choose a specific pregnancy state for each pool. The requested pregnancy state is found by finding libraries whose "KEYWORDS" field entry (see Figure 1S) contains the pregnancy stage requested as part of a longer string.

For example, if "pregnant" is requested, libraries whose "KEYWORDS" entry contains "pregnant" will be reported in that pool, while libraries without this term in their "KEYWORDS" annotation will not be reported in that pol.

This is unlike the CGAP tools, which do not provide options for choosing a pregnancy state. Because gene expression levels are altered during pregnancy, this enables the user to exclude libraries from individuals who are pregnant.

Step 1.10. Selection of the requested library name

Both CGAP's tools and new algorithms allow the user the option to specify a library name in each pool. The new tools provides options to choose whether the specified library name is found by either reporting only the library whose "LIBRARY" field entry (see Figure 1S) is identical to the name requested, or reporting all libraries which contain the requested phrase as part of a longer string in the "LIBRARY" field.

For example, choosing to report the exact match and entering "aorta endothelial cells results in the library named "Aorta endothelial cells" being reported in that pool. However, choosing to report a partial match and entering that same phrase will report in the library named "Aorta endothelial cells, TNF alpha-treated" also being presented in the pool.

This is different to the CGAP tools, which do not provide the option to report only the library with the exact name entered, and always look for a partial match. In the above example this results in both the two named libraries being presented in the pool, and the user would need to manually exclude one of them when the list of libraries is presented.

Step 1.11. Selecting the specified library group(s)

Using both CGAP's algorithms and the new tools, the user has the option to specify the library groups to be reported. Using the new tools, the requested library group(s) are found by finding libraries whose "KEYWORDS" field entry (see Figure 1S) contains the required phrase(s) within a longer string.

For example, if "CGAP" is requested, libraries whose "KEYWORDS" annotation contains "CGAP" (as in the example shown in Figure 1S) will be reported. Only if "Multiple Library Group" is requested will libraries whose "KEYWORDS" annotations contain both "CGAP" and "MGC" be presented, otherwise libraries containing only one of those phrases will be reported if one of those groups is selected alone.

This is different to the CGAP tools, which also searched the "KEYWORDS" field for the requested phrase within a longer string, but they presented libraries annotated with both "CGAP" and "MGC" in their "KEYWORDS" annotations if just one of these was requested. An example of such a library is "NCI_CGAP_GCB1", which has the "KEYWORDS" entry, "lymph node, B-cell, normal, flow-sorted, normalized, CGAP, EST, Soares normalized, unknown developmental stage, MGC".

Step 1.12. Selection of libraries which are larger than a specified minimum EST count

Both CGAP's tools and the new algorithms give the user the option to specify a minimum library size cut-off value. The new algorithms apply the specified minimum library size cut-off by finding libraries whose entry in an additional "Number of ESTs mapping onto transcripts in library" field is greater than or equal to the specified number of transcript-mapping ESTs. This new field was added by finding and recording the number of transcript-mapping ESTs in each library.

For example, a library whose "Number of ESTs mapping onto transcripts in library" entry is equal to 999 will not be presented in a search with a display cut-off value of 1,000.

This is unlike the CGAP tools, which search the "SEQS" field (see Figure 1S) in the same manner, therefore including ESTs which do not map onto transcripts. This will be commented on in A.3 “Conclusions” and is discussed in Appendix C.

Step 1.13. Selecting libraries which are smaller than a specified maximum EST count

The new tools provide the user with the option to specify a maximum library size display cut-off. The required maximum library size cut-off is applied by finding libraries whose entry in an additional “Number of ESTs mapping onto transcripts in library” field is less than or equal to the specified number of transcript-mapping ESTs.

For example, a library whose “Number of ESTs mapping onto transcripts in library” entry is equal to 1,001 wil lnot be reported in a search with a display cut-off value of 1,000.

This is unlike the CGAP tools, which do not provide the option for choosing a maximum library size cut-off. This enables the range of library sizes reported to be narrower, if desired, than would otherwise be the case.

Procedure 2: Reporting transcripts found in the libraries

This procedure reports, in a Boolean (present or absent) manner, the transcripts present in all the pools (currently a maximum of eight) on an existing library list produced using Procedure 1. In the example shown in Step 1.1, six pools of transcripts would be presented for the following tissues:

1. Pool 1: thyroid.
2. Pool 2: parathyroid.
3. Pool 3: vascular.
4. Pool 4: peripheral nervous system.
5. Pool 5: oesophagus.
6. Pool 6: muscle.

The databases required are the transcript database (Figure 2S) and expression table (Figure 3S). CGAP’s cDNA xProfiler carries out the same procedure. Furthermore, CGAP’s tool uses its own internally available flat file database that contains an entry for each library that lists the transcripts associated with it by the UniGene Cluster ID, and divides those transcripts into four groups according to whether they have a known or unknown cytogenetic location, and are unique or non-unique to that library. This enables one to report transcripts which are not expressed anywhere else in the body, and its function can be replicated using the new tool by setting up an additional pool containing all libraries other than those in the other pools.

The required transcripts are selected for reporting in each pool if their UniGene Cluster ID field entry “>>” (see Figure 2S) is identical to that in any expression records which contain the same UniGene Library ID as that shown in the “UNIGENE_LIB_ID” field (see Figure 1S) for a library found in that pool.

For example, if the library entitled “TEST1, Human adult Testis tissue”, which has a “UNIGENE_LID_ID” annotation of “40”, is in a pool, expression records with the same UniGene Cluster ID will be retrieved. This includes the transcript named “Discs, large

(Drosophila) homolog-associated protein 5", which has a ">>" entry of "77695", so this transcript would be presented in the pool.

Procedure 3: Filtering up to four pools on the list of transcripts

This procedure reports the transcripts which meet the criteria chosen using Equation (1) in up to four pools on an existing transcript list produced using Procedure 2. CGAP's cDNA xProfiler does not carry out this process at all, because it always uses two pools, but it always fitlers the transcripts by known, unknown, unique and non-unique. Four pools are available to allow for the majority of foreseeable situations.

$$\left(P1\begin{pmatrix}AND\\OR\\NOT\end{pmatrix}P2\right)\begin{pmatrix}AND\\OR\\NOT\end{pmatrix}\left(P3\begin{pmatrix}AND\\OR\\NOT\end{pmatrix}P4\right) \quad (1)$$

Where P1, P2, P3 and P4 are four pools. The user can choose which pools these refer to.

Step 3.1. Selection of pools to use

The pools are chosen by the user. The maximum is four. Any number of pools can be omitted (as indicated by the user).

For example, if the user wants to study three pools, the space that will not be used can be set to "ignore" so that the procedure runs on the three pools only. An application where this would be needed is if the aim is to find transcripts which are found only in a cancer of interest and not in related, connected or proximal tissues. For example, if the aim is to find genes which are expressed in rhabdomyosarcoma and not in normal muscle or the heart, a three-pool search would be set up as follows:

1. Pool 1: rhabdomyosarcoma.
2. Pool 2: muscle.
3. Pool 3: heart.

In this example, the cancer of interest is in one pool, with the normal tissue in which it is found (muscle) in another pool, with the third pool containing tissue of a similar structure and function (heart). This enables each to be studied separately whilst still comparing them in a single search.

Step 3.2. Creation of the filtered transcript list

The transcript list is filtered according to the criteria specifried using Equation (1). Selection of "AND" for an operator results in the selection of transcripts appearing in both pools (or groups of two pools) on each side. Selecting "OR" results in the selection of transcripts only appearing on one side. Choosing "NOT" results in the selection of transcripts which are found specifically on the left side and not on the right side.

For example, to find transcripts which are only found in rhabdomyosarcoma (as in the example named in Step 3.1), the options would be set as shown in Equation (2). and

$$(P1\ AND\ IGNORE)\ NOT\ (P2\ OR\ P3) \quad (2)$$

The aim of this is to report potential biomarkers or targets for rhabdomyosarcoma by selecting transcripts which are only expressed in this cancer and not in the host tissue (muscle) or a tissue related by structure and function (heart). Such transcripts can then be further investigated as potential biomarkers for diagnosis or prognosis, or targets for novel treatments.

A.3. Conclusions

The Boolean algorithm described here expands on and add additional functionality to CGAP's practice of comparing two pools of libraries, preventing, for example, thyroid being compared with a range of related, connected and proximal tissues whilst still being able to study each separately. Thus a wider range of investigations are possible within the same time frame.

Furthermore, the CGAP algorithms were found to contain serious flaws which resulted in libraries being presented which did not originated from the specified tissue and different transcripts being reported from those contained in the chosen libraries. Moreover, the Benjamini-Hochberg statistics used by CGAP were dependent on the proportion of the results displayed, instead of only indicating the statistical significance of differential expression between the two library groups. Finally, the library EST count annotations were incorrect. These sources of error are eliminated by reconfiguring the database and searching it in the ways described here.

A.4. Figures and Table

```
>>4
LIBRARY: NCI_CGAP_Pr4
UNIGENE_LIB_ID: 934
DESCR: mRNA made from prostate intraepithelial neoplasia (high-
grade), cDNA  made by oligo-dT priming.  Non-directionally
cloned.  Size-selected  on agarose gel, average insert size 600
bp.
KEYWORDS: prostate, high grade, non-normalized, microdissected,
adult, CGAP, EST, Krizman protocol 1, male, oligo-dT primed,
prostatic intraepithelial neoplasia
CLONES: 1536
SEQS: 667
LAB_HOST: DH10B
PRODUCER: Krizman Laboratory
TISSUE: Prostate prostatic intraepithelial neoplasia
TISSUE_SUPPLIER: W. Marston Linehan, M.D., Rodrigo Chuaqui,
M.D., Michael Emmert-Buck, M.D., Ph.D.
VECTOR: pAMP10
VECTOR_TYPE: plasmid (ampicillin resistant)
UNIQUE_TISSUE: prostate
UNIQUE_HISTOLOGY: neoplasia
UNIQUE_PREPARATION: microdissected
UNIQUE_PROTOCOL: non-normalized
RELATED_LIBRARY: 1
RELATED_LIBRARY: 2
RELATED_LIBRARY: 3
RELATED_LIBRARY: 65
```

Figure 1S. An example of entry in the library database. Each tabulated field entry begins with its heading, shown in capital letters, followed by the value associated with that field, after the colon

```
>>2
UNIGENE: Hs.2
SYMBOL: NAT2
TITLE: N-acetyltransferase 2 (arylamine N-acetyltransferase)
LOCUSID: 10
CYTOBAND: 8p22
OMIM: 612182
SEQUENCE: NM_000015
GO: 0006805|xenobiotic metabolic process
GO: 0008152|metabolic process
GO: 0044281|small molecule metabolic process
GO: 0005737|cytoplasm
GO: 0016746|transferase activity, transferring acyl groups
GO: 0016407|acetyltransferase activity
GO: 0005829|cytosol
GO: 0004060|arylamine N-acetyltransferase activity
MOTIF: Acetyltransf_2|PF00797|NM_000015|20|280|484.8|9e-143
HOMOLOG: Mm.14125|72
ALIAS: PNAT
ALIAS: NAT2
ALIAS: AAC2
SNP: 2420889|dbSNP|BC015878|C/T|coding|Tyr94Tyr|350|Validated
SNP: 2420889|dbSNP|BC067218|C/T|coding|Tyr94Tyr|385|Validated
```

Figure 2S. An example of an entry in the transcript database. Each tabulated field entry begins with its heading, shown in capital letters, followed by the value associated with that field, after the colon.

```
120   798   1
120   18323 1
120   1454  4
120   10421 4
120   9263  2
120   474   11
120   18472 4
120   18468 2
120   6762  7
120   525   3
120   10307 1
120   14131 1
120   7139  3
120   17292 2
120   16388 2
120   11917 1
```

Figure 3S. An example of a transcript in the expression database. The left-hand column shows the UniGene Cluster ID of the transcript mapped onto by the relevant ESTs. The central column shows the UniGene Library ID of the library in which the ESTs were detected. The right-hand column states the number of ESTs which map onto the transcript in the library.

Table 1S. Association of dependent tissues to parent tissues by the new algorithms.

Parent tissue	Dependent tissues
bone	bone marrow
brain	cerebellum
	cerebrum
endocrine	adrenal cortex
	adrenal medulla
	parathyroid
	pineal gland,
	pituitary gland
	thyroid
eye	retina
gastrointestinal tract	colon
	esophagus
	salivary gland
	stomach
lymphoreticular	lymph node
	spleen
	thymus
nervous	brain
	cerebellum
	cerebrum
	peripheral nervous system
pancreas	pancreatic islet
soft tissue	adipose tissue