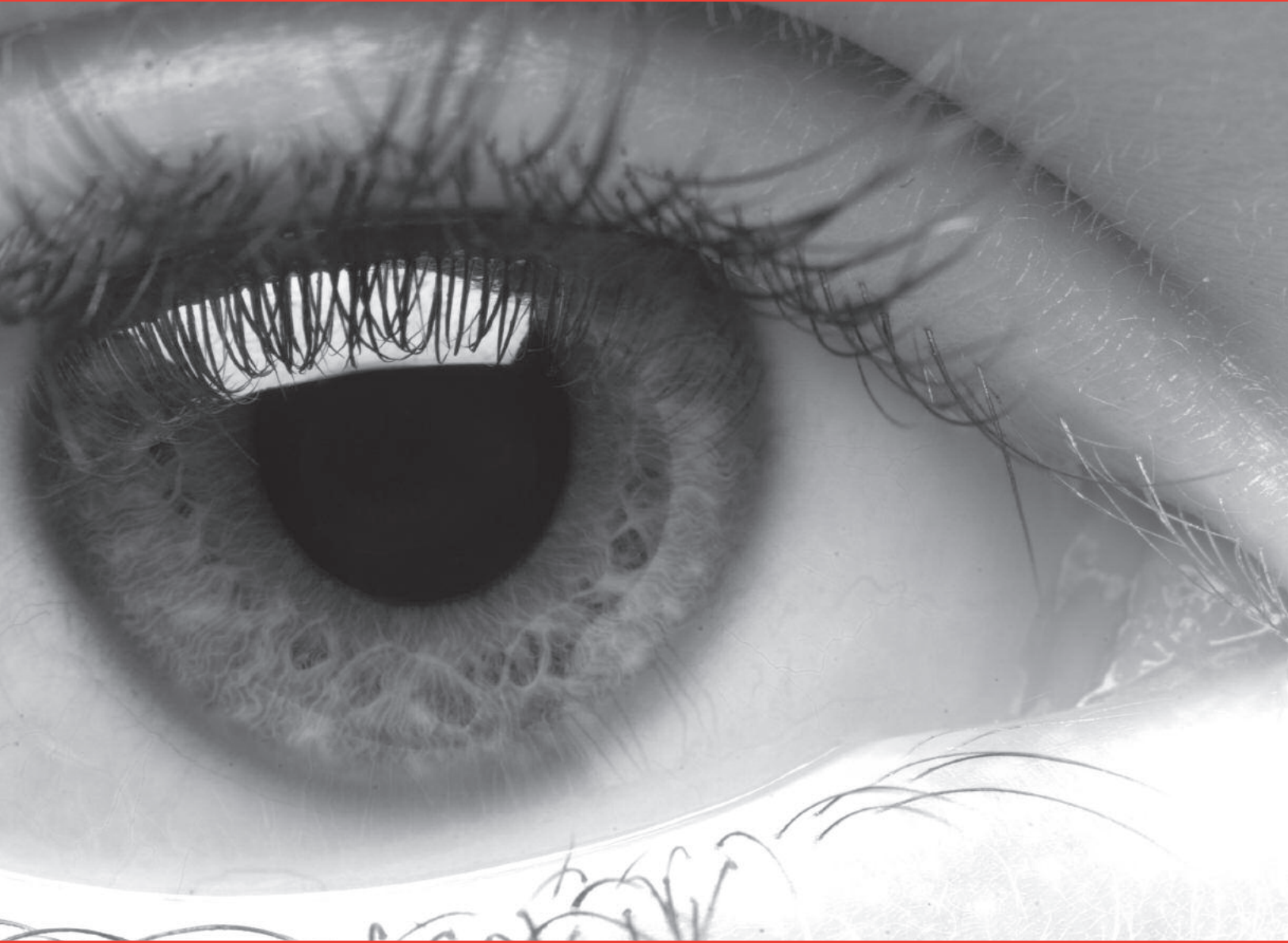


DECIPHERING THE TRANSCRIPTOME

By Austin Tanney and Gavin Oliver



Diagnostics
Sciences
Clinical Services
Clinical Technologies
Pharma Services

www.almacgroup.com

Deciphering the Transcriptome:

***DSA*TM**

By Austin Tanney and Gavin Oliver

The Transcriptome Based Approach

Unlike generic microarrays, DSA™ research tools provide a focused research approach to the chosen disease.

Microarray gene expression studies have, in recent years, become one of the key elements of biological research. The ability to analyse the expression levels of tens of thousands of genes in a single experiment has greatly facilitated researchers from a wide variety of backgrounds in making new and significant discoveries.

Currently available commercial microarray design tends to focus on the best-categorised and most commonly known genes from all body tissues. This may be advantageous in some situations, but it generally leads to a loss of specificity i.e. only a subset of the genes on the generic microarray will yield results in any given tissue-specific study. Likewise, many important transcripts solely expressed in the tissue of interest will not be represented.

The transcriptome of a given tissue or disease state contains a huge amount of transcribed, but not necessarily translated RNA. Recent research has increasingly shown the significant importance of tissue or disease specific splice variation, non-coding RNAs and inherent antisense transcripts.

Our approach has been to focus on the transcriptome of a disease and the result of this is the **DSA™** range of research tools (patent pending). We have undertaken to characterise the transcriptome of each disease and to create a microarray tailored to its study.

The transcriptome is hugely diverse – the DSA™ range addresses this diversity

The outcome of the Almac Diagnostics approach is the most comprehensive overview available of the transcriptome for a particular disease. With tens of thousands of transcripts not available on a leading generic array, each **DSA™** tool provides the technology to analyse gene expression at a level of specificity never before possible.

Product Design

Derivation of transcripts

The objective at the outset of design was to characterise as fully as possible the transcriptome of the disease under investigation.

In order to generate the best representation of the transcriptome possible, a three-pronged approach was taken, deriving data from in-house sequencing, public databases and through experimental investigation.

- **In-house sequencing:** for each disease type we have generated libraries from a representative range of diseased and normal tissues from the disease in question. We have carried out extensive in-house sequencing projects to generate expressed sequence tags (ESTs) for the tissue. These ESTs were subsequently assembled using CAP4 technology with a series of quality filters. The result of this is contiguous sequences of ESTs which have been sequenced several times from the disease setting with high quality values associated with the sequence. In addition to this we have also mined the databases of annotated genes using the remaining singlet sequences (ie those which do not fall into contigs) to derive those annotated genes which we may not have sequenced completely.
- **Public data mining:** In addition to in-house sequencing, we have also mined the public databases for all ESTs which have been sequenced from the disease setting in question. These have also been filtered and assembled in a similar manner to our own sequences and contiguous sequences derived. These are then pruned against our databases of in-house sequences for inclusion on the final product.
- **Experimental Investigation:** The third grouping of sequences is those which have been shown to be relevant to the tissue of interest through in-house experimental work. Those genes which were determined to be of interest and which were not contained in the previous two groupings were also included on the product.

Product content was garnered from all possible sources to ensure extensive and comprehensive coverage of the transcriptome.

This approach to product design has enabled us to populate our **DSA™** range with the widest possible selection of expressed transcripts.

Probeset design

DSA™ research tool probeset design is 3` focused i.e. the probeset is designed within the extreme 3` end of the transcript. The reason for this is twofold:

DSA™ tools are optimised for work with FFPE tissue – enabling use of archived samples.

Initially this approach aimed to provide an optimised platform for the analysis of formalin-fixed paraffin-embedded (FFPE) samples. Generally speaking, the majority of tissue banks use FFPE as the storage method and an array that is optimised for this provides significant advantages. RNA extracted from FFPE samples tends to have a shorter median length from 3` to 5` and the detection of these samples on generic arrays is rarely successful. By designing probesets specific to the 3` extremities of transcripts, a much higher detection rate is possible.

Probesets designed to detect the extreme 3` end of the transcript produce superior experimental data. The **DSA™** range takes advantage of this fact.

Subsequent to the initial product design, in-house studies have shown that the 3` biased design actually yields better detection rates regardless of storage media. Generally the further 5` the probeset is designed, the lower the detection rate. This is the case even in fresh frozen tissue samples. Probesets designed in the more 3` region of the transcript have better detection levels, higher signals and more consistent present calls.

Based upon the Affymetrix GeneChip technology, **DSA™** research tools provide multiple independent probes to each transcript. This ensures the most accurate and reproducible expression measurements possible on any microarray platform.

Annotation

Rationale

An experiment may demonstrate that Sequence X is upregulated tenfold in colorectal cancer, but is its upregulation causative or consequential of the observed disease state? Has anyone else reported similar findings or is this a novel discovery? Just what *is* Sequence X? Questions like these highlight the importance of **DSA™** research tool annotation. The information yielded by a microarray experiment is only as good as the annotation associated with the product. The more comprehensive the annotation information, the greater the value of the microarray.

Comprehensive annotation is as important as product content.

Explanation of Methodology

Almac Diagnostics have developed a custom pipeline in order to ensure the most comprehensive annotation – and thus the most valuable research tool. In-house development has enabled close observation and fine-tuning of the system in order to yield the best annotation information possible.

Target sequences corresponding to array probesets are BLASTed against a series of databases, at controlled stringency levels in an iterative 'BLAST and EXCLUDE' workflow. Once a sequence obtains a satisfactory 'hit' against a database, it is excluded from further BLASTing and proceeds to the annotation stage where the accession it hits against is used to obtain all publicly available annotation information. The structure of the workflow favours high-similarity hits against the best annotated databases and in cases where a good hit cannot be found, further investigation such as exon predictions and genomic alignments are performed. The end result is a catalogue of information for each sequence represented by the array.

*The **DSA™** range is annotated to a high standard to ensure maximum value of experimental results*

With over 40 fields of annotation information associated with each array sequence, **DSA™** research tools are amongst the best-annotated microarrays available.

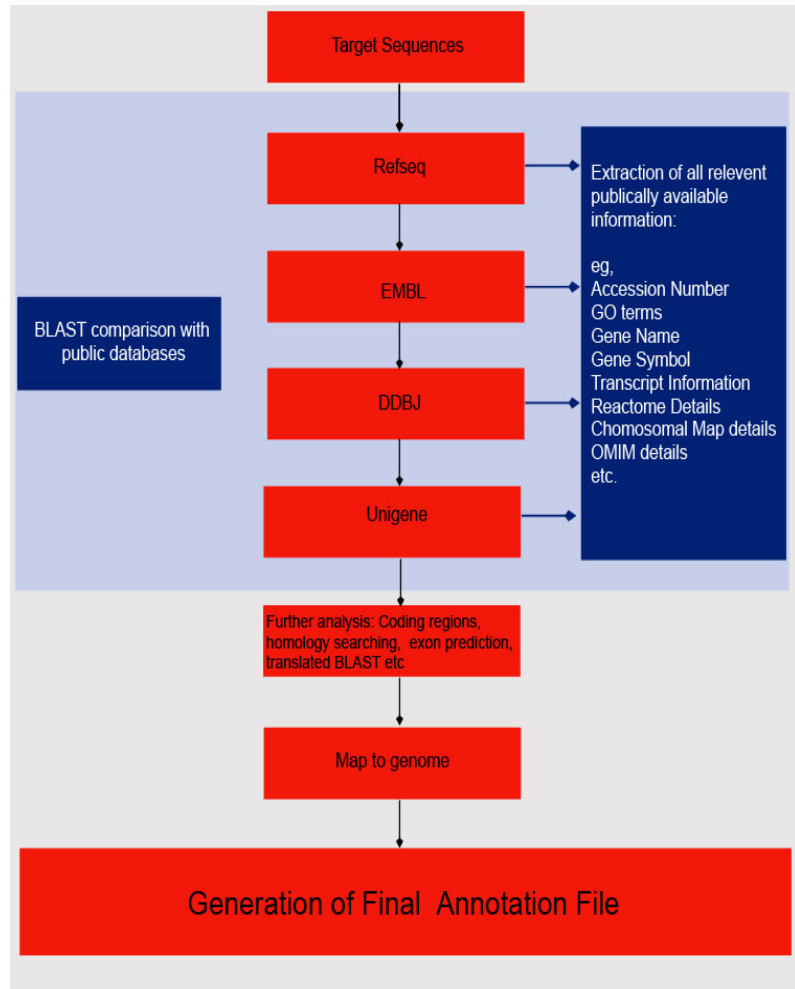


Figure 1: A diagrammatic representation of the DSA™ annotation pipeline.

Details of Updates

Across the globe, researchers from different disciplines are continually working to elucidate the roles and functions of the plethora of human genes and transcripts. With new sequence and annotation data being submitted to public repositories on a constant basis, the need for maintaining up-to-date annotation databases is paramount. In order to ensure this, all **DSA™** research tools are re-annotated on a weekly basis. The initial stages of our annotation pipeline cross-reference our in-house databases with their respective public server versions, performing complete sequence and annotation updates where new information is available. This ensures that we remain abreast of all scientific research pertaining to our product content.

Table 1: Details of the annotation of the Almac Diagnostics DSA™ research tools.

Field	Description
PROBE ID	Unique identifier of the probeset corresponding to the target sequence annotated.
Target Acc	Accession number of the public database sequence retrieved from the target sequence BLAST. The starting point for all annotation.
Transcript Information	A text field describing the exact characteristics of the alignment of the target sequence and database sequence.
Entrez gene Id	The unique accession provided by Entrez Gene. Serves as a link to annotation information in various biological databases.
RNA Acc	RNA accession number
RNA gi	RNA gene identifier
Prot Acc	Protein accession number if protein coding
Prot gi	Protein gene identifier number
Nuc Acc	Nucleic accession number
Nuc gi	Nucleic gene identifier
Unigene Id	Unigene cluster accession number
Refseq RNA	Refseq RNA accession number
Refseq Protein	Refseq protein accession number
Refseq Nucleic	Refseq nucleic accession number
Pubmed Id	Pubmed identifier for linking to related publications
Ensembl gene	Ensembl gene identifier
IPI	International Protein Identifier
Swissprot Acc	Uniprot accession number
Trembl Acc	Trembl accession number
Ensembl Peptide	Ensembl peptide identifier
H-InvDB Protein IDs	Human Invitational Database Ids
InterPro	Interpro accession numbers
Gene Ontology Biological Process	As title
Gene Ontology Cellular Component	As title
Gene Ontology Molecular function	As title
Type	Sequence type e.g. protein coding, regulatory etc.
Taxonomic Name	Taxonomic Name
Chromosome	Chromosome sequence maps to
Mapping	Positional mapping on chromosome
HGNC	Human Genome Nomenclature Committee identifier
OMIM	Online Mendelian Inheritance in Man identifier
Gene Symbol	Official Gene Symbol
Gene Name	Official Gene Name
Alias Gene Symbol	Gene Symbol alias
Alias Gene Name	Gene Name alias
EC	Enzyme Commission number
CDD	Conserved Domain Database identifier
CCDS	Consensus CDS accession
Annotation	Textual summary information
Interactions	Interactions as described at NCBI
Reactome Interactions	Interactions from the reactome database

Breakdown of Array Contents

Transcript Sources and Numbers

Almac Diagnostics aim to provide as complete as possible a representation of the transcriptome of the tissue under study. Consequently, the sequences utilised in the production of our research arrays are garnered from a range of sources. The major array content sub-groupings are as follows:

Public Contigs

Contiguous sequences generated from the assembly of publicly available disease-specific ESTs.

In-house Contigs

Contiguous sequences created from the assembly of sequences generated from Almac Diagnostics' disease-specific EST sequencing projects.

Public/In-house Contigs

Contiguous sequences created by assembling singlets that failed to cluster during the Public and In-House EST assemblies.

Annotated Transcripts Representative of Singlets

Transcripts derived from the RefSeq database. These sequences are representative of public or in-house singlet sequences that failed to form contigs at any stage of the assembly pipeline.

Public Database Sequences Representative of Singlets

These public database sequences are representative of public or in-house singlet sequences that failed to form contigs at any stage of the assembly pipeline and which were not represented by an annotated transcript in the RefSeq database.

Public Database Sequences Experimentally Determined to be of Interest

These are sequences obtained from public databases that have been demonstrated by laboratory experimentation to be of interest in the disease being studied.

Reverse Complement Sequences

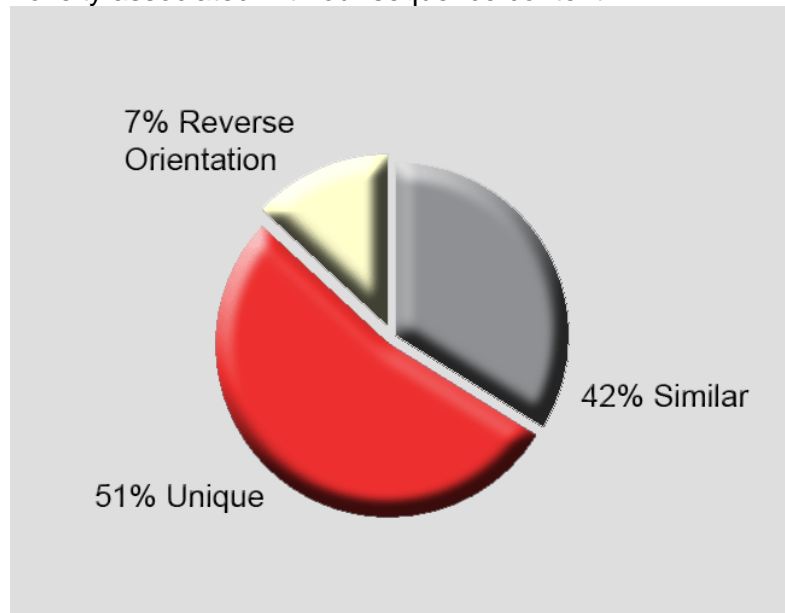
Select sequences from groups 1,2 & 3 exist in both standard and reverse-complemented forms. Where contiguous sequences generated in one orientation by our assembly processes have been shown to exist in an alternative (reverse) orientation in a public database, both forms have been included as **DSA™** content.

DSA™ Research Tool Content Analysis

DSA™ Research Tool Content vs Public Databases.

Almac Diagnostics' practice of generating sequences from in-house sequencing and assembly processes enables us to include transcripts that have not been previously shown to exist, or which show significant difference to previously reported forms. This allows us to detect transcripts undetectable by generic microarray products. By analysing DSA™ research tool content against publicly available human sequence databases, we can calculate the degree of novelty associated with our sequence content.

Figure 2: Contents of the Colorectal Cancer DSA™ research tool compared with the RefSeq database

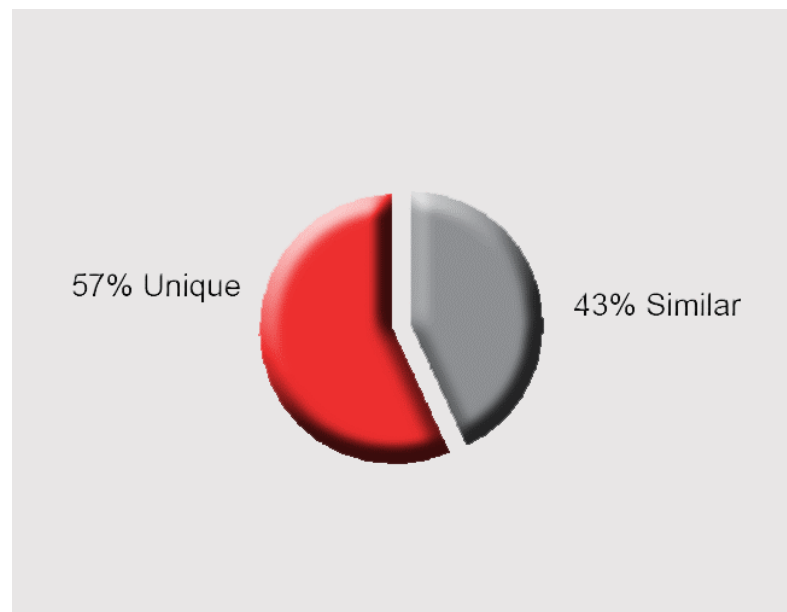


Microarray Comparisons

Comparison of DSA™ research tool content with Leading Generic Microarrays

Generic microarrays tend to focus exclusively on publicly available or extensively characterised sequences for their content, claiming “complete genomic coverage”. This method of sequence selection ignores the levels of variation between genome and transcriptome and lacks the focus of the DSA™ research tool approach. Thus a large number of potentially key transcripts for a given disease are rendered undetectable. Probe-level analysis of DSA™ research tools compared with leading generic microarrays highlights the abundance of data that is lost when generic products are utilised for focused studies.

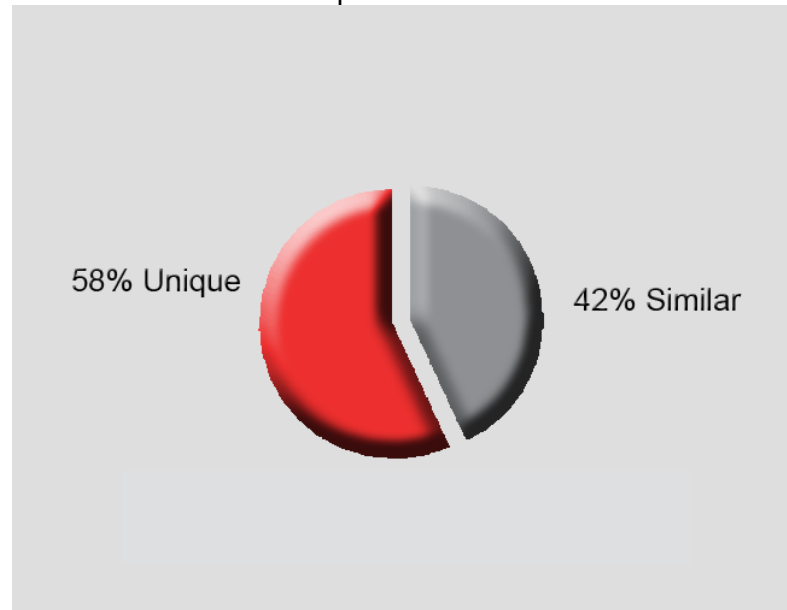
Figure 3: The Breast Cancer DSA™ tool compared with the content of a leading generic microarray.



Comparison of *DSA*TM Research Tools With Each Other

A belief in the *DSA*TM approach to microarray data analysis clearly implies a belief that the transcript pools produced by any two diseased tissues will be significantly different in their constitution. Similarly then, the content of two different *DSA*TM tools will reflect this difference. This inter-chip content analysis provides an indication of the disparity between disease transcriptomes.

Figure 4:
Comparison of the
Breast Cancer
***DSA*TM and**
Colorectal
Cancer *DSA*TM
tools.



Conclusion

Where research is focused on a disease setting, *DSA*TM research tools overcome the shortcomings of traditionally used generic microarrays. By tailoring the product content to the targeted disease transcriptome, extraneous elements are removed whilst critical information is maximised.

*DSA*TM research tools provide the most comprehensive means available for the study of the transcripts that really matter in a disease setting. To find out more about Almac Diagnostics and the *DSA*TM approach, visit <http://www.AlmacGroup.com>.

*DSA*TM, *Colorectal Cancer DSA*TM, *Breast Cancer DSA*TM are all trademarks owned by Almac Diagnostics Limited. The following Almac Diagnostics patents are pending in relation to the *DSA*TM technology referred to in this article: US 11/266/748, PCT EP05/011783, US 60/796,903.