

Lung Cancer DSA™: A platform for discovery of biomarkers in Lung cancer

Austin Tanney¹, Gavin R. Oliver^{1*}, Vadim Farztdinov¹, Richard D. Kennedy¹, Jude M. Mulligan¹, Ciaran E. Fulton¹, Susan M. Farragher¹, John K Field², Patrick G Johnston³, D Paul Harkin¹, Vitali Proutski¹, Karl A Mulligan¹.
¹Almac Diagnostics Ltd, 19 Seagoe Industrial Estate, Craigavon, BT63 5QD, UK. ²Roy Castle Lung Cancer Research Programme, The University of Liverpool Cancer Research Centre, 200 London Road, Liverpool, L3 9TA, UK.
³Centre for Cancer Research and Cell Biology, Queen's University of Belfast, 97 Lisburn Road, Belfast, BT9 7BL, UK. *E-mail: austin.tanney@almacgroup.com

Introduction



Non-small cell lung cancer (NSCLC) is the leading cause of cancer mortality worldwide but no reliable biomarkers are available to guide its management. Microarray technology may allow biomarkers to be identified but present platforms lack disease focus and are likely to miss potentially vital information in patient samples.

We have characterised the transcriptome of NSCLC and generated the first high-density disease specific transcriptome microarray. Built on the Affymetrix GeneChip® platform, the Lung Cancer DSA™ research tool allows for interrogation of ~60,000 transcripts relevant to NSCLC.

We present the design process and experiments demonstrating the array's utility.

Lung Cancer DSA™ unique content

Analysis of the content of the Lung Cancer DSA™ research tool against the 3 leading commercial human arrays (Figure 3) demonstrated significant unique content compared to each of the arrays.

Commercial Gene Expression Microarray	Lung Cancer DSA™ Research Tool Unique Transcripts
Affymetrix HG-U133 Plus2 Array	18635
Agilent Whole Human Genome Array	27777
Illumina Human 6 Array	31211
Comparison with content of 3 arrays combined	15541

Figure 3. Comparison of the Lung Cancer DSA™ content with the Affymetrix Plus2, Agilent Whole Human Genome and Illumina Human 6 arrays.

A core set of 15,541 Lung Cancer DSA™ research tool unique transcripts were not detectable by any of the 3 commercial arrays. Examination of these 15,541 transcripts (Figure 5) revealed that the large majority showed significant homology with major public sequence databases in sense and antisense orientation. 2% annotated only to the Human Genome while around 1% produced no alignment.

Technical assessment experiment

To assess the technical performance and biological relevance of the content of the Lung Cancer DSA™ research tool, five technical replicates of two RNAs extracted from a single patient matched normal and NSCLC frozen tissue were profiled on the arrays. Data generated by the Lung Cancer DSA™ was shown to be highly reproducible with extremely good correlation and low coefficient of variance (Figure 4).

Normal Tissue		Tumour Tissue	
Coefficient of Variance	Correlation	Coefficient of Variance	Correlation
5.2%	98.59%	5.3%	98.70%

Figure 4. Table showing the coefficients of variation and the subgroup average correlation coefficients calculated for the Lung Cancer DSA™ in the technical assessment experiment.

Detection of Lung Cancer DSA™ unique content

The results demonstrated that 35,625 transcripts were consistently detected by the Lung Cancer DSA™ research tool as being expressed in either the normal or tumor tissue. A significant number of the Lung Cancer DSA™ unique transcripts in both sense and antisense orientation were represented within this grouping. (Figure 5). These included antisense transcripts to well known genes previously implicated in cancer (Figure 6). 3,437 (55%) of the RefSeq sense-antisense (SA) transcript pairs were detected as being expressed in either the normal or tumor tissue.

Differential expression of Lung Cancer DSA™ unique content

Comparing the transcript expression levels between the normal and NSCLC lung tissue identified 2,148 of the unique transcripts as being differentially expressed and thus potentially important to the underlying biology of this disease (Figure 5).

Annotation Database	# Unique Transcripts	# Detected Unique Transcripts	# Differentially Expressed Unique Transcripts
RefSeq	647	155	43
Antisense to RefSeq	6157	3377	1301
EMBL	3302	982	274
Antisense to EMBL	2420	852	256
Unigene	2514	1021	306
Genome	353	9	4
Unannotated	148	2	0

Figure 5. Experimental detection of the Lung Cancer DSA™ unique content in matched normal/tumor lung tissue.

Target Accession Number	Probe ID	Database	Orientation	Gene Symbol	Fold Change	P value
NM_000546	LCHPRC.1183_s_at	RefSeq	Antisense	TP53	2.12	0.000147
NM_000059	LCHPRC.7_at	RefSeq	Antisense	BRCA2	2.73	0.000056
NM_130398	LCMXR.3025C1_at	RefSeq	Antisense	EXO1	4.22	0.000566
NM_078468	LC3P.8284C2_at	RefSeq	Antisense	BCCIP	1.80	0.00001
NM_002129	LCSS.2843_at	RefSeq	Antisense	HMG2	1.66	0.000061

Figure 6. Detection of probesets representing antisense to genes previously implicated in cancer by the Lung Cancer DSA™ research tool.

Gene ontology mining of Lung Cancer DSA™ unique content

To further investigate the relevance of the detected and differentially expressed unique content to the biology of NSCLC, the annotation associated with these transcripts was assessed by Gene Ontology mining. This clearly demonstrated association of these transcripts with the main cellular processes linked to cancer (Figure 7).

GO Process	Unique Detected Transcripts	Unique Detected RefSeq Antisense Transcripts	Unique Differentially Expressed Transcripts	Unique Differentially Expressed RefSeq Antisense Transcripts
Angiogenesis	26	18	13	9
Apoptosis	311	186	112	77
DNA Repair	119	65	32	21
Cell Migration	71	34	27	17
Proliferation	287	186	117	85
Immunology/Inflammation	196	118	83	55
Developmental Genes	12	8	10	7
Cell Cycle Control	384	239	139	111
Cell Signaling Pathway	55	40	33	25
Other	3851	2270	1382	910
Unknown	2461	820	770	293

Figure 7. Gene Ontology mining of the experimentally detected unique Lung Cancer DSA™ transcript annotation.

Comparison of the Lung Cancer DSA™ with other DSA™ research tools

This methodology serves as a template for the range of disease transcriptome focused microarrays we are developing. While defining content of the arrays we identified considerable differences between the transcriptomes of diseases (Figure 8). These differences emphasise both the importance and advantage of the transcriptome based approach in developing and using genomics tools.

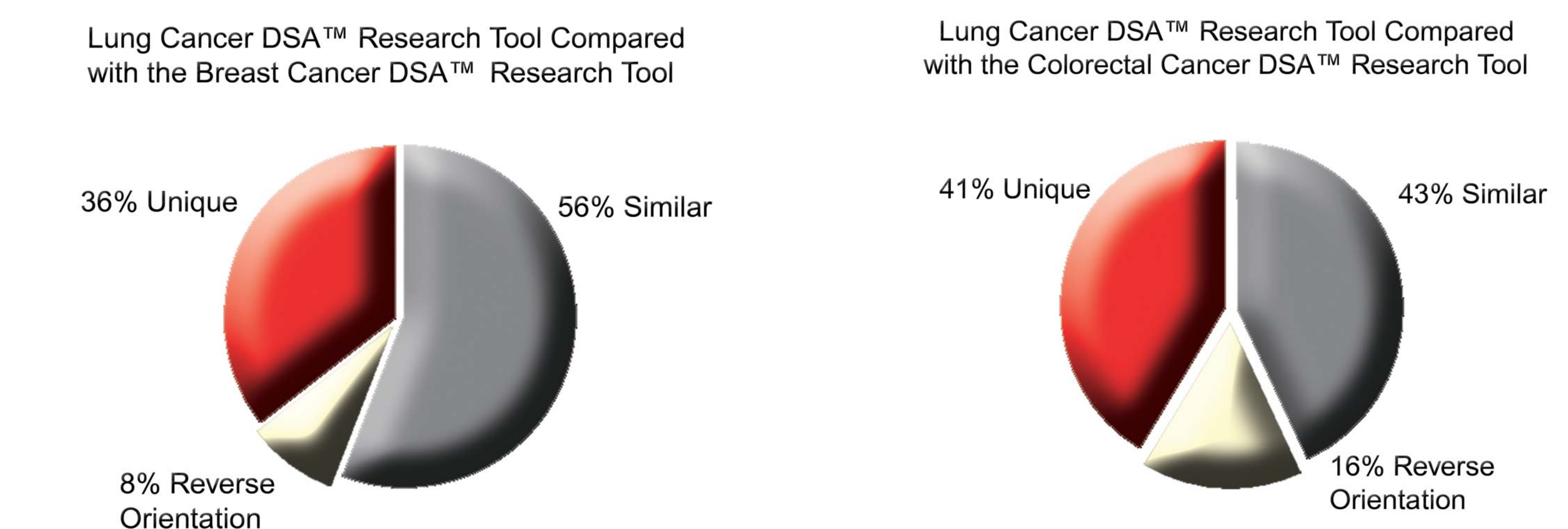


Figure 8. Comparison of the Lung Cancer DSA™ with other DSA™ research tools

Conclusions

- Generic microarray platforms lack the focus required for comprehensive disease research and discovery of biomarkers.
- The Lung Cancer DSA™ research tool addresses these shortcomings and provides a focused, comprehensive platform for NSCLC investigations.
- Based on the gold standard Affymetrix GeneChip® platform, the Lung Cancer DSA™ research tool provides reliable and reproducible data.
- The Lung Cancer DSA™ research tool detects many biologically relevant transcripts that are undetectable by the leading generic arrays.
- With ~60,000 transcripts the Lung Cancer DSA™ research tool provides the ideal platform for a range of applications from target and biomarker discovery to clinical diagnostics.

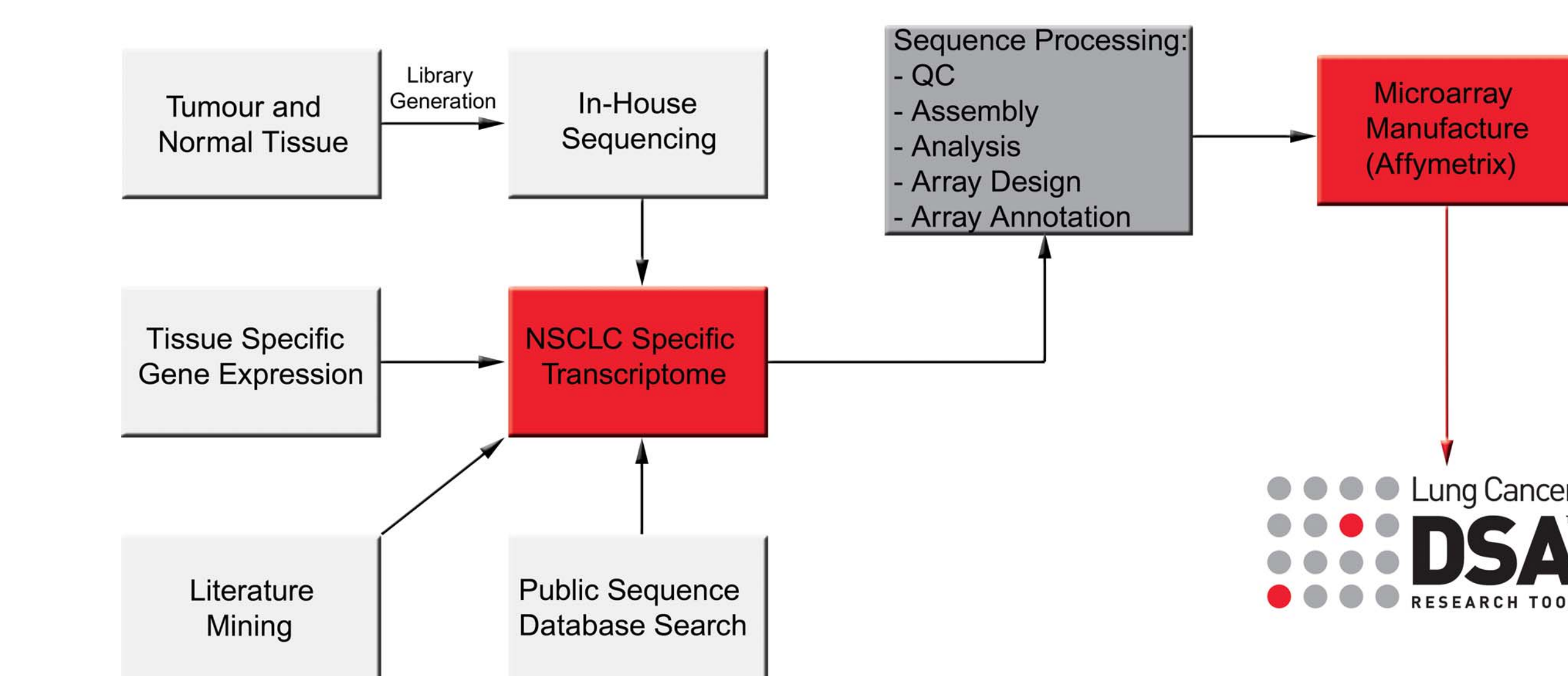


Figure 1. Schematic representation of the generation of the Lung Cancer DSA™ research tool

Generation of the Lung Cancer DSA™ research tool

The NSCLC transcriptome was characterised using three main sources of information: (i) in-house and publicly available ESTs from NSCLC and normal lung tissue (ii) gene expression data from in-house and public normal and NSCLC microarray profiles and (iii) scientific literature mining (Figure 1). Measures were taken to ensure 3 completeness of sequences and identify transcript variants. Final sequence data was submitted to Affymetrix who manufactured the array based on their GeneChip® platform.

Sequence content of the Lung Cancer DSA™ research tool

The Lung Cancer DSA™ research tool contains 59,927 probesets representing transcripts expressed in lung tissue. Sources from which the final sequences were derived are shown in Figure 2A. Analysis of the array content demonstrated that 42% of transcripts on the Lung Cancer DSA™ research tool have no significant homology with sequences from NCBI's RefSeq database in either orientation and 13% represent sequences transcribed in antisense orientation to annotated RefSeq transcripts (Figure 2B). Further analysis identified a total of 6,206 sense-antisense pairs represented on the Lung Cancer DSA™ research tool.

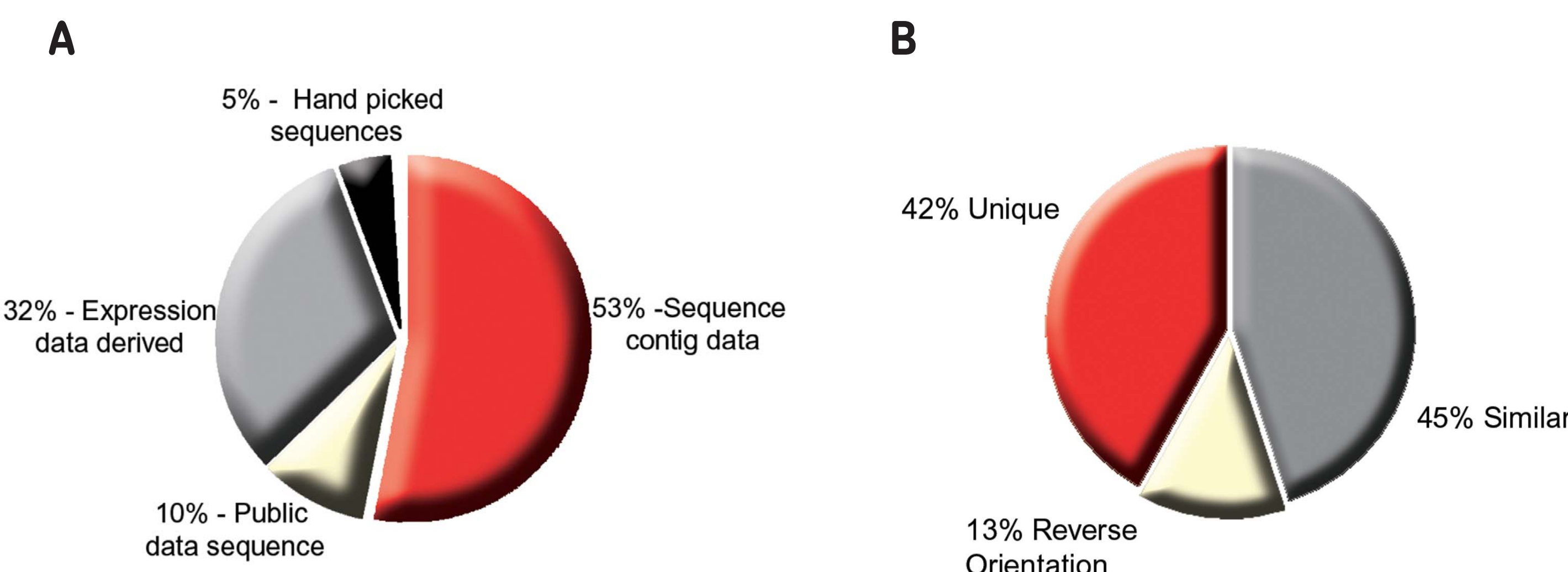


Figure 2 A) Origin of the content of the completed Lung Cancer DSA™ research tool. B) Comparison of the Lung Cancer DSA™ sequence content with the RefSeq mRNA database.