5 AN SIRNA SCREEN OF THE DRUGGABLE GENOME

The previous chapter presented a screen of kinases, phosphatases and associated genes for genes, that when knocked down, reduced the sensitivity of cells to TRAIL-induced apoptosis. This screen served as both a gene discovery experiment in its own right and as a pilot for larger screens. The results from this screen showed that while the methods developed were insufficiently sensitive to allow detection of the genes previously associated with sensitivity of cells to TRAIL-induced apoptosis other than the controls included on

each plate, novel genes connected to the regulation of the sensitivity of cells to TRAIL-induced apoptosis could be identified.

This chapter describes a screen of siRNAs targeting a further 6095 genes designated members of the "druggable genome", in order to identify further genes which play a role in regulating TRAIL-induced apoptosis. Like genes identified in the kinase and phosphatase screen, the involvement of genes targeted by siRNAs scoring highly in this screen was rigorously confirmed. An exploration of possible off target effects by examination of the seed sequences of highly scoring siRNAs is also described.

5.1 The druggable genome

Screening libraries of siRNAs targeting genome subsets, such as the kinase and phosphatase library screened in the previous chapter can be useful for identifying new genes in pathways. However, such screens are based on some hypothesis about genes likely to be involved in the process, and therefore risk missing genes in the pathway. Since these genes will be genes in unexpected gene families they are more likely to point to novel aspects of biology. The ideal solution is to screen libraries targeting each gene in the genome. However, whole genome libraries remain out of the reach of all but the largest research groups, pharmaceutical companies and specialised facilities. The cost of a whole genome screen is not just limited to the, already prohibitive, cost of the library itself, but also the cost executing the screen.

The first mention of the term 'Druggable Genome' in Medline is in 2002, in a review by Hopkins and Groom (Hopkins, Groom 2002) who use it to describe the set of genes containing protein domains which can bind small molecules (i.e. potential drugs), although Drews referred to a hypothetical set of proteins, related to disease genes, that could be targeted by pharmaceuticals in 2000 (Drews 2000). The number of genes classified as belonging to the druggable genome varies, with published estimates being 3,000-6,000 depending on the definition and the data set used. The druggable genome generally contains GPCRs, transcription factors, kinases, phosphatases, nucleotide binding proteins, proteases and more. Thus as well as the protein products of such genes being of interest to the pharmaceutical industry as possible drug targets, the druggable genome also contains many of the information carrying and processing gene families in the genome. This makes the druggable genome an attractive choice for RNAi screening – it is small enough for purchase and use to be with in the reach of a single academic group, yet contains many of the genes which could be of interest, and genes identified in this manner may be of therapeutic value.

As such, it represents a good compromise between an unbiased and a more targeted approach to screening.

The Qiagen Druggable Genome siRNA Set v2 contains siRNAs targeting 6,992 genes classified as being of "therapeutic value". This includes the 897 kinases and phosphatases screened in the previous chapter. The composition of the library by protein family is shown in Figure 5.1.

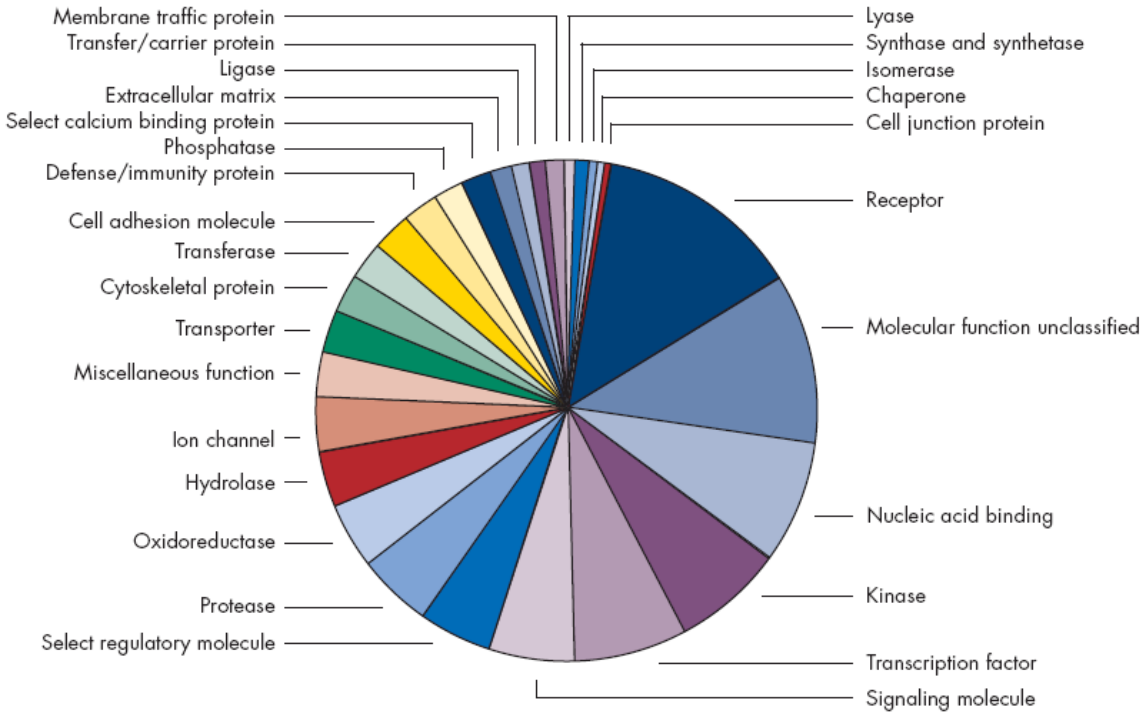

Figure 5.1 Composition of the Qiagen Human Druggable genome siRNA Set V2.0 by gene family
From Qiagen publicity material

5.2 Screen execution and initial data processing

In order to screen the library for siRNAs which affected sensitivity to TRAIL-induced cytotoxicity, siRNAs targeting 6095 genes (the genes screened in the previous chapter were not repeated) with 2 siRNAs per genes, from plates 1a – 77b of the Qiagen Druggable Genome Library v2, along with control siRNAs were transfected into HeLa cells in batches of 12 plates and the sensitivity of cells to 0.5µg/ml TRAIL determined. In assay development experiments, experiments were conducted using 1µg/ml TRAIL. Blind pseudo-screening using siRNAs gave a Z'-factor of 0.46 (Figure 3.12). In the screen reported in the previous screen comparison of the negative control with wells transfected with siCasp8 gave a Z'-score of -0.35 on a screen-wide basis. Despite the fact that this screen-wide score is calculated on the basis of plate-normalised values, it is still possible that plate-to-plate variation makes up some of this difference, since assay development pseudo-screens were

carried out on a single plate. While two wells per control are strictly insufficient to derive plate-by-plate Z'-factors, doing so gives a mean Z'-factor of 0.015 (median 0.22), showing a decrease even considering results within single plates. Except for the increase in the number of plates involved, the other difference between assay development experiments was a reduction in the concentration of TRAIL used from 1μg/ml to 0.25 μg/ml. In an attempt to counteract this decrease, the concentration of TRAIL used in this screen was increased to 0.5 μg/ml. In dosage curves of the effect of TRAIL on cells transfected with non-silencing siRNA, 0.5μg/ml TRAIL had a similar effect on survival to 1μg/ml (29% survival with 1μg/ml TRAIL compared to 31% survival with 0.5μg/ml TRAIL and 37% survival with 0.25μg/ml, Figure 3.11). The layout of siRNAs on the plate used was as described in Figure 4.1. The screen was initially carried out in duplicate. After both replicates were complete plate dynamic ranges were calculated as the ratio of the geometric mean of survival in siCasp8 control transfected wells to the geometric mean of survival in negative control transfected wells. Plates with a dynamic range of less than 2 were repeated. The repeated plate dynamic range was compared to the dynamic ranges of the two original replicates. The two of the three replicates with the largest dynamic range were used. In total each replicate of the screen took approximately 3 weeks to complete. The resulting 58,136 data points were analysed using the R/Bioconductor package cellHTS

It was previously observed that sensitivity of HeLa cells to TRAIL is dependent on the density of cells. Variation in density of cells could be due to two factors. Firstly variation could be due to inaccuracies in dispensing cells into the assay plates. Secondly variation in density of cells at time of treatment could be due to effects of particular siRNAs on the viability of cells. In order to investigate the relationship between pre-treatment viability and sensitivity of cells to TRAIL-induced cytotoxicity, pre-treatment viability was normalized to the plate median pre-treatment viability to account for plate to plate variations and the normalized viabilities divided into 20 quantiles. Post-treatment survivals were normalised to the plate-median survivals, and the median normalized survival for each viability quantile was calculated. Figure 5.2 shows the relationship between pre-treatment viability and post treatment normalized survival (blue line). There is no strong relationship between viability and survival when considering the higher viability quantiles. However, at lower viabilities there is a strong relationship between pre-treatment viability and sensitivity to TRAIL-induced cytotoxicity: cells in wells with lower pre-treatment viabilities are more sensitive to TRAIL-induced cytotoxicity. Since data is normalised on a per-plate basis, with the assumption that the median survival on the plate represents an estimation of base-line

sensitivity to TRAIL, wells where TRAIL sensitivity is high due to a low pre-treatment viability can affect plate normalisation factors. In order to prevent the observed trend affecting further analysis for this reason, wells in the bottom 20% for pre-treatment viability were removed from further analysis. Applying this cut-off entirely negated the relationship between pre-treatment viability and sensitivity to TRAIL-induced cytotoxicity (Figure 5.2, red line).

Data from each plate was normalized to the median of the survival in wells

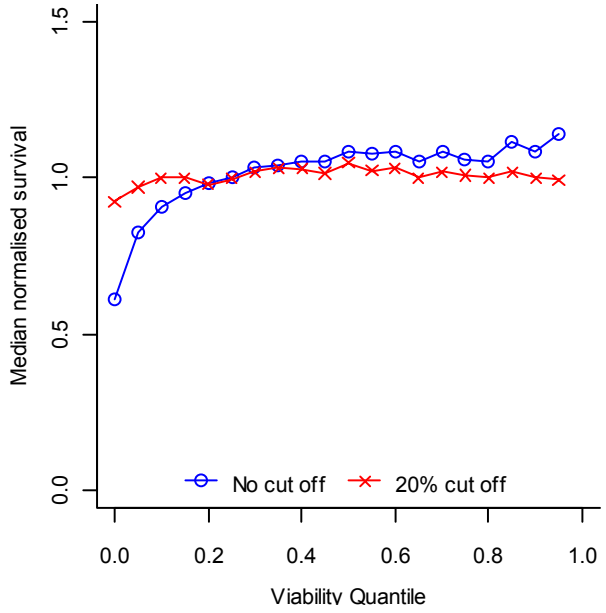

Figure 5.2 Relationship between pre-treatment viability and sensitivity to TRAIL-induced cytotoxicity
Viability of cells in each well prior to treatment was normalized to plate median viabilities. Normalised viabilities were divided into 20 quantiles. The median normalized post treatment survival was calculated for wells in each of these quantiles. The blue line represents the relationship between pre-treatment viability and post-treatment normalized survival for the entire data set. The red line represents the same relationship with the 20% of wells with the lowest pre-treatment viabilities are removed from the analysis.

containing "sample" siRNAs for each plate. Figure 5.3 shows effects of this normalisation. The raw data from the screen is very variable (Figure 5.3a), with the minimal survival value on some plates being higher than the maximum on other plates. Plate-to-plate variation is probably higher in this screen compared to the kinase and phosphatase screen due to the higher number of different batches of cells required to complete the screen, with each replicate of the kinase and phosphatase screen being completed with a single batch of cells, while each replicate of the screen presented here required several independent batches of cells, with for example, replicate 2 of the screen using cells from 10 flaks, defrosted from liquid nitrogen on 3 separate dates.

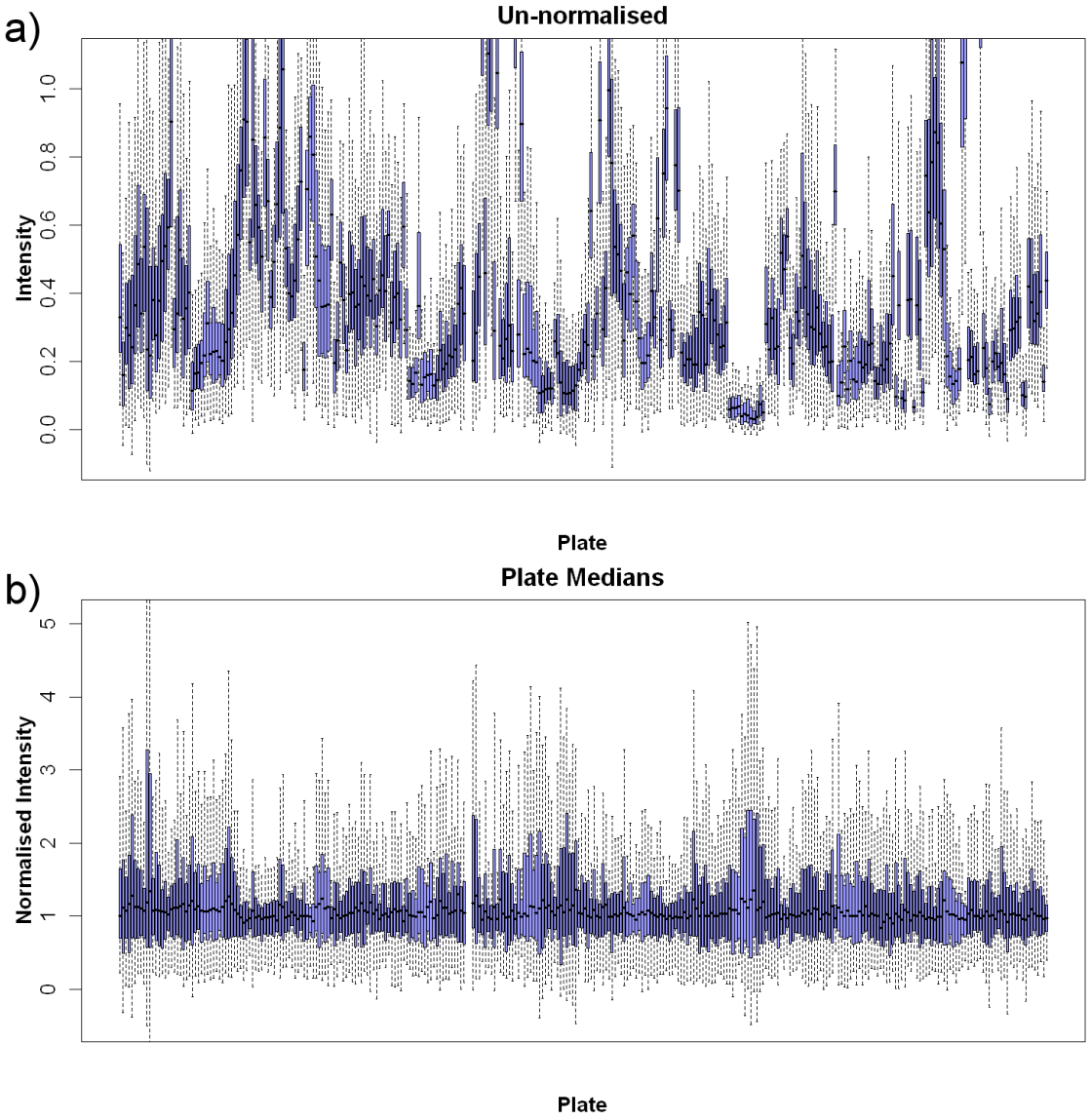

Figure 5.3 Normalisation of data from screen of druggable genome.
a) Box-plot of raw survival data from screen on a per plate basis. b) Box-plot of survival data normalized to plate median survival on a per plate basis

A correlation between effect size and variability can complicate interpretation of results. Such a correlation was observed in the data from the kinase and phosphatase screen. Various transformations, such as log transformation, can help to remove such correlations. The rank of the mean normalized survival for the replicates of each well was plotted against the standard deviation in order to examine a possible relationship between the two (Figure 5.4). A clear relationship between the mean and standard deviation can be observed in non-transformed data (Figure 5.4a): an increase in mean is accompanied with a large increase in standard deviation at the higher ranks. A relationship can also be observed between rank of mean and standard deviation for log-transformed data (Figure 5.4b): standard deviation is higher at lower ranks. However, the relationship does not appear to be as strong in log

transformed data. Observing the running median line, shown in red, indicates a stronger relationship, over a larger portion of the results for non-transformed data. On this basis, further analyses were conducted using log transformed values.

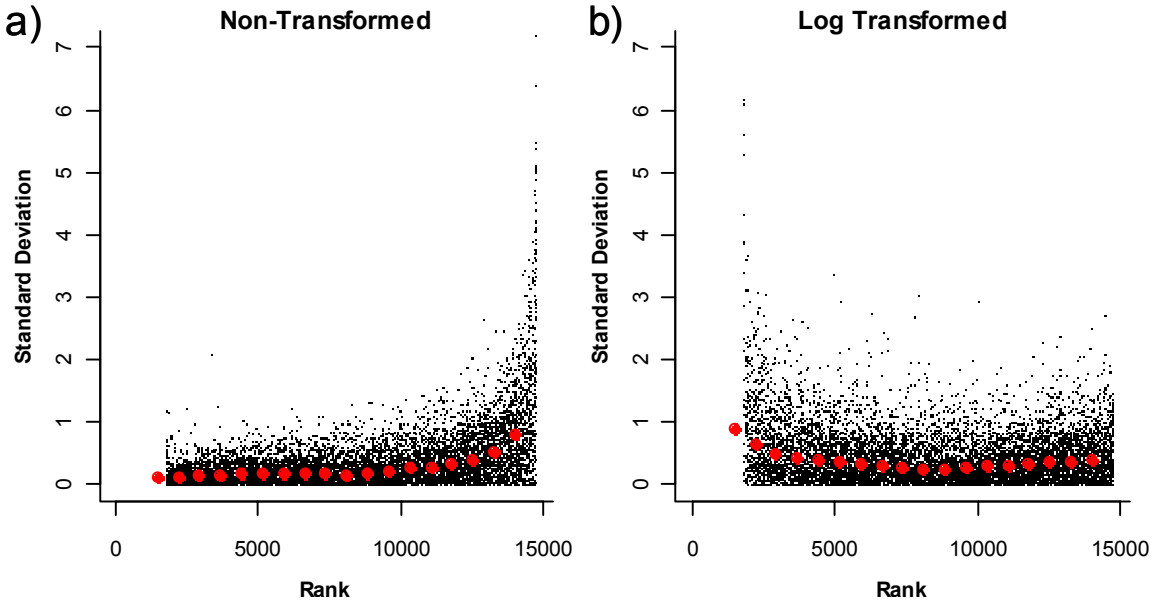

Figure 5.4 Relationship between standard deviation and rank of the mean for siRNAs.
For each siRNA the mean of the normalized data of the two replicates was calculated. The rank of this mean was then plotted against the standard deviation between the replicates for a) Non-transformed data and b) Log-transformed data. The red line in each plot represents the running median standard deviation.

Data were analysed using cellHTS, first excluding wells with a low pre-treatment viability, and then median normalizing plates with a log transformation, and using the minimum of replicates as a summary function. The HTML reports produced can be found on the included CD, or online at http://www.sanger.ac.uk/HGP/Chr22/RNAi/TRAIL_DG

5.3 Screen quality and analysis of controls

The processed screening data was used to assess the quality of the screen (Figure 5.5). The Pearson's correlation co-efficient, r, between the two replicates was 0.57 (Figure 5.5a), similar to that found in the kinase and phosphatase screen (0.65). As in the kinase and phosphatase screen the correlation between the two siRNAs targeting the same gene was dramatically lower at r = 0.075 (Figure 5.5b). Once again this demonstrates that the effect of a particular siRNA on the sensitivity of cells to TRAIL-induced cytotoxicity is fairly reproducible, while, the targeting of the same gene with different siRNAs does not give a reproducible effect.

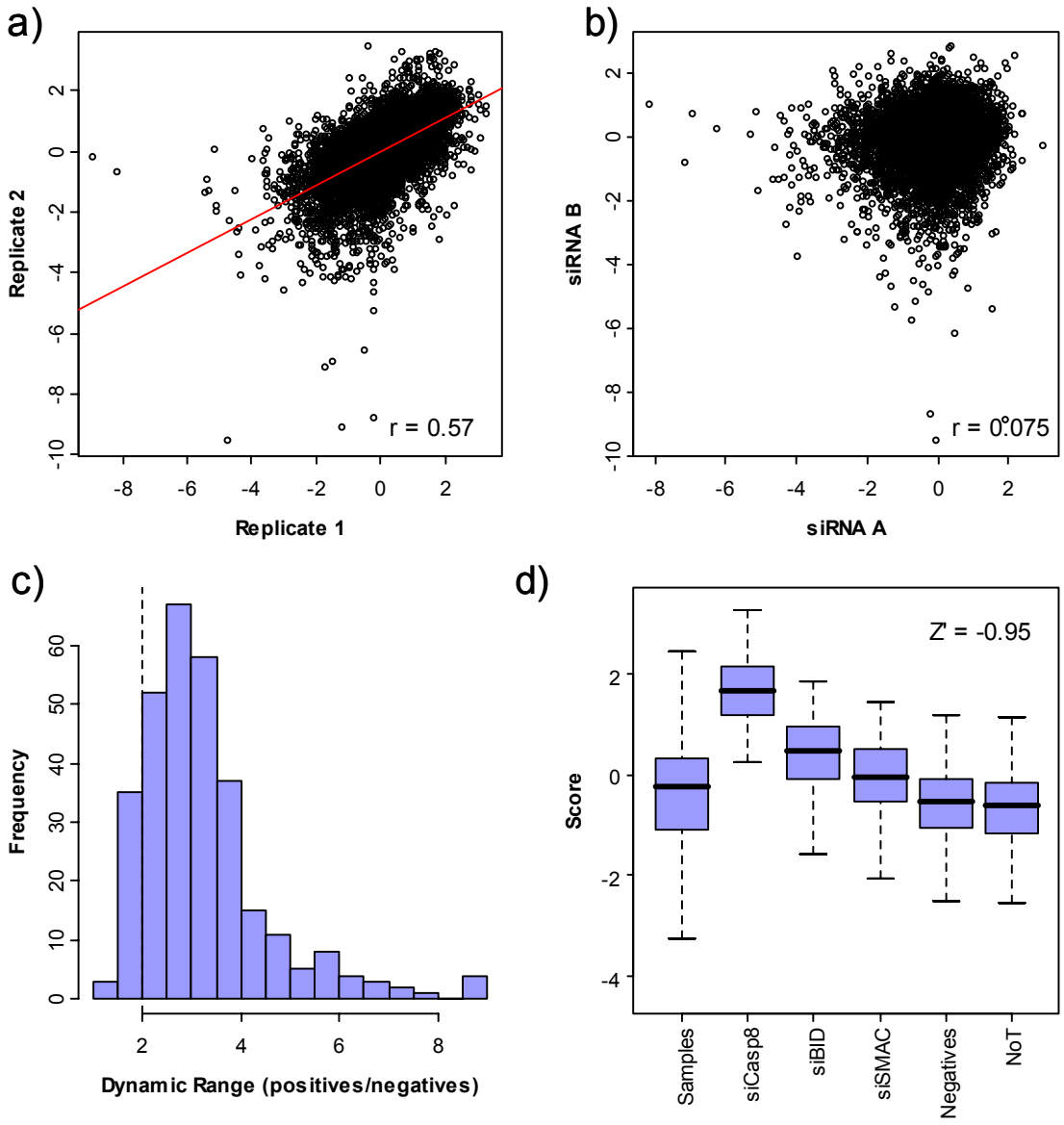

Figure 5.5 Assessment of screen quality and controls.
a) Normalized survival from replicate 1 plotted against normalised survival from replicate 2. Red line shows linear regression of replicate 2 on replicate 1. The Pearson's correlation co-efficient is shown in the bottom right corner. b) Plot showing normalised survival of the two siRNAs targeting the same gene. The Pearson's correlation co-efficient is shown in the bottom right corner. c) Histogram showing the distribution of plate dynamic ranges (see above for definition of plate dynamic range). Dashed line represents a dynamic range of 2. d) Box plot summarising the scores in different well types. Z'-factor between negative controls and siCasp8 is shown in top right corner. NoT = Untransfected

Figure 5.5c shows the distribution of plate dynamic ranges (the ratio of the geometric mean of survival of siCasp8 transfected wells and geometric mean of survival of negative control transfected wells). Despite increasing the concentration of TRAIL for this screen, the majority of plate dynamic ranges are still between 2 and 4 (70%, Figure 5.5c), with a minority having a dynamic range less than 2 (12%) and greater than 4 (18%).

The distribution of scores for different well types is shown in Figure 5.5d and Table 5-1. The median survivals, and consequently the median scores are higher for all of the positive controls than the negative controls. The difference in raw survival between siCasp8

Category	*Median Score*	*Median Survival*	*Survival MAD*
Samples	-0.23	28.2%	23.5%
siCasp8	1.68	80.5%	40.4%
siBID	0.48	48.4%	28.6%
siSMAC	-0.03	37.1%	21.0%
Negatives	-0.54	27.3%	18.1%
Untransfected	-0.59	27.9%	19.2%

Table 5-1 Summary statistics for different well types in screen of druggable genome

and the negative controls is larger than in the kinase and phosphatase screen. However the difference in score is much lower, due, both to the greater variance in this screen, and the log transformation of the data. Wells transfected with siBID and siSMAC are intermediate between siCasp8 and the negative controls. The Z'-factors for the difference between siCasp8, siBID, siSMAC and the negative controls are -0.95, -3.44 and -7.52 respectively. Again the increase in throughput has lead to a reduction in the Z' factors, due to the increase in the variability. This indicates that there would be little chance of finding hit with effects smaller than those of BID or SMAC, and a reduced chance of finding even hits with an effect size similar to Caspase-8.

The scores and median survivals of untransfected wells (-0.59 and 27.9% respectively) are remarkably similar to those for the negative controls (-0.54 and 27.3% respectively) and the spread of values is similar. The median score for sample wells is slightly higher than those for the negative controls (in contrast to the results of the kinase and phosphatase screen, where the negative controls had a higher median score), but the difference in terms of median survival is very small.

5.4 Screen Results

In the screen of kinases and phosphatase the scores from sample wells formed a distribution with an elongated left-hand tail and a foreshortened right-hand tail. Here the distribution of scores is closer to normal, but is skewed in the opposite direction due to the log transformation of the values (Figure 5.6). One interpretation of this would be that there are many siRNAs which are causing an increase in the sensitivity of cells to TRAIL-induced apoptosis. Under this hypothesis one would expect that if the results were analysed to find such siRNAs, that is, wells with a low proportion of cells surviving after TRAIL treatment were given a high score, that the distribution would be reversed, with a long tail of highly positive siRNAs. However, this is not the case with the resulting distribution of scores having a similar shape to the distribution seen here (data not shown). This suggests that the skew in the distribution of scores is an artefact of the analysis process.

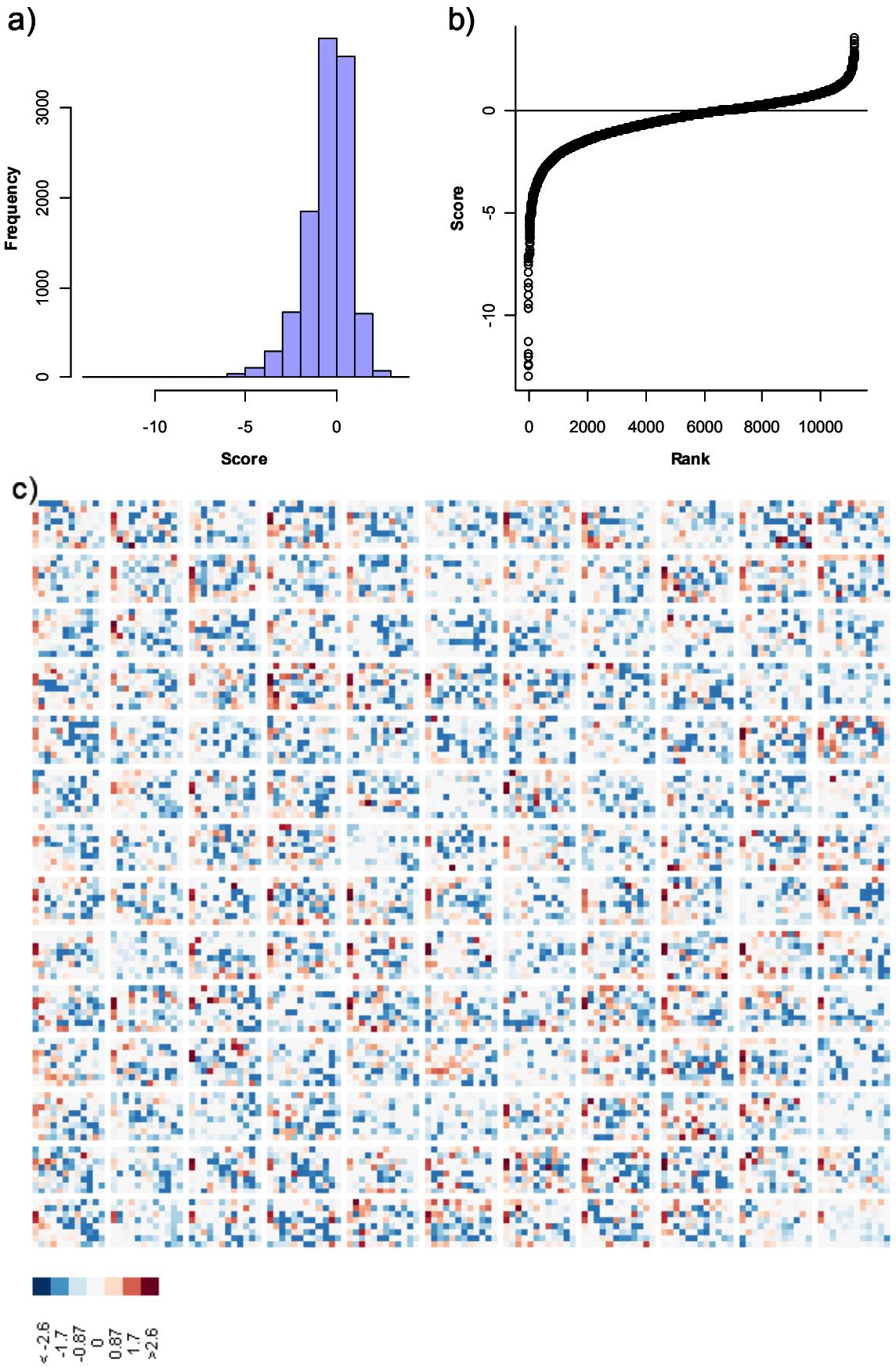

Figure 5.6 Results of siRNA screen of the druggable genome
a) Histogram of scores from sample wells b) Rank of siRNA score from sample wells plotted against score. c) Heat map of scores per plate. siRNAs with a highly positive score are shown in red, siRNAs with a highly negative scores are shown in blue. Plates are arranged row-wise.

GeneID	*Symbol*	*Description*	*Normalized Survival*		***score***
			Rep 1	*Rep 2*	
NM_003217	TEGT	Testis enhanced gene transcript (BAX inhibitor 1)	NA	2.572	**3.51**
NM_016368	ISYNA1	myo-inositol 1-phosphate synthase A1	2.468	NA	**3.43**
NM_001240	CCNT1	cyclin T1	2.361	NA	**3.28**
NM_002197	ACO1	aconitase 1, soluble	2.844	2.279	**3.11**
NM_003947	HAPIP	huntingtin-associated protein interacting protein (duo)	2.074	NA	**2.88**
NM_002337	LRPAP1	Low density lipoprotein receptor-related protein associated protein 1	2.018	NA	**2.8**
NM_005799	INADL	InaD-like (Drosophila)	2.03	2.026	**2.76**
NM_003554	OR1E2	olfactory receptor, family 1, subfamily E, member 2	1.962	NA	**2.72**
NM_013345	GPR132	G protein-coupled receptor 132	2.755	1.961	**2.68**
NM_004584	RAD9A	RAD9 homolog A (S. pombe)	NA	1.943	**2.65**
NM_003266	TLR4	Toll-like receptor 4	1.975	1.934	**2.64**
NM_000674	ADORA1	adenosine A1 receptor	1.893	NA	**2.63**
NM_002382	MAX	MAX protein	2.416	1.883	**2.57**
NM_006986	MAGED1	melanoma antigen, family D, 1	1.843	2.424	**2.56**
NM_000875	IGF1R	insulin-like growth factor 1 receptor	1.838	NA	**2.55**
NM_016953	PDE11A	phosphodiesterase 11A	1.865	1.865	**2.54**
NM_014379	KCNV1	potassium channel, subfamily V, member 1	1.832	1.911	**2.54**
NM_080674	C20orf86	chromosome 20 open reading frame 86	2.151	1.847	**2.52**
NM_002467	MYC	v-myc myelocytomatosis viral oncogene homolog (avian)	NA	1.842	**2.51**
XM_497793	LOC402037	similar to alpha tubulin	1.815	1.838	**2.51**
XM_089281	LOC149281	similar to RIKEN cDNA 2610205E22	NA	1.817	**2.48**
NM_001962	EFNA5	ephrin-A5	2.116	1.812	**2.47**
NM_138295	PKD1L1	polycystic kidney disease 1 like 1	1.754	1.881	**2.43**
NM_017522	LRP8	low density lipoprotein receptor-related protein 8, apolipoprotein e receptor	1.724	NA	**2.39**

Table 5-2 Top scoring siRNAs from screen of the druggable genome
Table shows the top 24 siRNAs ranked by score from the screen. NA indicates that result was removed due to low pre-treatment viability. The complete table is available in the file topTable.txt on the included CD or online at http://www.sanger.ac.uk/HGP/Chr22/RNAi/TRAIL_DG

Gene ID	Gene Symbol	Description	Score		Minimum Score	Maximum Score	Mean Score
			siRNA 1	siRNA 2			
NM_018558	GABRQ	gamma-aminobutyric acid (GABA) receptor, theta	1.59	1.45	1.45	1.59	1.52
NM_016368	ISYNA1	myo-inositol 1-phosphate synthase A1	1.81	1.45	1.45	1.81	1.63
NM_002357	MAD	MAX dimerization protein 1	1.61	1.43	1.43	1.61	1.52
NM_000875	IGF1R	Insulin-like growth factor 1 receptor	2.55	1.43	1.43	2.55	1.99
NM_054032	MRGX4	G protein-coupled receptor MRGX4	1.41	1.58	1.41	1.58	1.50
NM_003217	TEGT	Testis enhanced gene transcript (BAX inhibitor 1)	1.35	3.51	1.35	3.51	2.43
NM_001196	BID	BH3 interacting domain death agonist	1.33	2.01	1.33	2.01	1.67
NM_013231	FLRT2	fibronectin leucine rich transmembrane protein 2	1.78	1.33	1.33	1.78	1.55
NM_018337	ZNF444	zinc finger protein 444	1.31	2.07	1.31	2.07	1.69
NM_017949	CUEDC1	CUE domain containing 1	1.26	1.94	1.26	1.94	1.60
NM_002063	GLRA2	glycine receptor, alpha 2	1.23	1.25	1.23	1.25	1.24
NM_001279	CIDEA	cell death-inducing DFFA-like effector a	1.42	1.22	1.22	1.42	1.32
NM_198150	DKFZp313G1735	hypothetical protein DKFZp313G1735	1.20	2.11	1.20	2.11	1.66
NM_018527	NARG1L	NMDA receptor regulated 1-like	1.27	1.20	1.20	1.27	1.24
NM_198857	FLJ43855	similar to sodium- and chloride-dependent creatine transporter	1.48	1.18	1.18	1.48	1.33
NM_006340	BAIAP2	BAI1-associated protein 2	1.18	1.31	1.18	1.31	1.25
XM_377955	ANKIB1	ankyrin repeat and IBR domain containing 1	1.17	1.60	1.17	1.60	1.39
NM_018319	TDP1	tyrosyl-DNA phosphodiesterase 1	1.48	1.17	1.17	1.48	1.33
NM_020919	ALS2	amyotrophic lateral sclerosis 2 (juvenile)	1.16	1.17	1.16	1.17	1.17
NM_004821	HAND1	Heart and neural crest derivatives expressed 1	1.16	1.38	1.16	1.38	1.27
NM_018110	DOK4	Docking protein 4	1.15	1.49	1.15	1.49	1.32
NM_005856	RAMP3	Receptor (calcitonin) activity modifying protein 3	1.33	1.12	1.12	1.33	1.23
NM_005252	FOS	v-fos FBJ murine osteosarcoma viral oncogene homolog	1.11	2.14	1.11	2.14	1.63
NM_019839	LTB4R2	leukotriene B4 receptor 2	1.37	1.09	1.09	1.37	1.23

Table 5-3 Extract from table summarising results of screen of the druggable genome on a per gene basis

Genes are ranked on basis of the minimum of the scores from the two siRNAs. Where no score for an siRNA can be determined as the normalized survival for each of the replicates is NA, due to low pre-treatment survival, the score for the siRNA is taken to be NA, and NA is ranked as low. Full table is available as perGene.tab on included CD or online at http://www.sanger.ac.uk/HGP/Chr22/RNAi/TRAIL_DG/perGene.tab

Examination of the rank/score plot (Figure 5.6b) shows that the distribution of scores from this screen is essentially continuous, with a large section where the rate of increase in score with rank is constant. However, this section covers a smaller proportion of the total range of scores than did the equivalent section of scores from the kinase and phosphatase screen (Figure 4.6).

Figure 5.6c shows the spatial distribution of scores within the library. There are no obvious plate position effects. High and low scores are evenly distributed between and within plates and there are no obvious signs of edge effects.

siRNAs were ranked by their score in the screen, and genes ranked by the minimum of the score of the two siRNAs targeting the gene. Portions of these rankings are shown in Table 5-2 and Table 5-3.

In analysing the results of the kinase and phosphatase screen a cut off was established for hit selection based on capturing 95% of the siCasp8 positive controls. An equivalent cut off score for this screen would be 0.51 and this would also capture 6.7% of the negative controls. However, given that using this cut off to select "hit" genes was not successful in the case of the kinase and phosphatase screen, such a cut off was not be used to select "hit" genes here.

5.4.1 Analysis of genes previously associated with the TRAIL pathway

Examination of the results from siRNAs targeting genes known to be involved in the TRAIL apoptosis pathway can be used to give an assessment of the sensitivity of the screen. Raw survivals, scores and rank in the list of siRNAs for genes previously associated with the TRAIL pathway are given in Table 5-4. Without defining a cut-off it is not possible to say how many of these genes were "hits". Four genes are targeted by siRNAs with a score greater than 0.51 (BID, Casp8, MYC and DR4), and one is targeted by two (BID), while only one of the siRNAs targeting DR4 has a score. BID, Casp8 and DR4 are the genes that gave the largest effect when knocked down in assay development experiments. However, as previously noted, using such a cut-off, based on the results from controls did not prove a successful way of identifying hits previously.

Gene Symbol	Survival		Score		Rank	
	siRNA 1	siRNA 2	siRNA 1	siRNA 2	siRNA 1	siRNA 2
BAX	24%	72%	-0.32	-0.49	5910	6537
BID	109%	126%	1.33	2.01	396	76
CASP3	32%	73%	-0.86	-0.15	7756	5209
CASP8	137%	98%	0.87	0.43	1119	2376
DVL2	61%	79%	0.04	0.32	4177	2813
FADD	16%	NA	-0.74	NA	7365	NA
FBXO11	92%	NA	0	NA	4437	NA
MYC	66%	107%	2.51	0.44	2347	19
TCF4	17%	68%	-1.72	0.33	9590	2785
TNFRSF10A (DR4)	NA	40%	NA	0.64	NA	1641
TNFRSF10B (DR5)	36%	16%	0.07	-1.15	3967	8504
VPS16	19%	50%	-0.95	-0.86	7872	7739
Median	36%	72%	0	0.325	4437	2799

Table 5-4 Survivals, Scores and Ranks of siRNAs targeting genes previously associated with the TRAIL pathway in the screen of the druggable genome

Gene Set Enrichment Analysis (GSEA) is a method for testing if genes in a pre-defined gene set are enriched in high scores in a ranked list of genes (Subramanian et al. 2005). Enrichment scores are calculated by scanning a ranked list of genes and increasing the 'running' enrichment scores each time a member of the gene set is encountered and decreasing the score each time a gene which is not a member of the gene set is encountered. The amount by which the score is increased on encountering a member of the gene set depends on the value of the metric used to rank the genes. The score for the gene set is the maximum enrichment score reached during the walk across the ranked list. This score is normalised for the size of the gene set, and the significance of the normalized enrichment score calculated by using permutations of the data. In this way the statistic is roughly equivalent of a weighted Kolmogorov–Smirnov like statistic. This technique is usually applied to microarray gene expression data, comprising several arrays measuring expression levels in phenotypically positive or negative samples. In this case the phenotype labels are permutated (so called column wise permutation) and the enrichment score for each gene set re-calculated to assess the significance of the normalized enrichment scores. However, the screening data used here is not suitable for this process, comprising only four samples (two treated, two untreated). In such a case, significance can be assessed by permuting the gene list (so called row wise permutation) and recalculating the enrichment scores for each gene set.

The p-values calculated this way are then usually corrected using either a false discovery rate (FDR) estimation, or a family-wise error rate (FWER) estimation. Row-wise permutation has the disadvantage that it may under-estimate the variance of the enrichment score by breaking up any correlation between genes in the gene set, and so generally the

more conservative FWER estimation of significance rather than the FDR estimation is used for p-values calculated using row-wise permutations (A. Liberzon, GSEA team, personal communication). This algorithm is implemented in the software package GSEA-P.

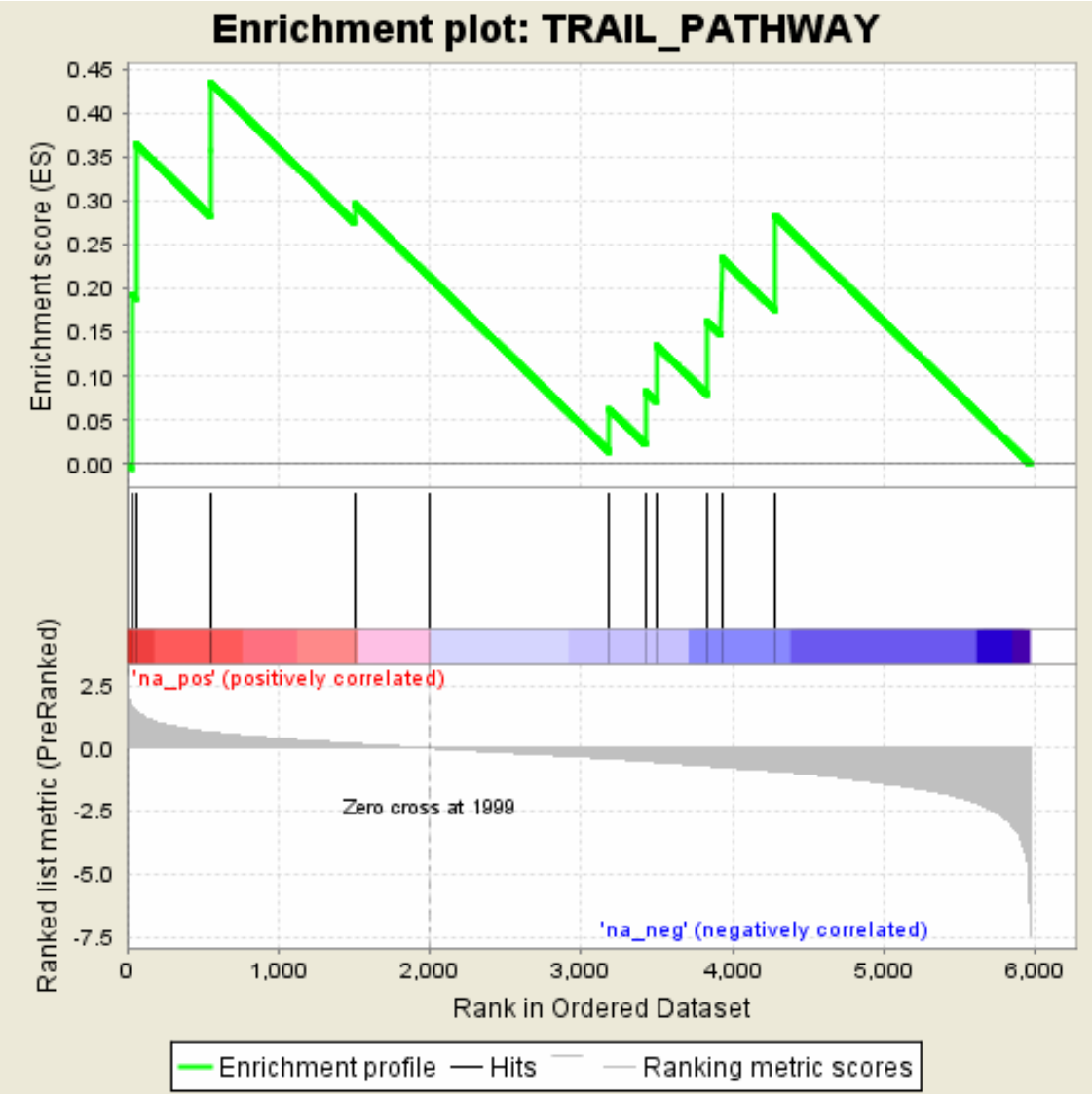

Figure 5.7 Enrichment plot for GSEA of genes previously associated with TRAIL pathway in the screen of the druggable genome

Top panel shows running enrichment score for genes previously associated with the TRAIL pathway across the ranked gene list from the druggable genome screen. The centre panel shows the position of the genes in the set in relation to the gene list and the lower panel shows the value of the ranking metric (mean score of siRNAs targeting gene).

The genes from the screen were ranked according to the average of the two siRNAs targeting the gene and GSEA-P was used to perform GSEA on the ranked list of screening results using the gene-set of genes previously associated with the TRAIL pathway as the set of genes for which enrichment is to be measured. Previously genes have been ranked on the basis of the minimum of the two siRNAs. In this case a conservative approach is taken to increase the confidence that identified genes are indeed involved in the process. Here, it is assumed that the genes are involved in the process, and so a more neutral summary of the

effect of knocking down the gene is used. The results of this analysis show that the set of previously associated genes is enriched at the top of the ranked list of genes (Figure 5.7). However, this result is only borderline significant with a nominal p-value of 0.083. The nominal p value is used rather than a corrected value since only one gene set was tested. Similar analysis of the kinase and phosphatase screen shows no enrichment for TRAIL pathway gene in the ranking of genes in that screen (data not shown).

The leading edge subset is the subset of genes that are higher in the list than the point at which the maximum enrichment score is reached and thus can be said to have contributed to the high score. In this case the leading edge subset consists of four genes: Casp8, BID, DR4 (TNFRSF10A) and MYC. This set of four genes includes three of the six "core-death pathway" genes in the set and it has been previously reported that knock-down of one of these six, DR5 (TNFRSF10B), is ineffective in preventing TRAIL-induced apoptosis in HeLa cells (Aza-Blanc et al. 2003). Of the six non-"core-death pathway" genes in the set 4 are non-confirmed hits from the Aza-blanc *et al* screen.

Thus, while it is not possible use the results here to define a sensitivity for the screen, the screen here can be shown to have been more sensitive for identifying genes previously associated with the TRAIL pathway than the kinase and phosphatase screen. This is probably at least in part due to the fact that the genes included in the screen described here from the TRAIL pathway are likely to induce a larger change in TRAIL sensitivity (e.g. Casp8, BID) compared with the TRAIL pathway genes present in the kinase and phosphatase screen and a smaller proportion of them are unconfirmed hits from the Aza-blanc screen. That is, this screen appears to be more sensitive because the sensitivity is being measured against a strong, better validated set of genes.

5.5 Confirmation of hits

In order to state that genes targeted by high scoring siRNAs are involved in TRAIL-induced apoptosis, it must be demonstrated that the effects are 1) reproducible, 2) specific and 3) related to TRAIL-induced apoptosis, rather than solely TRAIL-induced cytotoxicity. This requires that genes are targeted by multiple siRNAs that reproducibly affect TRAIL sensitivity (phenotypically active siRNAs), that these siRNAs knock-down the mRNA level of the targeted transcript to a greater extent than siRNAs which do not affect TRAIL sensitivity (phenotypically inactive siRNAs) and the activity of these siRNAs must also be reproduced in an assay that measures apoptosis rather than cytotoxicity.

Selecting genes based on a single high-scoring siRNA ultimately lead to the

identification of a larger number of confirmed high genes in the screen of kinases and phosphatases compared to selecting genes based on the score of both siRNAs targeting the gene. Therefore, here, the genes targeted by the 20 highest scoring siRNAs were selected for confirmation.

During confirmation of hits from the kinase and phosphatase screen, resources were wasted re-synthesising siRNAs which scored highly in the initial screen, but failed to repeat in confirmation experiments. Therefore, here both siRNAs targeting the selected genes from the library were retested as an initial filter. siRNAs were transfected into cells in triplicate and tested for sensitivity to 0.5μg/ml TRAIL using the alamarBlue assay Table 5-5. Transfection of 23 of the 42 tested siRNAs induced a reduction in sensitivity to TRAIL compared to transfection of the negative control which was significant at the 10% level (using a student's t-test on log transformed data, with p values corrected using the Hommel correction for multiple testing). This includes 71% of the siRNAs which were initially among the top 20 siRNAs (plus the top scoring MAD siRNA) and 38% of the second siRNAs targeting the same genes. In two cases the siRNA which was amongst the top 20 scoring siRNAs did not induce a significant change in TRAIL sensitivity in this test, while the second siRNA targeting the same gene did. Six genes were targeted by two siRNAs that significantly reduced the sensitivity to TRAIL.

At this point the four genes not targeted by any siRNA which significantly reduced the sensitivity to TRAIL-induced cytotoxicity were discarded. Where a gene was targeted by two siRNAs that significantly reduced the sensitivity to TRAIL induced cytotoxicity these two siRNAs were re-synthesised for use in future experiments. Where a gene was targeted by only one siRNA that significantly reduced the sensitivity of cells to TRAIL-induced cytotoxicity, this siRNA was re-synthesized, and in addition two further siRNAs targeting the gene were obtained.

Gene	*p-value siRNA A*	*p-value siRNA B*
TEGT	**0.01275**	**0.005579**
ISYNA1	0.498359	**0.0164**
CCNT1	0.286036	**0.002418**
ACO1	**0.006337**	**0.008985**
HAPIP	0.286036	**0.076998**
LRPAP1	**0.06912**	0.230341
INADL	0.498359	**0.081816**
OR1E2	0.498359	**0.097019**
GPR132	0.170937	**0.006844**
RAD9A	0.302649	**0.00775**
TLR4	0.241735	**0.002145**
ADORA1	0.261679	**0.054469**
MAX	**0.002621**	**0.01308**
MAGED1	0.923002	1
IGF1R	**0.054469**	**0.008152**
PDE11A	0.362945	**0.056278**
KCNV1	0.238429	0.498359
C20orf86	0.461725	0.462106
MYC	**0.010032**	**0.002418**
LOC402037	0.286036	0.286036
MAD	**0.006964**	**0.008152**

Table 5-5 Re-screen of siRNAs targeting genes targeted by the top 20 siRNAs from the druggable genome screen
siRNAs from the library targeting genes which were targeted by the top 20 scoring siRNAs from the druggable genome screen were transfected, along with siNeg, into HeLa cells in triplicate and these cells were tested for their sensitivity to TRAIL-induced cytotoxicity. p-values were calculated by one-tailed student's t-test on log transformed data and corrected using a Hommel correction for multiple testing. P-values less than 0.1 are highlighted in bold. Genes are ranked by the score of the highest ranking siRNA in the screen

In order to confirm the involvement of the selected genes in TRAIL apoptosis, and also to test the activity of novel siRNAs obtained, the effect of transfection of the siRNAs on TRAIL-induced Caspase-3/7 activity was measured using a luminescent Caspase-3/7 assay (Figure 5.8a). Transfection of siCasp8 reduced the level of TRAIL-induced Caspase-3/7 activity to 44% and 41% (for plates 1 and 2 respectively) compared to levels in cells transfected with the negative control. Transfection of 29 of 46 siRNAs significantly reduced the level TRAIL-induced Caspase-3 activity compared to the negative control. This includes 20 of the 22 siRNAs previously shown to have a significant effect on TRAIL-induced cytotoxicity. In total ten genes were targeted by at least two independent siRNAs that caused a significant reduction in TRAIL-induced Caspase-3/7 activity and two genes were targeted by three independent siRNAs that caused a significant reduction in TRAIL-induced Caspase-3/7 activity.

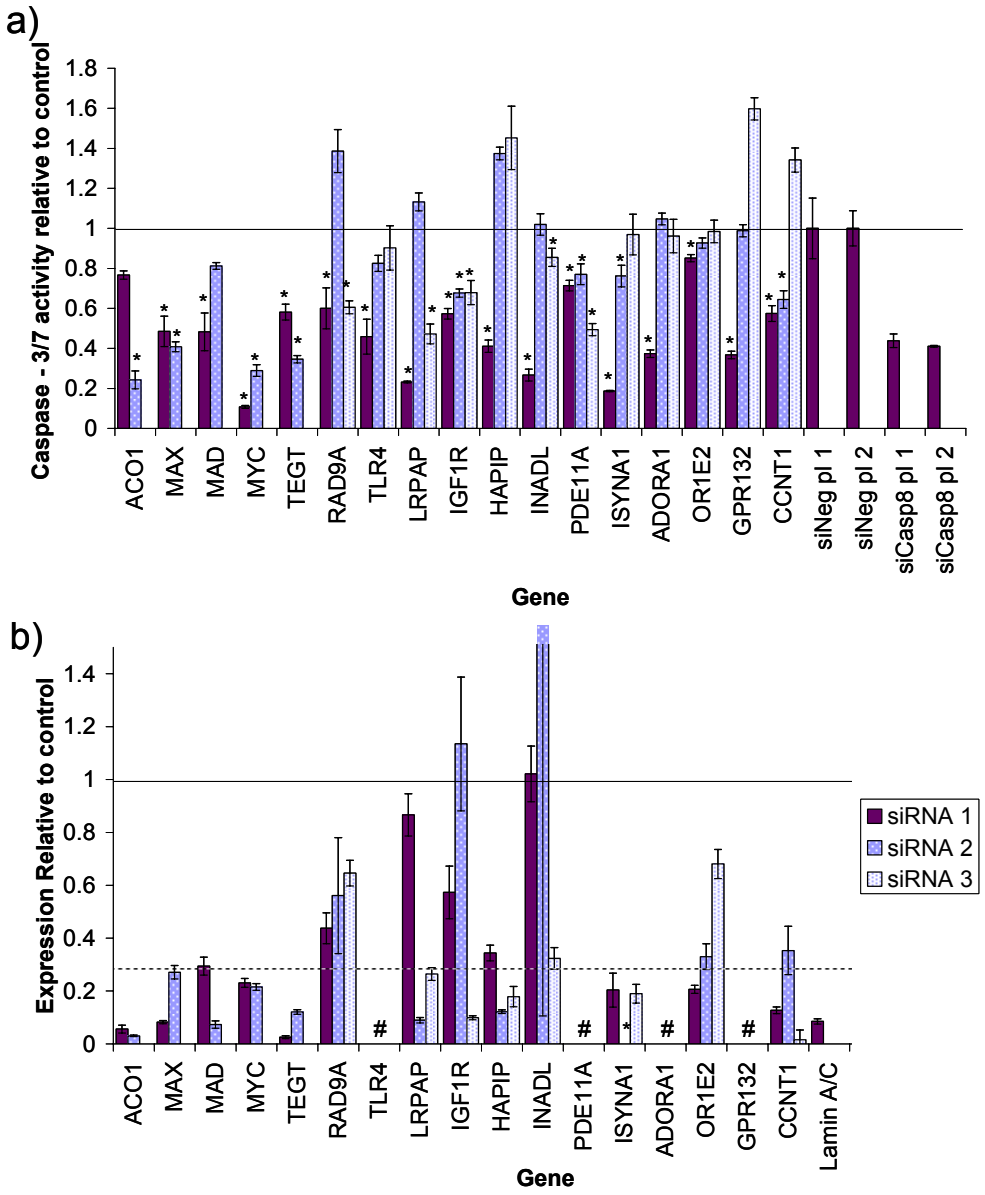

Figure 5.8 Confirmation of candidate hits from a screen of the druggable genome
a) Effects of transfection of siRNAs targeting candidate hit genes on TRAIL-induced Caspase-3/7 activity. siRNAs targeting candidate hit genes, Caspase-8, or a negative control were transfected into HeLa cells in triplicate. Cells were treated with 0.5μg/ml TRAIL for 6 hours, 48 hours after transfection. Caspase-3/7 activity was measured using Promega's Caspase-Glo 3/7 assay. The experiment was conducted over two plates. Each plate included siNeg and siCasp8 control transfections. siNeg pl1 and siCasp8 pl1 are controls from plate 1, and siNeg pl2 and siCasp8 pl2 are controls from plate 2. Results are normalised to control levels. Solid line represents 100% of control activity. * result is significantly different, at the 5% level, from the negative control of the same plate. P-values calculated using student's t-test. b) Effect of transfection of siRNAs targeting candidate hit genes on mRNA levels of targeted gene. cDNA was prepared by reverse transcription of RNA isolated from cells transfected with siRNAs targeting candidate hits or Lamin A/C. SYBR green qPCR was carried out in triplicate using primers designed to amplify from mRNA of genes targeted, GAPDH and ACTB. Primers were designed and tested as described in Methods. Primers were successfully designed for 12 of the 17 genes tested. Expression levels are shown relative to negative control and were calculated using a variation of the Pfaffl method to allow normalization to multiple housekeeping genes using GAPDH and ACTB to normalize samples (Hellemans et al. 2007). # = genes for which no primers were successfully designed. * no ISYNA1 transcript was detected in cells transfected with siISYNA1.2 Solid line represents 100% of negative control levels and dashed line represents 30% of control levels). Error bars represent 1 standard error of the mean.

Two possibilities for the observation that some siRNA do not reduce levels of TRAIL-induced Caspase-3/7 activity are that the siRNA does not reduce the levels of the target mRNA sufficiently, or that the target gene is not involved in TRAIL-induced cytotoxicity. To help distinguish between these possibilities levels of mRNA knock-down triggered by siRNA transfection were measured using qRT-PCR (Figure 5.8b). Efficient, specific primers were successfully designed for 12 out of the 17 candidate hit genes. The efficiency of knock-down of the Lamin A/C mRNA by a well characterised siRNA was used as a positive control. Transfection of siLaminA/C reduced levels of the Lamin A/C mRNA to 8.5% of levels in negative control transfected cells, demonstrating the efficiency of the siRNA transfection and qRT-PCR measurement. A total of 20 of the 33 siRNAs designed to target the hit genes tested reduced the level of the targeted mRNA to less than 30% of control levels.

Based on the ability of siRNAs to affect the level of TRAIL-induced Caspase – 3/7 activities and the ability of siRNAs to reduce levels of the intended target mRNA, candidate hit genes were categorised into one of the same four categories defined in the previous chapter. If siRNAs that significantly reduce the level of TRAIL-induced Caspase-3/7 activity are designated phenotypically active siRNAs, and the efficiency of an siRNA is the amount by which an siRNA reduces the intended target mRNA when transfected at a set concentration, then:

- **Confirmed hit genes** are genes targeted by at least two phenotypically active siRNAs that are more efficient than any phenotypically inactive siRNAs targeting the same gene.
- **Unconfirmed hit genes** are genes targeted by only one phenotypically active siRNA, where that siRNA is more efficient than phenotypically inactive siRNAs targeting the same gene, or the efficiency or the siRNAs is unknown.
- **Confirmed off-targets** are genes targeted by both phenotypically active and inactive siRNAs, where at least one phenotypically inactive siRNA is more efficient than the least efficient phenotypically active siRNA.
- **Unrepeatable genes** are genes that are not targeted by any phenotypically active siRNAs.

The results presented in Figure 5.8 along with the categorisations of the candidate hit genes are summarised in Table 5-6. In total six genes were categorised as confirmed hit genes, five as unconfirmed hit genes and five as off-targets. No genes were categorised as unrepeatable. One gene, ISYNA1, could not be categorised, due to the lack of a result for

the efficiency of siRNA 2 targeting ISYNA1. In this case the amplification of ISYNA1 transcript for the transfected cells was below the detection limit of the thermocycler used. This could be due to siRNA 2 targeting ISYNA1 knocking down the ISYNA1 mRNA sufficiently to make it undetectable, or due to a failure in the protocol.

Gene	*siRNA*	*Repeat*	*Caspase-3/7 significant*	*Caspase-3/7 Rank*	*>=70% KD*	*KD Rank*	*Conclusion*
ACO1	1	+	-	2	+	2	**Unconfirmed**
	2	+	+	1	+	1	
MAX	1	+	+	2	+	1	**Confirmed Hit**
	2	+	+	1	+	2	
MAD	1	+	+	1	+	2	**Off-Target**
	2	+	-	2	+	1	
MYC	1	+	+	1	+	2	**Confirmed Hit**
	2	+	+	2	+	1	
TEGT	1	+	+	2	+	1	**Confirmed Hit**
	2	+	+	1	+	2	
RAD9A	1	+	+	2	-	1	**Off-Target**
	2	N/A	-	3	-	2	
	3	N/A	+	1	-	3	
TLR4	1	+	+	1	N/A	N/A	**Unconfirmed**
	2	N/A	-	2	N/A	N/A	
	3	N/A	-	3	N/A	N/A	
LRPAP1	1	+	+	1	-	3	**Off-target**
	2	N/A	-	3	+	1	
	3	N/A	+	2	+	2	
IGF1R	1	+	+	1	-	2	**Confirmed Hit**
	2	N/A	+	2	-	3	
	3	N/A	+	3	+	1	
HAPIP	1	+	+	1	-	3	**Off-Target**
	2	N/A	-	2	+	1	
	3	N/A	-	3	+	2	
INADL	1	+	+	1	-	2	**Confirmed Hit**
	2	N/A	-	3	-	3	
	3	N/A	+	2	-	1	
PDE11A	1	+	+	2	N/A	N/A	**Confirmed Hit**
	2	N/A	+	3	N/A	N/A	
	3	N/A	+	1	N/A	N/A	
ISYNA1	1	+	+	1	+	2	
	2	N/A	+	2	N/A	1	
	3	N/A	-	3	+	2	
ADORA1	1	+	+	1	N/A	N/A	**Unconfirmed**
	2	N/A	-	3	N/A	N/A	
	3	N/A	-	2	N/A	N/A	
OR1E2	1	+	+	1	+	1	**Unconfirmed**
	2	N/A	-	2	-	2	
	3	N/A	-	3	-	3	
GPR132	1	+	+	1	N/A	N/A	**Unconfirmed**
	2	N/A	-	2	N/A	N/A	
	3	N/A	-	3	N/A	N/A	
CCNT1	1	+	+	1	+	2	**Off-Target**
	2	N/A	+	2	-	3	
	3	N/A	-	3	+	1	

Table 5-6 Summary of confirmation experiments and categorisation of candidate hit genes. KD: Knock-down

PDE11A was categorised as a "hit" despite the lack of measurements as to the

efficiency of the siRNAs targeting it. Since all three siRNAs targeting PDE11A are phenotypically active, any combination of ranking for the efficiencies would still lead to the categorisation of the gene as a "hit".

Three genes are categorised as off-target despite being targeted by two independent siRNAs (RAD9A, LRPAP1 and CCNT1). This is due to the genes being targeted by a third, phenotypically inactive, siRNA which is more efficient than either of the other two active siRNAs. This is unexpected, as it is assumed that two siRNAs will not share off-target effects and therefore if two independent siRNAs targeting the same gene have the same phenotype, that the effect is unlikely to be due to off-target effects. This was the case for one gene (PTP4A) from the kinase and phosphatase screen. One possibility is that the assay is sensitive to the non-specific effects of the siRNAs on the cell. If activating the siRNA pathway in general were affecting the assay, then it would be expected that all siRNAs would have the same effect. It is possible that some aspect of the particular siRNAs used here have a differential effect on non-specific responses. Several sequence motifs have been reported to stimulate innate immune responses to siRNAs, such as UGUGU(Judge et al. 2005) and GUCCUUCAA (Hornung et al. 2005). Of the siRNAs used for confirmation of hits in this study only siMYC.B contains either of these sequences (UGUGU). Further, generally only immune cells express TLR 7 and 8 which are necessary for the recognition of these sequences (Judge, Maclachlan 2008). However, it is known that poor quality or impure siRNA preparations can induce non-specific responses (Marques et al. 2006). Thus certain siRNAs (or siRNA preparations) maybe affecting the assay independent of the gene knock-down stimulated by them.

In light of this ambiguity, it is safer to designate these genes as potential off-target genes rather than confirmed off target genes. Only two genes (MAD and HAPIP) are therefore designated confirmed off-target.

5.6 Characterisation of hit genes

In order to further investigate the involvement of genes targeted by two phenotypically active siRNAs in the apoptotic pathway, the effect of transfection of these siRNAs on TRAIL-induced Caspase-8 and Caspase-9 was measured using luminescent caspase assays (Figure 5.9). siRNAs targeting the potential off-target genes are included in an attempt to further investigate the nature effects caused by these siRNAs.

Transfection of 10 out of the 20 siRNAs tested significantly reduced the level of TRAIL-induced Caspase-8 activity (Figure 5.9a). Interpretation of the data suffers from the

large variances observed in some cases. Despite this, data from the two siRNAs targeting each gene were in agreement in all but two cases (MAX and ISYNA1). In this case of MAX it is likely that this is due to the large amount of variation seen in the measurement of Caspase-8 activity in TRAIL-treated siMAX.1 transfected cells. There are four cases in which both siRNAs targeting a gene caused a significant reduction in TRAIL-induced Caspase-8 activity (MYC,RAD9A, INADL and LRPAP1).

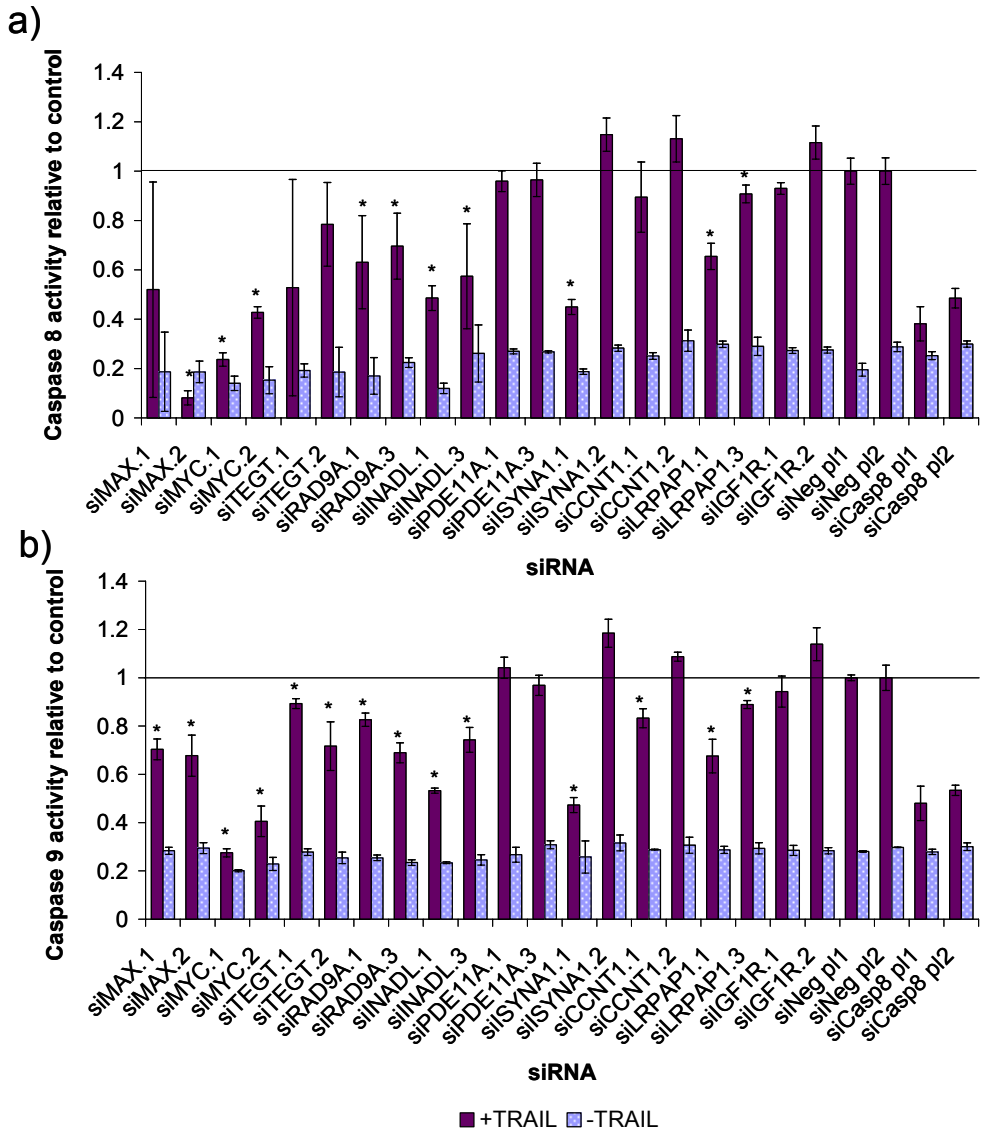

Figure 5.9 Effect of transfection of siRNAs targeting hit genes on TRAIL-induced Caspase activity
Cells were transfected with siRNAs targeting possible hit genes, siCasp8 or siNeg in sextuplet on two plates. 48 hours post-transfection, cells were treated with either 0.5μg/ml TRAIL or media for 6 hours. Levels of a) Caspase-8 or b) Caspase-9 were measured using Promega Caspase-Glo luminescent caspase assays. Results are expressed relative to the caspase activity levels in TRAIL-treated negative control transfected cells. Horizontal line represents negative control levels. Error bars represent 1 standard deviation, n = 3. * uct transfected, the effects should be similar, independent of the siRNA/shRNAmir transfected. Therefore it can be assumed that the differences between different siRNAs/shRNAmirs is the result of the differing effects of that construct on the pay.

Transfection of 14 out of 20 siRNAs significantly reduced the level of TRAIL-

induced Caspase-9 activity (Figure 5.9b). In all but two cases (ISYNA1 and CCNT1) data from the two siRNAs targeting each gene were in agreement. Both siRNAs targeting a gene significantly reduced the level of TRAIL-induced Caspase-9 activity in six out of ten genes (MAX, MYC, TEGT, RAD9A, INADL and LRPAP1). Four of these genes were targeted by 2 siRNAs that induced a significant reduction in TRAIL-induced Caspase-8 activation (MYC, RAD9A, INADL and LRPAP1). In the case of MAX, one siRNA significantly reduced the level of TRAIL-induced Caspase-8 activity, while the other showed a reduction that wasn't statistically significant, possibly due to the high level of variability. One gene (TEGT) was targeted by two siRNAs that significantly reduced levels of Caspase-9, but not Caspase-8 (although a non-significant reduction in TRAIL-induced Caspase-8 activity was observed).

In two cases genes were targeted by siRNAs, transfection of neither of which caused a significant reduction in either Caspase-8 or Caspase-9 (PDE11A and IGF1R). This information is summarised in Table 5-7.

Gene	*siRNA*	*Caspase-8*	*Caspase-9*	*Caspase-3*	*Confirmed?*
MAX	1	-	+	+	+
	2	+	+	+	
MYC	1	+	+	+	+
	2	+	+	+	
TEGT	1	-	+	+	+
	2	-	+	+	
RAD9A	1	+	+	+	-
	3	+	+	+	
INADL	1	+	+	+	+
	3	+	+	+	
PDE11A	1	-	-	+	+
	3	-	-	+	
ISYNA1	1	+	+	+	
	2	-	-	+	
CCNT1	1	-	+	+	-
	2	-	-	+	
LRPAP1	1	+	+	+	-
	3	+	+	+	
IGF1R	1	-	-	+	+
	2	-	-	+	

Table 5-7 Summary of effects of transfection of siRNAs from potential hits on TRAIL-induced caspase activity
+ signifies that a significant reduction in activity was observed, - that no significant reduction was observed. Confirmed indicates if gene was classed as a confirmed hit (+) or a potential off-target (-). See Table 5-6.

The evidence presented here suggests that siRNAs targeting MYC, RAD9A, INADL and LRPAP1 act to regulate the apoptosis pathway upstream of Caspase-8. It is also possible that MAX acts upstream of Caspase-8 since the lack of significance in the case of the effect of siMAX.1 on Caspase-8 activity could be due to the large amount of variation observed. Two of these (RAD9A and LRPAP1) were designated as potential off-targets in confirmation experiments (Figure 5.8 and Table 5-6). Despite the fact that both siRNAs

targeting these genes induce the same effect on TRAIL-induced Caspase activity, this does not demonstrate that the effects observed here are not due to the knock-down of off-target genes. While transfection of TEGT did not produce a significant decrease in Caspase-8, this could be due to the large variance. Indeed, no hypothesis test can demonstrate the truth of the null hypothesis – in this case that there is no difference between cells transfected with hit siRNAs and those transfected with the negative control. Therefore, all that can be concluded here is that the involvement on TEGT in regulation of TRAIL-induced Caspase-8 activity has not been demonstrated. Similarly, transfection of both PDE11A and IGF1R failed to produce significant reductions in TRAIL-induced activity of either Caspsae-8 or Caspase-9. Variability here was smaller than in the case of TEGT, however, it is possible, from these results, that knock-down of PDE11A or IGF1R cause a small, but real, reduction in the activity of Caspase-8 and -9 activity. Calculating the 95% confidence limit on the mean difference between Caspase-8 activity levels in the negative control transfected cells and in siIGF1R.1/2 transfected cells (-7%±9% for siIGF1R.1 and +12%±14% for siIGF1R.2) shows that there is a 95% confidence that these siRNAs do not cause a reduction in Caspase-8 levels greater than 16% and 2% (the lower bounds of the confidence intervals) respectively for siIGFR.1 and siIFGR.2 compared to the levels in the negative control. The equivalent figures for siPDE11A.1 siPDE11A.2 are 15% and 14% respectively. Confidence limits on Caspase-9 levels are similar (data not shown).

Results from the siRNAs for ISYNA1 and CCNT1 differ for the two siRNAs targeting these genes. CCNT1 was designated a potential off-target gene. These results do not prove that the effects of transfection with siRNAs targeting CCNT1 are off-target effects: the effect of transfection of siCCNT1.2 is weaker than siCCNT1.2 in the Caspase-3 assay (Figure 5.8a) and it is possible that a significant effect is not seen in the Caspase-9 assay due to a lower sensitivity. However, they do make it more likely that the results are due to off target effects. In the case of ISYNA1, siISYNA.1 clearly has a far larger effect than siISYNA.2 in all the caspase assays, arguing against siISYNA.2 being more efficient at knocking down the ISYNA1 transcript and so suggests that ISYNA1 should be categorised as an off-target gene.

5.7 Analysis of seed sequences

From rigorous confirmation experiments on genes targeted by the top 20 scoring siRNAs, six genes have been identified as confirmed hits (MAX, MYC, TEGT, IGF1R, INADL and PDE11A). A further 5 genes were categorised as unconfirmed hits (ACO1,

TLR4, ADORA1, OR1E2 and GPR132). Five genes were categorised as either potential or confirmed off-targets (MAD, RAD9A, LRPAP1, HAPIP and CCNT1). One gene (ISYNA1) could not be categorised, although evidence from experiments designed to determine the point at which targeted genes were acting in the apoptosis pathway suggested ISYNA1 might also be an off-target.

These results suggest that a large number of results seen in the screen might be due to off-target effects. An important factor in the determination of miRNA specificity is the sequence of the so-called "seed region": this is the region of the miRNA corresponding to positions 2-7 or 2-8 of the mature miRNA sequence or the equivalent region of the target sequence. Several lines of evidence suggest that at least some off-target effects of siRNAs might be due to matches between the seed region of the siRNA and the 3' UTR of transcripts, causing the siRNA to act as a miRNA in repressing these transcripts (Birmingham et al. 2006, Grimson et al. 2007, Lin et al. 2005, Lin et al. 2007, Sætrom et al. 2007) . This could lead to the knock-down of one or several unintended transcripts. Nielsen *et al* used the number and length of seed matches between an siRNA and a 3'UTR, along with the AU content and conservation of the surrounding sequences to predict the fold change which the siRNA would induce in the mRNA. In this way they accurately predicted which mRNAs would have the largest change in expression levels following transfection of the siRNA (Nielsen et al. 2007).

In a screen for genes which sensitize normally resistant cells to a Bcl-2/Bcl-XL inhibitor Lin *et al* found that the top three hits were due to off target effects. Further, they found that two of the siRNAs targeting these genes shared a 7nt seed sequence (a heptamer seed, see Figure 1.4 for definition of different types of seed sequence).

To determine if any of the hits in the screen presented here share seed sequences, the seed sequences were extracted and compared from the 40 siRNAs targeting each of the genes to which the top 20 scoring siRNAs from the screen were designed. Seed sequences that appear more than once are shown in Table 5-8. Three hexamer seed sequences (bases 2-7 of the siRNA guide strand or bases 15-20 of the target sequence) appeared more than once in the top 20 siRNAs from the screen, with one sequence appearing in four different siRNAs. In addition one hexamer seed sequence appeared both in an siRNA from the top 20 scoring siRNAs from the screen and also in the second siRNA targeting TEGT, which was later shown in confirmation experiments to induce a significant reduction in TRAIL sensitivity (Figure 5.8a). During confirmation experiments, siRNAs not used in the screen were obtained where only one of the siRNAs in the screen targeting a gene reproducibly gave

a phenotype (e.g. siINADL.3 and siPDE11A.3). These siRNAs were also checked, but in no case did any of these siRNAs contain a seed sequence also found in other siRNAs which targeted candidate hit genes If only siRNAs which gave a statistically significant result in confirmation experiments are considered, there are three hexamer seed sequences that appear in more than one siRNA in this collection, with one of these appearing in three confirmed siRNAs from the druggable genome screen and one siRNA from the kinase and phosphatase screen.

Only four of the 15 siRNAs targeting confirmed hit genes from the druggable genome screen contained one of the these seeds and in all but one case each, of these confirmed hit genes targeted by an siRNA containing one of the seeds was also target by an independent siRNA not containing a one of the seeds. This suggests that these are still valid hits.

Sampling was used to assess the significance of this observation. 5,000 samples of 20 siRNAs were drawn at random from the list of all siRNAs in the screen. The seed sequences of these siRNAs were determined and compared to determine if any seed sequences appeared multiple times. The probability of any seed appearing more than once in a random sample of 20 siRNAs was found to be 0.1828, the probability of two seed sequences appearing two or more times in a random sample of 20 siRNAs was 0.0128 and the probability of a seed sequence appearing 4 or more times in a random sample of siRNAs was less than 0.0002. This suggests that the observation that two seed sequences appear twice in the top 20 siRNAs from the screen is significant and the observation that one seed sequence appears in four siRNAs from the top 20 is highly significant. Using the same approach to assess the significance of finding the same seed sequences in more than one of the 14 siRNAs from the top twenty scoring siRNAs that were confirmed in confirmation experiments gives p-values of 0.08, 0.003 and less than 0.0002 for finding any seed sequence appearing in at least two siRNAs, finding any two seeds appearing in at least two siRNAs and finding any seed appearing in three or more siRNAs respectively. Two of the seed sequences that appeared multiple times in the top 20 siRNAs from the druggable genome screen also appeared in the siRNAs used to confirm hits from the kinase and phosphatase screen (Table 5-8). In addition two seed sequences that appeared only once in the top 20 siRNAs also appeared in the siRNAs used to confirm hits from the kinase and phosphatase screen.

These results suggest perhaps some of the effect of the ten siRNAs with these seed sequences on TRAIL sensitivity may be due to an off-target or effect or effects shared between siRNAs containing the same seed sequence. It is possible that this may be one of

the confirmed hit genes from the screen. While a match between the seed sequence and the 3' UTR of a transcript is partially predictive of the siRNA having a silencing effect on the transcript, many transcripts with a hexamer seed match will not be silenced. Both Birmingham *et al* and Neilsen *et al* showed that a heptamer seed match, with a match between bases 2 -8 of the siRNA guide strand and the 3' UTR of a transcript has a higher predictive value than a hexamer match, although the sensitivity is reduced (Birmingham et al. 2006, Nielsen et al. 2007). Further, since there are more possible heptamers than hexamers, a heptamer appearing more than once in the top 20 siRNAs would have an increased significance.

Seed Size	Seed Sequence (target)	Druggable Genome Screen				Kinase Screen		
		siRNA	*Rank*	*Confirmed*		*siRNA*	*Confirmed*	
				siRNA	*Gene*		*siRNA*	*Gene*
Hexamers								
	CAAGGT							
		siTEGT.B	1	+	+			
		siCCNT1.B	3	+	-			
	TGTCCA							
		siISYNA.B	2	+	+	siINPP5D.2	-	-
		siKCNV1.B	17	-	-			
	ACTTGA							
		siACO1.B	4	+	-	siSharpin.1	+	+
		siHAPIP.B	5	+	-			
		siINADL.B	7	+	+			
		siMAGED.A	14	-	-			
	AGATCA							
		siGPR132.B	9	+	-	siIKBKE.1	+	+
		siTEGT.A	376	+	+			
	GCATTA							
		siLOC402037.A	20	-	-	siPPP2CB.2	+	-
Heptamers								
	TCAAGGT							
		siTEGT.B	1	+	+			
		siCCNT1.B	3	+	-			
	GTGTCCA							
		siISYNA.B	2	+	+	siINPP5D.2	-	-
		siKCNV1.B	17	-	-			
	AACTTGA							
		siINADL.B	7	+	+			
		siMAGED.A	14	-	-			
	CACTTGA							
		siACO1.B	4	+	-	siSharpin.1	+	+
	GAGATCA							
		siGPR132.B	9	+	-			
		siTEGT.A	376	+	+			

Table 5-8 Repeated seed sequences from druggable genome screen results
Table shows seed sequences which appear more than once in siRNAs targeting genes targeted by the top 20 siRNAs from druggable genome screen , or together in one of the top 20 siRNAs from druggable genome screen and in one of the siRNAs targeting candidate hits from kinase and phosphatase screen. Sequences shown are the siRNA target sequences complimentary to bases 2-7 (Hexamers) or bases 2-8 (Heptamers) of the siRNA guide strand. siRNAs labelled A or B are siRNAs are from A or B plates of the library. Those numbered 1-3 are siRNAs used for hit confirmation. Rank show what rank the siRNA was in the screen results. 'Confirmed' shows whether the gene targeted by the siRNA was categorised as a confirmed hit.

Three heptamers appear more than once in the top 20 highest scoring siRNAs from the screen of the druggable genome. One of these appears in two siRNAs that were confirmed. There is one heptamer seed sequence that appears once in the top 20 scoring siRNAs from the screen of the druggable genome and once in the siRNAs used to confirmed hits from the kinase and phosphatase screen (both of which were confirmed), and one heptamer seed that appears once in the top 20 scoring siRNAs from the druggable genome screen and also appears in the second TEGT siRNA, which was shown in subsequent experiments to significantly reduce the sensitivity of cells to TRAIL-induced apoptosis (Figure 5.8 and Table 5-8). The probability of a heptamer seed appearing in a random sample of 20 siRNAs from the druggable genome library is 0.048, while the probability of two and three seeds appearing more than once in a random sample of 20 siRNAs from the library is 0.001 and less than 0.0005 respectively (based on 2,000 random samples of 20 siRNAs from the library of siRNAs used in the screen). The probability of finding a seed appearing in any two siRNAs if only the 14 confirmed siRNAs are considered is 0.03.

The region of an miRNA from base two to base eight is known as the 7mer-m8 seed (Figure 1.4). A different heptamer, shown to be effective at predicting miRNA targets is the 7mer-A1 site, which is the hexamer match flanked by an A in the target at the base that corresponds to base 1 of the siRNA guide strand (Lewis, Burge & Bartel 2005). Since all siRNA in the library have a U at this position, all the siRNAs that share a hexamer match also share a 7mer-A1 seed. Similarly, all those siRNAs that share a 7mer-m8 seed also share an octamer (8mer) seed (bases 1-8 of the siRNA guide strand, Figure 1.4). From here, on the set of seed sequences listed in Table 5-8 are referred to as "hit" seeds (hit hexamers/hit heptamers).

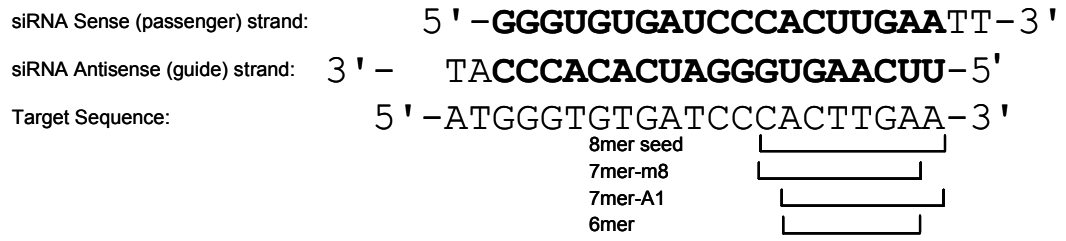

Figure 5.10 Different types of seed sequence as defined by (Lewis, Burge & Bartel 2005)
Seeds are shown in order of the predictive power of a match between the seed sequence of a miRNA and a UTR in predicting transcripts that will be affected by a miRNA.

Within each group of siRNAs from the top 20 scoring siRNAs sharing a common seed sequence, only one siRNAs targets a gene that was confirmed. However there are two cases of siRNAs targeting confirmed hit genes, one from this screen and one from the kinase

Seed	Size	Enrichment Score	Normalized Enrichment Score	Nominal P-value	FWER p-val	Rank at Max	Leading Edge
Hexamers							
TAATAA	70	0.542	2.691	<0.001	0	1942	44%
ACTTGA	14	0.846	2.674	<0.001	0	946	79%
CCTTAA	16	0.793	2.642	<0.001	0	1367	81%
AATTAA	44	0.566	2.567	<0.001	0	1798	45%
ACTGGA	10	0.877	2.510	<0.001	0	933	90%
TAGGAA	13	0.786	2.505	<0.001	0	621	54%
TAAAGA	20	0.685	2.482	<0.001	0.002	1659	45%
GAATAA	21	0.716	2.474	<0.001	0.002	1902	57%
AAGTTA	20	0.651	2.426	<0.001	0.004	1963	60%
TCACAA	12	0.792	2.413	<0.001	0.005	1723	75%
AAATGA	21	0.662	2.405	<0.001	0.005	2030	52%
AGATCA	35	0.588	2.378	<0.001	0.006	1728	60%
TTATAA	22	0.650	2.335	<0.001	0.009	1978	50%
AATATT	15	0.698	2.314	<0.001	0.015	3088	80%
AGATCT	17	0.666	2.304	<0.001	0.015	2808	71%
TGAATA	16	0.673	2.255	<0.001	0.026	2085	81%
CTGGAA	8	0.848	2.198	<0.001	0.047	341	50%
Heptamers							
CAATTAA	18	0.817	2.911	<0.001	0.000	1154	67%
GAGATCA	18	0.696	2.387	<0.001	0.001	600	50%
GAAAGAA	10	0.790	2.241	0.005	0.015	2173	90%
TTAATAA	17	0.624	2.227	<0.001	0.015	1555	41%
GTATTTA	16	0.671	2.211	<0.001	0.016	2928	81%
ATAGGAA	6	0.923	2.187	<0.001	0.021	467	67%
ACCTTAA	8	0.828	2.185	<0.001	0.021	1284	88%
TAATTAA	8	0.812	2.179	<0.001	0.024	1798	63%
TTAATTA	10	0.776	2.151	<0.001	0.037	2464	100%
ACAATTA	10	0.753	2.148	<0.001	0.039	2069	80%
AAGATCA	13	0.659	2.140	0.007	0.041	1643	69%
CTAATAA	20	0.595	2.139	<0.001	0.041	1043	45%
CTAATTA	7	0.855	2.136	<0.001	0.043	1042	71%

Table 5-9 Seed sequences enriched in high scoring siRNAs
Table shows results of applying GSEA to the ranked list of siRNAs from the druggable genome screen using "gene sets" composed of siRNA sequences which share a hexamer or heptamer seed. The size column refers to the size of the set (i.e. the number of siRNAs containing that seed), (normalized) enrichment score is the statistic calculated by GESA. Rank at max is the position in the ranked list of siRNAs that the enrichment score is maximal, and the Leading edge is the proportion of genes in the gene set that rank at or higher than this point.

and phosphatase, sharing a common seed sequence.

Increasing the length of the seed increases the predictive power of that seed in determining the identity of transcripts regulated by those siRNAs sharing this seed.

However, it reduces the sensitivity of the process. As such, although siRNAs which share a heptamer seeds are more likely to share some of the same off-target effects, using the hexamer seed increases the chance of finding siRNAs that share off-target effects.

5.7.1 Gene Set Enrichment Analysis of seed sequences

The process of RNAi screening is designed to enrich for siRNAs that produce a phenotype of interest. As such, it can be assumed that the screening process will enrich not only for siRNAs designed to target genes involved in the process of study, but also siRNAs which target genes involved in the process through 'off-target' effects (Lin et al. 2007). In the previous section, seed sequences were found multiple times in siRNAs targeting the genes targeted by the top 20 siRNAs. However, defining the top 20 siRNAs as a cut off is arbitrary. In order to determine if any seeds had been enriched generally in high scoring siRNAs, gene set enrichment analysis (GSEA, see 0) was applied to the results of the druggable genome screen. The hexamer and heptamer seed sequence of every siRNA used in the screen was determined and 'gene sets' were constructed where every set was composed of all siRNAs from the screen that shared a particular seed sequence. GSEA was then applied to the ranked list of all siRNAs from the screen using these gene sets in order to identify sets of siRNAs, all containing the same seed, which were enriched at the top end of the ranked list of siRNAs. A FWER significance cut-off of 0.05 was applied to determine which sets were enriched in the high-scoring siRNAs. The output from this analysis of 'gene sets' based on both hexamer and heptamer seeds is shown in Table 5-9. The analysis shows that 17 hexamer and 13 heptamer seeds are enriched in the high-scoring siRNAs, (defining a set of seed sequences, referred to from here on as "enriched" seeds. These are distinct from, but overlapping with, the "hit" seeds). Two of the enriched hexamers and one of the enriched heptamers are among the seed sequences in Table 5-8. Three of the enriched hexamer seeds and one of the enriched heptamer seeds which are not among the hit seeds are found in top 20 siRNAs: The hexamer CCTTAA is found in siMYC.A, TAAAGA is found in the siRNA siMAD.A and the hexamer/heptamer (C)TAATAA is found in the siRNA siADORA.B

The enrichment of these seed sequences in the high scoring siRNAs shows siRNAs containing these seed sequence have an increased probability of inducing the phenotype of interest when transfected. There are two possible explanations for this. Firstly it is possible that these seed sequences are highly efficient sequences, and that siRNAs containing these sequences are, therefore, more likely to knock-down the intended target sufficiently to induce the phenotype of interest. The second possibility is that the enriched seed sequences specify

siRNAs containing them to knock-down off-target transcripts, some of which could be among the hits from the screen, which are involved in the process of interest. That is, as predicted by Lin *et al*, the screening process itself is enriching for off-target effects. However, it is important to note that in no case is the "rank at max" close to the very top of the list, nor in most cases is the proportion of siRNAs containing this seed which appear at or above the "rank at max" 100%. This suggests that the score of an siRNA in the screen is not purely determined by its seed sequence.

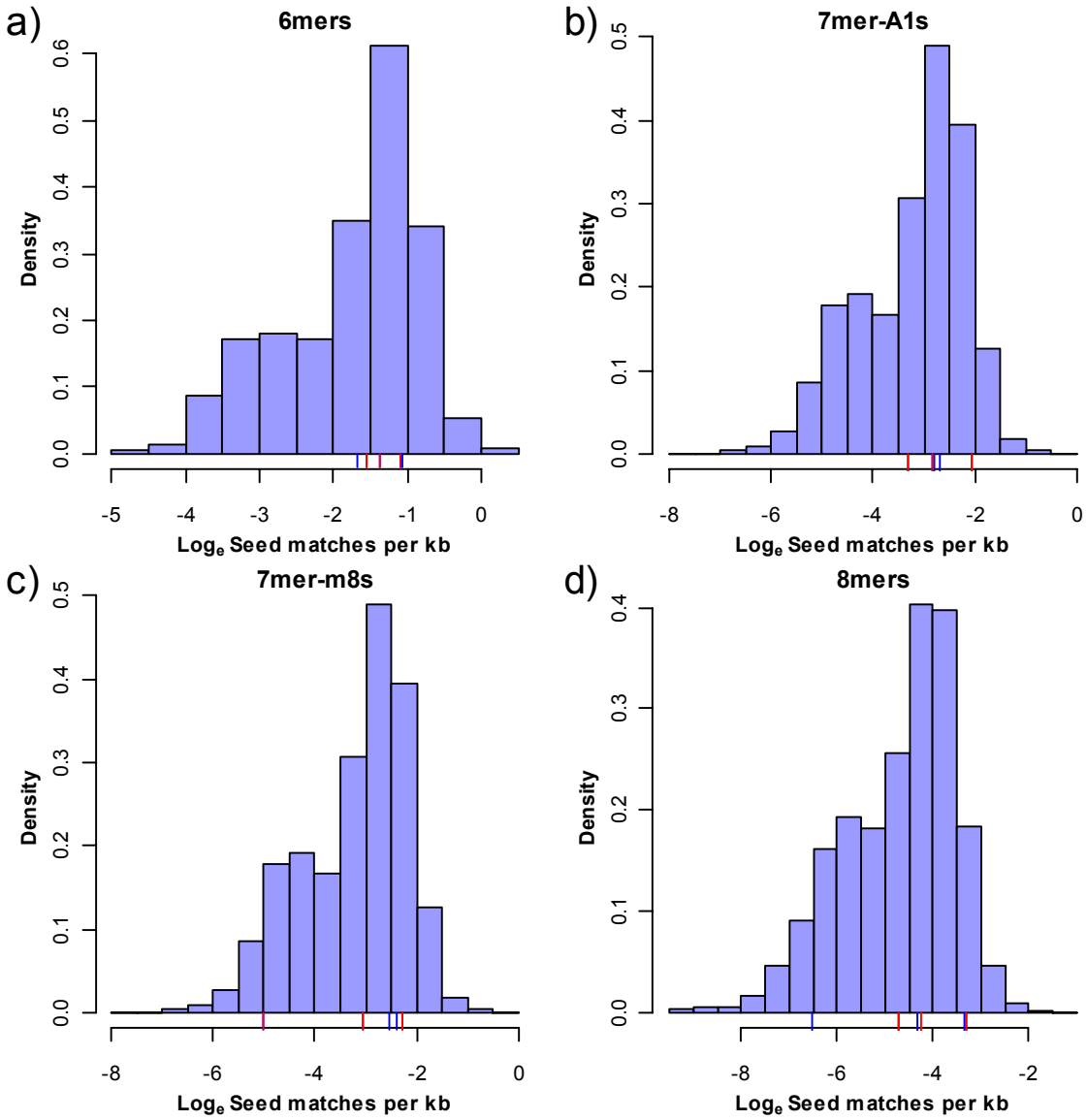

Figure 5.11 Average frequency of 'Hit' seed sequences in all 3' UTRs compared to average frequency of all possible seed sequences

3'UTR sequences were obtained from ensembl (release 46). Each 3' UTR was searched for matches to every possible 6nt (a), 7nt (b/c) and 8nt (d) sequence. Frequency of each sequence per kb of each UTR was calculated and averaged (using a script written by Dr. A. Enright). The distributions of log average sequence frequency were plotted. The average frequency of each of the a) hit 6mer seeds, b) hit 7mer-A1seeds, c) hit 7mer-m8 seeds and d) hit 8mer seeds is shown as a blue tick beneath the histograms, seeds present in two or more confirmed siRNAs are shown in red.

5.7.2 Frequency of hit and enriched seed sequences in 3' UTR sequences

siRNAs containing seed sequences that are found more frequently in the 3'UTR of transcripts are likely to have a larger number of off-target effects. Further, finding multiple matches between a 3' UTR and an siRNA seed sequence has a higher predictive value for off targets than the presence of a single match (Birmingham et al. 2006). Thus it is the total frequency of seed matches in a 3' UTR that is important, rather than just the presence or absence of such matches.

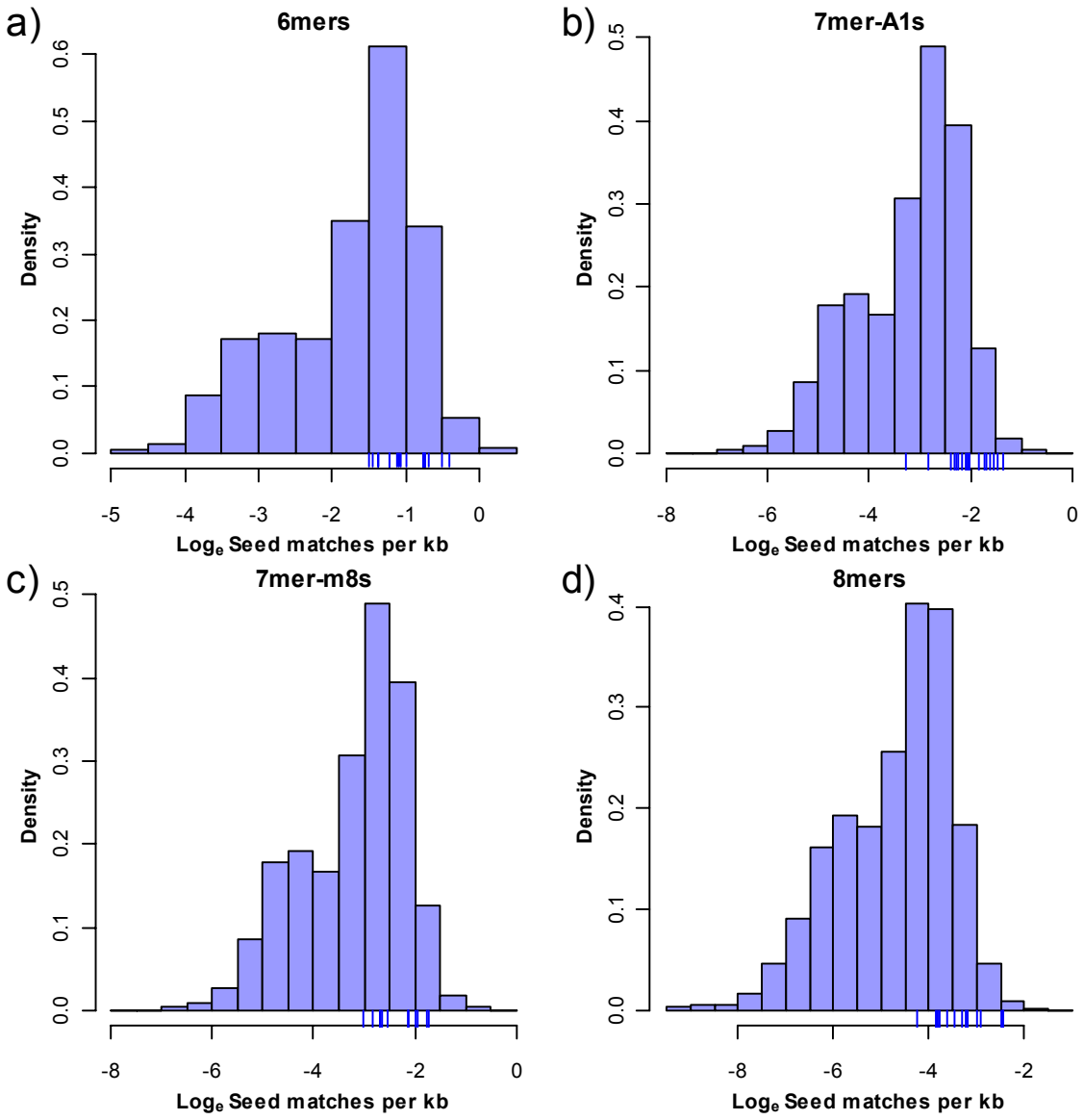

Figure 5.12 Frequencies of 'enriched' seed sequences in all 3' UTRs compared to the all possible seed sequences.
Distributions of frequencies of all possible sequences were calculated as described for Figure 5.11. The frequencies of a) enriched hexamers seeds (6mers), b) enriched 7mer-A1 heptamer seeds, c) enriched 7mer-m8 heptamer seeds and d) enriched octamer (8mer) seeds are indicated by tick marks under the histograms.

To examine the possibility that the hit/enriched seed sequences are found at an unusually high frequency in the 3' UTRs of transcripts, the average frequency of every possible six nucleotide, seven nucleotide and eight nucleotide sequence in each of 3' UTRs contained in the complete set of ensembl 3' UTRs was calculated (using a script written by Dr. A. Enright) and their distributions plotted (Figure 5.11). The average frequency of each of the hit hexamer seeds (Figure 5.11a), both the 7mer-A1 and 7mer-m8 hit heptamer seeds (Figure 5.11b and Figure 5.11c) and the hit octamer seeds (Figure 5.11d) were marked on the plots.

The distributions of average log seed match frequencies for all possible seeds form bimodal normal distributions, with a large peak on the right-hand, higher-frequency end of the plot and a smaller peak on the left-hand, lower frequency end of the plot. The significance of the minor peak is unclear. One possibility is that the sequences in this peak have some biological function and are therefore selected against by evolution (e.g. they are miRNA targets). Another alternative is that the sequences in the minor peak could contain a different base composition to that generally found in 3' UTRs. The hit seeds all fall towards the higher end of the total range of frequencies. However, hit seed sequences are mostly located towards the centre of the major peak. That is, if those sequences which are unusually under-represented are not considered, the hit seed sequences have neither an unusually high or low frequency in 3' UTRs compared to average frequencies of all possible seed sequences. An exception to this is the heptamer/octamer sequence GCGATCA(A), which is clearly in the part of the distribution containing under-represented sequences.

The frequencies of the enriched seed sequences were plotted on the same distributions (Figure 5.12). Unlike the hit seed sequences, the enriched seed sequences cluster towards the higher end of the major peak. This effect becomes stronger as the length of the sequence increases. This suggests that those seed sequences that are enriched in the high scoring siRNAs have a larger number of off-target effects, increasing the chance that they will knock-down a larger number of genes connected to the TRAIL-induced cytotoxicity pathway.

5.7.3 Occurrence of hit seed sequences in genes previously associated with the TRAIL pathway

It is possible that siRNAs which cause a large change in TRAIL-induced cytotoxicity and contain a hit seed may be exerting their effect in part by reducing the levels of genes demonstrated to be associated with the TRAIL pathway either here, or previously, (it is likley

that part of the effect may still be due to reduction of the intended target).

Seed Sequence	Gene hit at least once	Genes hit at least twice
6mers		
CAAGGT*	INADL, TEGT, IRAK1, PRKAA2, FBXO11, MAPK10	INADL
TGTCCA	INADL, **DIABLO**, **BID**, IRAK1, **FADD**	INADL
ACTTGA*	INADL, IGF1R, **BID**, **TNFRSF10B**, PRKAA2, **TNFRSF10A**, PRKCD, PRKCQ, FBXO11	INADL, **TNFRSF10B**, **TNFRSF10A**, PRKCQ, FBXO11
AGATCA*	INADL, TEGT, **TNFRSF10B**, PRKAA2, PRKCD, PRKCQ, FBXO11	INADL, **TNFRSF10B**, PRKAA2
GCATTA	INADL, **TNFRSF10B**, PRKAA2, IKBKE, **TNFRSF10A**, WDFY4, FBXO11	**TNFRSF10B**, PRKAA2, IKBKE, **TNFRSF10A**, FBXO11
7mer-A1		
CAAGGTA*	INADL, TEGT	
TGTCCAA	IRAK1	
ACTTGAA*	**BID**, PRKAA2, **TNFRSF10A**, PRKCD, PRKCQ, FBXO11	**TNFRSF10A**, FBXO11
AGATCAA*	PRKCQ	
GCATTAA	**TNFRSF10B**, PRKAA2, **TNFRSF10A**, WDFY4, FBXO11	
7mer-m8		
TCAAGGT*	INADL, TEGT	
GTGTCCA	IRAK1, **FADD**	
AACTTGA	**TNFRSF10B**, **TNFRSF10A**, PRKCQ	
CACTTGA*	INADL, IGF1R, **TNFRSF10B**, **TNFRSF10A**	INADL, **TNFRSF10A**
GAGATCA*		
8mer		
TCAAGGTA*	TEGT	
GTGTCCAA	IRAK1	
AACTTGAA	**TNFRSF10A**	
CACTTGAA*	**TNFRSF10A**	
GAGATCAA*		

Table 5-10 Genes associated with the TRAIL pathway with 3' UTR matches to "hit" seeds
Human 3' UTR sequences were retrieved from ensembl 46. The UTRs of genes previously associated with the TRAIL pathway, or confirmed hits from the two screens presented here, were searched for matches to seed sequences which were either repeated in the siRNAs targeting the top 20 siRNAs, or appeared in both these siRNAs and the siRNAs used to confirm the hits from the kinase and phosphatase screen. Genes in bold are 'core' TRAIL genes, which were not associated with the TRAIL pathway through RNAi screening. *Seeds found in two or more confirmed siRNAs.

In order to discover if siRNAs which cause a large change in TRAIL-induced cytotoxicity and contain hit seed sequences may be reducing the level of genes previously

associated with the TRAIL pathway, the 3' UTRs of these genes were searched for matches to the hit seed sequences. Genes shown to be involved in TRAIL mediated apoptosis, either previously or in this work, with either one or two matches to hit seed sequences in their 3'UTR are shown in Table 5-10. All of the hit hexamer seed sequences are found twice in the UTRs of at least one of the genes associated with the TRAIL pathway, other than the one the siRNA containing the seed is designed to target. All of the hit 7mer-A1 and 4 of the 5 hit 7mer-m8 seeds are found in at least one of the UTRs of genes associated with the TRAIL pathway, other than the one the siRNA containing the seed is designed to target. For each type of heptamer, one seed is found twice in the UTR of two of these genes. Similarly, matches to all but one of the 8mer seeds are found in the UTR of at least one of these genes. In order to assess the significance of this, the 3' UTRs of these genes were searched with all the seed sequences appearing in the library, and the percentage of seeds that were found either once or twice in at least one of these genes recorded (Table 5-11). More than 75% of hexamer or either of the types of heptamer seed are found one or more times in the 3'UTR of at least one of the TRAIL genes. Large numbers of seed sequences are also found twice at least one of the TRAIL pathway genes (72% for 6mer seeds, 24% for 7mer-A1 seeds and 22.7% for 7mer-m8) seeds.

The gene set used above includes genes associated with the TRAIL pathway through the use of RNAi screening. Removing these genes produces a smaller set of genes (first column, Table 1-2), the 'core' genes. Genes from this 'core' set, which contain hit seed sequences are highlighted in Table 5-10, and the percent of all seed sequences from the library found in the 3' UTRs of one or more of these genes is shown in Table 5-11. A large number of seed sequence are found in the 3' UTRs of 'core' TRAIL genes, with 80% of hexamer seeds, 39% of 7mer-A1 seeds and 43% of 7mer-m8 being found in at least one.

Seed Type	At least one match per UTR		At least two matches per UTR	
	All genes	*Core genes*	*All genes*	*Core genes*
6mer	94.7%	80.1%	72.7%	31%
7mer-A1	76.8%	39.3%	23.9%	4.25%
7mer-m8	80.0%	42.5%	22.7%	4.45%
8mer	41.6%	13.5%	4.05%	0.476%

Table 5-11 Percentage of screen seeds found in the 3' UTR TRAIL genes
3' UTRs of genes associated with TRAIL-induced apoptosis, either here or previously (all genes) or genes with a well established role in TRAIL-induced apoptosis (Core genes, see Table 1-2) were searched for matches to all the seeds found in siRNAs used in the druggable genome screen.

This finding suggests that either a large number of the siRNAs used in the screen are knocking-down genes previously associated with the TRAIL pathway, or that the seed matches are not sufficient to specify off-target effects. Under either of these hypotheses the finding that high scoring siRNAs contain seeds which are found in the 3' UTR of genes

involved with the TRAIL pathway is not significant, as many other siRNAs in the library that did not score so highly also have seed sequences which are found in these 3' UTRs. However many of the hit seed sequences are found in the 3' UTR of not just one, but several genes associated with the TRAIL pathway. One explanation for the overrepresentation of these seed sequences in high scoring siRNAs could be that in part their effect on TRAIL sensitivity is due to the additive effect of small reductions in levels of transcripts for a number of genes involved in the TRAIL pathway.

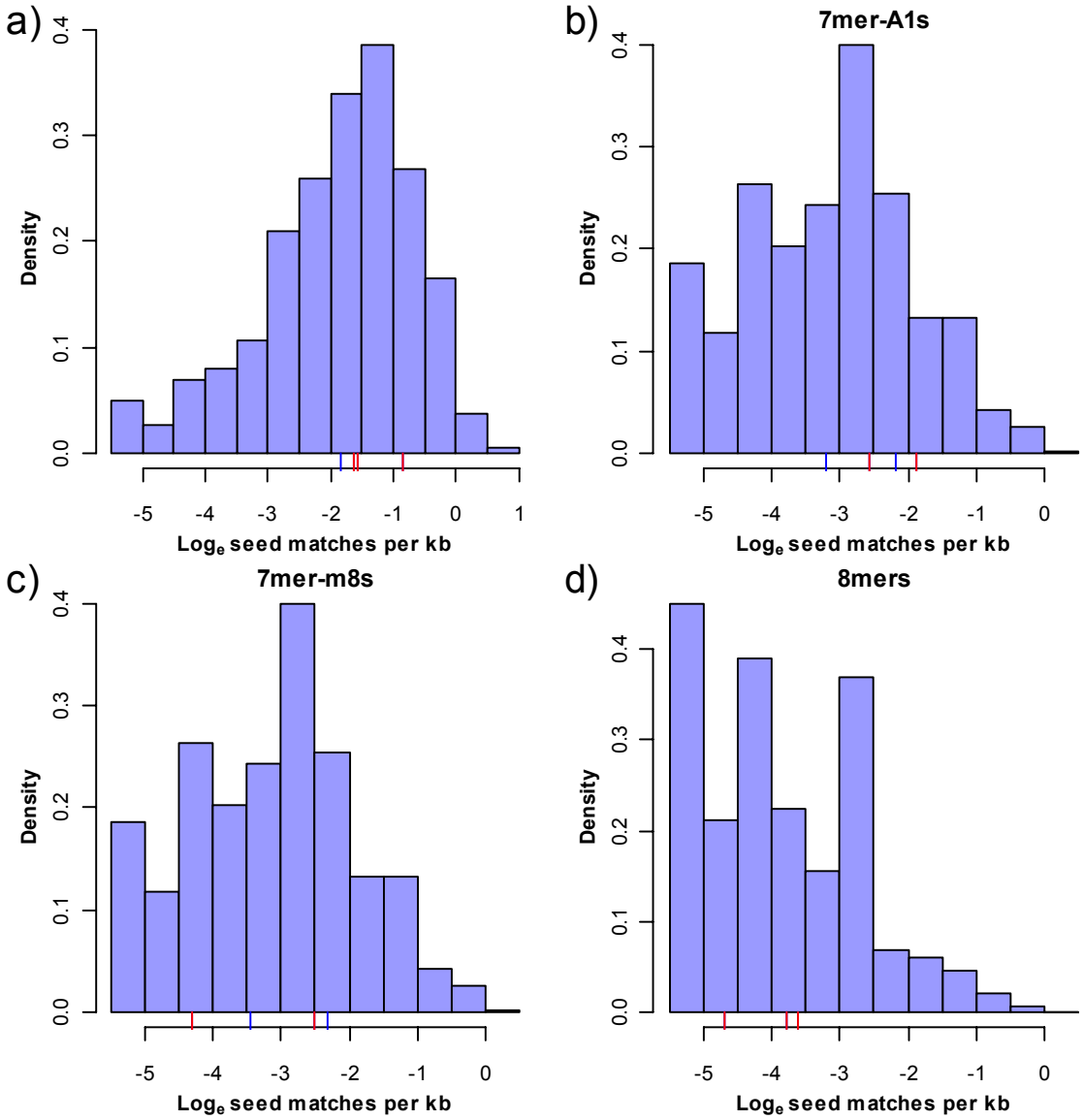

Figure 5.13 Average frequencies of "hit" seed sequences compared to all possible seed sequences in 3' UTRs of genes previously associated with the TRAIL pathway
Average frequencies of all possible seed sequences calculated as described for Figure 5.11, expect using 3' UTRs associated with the TRAIL pathway, either in this work or previously. Frequency of a) hit hexamers seeds(6mers), b) hit 7mer-A1 heptamer seeds, c) hit 7mer-m8 heptamer seeds or d) octamer (8mer) hit seeds are shown by tick marks under the histograms. Seeds found in two or more confirmed siRNAs are shown in red.

If siRNAs containing the hit seeds are causing a reduction in the level of transcripts

previously associated with the TRAIL pathway, then the average frequency of hit seed sequences in each of the 3' UTRs of these genes should be higher than other possible seed sequences of the same length.

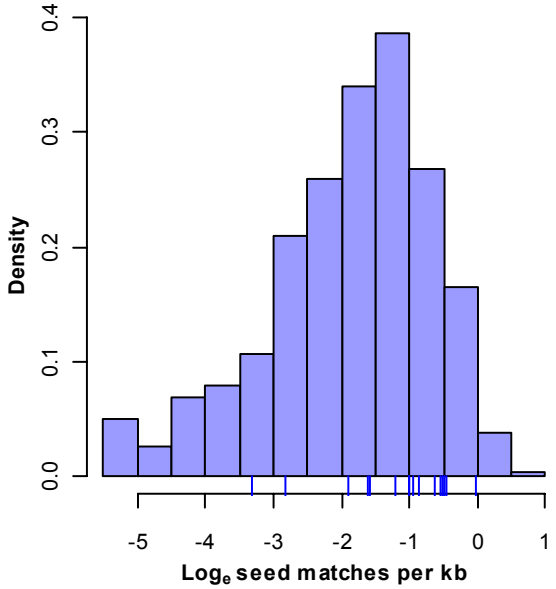

Figure 5.14 Frequency of "enriched" hexamer seeds compared to all possible hexamer seeds in 3'UTRs of genes previously associated with the TRAIL pathway
Average frequency of all possible hexamer seeds were calculated as for Figure 5.. Tick marks show frequency of "enriched" seed sequences

The distribution of average frequencies of all possible 6nt (Figure 5.a), 7nt (Figure 5.b/Figure 5.c) and 8nt (Figure 5.d) sequences in each of the 3' UTRs of genes associated with the TRAIL pathway either here or in previous work was calculated. The average frequencies of hit seed sequences were examined relative to these distributions. In all cases the average frequencies of the hit seed sequences were contained with the central portion of the distribution of all possible sequences. This was also the case when using only "core" genes (data not shown). This shows that the 3' UTRs of genes previously associated with the TRAIL pathway are not enriched for matches to the hit seed sequences, and therefore does not support the idea that siRNAs containing these seed sequences are more likely to knock-down genes previously associated with the TRAIL pathway than siRNAs containing other seed sequences. Examining the average frequencies of enriched seeds gives a similar result, with the possible exception of the enriched hexamers, which do cluster slightly to the right of the main peak, although the effect is weak (Figure 5.14, data not shown).

5.7.4 Identifying possible off-target transcripts for hit and enriched seeds

If the seed region of an siRNA is important in determining the specificity of the siRNA with regards to off-target effects, then, by searching for matches to seeds in 3' UTR

sequences, it may be possible to identify novel candidate transcripts that are unintended targets of these siRNAs. Lin *et al* used the two heptamer seed sequences that occurred in their top three hits to search transcripts that might be responsible for the effect observed (Lin et al. 2007). They found that 3,312 and 2,503 genes contained each of the heptamer seeds. To further reduce the number of candidates they looked at the overlap between these sets. The overlapping set contained 343 genes, including Mcl-1, a member of the Bcl-2 family. Bcl-2 proteins were known to be involved in the process being studied. It was confirmed that the siRNAs containing these seeds did knock-down Mcl-2, and new siRNAs targeting Mcl-2 scored well in their assay.

Ensembl human 3' UTR sequences (www.ensembl.org, release 46) were searched for matches to hit and enriched seed sequences. Matches to at least one of the five hit hexamers were found in 9,523 UTRs and matches to at least one of the 17 enriched hexamers were found in 14,827 UTRs (around 50% of the genes in the genome). There are 171 UTRs that contain matches to all 5 hit hexamer seeds and 54 that contain hits to all 17 enriched hexamers. These lists of hit can be reduced in several ways. The first is to require that each UTR contains two matches to each seed. There are 10 UTRs which contain two matches to all of the 'hit' hexamer seed sequences and 3 UTRs which contain two matches to all 17 of the enriched hexamer seeds. The second method for reducing the number of possible candidates is to use heptamers (either the hexamer with a flanking A – the 7mer-A1 site or the 7mer-m8 seed) or octamers (7mer-m8 with a flanking A) in the search. There are no UTRs which contain matches to all hit or enriched 7mer-A1 seeds or the 8mer seeds. There is one UTR hit by all of the hit 7mer-m8 seeds (the UTR of CCDC93) and two hit by all of the enriched 7mer-m8 seeds (the UTRs of FZD3 and THAP2).

Longer 3' UTRs can be expected to contain more matches to a given set of seed sequences than shorter sequences by chance. In order to correct for this, the total number of matches to any seed within one of the seed sets (hit hexamers/enriched 7mer-A1 heptamer seeds etc.) was normalised to the length of the UTR in kilobases. For each seed type UTRs were ranked according firstly to the number of independent hit or enriched seed sequences contained within the UTR and then by the frequency of matches to any of the hit or enriched seed sequences (Table 5-12 and Table 5-14). Table 5-13 presents the same analysis for seeds found in at least two confirmed siRNAs. These lists provide candidates for genes which may be involved in regulation of sensitivity to TRAIL-induced cytotoxicity but that were missed in the primary screen. For example, the 3' UTR of the gene AKAP11 contains

Rank	6mers			7mer-A1s			7mer-m8s			8mers		
	Gene Symbol	Seeds Matched	Match Frequency	Gene Symbol	Seeds Matched	Match Frequency	Gene Symbol	Seeds Matched	Match Frequency	Gene Symbol	Seeds Matched	Match Frequency
1	AKAP11	5	5.87	PPP3R2	5	1.79	CCDC93	5	1.05	Q9HCM6	3	1.22
2	NP_849153	5	5.02	AKAP11	4	3.67	CES3	4	2.38	IPO9	3	0.36
3	INADL	5	4.89	AGPAT5	4	2.20	CD300E	4	1.81	C3orf1	2	3.87
4	RNUXA	5	4.79	EDG2	4	1.85	TPMT	4	1.70	C10orf137	2	3.00
5	AGPAT5	5	4.40	ETS1	4	1.67	ZBTB43	4	1.25	Q86TT0	2	2.78
6	CRNKL1	5	4.38	KLK2	4	1.59	Q6ZT99	4	1.18	SLC25A20	2	2.57
7	KLK2	5	3.97	NP_001004299	4	1.56	CD59	4	0.98	MOSC1	2	2.06
8	NP_001009555	5	3.88	SOX1	4	1.39	HIVEP3	4	0.98	IRF2	2	2.01
9	CTTNBP2NL	5	3.73	ZNRF3	4	1.29	ATRN	4	0.94	Q4G197	2	1.96
10	EDG2	5	3.71	C1orf151	4	1.27	AKAP12	4	0.90	KCNV1	2	1.85
11	ATE1	5	3.69	FNDC3B	4	1.24	WHSC1	4	0.88	LOH12CR1	2	1.77
12	VAV3	5	3.68	TPD52	4	1.23	SEC31B	4	0.85	TNFRSF10A	2	1.75
13	CDV3	5	3.61	NP_001032309	4	1.19	IPO9	4	0.85	PNPLA3	2	1.69
14	XK	5	3.54	ZNF275	4	1.17	CXorf39	4	0.82	TOP1	2	1.68
15	SLC6A20	5	3.40	BRWD1	4	1.16	KCTD12	4	0.80	ARF1	2	1.67
16	Q6ZTR4	5	3.35	NP_001017980	4	1.16	MIB1	4	0.64	LEF1	2	1.65
17	ETS1	5	3.34	ZKSCAN1	4	1.11	AFF2	4	0.54	CPOX	2	1.62
18	MTRF1L	5	3.31	SPATA17	4	1.07	TNRC6B	4	0.40	NP_001013646	2	1.52
19	LDLRAD3	5	3.29	ALS2CR13	4	1.00	EIF2C3	3	3.96	TMEM155	2	1.47
20	PPP3R2	5	3.22	CD47	4	0.95	AKAP11	3	2.94	AKAP11	2	1.47

Table 5-12 Top 20 3' UTRs containing matches to "hit" seed sequences

Ensembl 3'UTRs were searched for matches to hit seed sequences. Both the number of independent seeds which matched the 3' UTR (seeds matched) and the total number of matches to any of the hit seeds were calculated. The total number of matches to any of the hit seeds was normalised to the length of the UTR (match frequency, in matches per kb). Shown are the top 20 3'UTRs ranked first by number of seeds matched and then by the match frequency.

Rank	6mers			7mer-A1s			7mer-m8s			8mers		
	Gene Symbol	Seeds Matched	Match Frequency	Gene Symbol	Seeds Match	Match Frequency	Gene Symbol	Seeds Match	Match Frequency	Gene Symbol	Seed Matches	Match Frequency
1	NP_612420.1	3	17.44	CROT	3	2.61	EIF2C3	3	3.96	AKAP11	3	2.20
2	KRBA2	3	8.35	HISPPD2A	3	2.48	K0415	3	0.78	LCP2	2	3.27
3	ZNF701	3	7.47	AKAP11	3	2.20	CCDC93	3	0.63	TMEM156	2	3.22
4	NOB1	3	6.52	AGPAT5	3	1.76	GSN	2	5.62	Q86TT0	2	2.78
5	POLR1E	3	6.06	RUNDC1	3	1.52	GPR34	2	4.05	CMBL	2	1.70
6	ZNF761	3	5.95	EDG2	3	1.39	ARPC2	2	3.58	ZNF696	2	1.68
7	C10orf58	3	5.57	O94914	3	1.22	RYK	2	2.94	HISPPD2A	2	1.65
8	ZNF600	3	5.11	KLK2	3	1.19	GJB6	2	2.74	KIAA0753	2	1.41
9	DBF4B	3	4.95	COL5A1	3	1.18	C18orf19	2	2.71	Q8N849	2	1.41
10	IL12RB2	3	4.93	FLJ43980	3	1.17	LONRF2	2	2.65	TSR1	2	1.34
11	SNX21	3	4.77	THOC5	3	1.16	FAM29A	2	2.25	GDF8	2	1.29
12	RPIA	3	4.67	CTSB	3	1.15	AKAP11	2	2.20	CINP	2	1.22
13	ZNF397	3	4.65	PPP1R15B	3	1.11	TCF7L1	2	2.18	ELAVL4	2	1.15
14	Q8N1I6	3	4.46	PPP3R2	3	1.07	C9orf57	2	2.17	SUV39H2	2	1.11
15	CMKLR1	3	4.43	C6orf107	3	0.99	ANGPTL7	2	2.09	DAP	2	1.10
16	Q6ZR34	3	4.38	C1orf151	3	0.96	MGC12966	2	2.08	FAM98B	2	0.95
17	PIPOX	3	4.38	FNDC3B	3	0.93	BTN3A3	2	1.99	PARP11	2	0.92
18	Q8N9A9	3	4.38	Q6ZSF1	3	0.92	ZNF784	2	1.98	EDG7	2	0.89
19	MRP63	3	4.35	PLCB1	3	0.91	Q8N1N1	2	1.96	SLC26A4	2	0.85
20	RASGEF1B	3	4.27	FLJ27459	3	0.89	C1orf167	2	1.93	TCF12	2	0.84

Table 5-13 Top 20 3' UTRs containing matches to "hit" seed sequences found in two or more confirmed siRNAs.
Ensembl 3'UTRs were searched for matches to hit seed sequences that are found in two or more confirmed siRNAs. Both the number of independent seeds which matched the 3' UTR (seeds matched) and the total number of matches to any of the hit seeds were calculated. The total number of matches to any of the hit seeds was normalised to the length of the UTR (match frequency, in matches per kb). Shown are the top 20 3'UTRs ranked first by number of seeds matched and then by the match frequency.

Rank	6mers			7mer-A1s			7mer-m8s			8mers		
	Gene Symbol	Seeds Matched	Match Frequency	Gene Symbol	Seeds Matched	Match Frequency	Gene Symbol	Seeds Matched	Match Frequency	Gene Symbol	Seeds Matched	Match Frequency
1	NR1D2	17	15.48	GABRA4	16	3.65	THAP2	13	5.38	THAP2	11	3.97
2	DCDC2	17	14.12	CPEB4	15	6.00	FZD3	13	3.11	FZD3	9	1.60
3	ARHGAP5	17	14.02	CXorf23	15	5.62	MAN1A1	12	5.36	TMEM106B	8	1.70
4	CENTB2	17	12.89	FZD3	15	5.60	LRRC8C	12	3.96	RAB22A	8	1.41
5	C1orf96	17	12.80	ARL5B	15	4.98	ACOT11	12	3.59	ZNRF2	7	4.76
6	ATP11B	17	12.76	CKAP5	15	3.88	CRSP2	11	4.86	EDG2	7	3.71
7	CXorf23	17	12.60	OGT	15	3.35	SLC30A1	11	4.45	SLAIN2	7	2.52
8	GABRA4	17	12.23	SLC1A2	15	3.30	TFEC	11	4.08	PPP3R2	7	2.51
9	FZD3	17	11.37	ACVR2A	14	6.76	SLC4A7	11	3.96	TMEM48	7	2.50
10	CPEB3	17	11.20	FLRT3	14	6.23	FSD1L	11	3.53	SLC4A7	7	2.23
11	LGALS8	17	10.79	MOBKL1A	14	6.05	REEP3	11	3.53	RAB11FIP2	7	2.21
12	NP_689969	17	10.72	LGALS8	14	5.19	NP_001017980	11	3.48	TPBG	7	1.95
13	PDE7B	17	10.72	CPD	14	5.17	EXOC5	11	3.43	ZBTB41	7	1.94
14	VAPA	17	10.63	TMEM106B	14	5.11	MIB1	11	3.34	EXT1	7	1.91
15	PGM2L1	17	10.57	PGM2L1	14	4.73	SYNCRIP	11	3.30	SOCS4	7	1.76
16	RBM12	17	10.56	PLCXD3	14	4.33	NP_872329	11	3.28	LRRIQ2	7	1.63
17	PDE10A	17	10.28	NHLRC2	14	4.27	CRB1	11	3.09	CRB1	7	1.63
18	GRIN3A	17	10.21	CRB1	14	4.07	VAPA	11	2.61	PLCXD3	7	1.55
19	ANGPT2	17	9.86	BRWD1	14	3.99	OGT	11	1.93	ARL5B	7	1.40
20	LANCL3	17	9.85	TLOC1	14	3.95	ENAH	11	1.92	KIAA2022	7	1.37

Table 5-14 Top 20 3' UTRs containing matches to "enriched" seed sequences

Number of matches and match frequency for each of the types of seed sequence was calculated as for Table 5-12 except using enriched seed sequences rather than hit seed sequences.

matches to all of the hit hexamer seeds, four of the five 7mer-A1 heptamer seeds, three of the five hit7mer-m8 heptamer seeds and two of the five hit octamer seed sequences. In each case, the frequency of these matches with in the 3' UTR of AKAP11 is high enough to place the gene in the top twenty for each of the seed types (Table 5-12). AKAP11 is an A-kinase anchor protein. Such proteins are involved in controlling the localization of protein kinase A. AKAP11 has been shown to form a complex with protein kinase A and GSK3β, thereby allowing protein kinase A to regulate the activity of GSK3β (Tanji et al. 2002). Regulation of GSKβ has been shown to be important for the regulation of TRAIL sensitivity in MYC over-expressing cells (Rottmann et al. 2005). Given the number of genes that have been associated with TRAIL-induced apoptosis, it is fairly likely that any collection of genes of the size of the collection in Table 5-12 there will be related to TRAIL-induced apoptosis in this sort of way. Interestingly, however, the 3' UTR of INADL is also found in the list of top 20 3'UTRs containing matches to the hit hexamers. INADL is one of the genes selected for confirmation from the screen and was confirmed in the confirmation experiments (Table 5-2 and Table 5-6).

These lists of possible effectors of the off-targets effects contain too many genes with a zero score to be able to test for enrichment of gene sets using GSEA. While it is not possible to test for enrichment of previously identified TRAIL pathway genes in high-scoring genes, it is possible to test for enrichment of these genes within genes meeting a certain significance criteria using a chi-squared test. The lists of genes containing a match to one or more of the hit or enriched seed sequences within their 3' UTRs were tested for enrichment of genes previously associated with the TRAIL-induced apoptosis pathway. The lists of genes containing a match to one or more of the hit hexamer or hit 7mer-A1 heptamers have a significant enrichment of genes previously associated with the TRAIL-induced apoptosis pathway ($p=0.044$ and $p=0.0201$). When only seeds that appear in one or more of the confirmed siRNAs are considered, only the enrichment of 7mer-A1 seeds remains significant ($p=0.0343$), but this could be due a reduction in the sample size of genes containing a match to the seed sequences. The list of genes whose 3' UTRs contain a match to one or more of enriched hexamers has a highly significant enrichment of genes previously associated with TRAIL-induced apoptosis ($p=0.0016$). Lists of genes whose 3' UTRs contain matches to other seed types (e.g. enriched heptamers, hit octamers etc) were not significantly enriched in genes previously associated with TRAIL-induced apoptosis at the 5% level.

5.7.5 Micro RNA seeds

mirBase is a database of miRNA sequences. In order to determine if any of the hit or enriched seeds from the screen are shared with natural miRNA sequences, the sequences of all human miRNAs were obtained from mirBase and searched for matches to the hit and enriched seeds from the screen. Four of the seeds from the screen are also found in natural miRNAs (Table 5-15). It is interesting to note that ACTTGA is the seed sequence of the human miRNA miR-26a, as this seed sequence appears four times in the top 20 siRNAs, and is the seed which scored second highest in the analysis of seed enrichment. ACTTGA is also found 9 times in the 3'UTR of DR5 (TNFRSF10B), 4 times in the 3' UTR of DR4 (TNFSRF10A) and 3 times in the 3'UTR of BID, all important genes in the TRAIL-induced apoptosis pathway. Counting the number of ACTTGA sites per kilobase of UTR in genes previously associated with TRAIL-induced apoptosis and in all 3'UTRs as a whole (which is not the same as the average frequency with which it appears in each 3'UTRs as calculated above) reveals an enrichment for this seed sequence in the 3' UTRs of genes previously associated with the TRAIL pathway. The frequency of the seed sequence ACTTGA is three times higher in UTR sequence from genes associated with the TRAIL pathway than it is in all UTR sequence. This could suggest that miR-26a could be involved in controlling the sensitivity of the cell to TRAIL.

Seed	*miRNA*
ACTTGA	has-miR-26a
GCATTA	has-miR-155
ACTGGA	has-miR-145
TAGGAA	has-miR-384

Table 5-15 "Hit" and "enriched" seed sequences also found in natural human miRNAs

5.8 Discussion and conclusions

A screen of siRNAs targeting genes in the druggable genome was conducted in order to identify genes which affected the sensitivity of cells to TRAIL-induced apoptosis. Genes targeted by the 20 highest scoring siRNAs were selected for confirmation. An initial re-screen of these 20 genes eliminated 3 genes for which neither siRNA from the library caused a significant reduction in sensitivity to TRAIL-induced cytotoxicity. The 17 remaining genes were examined in more careful confirmation experiments. Of the 17 genes, six were categorised as confirmed hits, five as unconfirmed hits, and five as 'off-targets'. In one case a categorisation could not be given, but later experiments suggested that this was also an off-target gene. Genes for which two siRNAs had a significant effect on TRAIL-induced caspase-3 activity were tested for their ability to affect the TRAIL-induce activity of other

caspases.

The seed sequences targeting potential hits, both from this screen and from the kinase and phosphatase screen, were examined. It was found that five hexamer and five heptamer (7mer-m8) seeds were found more than once in this set of siRNAs (hit seeds). A further 15 hexamer and 12 heptamer (7mer-m8) seeds were found to be enriched in high scoring siRNAs by GSEA (enriched seeds).

5.8.1 The Screen

Before analysis of the screening data was conducted, the relationship between pre-treatment viability and post-treatment survival was examined. A link between cell density and TRAIL sensitivity was observed previously in assay development expriments. It was observed that there was a link between pre-treatment viability and post-treatment survival in the data from the screen. While removing some of the data from the analysis might mean the loss of potentially interesting siRNAs that reduce both cell viability and TRAIL sensitivity, the relationship observed is likely to interfere with analysis further down the line. One of the consequences of removing such siRNAs from the screening data is normalised survivals for many siRNAs based on only available for one replicate. It was decided that in these cases the score for the siRNA should be based on this remaining data point.

A strong relationship was observed between the mean normalized survival of cells transfected with an siRNA and the standard deviation of survival between the two replicates. This relationship was also observed for the data from the kinase and phosphatase screen. However, unlike the data from the kinase and phosphatase screen, log transformation of the data in this screen showed a weaker (although clearly present) relationship. Thus the data from this screen was log transformed during the analysis procedure. The difference between this and the previous screen could simply be one of size, with the reduction in apparent relationship more clear here due to the increased number of data points.

Analysis of the positive control wells in the previous screen showed that there was a drop both in the dynamic range between siCasp8 and siNeg controls, and the Z'-factor between siCasp8 transfected and siNeg transfected wells when compared to assay development experiments. In order to try and reverse this reduction, the screen presented here was conducted using a higher concentration of TRAIL ligand (0.25μg/ml in the kinase and phosphatase screen and 0.5μg/ml in this screen). However, even with this increase in the concentration of the TRAIL ligand and also the quality control threshold used here (all plates with a dynamic range of less than 2 were repeated, albeit only once) the dynamic ranges of

plates used in the analysis of this screen followed a similar distribution to those from the kinase and phosphatase screen, and further, the Z' –factors between siNeg transfected cells and positive control transfected cells were lower than for the kinase and phosphatase screen (Z' factor for siCasp8, siBID and siSMAC were -0.95, -3.44 and -7.52 respectively here, compared with -0.35, -1.05 and -4.08 for the kinase and phosphatase screen). This suggests that the decrease in relative effect for the positive controls is due to the increase in throughput rather than the concentration of TRAIL. This effect could well be due to the increase in variation when throughput increases, possibly connected with the use of larger numbers of independent cell batches.

In many ways the results from this screen were similar to the results for the kinase and phosphatase screen. The correlations between replicates and between siRNAs targeting the same gene were similar for this screen to those from the previous screen, with the correlation between replicates ($r = 0.57$) being higher than the correlation between siRNAs targeting the same gene ($r = 0.075$). This again suggests that the screening process is more reliable for identifying siRNAs with an effect on TRAIL-induced apoptosis than for selecting genes involved in the process. This was reflected in the findings of the re-screen of siRNAs targeting genes targeted by the top 20 scoring siRNAs from the screen. 71% of siRNAs which had scored highly in the screen gave a statistically significant reduction in TRAIL sensitivity when re-tested in triplicate. In contrast only 38% of second siRNAs targeting these same genes gave a statistically significant reduction in TRAIL-induced cytotoxicity (or conversely 62% of these siRNAs recapitulated the lack of large effect seen in the screen). Six of the genes initially selected for verification by virtue of being targeted by siRNAs in the top 20 were eventually confirmed to be hits (using the definition of a confirmed hit given above). This is 30% of the genes targeted by the 20 top scoring siRNAs, exactly the same proportion of genes selected the same way for confirmation from the kinase and phosphatase screen that were eventually designated confirmed hits. It should, however, be noted that the criteria for confirmation here was slightly higher as siRNAs had to show a significant effect on TRAIL-induced Caspase-3 activity before they were classified as hits. This would suggest that the druggable genome screen has the same accuracy as the kinase and phosphatase screen.

Defining a threshold score for the designation of a hit based on the scores of positive controls proved to be of limited use in the analysis of the data from the kinase and phosphatase screen. Further, if genes affect the sensitivity of cells to TRAIL in a quantitative rather than qualitative way, which would seem at least plausible given that siRNAs affect

TRAIL sensitivity in a quantitative way, defining a binary boundary between hit and non-hits makes little sense. This presents a problem for defining the sensitivity of a screen in terms of the number of genes previously associated with the TRAIL pathway that are identified in the screen as hits. In order to provide some sort of measure of the success of the screen in identifying genes previously associated with the TRAIL pathway, Gene Set Enrichment Analysis (GSEA) was applied to the screening results. GSEA assesses the enrichment of a set or sets of genes within the high scoring portion of a ranked list of genes. The set of genes previously associated with the TRAIL pathway was enriched in the high scoring siRNAs, but with a low level of significance ($p = 0.087$). The enrichment score found was due to the effects of siRNAs targeting Caspase-8, BID, DR4 and MYC: genes which would be expected to have large effects on the sensitivity of cells to TRAIL-induced apoptosis. A similar analysis of the data from the kinase and phosphatase screen failed to identify an enrichment of genes previously associated with the TRAIL pathway in the high scoring siRNAs. This can at least partly be explained by the observation that the genes previously associated with the TRAIL pathway that are targeted by siRNAs in the kinase and phosphotase screen are mainly un-confirmed genes from the Aza-blanc *et al* screen, while the druggable genome screen contained siRNAs targeting well established genes in the TRAIL pathway, knock-down of which is known to have a large effect on TRAIL sensitivity. This identifies a further problem with the use of previously associated genes in assessing the sensitivity of a screen: the quality of the set of genes previously associated.

Controls demonstrate that in the screen, siRNAs targeting genes with known larger effects score more highly than genes known to have a smaller effect on TRAIL-induced apoptosis. However, it is impossible from the data presented to tell if efficiency of this screen in terms of whether the selection of genes using the screening data is better than selecting genes at random. Assuming that the results of the screen were no better than random, this would suggest that 30% of all genes included in the screen were involved in the TRAIL-induced apoptosis pathway. This is not so improbable as might be initially assumed. In a network model of cellular signalling, many or most genes have some effect on the network output. This could be tested by selecting a number of genes at random from the list of genes included in the screen and applying the same confirmation process to them as was applied to the candidate hits selected.

Two genes were identified where the effects of one of the siRNAs targeting them could be shown to be due to off-target effects. One of these was MAD, the binding partner of one of the confirmed hits MAX. MAD competes with MYC for MAX binding and

antagonises the effects of MYC (Luscher 2001). Worryingly, confirmation experiments identified three genes where the effects of siRNAs targeting them on TRAIL sensitivity could be ascribed to off-target effects, despite the fact that two independent siRNAs targeting each gene caused a significant reduction in TRAIL-induced cytotoxicity, as a third siRNA targeting each gene reduced the transcript levels further than either of the other two, without causing a reduction in TRAIL-induced cytotoxicity. Current best practice for RNAi experiments suggests that phenotypes should be confirmed by the use of at least two siRNAs targeting the same gene although some in the field argue for the use of three siRNAs (Echeverri et al. 2006). This finding supports the use of at least three siRNAs in confirming a phenotype, particularly when one of the siRNAs is from a screen, since screening can be shown to enrich for off-target effects (see below). This finding suggests that those confirmed hits which are targeted only by both library siRNAs (MYC, MAX and TEGT from this screen and Sharpin from the kinase and phosphatase screen) require further confirmation. Indeed, both siRNAs targeting TEGT are among the set of "hit" seeds that appear multiple times in the siRNAs which target genes selected for confirmation, suggesting that it is possible that the effect of both siRNAs targeting TEGT is due, at least in part, to off-target effects. This raises the question of how many siRNAs are necessary to confirm a hit. In theory it is possible even for genes targeted by three siRNAs to be shown to be off-targets due to a fourth siRNA not giving a phenotype. Conversely the finding that a third siRNA does not induce a phenotype because it is not as efficient at silencing the targeted transcript does not add any evidence either way. This suggests that real confirmation of the involvement of a gene in a process must come from a rescue experiment, where a non-silenceable form of the target gene is reintroduced into the cell, or, alternatively the involvement of the gene is confirmed using some other, non-RNAi, technique such as a small molecule inhibitor.

The increase in variability seen in the results of this screen compared to the smaller screen and assay development experiment clearly presents difficulties in the interpretation of data. Part of variance may be attributable to differences between the increased number of batches of cells used. It is possible that this could be decreased by sub-cloning the cells used in the experiment. This would allow the selection of a clone with a consistently lower level of surviving cells in the negative controls. If no such clone of HeLa cells were found, it may have been beneficial to test other TRAIL sensitive cell lines in the assay. Further, the power of the screen may have been increased by optimising the assay using a less powerful positive control. For example 2.5pmol of siCasp8 was sufficient to see an almost complete abrogation

of TRAIL sensitivity. However, high quantities may have been required to reliably distinguish siRNAs with a smaller effect on the sensitivity of cells from negatives. In both these screens the dynamic range between positive and negative controls was used to assess the quality of each plate. However, dynamic range does not account for the variance seen in the controls, only the average value. The Z'-factor is a better measure of the distance between two distributions, but relies on measures of variance which are not meaningful with only two replicates. In future it may be beneficial to include more replicates of positive (and possibly negative) controls on each plate. In order to make room for this, it would be necessary to include fewer controls in total. Finally, it is pertinent to return to the question of whether it would have been beneficial to perform the screen in triplicate rather than duplicate. While for an individual datum point, taking the minima of two points is more conservative, on average than taking the mean of three, it is not clear that this is the case when considering the ranking of all genes. Further, it could be argued that one of the issues with the screen is that the results were overly conservative – that is there was a high number of false negatives – arguing that a less conservative approach may have been beneficial. However, performing a third replicate on the druggable genome would have involved the expenditure of significantly more resources, leaving fewer resources available for confirmation and follow up work. One possibility would have been to perform three replicates, but not repeat poor-quality plates. It is clear that a third high-quality replicate in the kinase screen would have allowed the question of whether the use of the extra resources is justified, given the extra expense, to be addressed.

5.8.2 Hit and enriched seeds

The seed sequences of the top siRNAs targeting the genes targeted by the top 20 siRNAs and the siRNAs targeting candidate hits from the kinase and phosphatase screen were examined to look for sequences that appeared more than once. Five hexamer seeds appeared more than once in this set and one appears in four separate siRNAs. Three of these five sequences are found in two or more siRNAs that showed a statistically significant effect on TRAIL sensitivity in confirmation experiments. Five heptamer (7mer-m8, see Figure 1.4) seeds also appeared more than once in this set of siRNAs (3 if only confirmed siRNAs are considered). Together these seed sequences make up the set of "hit" seed sequences (Table 5-8). GSEA was used to look for seed sequences that were enriched in siRNAs that scored highly in the screen. This analysis identified 17 hexamer and 13 heptamer seed sequences where siRNAs containing these seeds were enriched in the high-scoring siRNAs. These seed

sequences define the set of "enriched" seeds. Two of the hit hexamer seed sequences and one of the hit heptamer seed sequences were also in the set of enriched seeds (Table 5-9). Because all of the siRNA guide strands in the library start with a U at the 5' end, all siRNAs that share a hexamer seed also share a 7mer-A1 heptamer seed and all siRNAs that share a 7mer-m8 heptamer seed also share an octamer (8mer) seed (Figure 1.4). These were generally found in siRNAs targeting genes categorised as unconfirmed or off-target.

This suggests that as hypothesised by Lin *et al*, the process of screening enriches for siRNAs with off-target effects which affect the process being studied (Lin et al. 2007). This is not unexpected, since screening selects for siRNAs which have a phenotypic effect on the assay. However, for the purpose of the assay there is no difference between an off-target and an on target effect. Therefore, if the library contains siRNAs which off-target genes involved in the process, then these siRNAs will inevitably be enriched in the top scoring siRNAs for the screen. If off-target effects are specified by the seed sequence of the siRNA, then since the library contains siRNAs contain around 2000 different hexamer seeds out of 4096 possible hexamer seeds, it is likely that multiple siRNAs which off-target any gene will be found in the library. It is important to note that not all siRNAs containing these sequences score highly in the screen, this suggests either the seed sequence is not the sole determinant of off-target specificity, that the off-target effects are mostly weak, or likely both.

As the length of a seed match increases its positive predictive power increases but the sensitivity of using it to predict off-target effects decreases. For example Birmingham *et al* took 84 mRNAs that were significantly down regulated by an siRNA and 84 which were not. They found that 84% of the down regulated transcripts contained a match to the siRNA hexamer seed, while 17% of the negative set had a match. These numbers were 69% and 8% respectively for heptamer matches (Birmingham et al. 2006). A similar effect is seen with increasing numbers of matches. Neilsen *et al* find that the effect of the number of matches is multiplicative and that a single octamar match has the same effect as two heptamer matches (Nielsen et al. 2007).

In order to examine what is causing siRNAs containing hit or enriched seeds to score highly in the screen, the average frequency at which these seed sequences appear in each of the 3' UTRs for all the human genes in ensembl was calculated and compared to the distribution of the average frequency of all possible seed sequences (Figure 5.11 and Figure 5.12). Hit seed sequences clustered in the central part of the main peak of the distribution of the average frequencies of all possible seed sequences. This suggests that hit seed sequences do not cause a larger number of off-target effects than would be expected by chance. In

contrast, enriched seeds tended to cluster towards the upper part of the frequency distribution of all sequences. This suggest that siRNAs contain these sequences may cause the off-target reduction of a higher number of transcripts than could be expected by chance. Such promiscuous siRNAs are likely to score highly in any screen as it is likely that the sum total effects of knocking down a large number of genes, even by a small amount, is likely to affect many processes. It suggests that siRNA design algorithms which do not already do so should take account of the frequency of seed sequences in 3' UTRs.

At least two matches to the 3' UTRs of at least one genes associated with the TRAIL pathway, either here or previously, were found for all of the hexamers. At least one match to the 3' UTR of at least one of these genes was found for all the 7mer-A1 heptamer seeds, four of the five 7mer-m8 seeds and four of the five octamer seeds (Table 5-10). Surprising however, a very high proportion of all seed sequences were found once in the 3' UTR of at least one of the genes associated with TRAIL sensitivity (94.7% of hexamers, 76.8% of 7mer-A1 heptamers, 80.0% of 7mer-m8 heptamers and 41.6% of octamers). These numbers remained high even when considering the percentage of all seed sequences that match twice in the 3' UTR of one of these genes, or if a more restrictive, higher confidence set of TRAIL genes was used (Table 5-11). These numbers are for seeds found in at least one of the 3' UTRs, while matches to the hit seeds are found in the 3' UTR of several of these genes. It was also found that the average frequency of matches of hit seed sequences to the 3' UTRs of genes associated with TRAIL-induced apoptosis was not any higher than seen for other seed sequences. Thus the siRNAs containing hit sequences did not score highly because they targeted a higher number of the genes previously associated with TRAIL sensitivity than other seed sequences. Performing the same analysis for enriched seeds showed that in general this also held true for the enriched heptamer and octamer seeds. However, the average frequencies of enriched hexamer seeds were clustered in a position slightly to the higher end of the distribution suggesting that these siRNAs may be hitting a larger number of genes associated with the TRAIL pathway than other seed sequences. This is not to say that the 3' UTRs of genes associated with the TRAIL pathway are unusual, rather, that there is a high likelihood of finding a match to any seed sequence in any collection of this many 3' UTRs. Indeed the mean frequency of all hexamers in 3'UTRs is 0.24 matches per kb, meaning a match to any hexamer would be expected in one in every four 1 kb 3' UTRs (Figure 5.11)

Four of the seeds from the hit and enriched seeds are also found in natural miRNAs. The seeds GCATTA and ACTGGA are the hexamer seed of miR-155 and miR-145

respectively. Deregulation of the expression of these miRNAs has been shown in B-cell lymphomas and colorectal neoplasia respectively (Eis et al. 2005, Michael et al. 2003). The seed ACTTGA is the seed sequence of miR-26a. The sequence ACTTGA is found in four of the top 20 scoring siRNAs, one of the siRNAs from the kinase and phosphatase screen and is also the second most enriched seed in high scoring siRNAs. This seed is found nine times in the 3'UTR of DR5, four times in the 3'UTR of DR4 and three times in the 3' UTR of BID. Indeed, per kb of 3' UTR sequence associated with TRAIL genes this seed is three times more frequent than found in total 3' UTR sequence. The frequency of a seed per kb of 3' UTR is subtly different from the average frequency in each 3' UTR. In the situation where a smaller number of UTRs have very high frequencies of seed matches, the former will be higher than the latter, as the latter averages out values from a small number of highly enriched UTRs. Repeating the analyses in 5.7.2 and 5.7.3 using total frequency over all 3' UTRs gives the same results as for average frequency in each 3' UTR except for this seed (data not shown). This suggests that possibly miR-26a is involved in control of sensitivity to TRAIL-induced cytotoxicity. Experiments to determine if siRNAs contain this seed do affect the level of DR4, DR5 and BID transcript / protein will show if this is indeed a possible explanation for the effect of these siRNAs. Experiments to examine the correlation between miR-26a expression and TRAIL sensitivity may also throw light on a possible role for this miRNA in controlling sensitivity to TRAIL-induced cytotoxicity.

Thus it seems likely that siRNAs containing enriched seeds score highly due to a larger number of weak off-target effects rather than particularly targeting genes involved in TRAIL-induced apoptosis. That is the screening process is enriching for genes with a large number off-target effect, rather than specifically enriching for siRNAs with off-target effects on genes involved in TRAIL-induced cytotoxicity. In contrast siRNAs containing the seed sequences ACTTGA may score highly through stronger off-target effects against core genes in the TRAIL pathway as well as the through effect on the intended target. The reason for the high score of siRNAs containing the other "hit" sequences remains unknown.

Despite this evidence that part of the effect of siRNAs containing these sequences may be attributable to off-target effects, several of the genes targeted by these siRNAs were confirmed through the action of independent siRNAs that do not contain over-represented seed sequences. One explanation for this maybe that while part of the effect elicited by the siRNA in question maybe due to off-target effects, part of the effect is also due to knock-down of the intended target. In support of this hypothesis, in all cases where a gene is targeted by two phenotypically active siRNAs, one of which contains an over-represented

seed, the siRNA with the over-represented seed causes a stronger phenotype, irrespective of the relative efficiency of the siRNAs in knocking-down the transcript of the targeted gene. The one hit for which this is not the case is TEGT. In this case both siRNAs targeting this gene contain hit seeds.

It may be possible to use the hit/enriched seed sequences to identify novel genes in the TRAIL pathway. Lin *et al* used the seed sequences from siRNAs scoring highly by off-target effects to identify Mcl-1 as a regulator of sensitivity to Bcl-2/Bcl-XL inhibitor ABT-737 (Lin et al. 2007). Transcripts with matches to heptamer/octamer or multiple hexamer seeds are possible off-targets for an siRNA containing that seed. However, each siRNA will have many off-target effects, only a small number of which may be involved in TRAIL-induced apoptosis. The chance of a gene being involved in the TRAIL pathway is increased by finding matches to multiple different hit/enriched seed sequences. There are many transcripts whose 3' UTRs contain matches to all of the hit hexamers or to all the enriched hexamers. These numbers can be reduced by requiring multiple hits per 3' UTR, which increases the probability that a transcript is affected by siRNAs containing the seed. Therefore transcripts were ranked first by the number of independent hit (Table 5-12) or enriched (Table 5-14) seeds with matches in the 3' UTR and then by the frequency of these matches as ranking purely by match frequency means that the highest ranking genes are ones with very short 3' UTRs containing one seed match This examination produces a large number of candidate genes. Further investigation will involve devising some form of measure of the significance of finding matches to multiple seeds in a 3' UTR, possibly similar to the method proposed by Nielsen *et* al (Nielsen et al. 2007) to give a smaller number of these genes which can be examined by experimental investigation. One such gene may be AKAP11 which has previously been shown to be involved in the regulation of MYC levels, which known to be involved in the regulation of TRAIL sensitivity.

5.8.3 The Hits

Six genes selected for confirmation from the screening results were categorised as hits by the definition above. That is they are targeted by at least two phenotypically active siRNAs, which are more efficient at reducing the level of the targeted transcript than any siRNA tested which targets the same transcript but was not found to be phenotypically active. In one case – MYC - one of these two siRNAs contained a seed sequence which was enriched in high scoring siRNAs. In a second case – TEGT – both of the siRNAs targeting this gene contained suspect seed sequences.

5.8.3.1 MYC and MAX

The MYC protein (also known as c-MYC) is a multifunction transcription factor, and a prototypical proto-oncogene. It has many roles connected with tumorigenesis including increased proliferation and regulation of both intrinsic and extrinsic apoptotic pathways and is up-regulated in many human cancer types (Reviewed: Nilsson, Cleveland 2003). It is well established that MYC is involved in sensitivity to TRAIL-induced apoptosis (Wang et al. 2004/5, Ricci et al. 2004, Rottmann et al. 2005, Wang et al. 2005). MYC suppresses the transcription of the cFLIP apoptosis inhibitor (Ricci et al. 2004), promotes the transcription of DR5(Wang et al. 2004/5). It has also recently been shown that MYC can inhibit the pro-survival functions of the NF-κB subunit RelA which is itself activated by TRAIL signalling (Ricci et al. 2007). Finally, a new model hypothesizes that MYC is involved in the "priming" of the mitochondrial pathway, thereby prompting this pathway to amplify the pro-apoptotic TRAIL signals (Nieminen, Partanen & Klefstrom 2007).

MAX is MYC's dimerization partner, and is required for both the transcriptional activating and suppressing functions of MYC (Reviewed: Luscher 2001). It is therefore unsurprising that it was also isolated from the screen.

Both siRNAs targeting both MAX and MYC caused a significant reduction in TRAIL-induced Caspase-3/7 activity (a condition of being classified a hit) and TRAIL-induced Caspase-9 activity. Both of the MYC and one of MAX siRNAs caused a significant reduction in Caspase-8 activity. The failure of one of the MAX siRNAs to cause a reduction could be due the large variation observed in the experiment (Figure 5.8a and Figure 5.9). The size of the effect of MYC knock-down on TRAIL-induced Caspase-8 activity suggests that the effect is direct – i.e. it is not the product of a feedback activation of Caspase-8 by other Caspases. This suggests that MYC does have effects on the TRAIL pathway at points other than reducing the inhibition, or priming, of the mitochondrial apoptosis pathway (Luscher 2001, Ricci et al. 2007).

5.8.3.2 IGF1R

The IGF1R (Insulin-Like Growth Factor Receptor) protein is a tyrosine kinase that regulates a number of pathways connected to cancer cell survival and proliferation (Reviewed: Tao et al. 2007). It has been shown that IGF1R signalling activates both Ras/Raf/ERK and AKT signalling pathways, both of which have been shown to be involved in the control of sensitivity to TRAIL (Chen et al. 2001, Frese et al. 2003, Nesterov et al. 2004, Thakkar et al. 2001, Wang et al. 2005). Indeed it was recently shown that

treatment of colon carcinoma cells with the ligand for IGF1R – IGF1 – increased the sensitivity of these cells to TRAIL-induced apoptosis (but protected against TNFα-induced apoptosis). This effect was dependent on the AKT pathway, but not the Ras activated ERK or p38 MAPK pathways and could be enhanced by blocking NF-κB (Remacle-Bonnet et al. 2005). This is surprising since previous reports have shown that activation of the AKT protects rather than sensitizes cells to TRAIL-induced apoptosis (Chen et al. 2001, Thakkar et al. 2001).

Here three siRNAs targeting IGF1R showed a significant reduction in TRAIL-induced Caspase-3/7 activity (Figure 5.8a). The finding that this effect was induced by all three siRNAs targeting IGF1R, in addition to the fact that none of these siRNAs contain any of the hit or enriched seed sequences adds confidence to the conclusion that the expression of IGF1R is involved in the sensitivity of cells to TRAIL-induced apoptosis. It is interesting to note that knock-down of IGF1R did not have such a large effect on Caspase-8 and Caspase-9 activity (Figure 5.9) as previous reports of the effect of AKT on TRAIL sensitivity have suggested that it acts to control BID cleavage (Chen et al. 2001, Thakkar et al. 2001), although it is possible that a small effect on Caspase-9 has an amplified effect on Caspase-3/7 activity levels. These results support a role for IGF1R/AKT in positively regulating Caspase-3 activation in TRAIL-treated cells in concurrence with the results of Remacle-Bonnet *et al.*

5.8.3.3 PDE11A

As with siRNAs targeting IGF1R, all three siRNAs targeting PDE11A caused a significant reduction in the level of TRAIL-induced Caspase-3/7 activity (Figure 5.8a). None of the siRNAs targeting PDE11A contained any of the hit or enriched seeds, thus allowing a high level of confidence in this gene. Again like IGF1R, knock down of PDE11A did not cause a significant reduction in the levels of TRAIL-induced Caspase-8 or Caspase-9 activity (Figure 5.9).

The PDE11A gene encodes for a duel specificity phosphodiesterase protein (Fawcett et al. 2000) and is widely expressed in many normal tissues and some carcinoma cell types (D'Andrea et al. 2005). Phosphodiesterases are involved in the linearization of cyclic AMP and GMP molecules which are important secondary messenger molecules in cellular signalling. PDE11A is unusual in that it hydrolyses both cAMP and cGMP (Fawcett et al. 2000). cAMP levels in the cell are important as they regulate the activity of Protein Kinase A and the cAMP response element binding (CREB) transcription factor which are both

activated by high levels of cAMP, and are therefore repressed by the action of phosphodiesterases. The gene for the anti-apoptotic protein Bcl-2 contains cAMP response element (CRE) in its promoter region and it has been shown that AKT induces expression of Bcl-2 through the action of the CREB transcription factor on this promoter. Since CREB activity is dependent on cAMP its activity could be reduced through the activity of PDE11A. This would lead to a reduction in the levels of Bcl-2 transcription and thus a reduction in the protection Bcl-2 provides against apoptosis (Pugazhenthi et al. 2000). It was noted above that AKT has been shown to have both pro- and anti-apoptotic roles (Chen et al. 2001, Remacle-Bonnet et al. 2005, Thakkar et al. 2001). Reducing PDE11A expression could serve to alter the balance in away from pro-apoptotic signals.

5.8.3.4 INADL

Two of three siRNAs targeting INADL caused a significant reduction in TRAIL-induced Caspase-3/7 activity. A third siRNA did not significantly alter TRAIL-induced Caspase-3/7 activity (Figure 5.8), however cells transfected with this siRNA had a higher level of INADL expression than cells transfected with either of the other siRNAs targeting INADL. siINADL.1 was one of the top 20 scoring siRNAs from the screen and contains the hit seed sequence ACTTGA shared between three other siRNAs from the top 20 scoring siRNAs as well as the siSharpin.1 siRNA used to confirm Sharpin as a hit in the kinase and phosphatase screen. This seed is also found in the miRNA miR-26a. siINADL.1 also shares its heptamer seed with one other top 20 scoring siRNA. This evidence suggests that at least part of the effect of siINADL.1 on TRAIL-induced cytotoxicity/Caspase-3/7 activity could be due to off-target effects. This is supported by the fact that siINADL.1 has a minimal effect on INADL transcript levels, although the fact that siINADL.2 increases the level of INADL transcript suggests that there is possibly some problem with the qPCR data here (Figure 5.9b). Set against this, siINADL.3, an independent siRNA that was not selected from the screen and does not contain a hit or enriched seed, did significantly reduce both the level of INADL transcript and TRAIL-induced Caspase-3/7 activity (Figure 5.9). Also the 3' UTR of INADL contains matches to all 5 hit seeds with the third highest frequency of matches to any of these seeds. This suggests that this gene at least warrants further investigation, even if the authenticity of its involvement in the regulation of TRAIL-induced apoptosis is currently unclear.

INADL is the human homolog of the *Drosophila* gene Inactivation No Afterpotential D (*Ina-D*). It contains 9 distinct PDZ domains (Pfam, Finn et al. 2006), which are domains

involved in protein/protein interaction. INADL is found at tight junctions in polarised cells (Lemmers et al. 2002) and has shown to be involved in directional migration of epithelial cells (Shin, Wang & Margolis 2007). *Ina-D* the *Drosophila* homolog of INADL is a scaffold protein involved in the organisation of signalling complexes at the cell membrane (Tsunoda, Zuker 1999). Phosphorylation of *frizzled* by atypical protein kinase C (*aPKC)* requires the *Ina-D* in drosophila cells, and it is hypothesised that *Ina-D* serves to anchor *aPKC* to *frizzled*. How INADL might affect the TRAIL pathway remains unclear.

5.8.3.5 TEGT

TEGT was isolated as a transcript homologous to a transcript expressed in the rat testis (Walter et al. 1994, Walter et al. 1995). The same gene was also isolated in a screen of human cDNAs in yeast which rescued Bax induced cell death (Xu, Reed 1998) . TEGT was shown interact with Bax and protect against apoptosis induced by several triggers of the intrinsic apoptosis pathway, but not FAS (Xu, Reed 1998). TEGT has been shown to up-regulated in breast cancer cells, and that its knock-down by RNAi leads to spontaneous apoptosis (Grzmil et al. 2006). Surprisingly expression of the *Arabidopsis* homolog of TEGT in human cells triggers a cell death which is blocked by overexpression of XIAP, an effect which could be due to a dominant negative effect on endogenous TEGT (Yu et al. 2002). TEGT is isolated here as a gene whose knock down protects cells from TRAIL-induced apoptosis. However, several other genes have been isolated that have both pro and anti-apoptotic functions.

TEGT was targeted by two siRNAs which cause a significant reduction in TRAIL-induced Caspase-3 activity. Both of these siRNAs are from the library and both contain seeds from the hit seed set. Taking into account the known biological functions of TEGT, caution must be taken in concluding that TEGT is a positive regulator of TRAIL-induced apoptosis until further confirmation experiments have been carried out.

5.8.4 Conclusions

An siRNA screen was executed to identify genes from the druggable genome which are involved in the regulation of TRAIL-induced apoptosis. The screen accurately identified a number of siRNAs which reproducibly affected the sensitivity of cells to TRAIL. Confirmation experiments allowed for the genes targeted by some of these siRNAs to be identified as novel regulators of TRAIL –induced apoptosis. These genes are from diverse cell pathways.

A screen of this size is a major undertaking, requiring a large amount of resources. The majority of the resources consumed by the screen are in the form of assay reagents and consumables such as plates, pipette tips and media rather than in the form of the siRNA library itself. It is also a major undertaking in terms of time. However, the cost, in both time and materials, of the confirmation and follow up work is as much, if not more than that of the actual screen. Several decisions were taken to save time and materials such as performing only two replicates of the screen. It is worth considering the question of if these decisions were necessary, the cost of performing the screen can be justified. While expensive, time consuming and possibly imperfect, the screen, together with the accompanying follow-up work did identify several genes that are unlikely to have been identified by other means.

Worryingly several genes were targeted by two siRNAs that significantly reduced the level of TRAIL-induced Caspase-3/7 activity but were classified as off-targets due to the lack of phenotypic activity of a third siRNA which reduced the level of targeted transcript more efficiently. This suggests that screening hits should be confirmed by more than two siRNAs, and that results from RNAi experiments should ultimately be confirmed by non-RNAi experiments.

That results should be confirmed by more than two siRNAs, or at least an additional two siRNAs from those used in the screen is also suggested by examination of the seed sequences of high scoring siRNAs. Several seed sequences appear more than once in siRNAs targeting the genes targeted by the top 20 siRNAs, or in siRNAs used to confirm hits from the kinase and phosphatase screen, and additional seed sequences were found that were enriched in high scoring siRNAs. This suggests that process of screening is enriching for siRNAs with relevant off-target effects as well as relevant on-target effects. Analysis of the average frequency of these seeds in 3' UTRs suggests that siRNAs that contain "enriched" seeds may induce more off-target effects than other seed sequences. In contrast "hit" seed sequences were not found at a higher average frequency in 3' UTRs generally or specifically the 3' UTRs of genes associated with the TRAIL pathway with the exception of the seed ACTTGA, many copies of which is found in the 3' UTRs of several genes associated with the TRAIL pathway and is also the seed sequence of the human miRNA miR-26a.

A large number of the seed sequences found in the library match the 3' UTR of genes associated with TRAIL-induced apoptosis. This holds even when a smaller, higher quality list of TRAIL associated genes is used. This suggests that many of the siRNAs in the library have the potential to affect the level of genes previously associated with the TRAIL pathway. In this case the effect of an siRNA on TRAIL-induced cytotoxicity would be the

sum of all of its off-target effects and its on-target effects. The highest scoring siRNAs would be those that affected TRAIL-induced cytotoxicity through both on- and off-target effects. This is supported by the confirmation of some genes targeted by siRNAs containing these seeds, suggesting that at least some of the effect of the original siRNA was due to on-target effects.