Evolution of the underlying model

The value returned by a probe is the consequence of the interaction between the probe's 25-base sequence and the proportions of the various sequences in the biological material to which the microarray is introduced. The simple model give in the paper has its genesis in a rather complex model, the ingredients of which we now define:

$I_{i,j}$	The observed intensity for probe i in a probeset intended to measure the expression level of gene G for CEL file j.
K_j	The file-specific scaling required because the quantity of genetic material avaiable for binding varies considerably from one CEL file to another
θ_i	The extra propensity of probe i (as opposed to an average probe) to stick to a random selection of genetic material. We choose the values so that $\sum_{\forall i} \log(\theta_i) = 0$.
γ_j	The proportion of fragments for CEL file j that correspond to gene G.
$\phi_j(M)$	The proportion of fragments for CEL file j that would bind with the motif M.
r_G	A multiplier, assumed the same for all probes, that applies to the binding of a probe to the gene for which it is intended.
r_M	A motif-specific multiplier that applies to the binding of a probe with material containing the reverse complement of motif M.
$\delta_i(M)$	A multiplier that takes the value 1 if probe i contains motif M, and is otherwise zero.

The algebra is greatly simplified if the usual error term is omitted until a final model is reached. Initially we have:

$$I_{i,j} = K_j \theta_i \left\{ \gamma_j r_G + (1-\gamma_j) \sum_{\forall M} r_M \delta_i(M) \phi_j(M) \right\}$$

If gene G is unexpressed then $\gamma_j = 0$ so that:

$$I_{i,j} = K_j \theta_i \left\{ \sum_{\forall M} r_M \delta_i(M) \phi_j(M) \right\}$$

Taking logarithms:

$$\log(I_{i,j}) = \log(K_j) + \log(\theta_i) + \log \left\{ \sum_{\forall M} r_M \delta_i(M) \phi_j(M) \right\}$$

We now standardise by summing over i, subtracting the overall mean and scaling to a unit variance. The latter makes comparisons across CEL files and across platforms easier. The resulting standardised values are given by:

$$S_{ij} = \log(\theta_i) + \log \left\{ \sum_{\forall M} r_M \delta_i(M) \phi_j(M) \right\}$$

Consider a particular 5-base motif, M_0, say, which typically occurs in tens of thousands of probes. Rewrite the relation as:

$$S_{ij} = \log\{r_{M_0} \phi_j(M_0)\} + \log(\theta_i) + \log \left\{ \sum_{\forall M \neq M_0} [r_M \phi_j(M) / r_{M_0} \phi_j(M_0)] \delta_i(M) \right\}$$

We now average over all probes containing M_0. Each probe will have a different relation, but all will have the same first term and the average of the remaining terms will be near zero. An approximation for the average, $\overline{S}_{M_0}$, is therefore:

$$\overline{S}_{M_0} = \log\{r_{M_0}\} + \log\{\phi_j(M_0)\} + \epsilon$$

In practice some of the probes containing M_0 will correspond to expressed genes, so that these probe values may be dominated by their $\gamma_j r_G$ component. However, because the average is being taken over so many probes, it is reasonable to assume that both

the overall contribution from those genes and the contribution from the $\log\{\phi_j(M_0)\}$ term can be jointly represented by the sum of a constant, μ, and an error, ϵ, giving the simple linear model

$$\overline{S}_M = \mu + \lambda_M + \epsilon.$$

In this formulation, each 5-base motif would have its own parameter, λ_M; but, in what follows, we write the 1024 individual parameters as linear combinations of a much smaller number of parameters that describe the base sequences in the various motifs, and thereby provide insights into the factors influencing probe values.

F-tests for the comparison of nested models

We illustrate the procedure using Model 4 and Model 4a. There are 1024 observations, while Model 4 contains 3 mononucleotide parameters, 15 dinucleotide parameters, and dummy parameters for GGGG and CCGCCTCCC. Model 4a includes one further parameter (the dummy for CCGCC). The ANOVA table is in outline as follows:

Source of variation	Degrees of freedom
Model 4	20
The CCGCC dummy	1
Residual	1002
Total about the mean	1023

Assuming Model 4a provides a reasonable fit to the data, the residual mean square provides an estimate of the experimental error variance (σ^2). If the CCGCC dummy parameter is not required, then the sum of squares corresponding to that term is also an estimate of σ^2, and the ratio of the two estimates is an observation from an F-distribution with the corresponding degrees of freedom (1 and 1002).

In practice, the estimate of σ^2 provided by the residual mean square is biased upwards, since the addition of further dummy parameters materially improves the fit. However, since Model 4 explains 85% of the variation in the 1024 group averages, the bias is very small. The F-test is, in any case, being used only as an indicator of dummy variables that may need to be added to the current model.

Further details concerning the detection of apparently relevant motifs

Model 5 explained an average 86% of the variation in the motif averages. Models 5a-5c were significant improvements on Model 5 (using the 1% level) only on 2.7%, 0.04% and 0.33% of occasions. Persistent outliers to Model 5 were ACCGC 69%; CCCGC 65%; GCCCC 61%; CGCCC 44%.; TCCGC 40%. Taking these together with CCGCC, which is already present in model 4a, we have the components of the 8-base motif CCCGCCCC, and also ACCGC and TCCGC. The next models considered are therefore:

Model 6: $v = M_6 + \epsilon +$ where $M_6 = M_4 + \beta_{CCCCCTC}d_{CCCCCTC} + \beta_{CCCGCCCC}d_{CCCGCCCC} + \beta_{(AT)CCGC}d_{(AT)CCGC}$

Models 6a to 6d: $v = M_6 + \epsilon +$ one of $\beta_{CCCGC}d_{CCCGC}$ to $\beta_{GCCCC}d_{GCCCC}$

Models 6e and 6f: $v = M_6 + \epsilon +$ either $\beta_{ACCGC}d_{ACCGC}$ or $\beta_{TCCGC}d_{TCCGC}$

Here the dummy variable $d_{(AT)CCGC}$ takes the value 1 for the motifs ACCGC and TCCGC and is otherwise zero.

Model 6 explained an average 87% of the variation in the motif averages. At the 1% level only d_{CCGCC} (on 4.3% of occasions) and d_{GCCCC} (3.2%) provided significant improvements on Model 6. On this occasion the more prominent outliers were those that were consistently over-estimated: GCCGG 80%; CCGGG 66%; AGGCC 62%; and GGCCA 43%. They suggest examination of the following models

Model 7: $v = M_7 + \epsilon +$ where $M_7 = M_6 + \beta_{GCCGGG}d_{GCCGGG} + \beta_{AGGCCA}d_{AGGCCA}$

Models 7a: $v = M_7 + \epsilon + \beta_{GCCGG}d_{GCCGG}$

Models 7b: $v = M_7 + \epsilon + \beta_{AGGCC}d_{AGGCC}$

Model 7 explains on average 88% of the variation in motif averages and Models 7a and 7b are not significantly better.

Table 1: GSEs having many outliers for Model 7.

GSE	$\overline{R^2}$	No. files	Motifs with outlier values of 0.5 (s.d.) or more	Apparently underlying motifs
7451	51	20	AGATC (3); GATCT (9); ATCTC (11); TCTCC (14);	
			CTCCC (15); TCCCC (10)	AGATCTCCCC
			CCAGC (10); CAGCA (5);	CCAGCA
			TCCCA (3)	TCCCA
9844	54	38	GATCT (26); ATCTC (36), TCTCC (38); CTCCC (38); TCCCC (37)	GATCTCCCC
			CCAGC (16);	CCAGC
			TCCCA (11); TCCCT (8);	TCCCA; TCCCT
11524	77	30	ATCTC (7); TCTCC (24); CTCCC (24); TCCCC (18)	ATCTCCCC
			TCCCT (2)	TCCCT
7538	79	24	ATGCG (3); TTGCG (12); TGCGG (16); GCGGT (16);	
			CGGTC (16); GGTCT (12); GTCTC (8)	(A/T)TGCGGTCTC
3678	82	14	TGCGG (4); GCGGT (8); CGGTC (5); GGTCT (2)	TGCGGTCT
2125	83	45	TCTCC (10); CTCCC (4); TCCCC (4)	TCTCCCC
5850	81	12	TTTTT (5)	TTTTT
			TGCGG (1); GCGGT (1)	TGCGGT
2634	82	17	ACGCC (3); TCGCC (11); CGCCG (17); GCCGC (9);	(A/T)CGCCGC
			ACTGG (3); CTGGC (7);	ACTGGC
6982	82	2	CCGTC (2); TCTCC (2); TCGCC (2); CCGCT (2)	
			CTCCT (2); TGCCT (2)	
6519	83	12	TCCCC (9); CTCCT (7); TCTCC (3)	
10270	89	48	TTTTT (14)	TTTTT
10406	86	24	CCGTC (12); TCGCC (2); CCGCT (2)	
8121	87	75	CTCCT (14)	CTCCT
10609	86	91	TCGCC (7); ACGCC (5); CCGCT (5); TGCCT (5);	
			GCCGC (5): CCGTC (3)	
9692	86	45	CTCCT (7)	CTCCT
5816	86	58	TCTCC (6); CTCCC (6)	TCTCCC

Although Model 7 explains on average 88% of the variation in motif averages, there are a few experiments where the value of R^2 is far lower. We can identify the nature of the differences from the normal by examining the motifs that are severely under-estimated in these experiments. (All extreme outliers are under-estimates). The results are summarised in Table 1.

For our next model, Model 8 we simply added dummy variables to take account of these outliers, so that the resulting model provided an excellent fit ($R^2 > 80\%$) for nearly every CEL file examined.

We next turned our attention to other Affymetrix platforms, commencing with the final model chosen for the Plus2 platform, proceeding in a corresponding fashion, and adding further terms as the results dictated. The result was a cumbersome model with 40 dummy variables, many of which appeared generally unnecessary. To focus on the motifs of greatest relevance, we tested each parameter to determine whether its removal resulted in a worsening of the fit of the model that was statistically significant at the 0.01% level (an arbitrary choice, but necessarily extreme given the number of tests performed). Most parameters provided at least one result that was significant at that level, so a further requirement was that the significance level should be achieved for at least 10% of the CEL files for at least one of the nine platforms examined. This resulted in the model with 26 dummy variables listed in Table 3 of the paper.

The groups of experiments noted in Figure 2 of the paper

Group	Number of files	GSEs
A	47	2735, 3678, 4217, 4218, 5850, 6021, 6022, 7364, 7538, 9250, 9757, 9758, 9759, 9761, 9819, 9890, 10270
B	68	2125, 4824, 5816, 6573, 6695, 6798, 6872, 6969 7161, 7846, 8302, 8316, 8832, 8853, 9361, 9686 9834, 9835, 10070, 10479, 10575, 10709

Apart from their unusual values groups A and B are also distinguished by the size of the 39 experiments involved: they average just 3 CEL files per experiment, compared to an average of 34 CEL files for the remaining 290 experiments considered (though that figure is inflated by the several thousand CEL files from GSE2109, a major cancer study).

Full 10-platform version of Table 3

Entries are the mean parameter estimates for the 26 dummy variables forming part of a multiple regression model fitted to each of more than 28,000 CEL files. The units are standard deviations (sd) of logarithms of the raw data. Values of 0.25 sd or greater are shown in bold type.

Parameter	Human			Mouse	Arabidopsis	Barley	Rice	Soybean	Drosophila	
Array	U133A	U133A_2	U133+2	430_2	ATH1	Barley1	Rice	Soybean	DrosG	Dros_2
CCCCC	**0.56**	**0.41**	**0.55**	**0.56**	**0.32**	**0.37**	**0.57**	**1.01**	0.10	0.24
CCCGCCCC	**0.42**	**0.33**	**0.40**	**0.34**	**0.37**	**0.26**	**0.33**	**0.38**	**0.32**	**0.41**
CCGCCTCCC	**0.33**	**0.25**	**0.46**	**0.42**	**0.34**	0.17	**0.30**	**0.43**	0.18	**0.30**
TCGCCGCT	**0.25**	0.19	**0.25**	**0.28**	**0.27**	0.22	**0.25**	**0.33**	**0.36**	**0.26**
CCCCG	**0.32**	0.21	**0.26**	0.19	**0.29**	**0.26**	**0.25**	0.23	0.23	**0.28**
GGGG	**0.33**	0.09	**0.33**	**0.37**	-0.03	0.18	0.16	**0.44**	0.23	0.08
(AT)CCGC	0.20	0.17	0.24	0.23	0.21	0.12	0.20	**0.26**	0.23	**0.26**
GCCCG	0.15	0.14	0.10	0.17	0.19	0.18	0.20	0.14	**0.40**	**0.31**
AGGCCA	-0.21	-0.17	-0.20	-0.18	-0.17	-0.10	-0.11	-0.18	-0.14	-0.13
CCCCTC	**0.30**	0.20	**0.28**	0.21	0.10	0.07	0.16	-0.04	0.08	
CTGCCT	0.20	0.15	0.19	0.20	0.12	0.13	0.16	0.17	0.15	0.12
CTGGCC	-0.15	-0.11	-0.16	-0.15	-0.18	-0.14	-0.14	-0.18	-0.08	-0.13
AACCC	-0.19	-0.12	-0.16	-0.19	-0.09	-0.09	-0.14	-0.07	-0.21	-0.10
TCGCTC	0.14	0.08	0.12	0.13	0.19	0.13	0.15	0.09	0.17	0.11
GGGGG	0.15	0.06	0.13	0.14	-0.04	0.10	0.10	-0.22	0.22	0.06
ACGCCA	0.13	0.10	0.14	0.14	0.16	0.05	0.12	0.12	0.12	0.15
NotAorT	-0.15	-0.14	-0.14	-0.12	-0.17	-0.06	-0.10	-0.10	-0.14	-0.07
TCCCC	0.21	0.12	0.20	0.12	0.10	0.11	0.02	0.11	-0.05	0.06
TCCCT	0.20	0.14	0.20	0.20	0.07	0.05	0.04	0.12	0.04	0.05
TGGGG	-0.15	-0.13	-0.15	-0.11	-0.12	-0.17	-0.11	-0.09	-0.07	0.02
GCTCCTCG	0.15	0.06	0.13	0.14	0.11	0.10	0.11	0.14	0.12	0.01
GGTTGCCC	0.12	0.09	0.08	0.09	0.10	0.07	0.08	0.13	0.17	0.06
GAACCA	-0.19	-0.10	-0.13	-0.12	-0.09	-0.04	-0.08	-0.05	-0.11	-0.05
GGTGCT	0.05	0.03	0.04	0.07	0.18	0.13	0.15	0.11	0.13	0.12
GCCCTCCG	0.12	0.06	0.11	0.12	0.06	0.09	0.09	0.16	0.07	0.05
GTGGTTC	0.06	0.06	0.06	0.07	0.15	0.06	0.10	0.17	0.10	0.06
Median R^2	89%	88%	91%	91%	86%	91%	90%	90%	94%	79%
No. of files	4753	2002	10000	1556	2288	1072	1356	3049	997	1186
No. of GSEs	114	69	322	107	160	19	51	62	62	

GSEs with consistently high or low estimates for the dummy parameters

Let $p_m(E)$ be the proportion of CEL files within experiment E for which β_m (for motif m) is significant at the 0.01% level. Let $M(p_m)$ be the median of the $p_m(E)$ values across the various experiments. Let N_E be the number of CEL files in experiment E. The criterion for inclusion in the tables that follow is that $N_E \times |p_m(E) - M(p_m)| \geq 10$, which implies that at least 10 CEL files in experiment E had atypical values for β_m. Because of the criterion for inclusion, the experiments with the largest numbers of experiments are the most likely to appear in a table and those involving less than 10 experiments cannot appear. In calculating the median, however, each experiment is equally weighted. These tables should be regarded as indicative rather than definitive.

HGU133A

Experiment	Estimate relatively rarely important	Estimate relatively commonly important
GSE473	CCCCG	GGTTGCCC, ACGCCA, GAACCA, TCCCC
GSE474		TCCC
GSE1133	(AT)CCGC, CCCCG	GCCCTCCG, GAACCA, TCCCC
GSE1159	GCTCCTCG, GAACCA, TCCCC	AGGCCA, ACGCCA, NotAorT
GSE1295		TCGCTC, ACGCCA, TCCCC
GSE1297	GAACCA	TCCCC
GSE1420	CCCCG	TCCCT
GSE1456	(AT)CCGC	AGGCCA, GAACCA
GSE1460	(AT)CCGC, AGGCCA, GCTCCTCG, CCCCG, GAACCA	GGTTGCCC
GSE1462	GCTCCTCG	
GSE1561		GGTTGCCC, GAACCA, TCCCC
GSE1577	(AT)CCGC, AGGCCA, CTGCCT, GCTCCTCG,CCCCC, CCCCG, GAACCA	
GSE1615	GGGG	
GSE1650	(AT)CCGC, GGGG	GCCCTCCG, TCCCC
GSE1722	(AT)CCGC, GCTCCTCG	GGTTGCCC
GSE1729		ACGCCA
GSE1786	(AT)CCGC, CTGCCT, GCTCCTCG, GAACCA	TCCCC
GSE1869		TCCCC
GSE1922		GGTTGCCC, ACGCCA
GSE1935		TCCCT, TCCCC
GSE2004	(AT)CCGC	
GSE2018	GAACCA	ACGCCA
GSE2044		TCCCC
GSE2113	(AT)CCGC	GCCCTCCG, GAACCA
GSE2189		GGTTGCCC, TCCCC
GSE2240		TCCCC
GSE2280	(AT)CCGC, GGGG	
GSE2328	(AT)CCGC, GGGG, CCCCG	GGTTGCCC, ACGCCA
GSE2351	CCCCTC, CCCCG	ACGCCA, TCCCC
GSE2361	GGGG	GCCCTCCG, GAACCA, TCCCC
GSE2443	(AT)CCGC, CTGCCT, TCGCCGCT, CCCCG, GAACCA	TCCCT, TCCCC
GSE2450		ACGCCA
GSE2485	(AT)CCGC, CTGCCT, TCGCCGCT, GCTCCTCG, CCCCC, CCCCG, GAACCA	TCCCT
GSE2742	GCTCCTCG	
GSE2990	GCTCCTCG, GGGG, TCCCC	AGGCCA, ACGCCA

Experiment	Estimate relatively rarely important	Estimate relatively commonly important
GSE3167	(AT)CCGC, CCCCG	GGTTGCCC, TCCCT, TCCCC
GSE3218	(AT)CCGC	ACGCCA
GSE3284	GGGG, (AT)CCGC, CCCCG	
GSE3307		AGGCCA, GAACCA
GSE3419		ACGCCA, TCCCT, TCCCC
GSE3494	GGTTGCCC, GAACCA, TCCCC	(AT)CCGC, AGGCCA, GCTCCTCG, GGGGG NotAorT
GSE3524	GAACCA	ACGCCA, TCCCC
GSE3780	(AT)CCGC	
GSE3790	(AT)CCGC, GGTTGCCC	AGGCCA, GCTCCTCG, NotAorT, GAACCA TCCCC
GSE3823	(AT)CCGC, AGGCCA, GCTCCTCG, CCCCG	GGTTGCCC
GSE3846		GCCCTCCG, TCCCT
GSE3860	(AT)CCGC, GGTTGCCC	
GSE3910	(AT)CCGC, CTGCCT, TCGCCGCT, CCCCG, GAACCA, AGGCCA, GCTCCTCG	TCCCT, TCCCC
GSE3911	CCCGCCCC, (AT)CCGC, CTGCCT, TCGCCGCT, GCTCCTCG, CCCCG, GAACCA	TCCCT, TCCCC
GSE3912	(AT)CCGC, CTGCCT, TCGCCGCT, CCCCG, GAACCA, CCCCC, AGGCCA, GCTCCTCG	TCCCT, TCCCC
GSE4045		GAACCA, TCCCC
GSE4127	GCTCCTCG	GGTTGCCC
GSE4271	GGGG	GGTTGCCC, GAACCA, TCCCC
GSE4475	(AT)CCGC, GGTTGCCC, GGGG, TCCCC	AGGCCA, GGTTGCCC, NotAorT, GAACCA
GSE4636	GAACCA	
GSE4698		GCCCTCCG, GGTTGCCC, GAACCA
GSE4824	(AT)CCGC, GGTTGCCC, GGGG, GAACCA	GGTTGCCC
GSE4917	CCCCG	CTGGCC, GCCCTCCG, ACGCCA, TCCCT GGGGG, TCCCC
GSE4922	GGTTGCCC, GAACCA, TCCCC	(AT)CCGC, AGGCCA, GCTCCTCG, GGGGG NotAorT
GSE5258	(AT)CCGC, GGTTGCCC, GGGG, TCCCC	AGGCCA, GCTCCTCG, NotAorT, GAACCA

There are 60 experiments listed in the table above. There were 18 other experiments that contained at least 10 CEL files. These were: GSE1000 (10 CEL files), GSE1140 (14), GSE1318 (10), GSE1364 (21), GSE1455 (18), GSE1648 (11), GSE1937 (12), GSE2060 (12), GSE2152 (22), GSE2225 (18), GSE2395 (20), GSE2487 (10), GSE3183 (15), GSE3585 (12), GSE3772 (10), GSE4646 (23), GSE4885 (12), and GSE5090 (17).

The following experiments contained between 100 and 350 CEL files: GSE1133, GSE1159, GSE1456, GSE2351, GSE2990, GSE3218, GSE3307, GSE3494, GSE3790, GSE3846, GSE3912, GSE4271, GSE4475, GSE4922 and GSE5258.

HGU133A_2

Experiment	Estimate relatively rarely important	Estimate relatively commonly important
GSE10040	GGGG	ACGCCA
GSE10087		GGTTGCCC, GCTCCTCG
GSE10174	CCGCCTCCC, CCCCTC, CCCGCCCC, AGGCCA, CTGCCT, TCGCCGCT, GGGG, CCCCC, CCCCG, (AT)CCGC	TCCCC
GSE10240	(AT)CCGC	GGTTGCCC
GSE10474	(AT)CCGC, GGGG	GCCCTCCG, GCTCCTCG
GSE10797	CCGCCTCCC, CCCCTC, CCCGCCCC, AGGCCA, CTGCCT, TCGCCGCT, GGGG, CCCCC, CCCCG, (AT)CCGC	GGGGG
GSE10804		GGTTGCCC
GSE10841	GGGG, CCCCG	CTGGCC, GGTTGCCC, GCTCCTCG, TCCCT
GSE10911		GGGGG
GSE11011	CCCCG	GGTTGCCC
GSE11630	GGGG	
GSE11792	GGGG	
GSE11889	AGGCCA, NotAorT, CCCCG	GCCCTCCG, GCTCCTCG, ACGCCA, TCCCT, GGGGG, TCCCC
GSE11903	GGGG	GCCCTCCG, GGTTGCCC, GCTCCTCG, ACGCCA
GSE11904	CCGCCTCCC, CCCCTC, (AT)CCGC, AGGCCA, CTGCCT, TCGCCGCT, GGGG, CCCCC, CCCCG	
GSE12109	GGGG	
GSE12211	(AT)CCGC	GGTTGCCC
GSE12438	(AT)CCGC, GGGG	GGTTGCCC
GSE12626	CCGCCTCCC, CCCCTC, (AT)CCGC,ACGCCA GGGG, CCCCG	GGTTGCCC, AGGCCA, NotAorT
GSE12666	GGGG	
GSE12682	CCCCG	CTGGCC, GCCCTCCG, GCTCCTCG, ACGCCA TCCCT, GGGGG, TCCCC
GSE12868	CCCCG	CTGGCC, GCCCTCCG, GCTCCTCG, ACGCCA, TCCCT, GGGGG, TCCCC
GSE13009		ACGCCA, TCCCC
GSE13162		TCCCC, CTGGCC
GSE13267		CTGGCC, ACGCCA, TCCCC
GSE13996	GGGG, CCCCG	ACGCCA
GSE14034	(AT)CCGC, GGGG	GCCCG, GGTTGCCC
GSE14098	CTGCCT	
GSE14107	(AT)CCGC, GGGG	GGTTGCCC, GCCCG
GSE14210	(AT(CCGC, GGGG,CCCCG	TCCCC, CTGGCC, GCTCCTCG, ACGCCA TCCCT, GGGGG, NotAorT
GSE14317		GCTCCTCG, GGTTGCCC, ACGCCA
GSE14323	GGGG, CCCCG	CTGGCC, TCCCC, GCCCTCCG, GCTCCTCG ACGCCA, TCCCT
GSE14330		GGTTGCCC
GSE14520		TCCCC

GSE12626 contained 465 CEL files and relates to a "*Genetic analysis of radiation-induced changes in human gene expression*." Two other GSEs (14210 and 14323 contain between 100 and 20 CEL files). Apart from the 35 experiments given above, the following experiments also contained at least 10 CEL files:

GSE10433 (12 CEL files), GSE10935 (12), GSE12100 (12), GSE13046 (16), GSE13577 (20), GSE14256 (10), GSE14325 (10), GSE14335 (10), and GSE1419 (16).

HGU133Plus2

Experiment	Estimate relatively rarely important	Estimate relatively commonly important
GSE2125	(AT)CCGC, CTGCCT, ACGCCA, GGGG, CCCCG	TCCCT
GSE2634	CCCCTC, AGGCCA, GGGG, CCCCC, CCCCG	GGTGCT
GSE2677	TCCCC	
GSE2817		GGGGG
GSE2842	TCCCC	
GSE3062		GGTTGCCC
GSE3077		GCCCTCCG, TCCCT
GSE3284	TCCCC	
GSE3325		TCCCT, GGGGG
GSE3678	CCGCCTCCC, CCCCTC, CCCGCCCC, (AT)CCGC, AGGCCA, CTGCCT, TCGCCGCT, GCTCCTCG, ACGCCA, CCCCC, CCCCG, TCCCC	
GSE3744		CTGGCC, TCCCT
GSE4036	GCTCCTCG, TCCCC	
GSE4183		CTGGCC
GSE4217	CCGCCTCCC, CCCCTC, CCCGCCCC, (AT)CCGC, AGGCCA, CTGCCT, TCGCCGCT, GCTCCTCG, ACGCCA, CCCCC, CCCCG, TCCCC	
GSE4218	CCGCCTCCC, CCCCTC, CCCGCCCC, (AT)CCGC, AGGCCA, CTGCCT, TCGCCGCT, GCTCCTCG, ACGCCA, CCCCC, CCCCG, TCCCC	GTGGTTC
GSE4237	ACGCCA	CTGGCC, TCCCT
GSE4488	ACGCCA	GCCCTCCG, GAACCA
GSE4600		CTGGCC, TCGCTC, GCCCTCCG, TCCCT
GSE4773	GGGG, TCCCC	GCCCTCCG, GGTTGCCC
GSE4780		CTGGCC
GSE4984	GCTCCTCG	
GSE5040	GCTCCTCG, TCCCC	
GSE5110	GCTCCTCG, ACGCCA	
GSE5116	CCCCG	
GSE5264		GCCCTCCG, TCCCT
GSE5281		CTGGCC, GGGGG, TGGGG
GSE5460		CTGGCC, TCCCT
GSE5547	CCCCG	TCCCT
GSE5563	GCTCCTCG	GGGGG
GSE5679	ACGCCA	GCCCTCCG
GSE5787	GCTCCTCG	
GSE5790	GGGG, CCCCG	
GSE5809	ACGCCA	
GSE5816	(AT)CCGC, ACGCCA, GGGG	GGTTGCCC
GSE5823		CTGGCC, GGGGG
GSE5850	CCGCCTCCC, CCCCTC, CCCGCCCC, (AT)CCGC, AGGCCA, CTGCCT, TCGCCGCT, GCTCCTCG, ACGCCA, CCCCC, CCCCG, TCCCC	
GSE5968		CTGGCC, GCCCTCCG, TCCCT

Experiment	Estimate relatively rarely important	Estimate relatively commonly important
GSE6004	CCCCG	CTGGCC, GCCCTCCG, TCCCT
GSE6013		GGTTGCCC
GSE6034	TCCCC	
GSE6207	(AT)CCGC, ACGCCA	GGTTGCCC, TCCCT
GSE6338		CTGGCC
GSE6351	ACGCCA	GCCCTCCG, GAACCA
GSE6519	CCCCG	CTGGCC, TCCCT
GSE6565	GCTCCTCG	
GSE6575		GAACCA
GSE6728	GGGG	
GSE6791		GAACCA
GSE6798	(AT)CCGC, CTGCCT, GCTCCTCG, ACGCCA	
GSE6872	(AT)CCGC, TCGCCGCT, ACGCCA, CCCCG	
GSE6960	GCTCCTCG, TCCCC	
GSE6962	GCTCCTCG	
GSE6969	(AT)CCGC, TCGCCGCT, ACGCCA, v	
GSE6972	GCTCCTCG	
GSE7011		TCCCT
GSE7023		CTGGCC, TCCCT
GSE7116		GCCCTCCG, GGGGG
GSE7158		CTGGCC, TCGCTC, GCCCTCCG
GSE7161	(AT)CCGC, ACGCCA	
GSE7216	(AT)CCGC, ACGCCA	GCCCTCCG
GSE7224		TCCCT
GSE7247		CTGGCC, TCGCTC, GCCCTCCG
GSE7305	GCTCCTCG, TCCCC	
GSE7392		CTGGCC, TCCCT
GSE7400		GGGGG, TCCCC
GSE7440	(AT)CCGC, CTGCCT, TCGCCGCT, ACGCCA	TCCCT
GSE7451	CCCCTC, CCCGCCCC, (AT)CCGC, AGGCCA, CTGCCT, TCGCCGCT, GCTCCTCG, ACGCCA, NotAorT, CCCCG	
GSE7462		GCCCTCCG, TCCCT
GSE7476		CTGGCC, GGGGG
GSE7486	TCCCC	
GSE7500	GCTCCTCG	CTGGCC

Experiment	Estimate relatively rarely important	Estimate relatively commonly important
GSE7509		TCCCT
GSE7538	CCGCCTCCC, CCCCTC, CCCGCCCC, (AT)CCGC, AGGCCA, CTGCCT, TCGCCGCT, GCTCCTCG, ACGCCA, CCCCC, CCCCG, TCCCC	GTGGTTC
GSE7553		GCCCTCCG
GSE7562	GCTCCTCG	
GSE7568	ACGCCA	
GSE7578	ACGCCA	CTGGCC
GSE7835	GGGG	AACCC
GSE7846	(AT)CCGC, GCTCCTCG, ACGCCA	
GSE7874	TCCCC	GAACCA
GSE7879		GCCCTCCG, TCCCT, AACCC
GSE7888		CTGGCC, TCGCTC, GCCCTCCG, TCCCT, AACCC
GSE7890		CTGGCC, GGTTGCCC
GSE7904		CTGGCC, TCCCT
GSE8023		TCCCT
GSE8049	ACGCCA	
GSE8066	GCTCCTCG	
GSE8121		GCCCTCCG, TCCCT
GSE8192	GCTCCTCG, TCCCC	
GSE8332	GGGG	
GSE8507		GCCCTCCG
GSE8514	GGGG	GAACCA
GSE8565	GCTCCTCG, GGGG	
GSE8586		GCCCTCCG
GSE8597	GCTCCTCG	
GSE8646	GGGG	GCCCTCCG, TCCCT
GSE8665		GAACCA
GSE8668		GCCCTCCG
GSE8671		CTGGCC
GSE8685	TCCCC	
GSE8687	ACGCCA, GGGG	
GSE8717		GGGGG, TGGGG
GSE8742	GCTCCTCG, TCCCC	
GSE8961		CTGGCC, GCCCTCCG, TCCCT

Experiment	Estimate relatively rarely important	Estimate relatively commonly important
GSE9101		GCCCTCCG, AACCC
GSE9103		GCCCTCCG, GAACCA
GSE9150	TCCCC	
GSE9171		GCCCTCCG, TCCCT, GGGGG
GSE9200		GCCCTCCG, TCCCT, GGGGG
GSE9250	CCGCCTCCC, CCCCTC, (AT)CCGC, AGGCCA, CTGCCT, TCGCCGCT, GCTCCTCG, ACGCCA, GGGG, CCCCC, CCCCG, TCCCC	
GSE9254		TCGCTC
GSE9438		TCCCT
GSE9526		CTGGCC, GCCCTCCG, TCCCT
GSE9599		GAACCA
GSE9647	GGGG	GCCCG
GSE9686	(AT)CCGC, GCTCCTCG, ACGCCA	GGTTGCCC
GSE9692		GCCCTCCG, TCCCT, GAACCA
GSE9709	ACGCCA, GGGG	AACCC
GSE9757	CCGCCTCCC, CCCCTC, CCCGCCCC, (AT)CCGC, AGGCCA, CTGCCT, TCGCCGCT, GCTCCTCG, ACGCCA, NotAorT, CCCCC, CCCCG, TCCCC	
GSE9758	CCGCCTCCC, CCCCTC, CCCGCCCC, (AT)CCGC, AGGCCA, CTGCCT, TCGCCGCT, GCTCCTCG, ACGCCA, NotAorT, CCCCC, CCCCG, TCCCC	
GSE9759	CCGCCTCCC, CCCCTC, CCCGCCCC, (AT)CCGC, AGGCCA, CTGCCT, TCGCCGCT, GCTCCTCG, ACGCCA, NotAorT, CCCCC, CCCCG, TCCCC	
GSE9761	CCGCCTCCC, CCCCTC, CCCGCCCC, (AT)CCGC, AGGCCA, CTGCCT, TCGCCGCT, GCTCCTCG, ACGCCA, NotAorT, CCCCC, CCCCG, TCCCC	
GSE9768	GCTCCTCG, TCCCC	
GSE9819	(AT)CCGC, AGGCCA, CTGCCT, GCTCCTCG, ACGCCA	
GSE9826	TCCCC	
GSE9844	CCCGCCCC, (AT)CCGC, AGGCCA, CTGCCT, TCGCCGCT, GCTCCTCG, ACGCCA, NotAorT, CCCCG	TCCCT
GSE9890	CCGCCTCCC, CCCCTC, CCCGCCCC, (AT)CCGC, AGGCCA, CTGCCT, TCGCCGCT, GCTCCTCG, ACGCCA, GGGG, CCCCC, CCCCG, TCCCC	
GSE9894	GGGG	

Experiment	Estimate relatively rarely important	Estimate relatively commonly important
GSE10070	(AT)CCGC, GCTCCTCG, ACGCCA, TCCCC	GGTTGCCC
GSE10270	CCGCCTCCC, CCCCTC, CCCGCCCC, (AT)CCGC, AGGCCA, CTGCCT, TCGCCGCT, GCTCCTCG, ACGCCA, CCCCC, CCCCG, TCCCC	GGGGG
GSE10311		TCGCTC, GCCCTCCG, TCCCT
GSE10315	GGGG	
GSE10406		CTGGCC, TCGCTC, TCCCT
GSE10410	TCCCC	
GSE10575	(AT)CCGC, ACGCCA , GGGG, TCCCC	
GSE10609		CTGGCC, TCCCT
GSE10700	GGGG, TCCCC	
GSE10709	(AT)CCGC, ACGCCA	GTGGTTC
GSE11510	GGGG	TCCCT
GSE11524	(AT)CCGC, CTGCCT, TCGCCGCT, ACGCCA	TCCCT
GSE11550		CTGGCC
GSE11552		CTGGCC

There were 198 experiments for which there were data for 10 or more CEL files. Of the 198, 141 appeared in the table above. These included GSE5460 (127 CEL files), GSE8332 (184 CEL files) and GSE8507 (141 CEL files). Most of the large experiments are included amongst the remaining 57 GSEs (with 10 or more CEL files) that were examined and were not judged to have unusual characteristics were numbers:

2109, 3202, 3526, 4888, 5350, 5372, 3062, 4107, 4498, 5058, 5060, 5081, 5675, 5764, 6054, 6088, 6269, 6281, 6364,

6532, 6764, 6885, 7127, 7152, 7153, 7268, 7307, 7434, 7586, 7621, 7967, 8052, 8527, 8581, 8596, 8658, 8702, 8762,

8977, 9086, 9089, 9090, 9091, 9195, 9196, 9264, 9452, 9517, 9762, 9770, 9832, 9865, 9891, 9899, 10358, 10586, 11525

Mouse430_2

Experiment	Estimate relatively rarely important	Estimate relatively commonly important
GSE1074	AGGCCA, CTGGCC	GCCCTCCG, GGGGG, AACCC
GSE1435	CCGCCTCCC, CCCCTC, AGGCCA, CTGCCT, GCTCCTCG, CCCCC	
GSE1479	ACGCCA	
GSE1871		GCCCTCCG
GSE1999	CTGGCC	
GSE2019	CCGCCTCCC, CCCCTC, (AT)CCGC, AGGCCA, CTGCCT, CTGGCC, GCTCCTCG, ACGCCA, TCCCT, GGGG, CCCCC	
GSE3100	TCCCT	
GSE3203	(AT)CCGC, CTGGCC, ACGCCA	GGTTGCCC, TCCCC
GSE3414	ACGCCA	GCCCTCCG
GSE3440	CTGGCC, TCCCT	GAACCA, AACCC
GSE3463	(AT)CCGC, CTGCCT, CTGGCC, ACGCCA, TCCCT	
GSE3653	AGGCCA, CTGGCC, GCTCCTCG, TCCCT	GGGGG, GCCCG
GSE3822	(AT)CCGC, CTGGCC, ACGCCA	
GSE4034	CTGGCC, ACGCCA, TCCCT	CCCCG
GSE4035		CCCCG
GSE4051	CCGCCTCCC, CCCCTC, (AT)CCGC, AGGCCA, CTGCCT, CTGGCC, GCTCCTCG, ACGCCA, TCCCT, CCCCC	GGGGG, CCCCG
GSE4098	CCCGCCCC, (AT)CCGC, AGGCCA, CTGCCT, CTGGCC, TCGCCGCT, GCTCCTCG, ACGCCA, NotAorT	TCCCC
GSE4189		GCCCTCCG
GSE4307	AGGCCA, CTGGCC, GCTCCTCG	AACCC, GCCCG
GSE4308	CTGGCC, GCTCCTCG	GCCCG
GSE4309	AGGCCA, CTGGCC, GCTCCTCG	AACCC, GCCCG
GSE4411	CTGGCC, GCTCCTCG, TCCCT	
GSE4481		GCCCTCCG, TCCCC
GSE4758	ACGCCA	GGGGG, CCCCG
GSE4774	CTGGCC, GCTCCTCG, TCCCT	GCCCG

Experiment	Estimate relatively rarely important	Estimate relatively commonly important
GSE5035	CTGGCC, ACGCCA, TCCCT	CCCCG
GSE5037	CCGCCTCCC, (AT)CCGC, CTGCCT, CTGGCC, GCTCCTCG, ACGCCA	CCCCG
GSE5128	GCTCCTCG	
GSE5198		GGTTGCCC, ACGCCA, TCCCT, GAACCA, GCCCG
GSE5202		GCCCTCCG
GSE5245	(AT)CCGC, CTGCCT, CTGGCC, ACGCCA	GGGGG, TCCCC
GSE5296	TCCCT	TCGCTC, GGGGG
GSE5324	AGGCCA, TCCCT	TCGCTC, CCCCG, GCCCG
GSE5500	GCTCCTCG, TCCCT	GGGGG
GSE6065		TCGCTC, GCCCTCCG
GSE6210	CTGGCC, ACGCCA	
GSE6223	CCGCCTCCC, (AT)CCGC, CTGCCT, TCGCCGCT, GCTCCTCG, ACGCCA, TCCCT	GGTTGCCC, GCCCG
GSE6290	(AT)CCGC, ACGCCA	GGTTGCCC, GCCCG
GSE6397		GCCCTCCG
GSE6398		GCCCTCCG, GGTTGCCC
GSE6399		GCCCTCCG
GSE6514		TCGCTC, GCCCTCCG
GSE6589	(AT)CCGC, ACGCCA	GGTTGCCC, GCCCG
GSE6595	ACGCCA	GCCCTCCG, GGTTGCCC, CCCCG
GSE6623	CCGCCTCCC, CCCCTC, (AT)CCGC, AGGCCA, CTGGCC, ACGCCA, TCCCT, CCCCC	
GSE6881	CTGGCC, TCCCT	TCGCTC, CCCCG, GCCCG
GSE6882	TCCCT	TCGCTC, CCCCG, GCCCG
GSE6916	TCCCT	TCGCTC, CCCCG, GCCCG
GSE6959	ACGCCA	TCCCC

Arabidopsis ATH1-12501

Experiment	Estimate relatively rarely important	Estimate relatively commonly important
GSE431		GGGG
GSE630	ATCCGC, CCCCG, GCTCCTCG, CCCCC ACGCCA, GGGG	
GSE680	ATCCGC, ACGCCA, CCCCC,CCCCG	AGGCCA, GCCCTCCG, GGGGG
GSE911	CCCCC	
GSE1051	CCCCG	
GSE1491	ATCCGC, ACGCCA, CCCCC, CCCCG	
GSE2169	ATCCGC, TCGCTC, CCCCC, CCCCG	GGGG
GSE2473	CCGCCTCCC, ATCCGC, ACGCCA, CCCCC, CCCCG, TCGCTC, GCTCCTCG	AGGCA
GSE3326	ATCCGC, CCCCC	
GSE3350		AGGCCA, CTGGCC, GGTTGCCC, GGGG
GSE3416	CCCCC	
GSE4733	ACGCCA	AGGCCA, CTGGCC
GSE4847	CCCCC	
GSE5520	ATCCGC, ACGCCA, CCCCC, CCCCG	GGTTGCCC
GSE5525		GCTCCTCG
GSE5530		AGGCCA
GSE5533	GGTGCT, GTGGTTC, CCCCG	
GSE5612		AGGCCA, GCTCCTCG
GSE5613		CCCCTC, AGGCCA, CTGGCC, GCTCCTCG
GSE5615		CCCCTC, GGGG
GSE5616		GGGG
GSE5617	GGTGCT	AGGCCA, CTGGCC, GCTCCTCG, CTGCCT ACGCCA
GSE5620	ACGCCA, CCCCC, ATCCGC,GCTCCTCG CCCCG	GGTTGCCC
GSE5621	CCCCC, ACGCCA, CCCCG	
GSE5622	TCGCTC, ACGCCA, CCCCG	GGGG, GGTTGCCC
GSE5623	ATCCGC, TCGCTC, ACGCCA, CCCCC CCCCG	GGTTGCCC, GGGG
GSE5624	ATCCGC, TCGCTC, ACGCCA, CCCCC CCCCG	GGGG
GSE5625	ATCCGC, ACGCCA, CCCCC, CCCCG	GGGG
GSE5626	ATCCGC, TCGCTC, ACGCCA, CCCCC CCCCG	GGGG
GSE5627	ATCCGC, ACGCCA, CCCCC, CCCCG	GGTTGCCC
GSE5628	ATCCGC, ACGCCA, CCCCC, CCCCG	
GSE5629		AGGCCA

Experiment	Estimate relatively rarely important	Estimate relatively commonly important
GSE5630	GCTCCTCG, ACGCCA, CCCCC	
GSE5631	ATCCGC, CCCCC, CCCCG	GGTTGCCC, CTGGCC
GSE5632	GCTCCTCG, ACGCCA, GGGG, CCCCG	AGGCCA, CTGGCC, GGTTGCCC
GSE5633	GCTCCTCG, ACGCCA, GGGG, CCCCG	
GSE5634	GGTGCT, CCCCG	AGGCCA, CTGGCC
GSE5636		AGGCCA, GCTCCTCG, CTGGCC, ACGCCA
GSE5637		AGGCCA, GCTCCTCG, CTGGCC, ACGCCA
GSE5638		AGGCCA, GCTCCTCG, CTGGCC, ACGCCA
GSE5685	GGTGCT, GTGGTTC, CCCCG	CTGCCT, GCTCCTCG, ACGCCA, GGGG
GSE5686	GCTCCTCG, CCCCC	
GSE5688	ATCCGC, ACGCCA, CCCCC, CCCCG	GGTTGCCC, CTGGCC
GSE5696	GTGGTTC	AGGCCA, CTGGCC, GCTCCTCG
GSE5701	ATCCGC, CCCCG	CTGGCC, GGTTGCCC
GSE5728	GGTGCT	
GSE5730	CCCGCCCC, ATCCGC, TCGCCGCT, TCGCTC, NotAorT, CCCCC, ACGCCA, CCCCG	TCCCC
GSE5738	GTGGTTC	CCCCTC, CTGCCT, CTGGCC, GGTTGCGG ACGCCA, GCTCCTCG
GSE5746	ATCCGC, ACGCCA, CCCCG	
GSE5748	ATCCGC, TCGCTC, CCCCG	
GSE5749	ATCCGC, GTGGTTC, CCCCC, CCCCG	AGGCCA, CTGCCT, CTGGCC, GGTTGCCC, GGGG
GSE5751	CCCCG	CCCCTC, CTGCCT
GSE5756		GGGG
GSE5757		GGGG
GSE5758		GGGG
GSE6150		CTGGCC
GSE6151	TCGCTC, CCCCG	AGGCCA, CTGCCT, CTGGCC, GGTTGCCC ACGCCA
GSE6160	ATCCGC	CTGGCC
GSE6161		AGGCCA, CTGCCT, CTGGCC
GSE6174	CCCCC	
GSE6176	CCCCC, GTGGTTC, CCCCG	GCTCCTCG
GSE6179		CTGCCT
GSE6203	CCGCCTCCC, ATCCGC, TCGCCGCT, GGTGCT, TCGCTC, GTGGTTC, CCCCC, CCCCG	GGGG
GSE6556		CCCCTC
GSE6825		AGGCCA, GCTCCTCG
GSE6828		CCCCTC, CTGCCT

There are 66 experiments listed above. There were a further 18 experiments that contain 10 or more CEL files. These are GSE631 (12 files), GSE2848 (12), GSE3056 (10), GSE5522 (12), GSE5684 (12), GSE5698 (12), GSE5737 (12), GSE5745 (12), GSE5752 (17), GSE5753 (16), GSE5754 (17), GSE5755 (17), GSE5770 (12), GSE6158 (12), GSE6169 (10), GSE6177 (26), GSE6826 (12) and GSE6832 (12). The two largest experiments were GSE5632 (66 CEL files) and GSE5630 (60).

Rice

Experiment	Estimate relatively rarely important	Estimate relatively commonly important
GSE4471		GGGGG
GSE6737		CTGGCC, GCCCTCCG, TCGCTC
GSE6893	GCTCCTCG	CTGGCC, GGTGCT, GCCCG
GSE6901		GCCCG
GSE10373	GGGG	TCGCTC
GSE10857	GCTCCTCG	CCCCTC, GGGGG
GSE11025	ATCCGC	
GSE11966	CCGCCTCCC, ATCCGC	
GSE13735		GCCCG
GSE14304	GGGG, GCTCCTCG	CTGGCC, GGTTGCCC, GCCCG, GGTGCT, TCGCTC, GCCCTCCG, NotAorT
GSE14692	ATCCGC	
GSE15071	CCGCCTCCC, CCCGCCCC, ATCCGC, CTGCCT, GGTGCT, TCGCTC, GCTCCTCG, NotAorT, CCCCC, CCCCG	
GSE16108		CTGGCC, GCCCTCCG, ACGCCA, GGGGG, CCCCTC, TCCCC
GSE16341	CCGCCTCCC, ATCCGC, CTGCCT, GGTGCT GCTCCTCG, CCCCC, CCCCG	AGGCCA, GGGGG
GSE16793		CTGGCC, GGTGCT
GSE17245		GCCCTCCG, TCGCTC
GSE18361	ATCCGC	
GSE19024	CCGCCTCCC, ATCCGC	CCCCTC, CTGGCC, GCCCTCCG, GGGGG GCCG, GGTTGCCC, TGGGG
GSE19239	ATCCGC	
GSE22564	ATCCGC, CTGGCC, TCGCCGCT, GGGG NotAorT, CCCCC, CCCCG	GGTGCT, GCTCCTCG, GCCCG
GSE24048	CCGCCTCCC, ATCCGC	
GSE24228	ATCCGC	

There are 22 Rice experiments in the table above. There are a further three experiments containing at least 10 CEL files: GSE7951 (13 CEL files), GSE12069 (14), and GSE15046 (12). Two GSEs,19024 and 22564, each contain about 200 CEL files.

Soybean

Experiment	Estimate relatively rarely important	Estimate relatively commonly important
GSE6414	ATCCGC, CTGCCT, CTGGCC, TCGCCGCT GCCCTCCG, GTGGTTC, NotAorT	TCCCT, TCCCC
GSE7124	CCCCTC, CTGGCC, GCTCCTCG	GGTGCT, TCGCTC, ACGCCA, NotAorT, GGGGG, GCCCG, CCCCG
GSE7511	ATCCGC, CTGGCC, TCGCCGCT, GCCCTCCG, NotAorT	TCCCT, TCCCC
GSE7881	ATCCGC, CTGGCC, TCGCCGCT, GCCCTCCG, NotAorT	TCCCT, TCCCC
GSE8112	ATCCGC, CTGGCC, TCGCCGCT, GCCCTCCG, GTGGTTC, NotAorT, CCCCG	TCCCT, TCCCC
GSE8432		ACGCCA
GSE9687	CCCCTC, CTGGCC, NotAorT	AGGCCA, CTGGCC, GGTGCT, GGGGG, GCCCG, NotAorT, CCCCG
GSE11611	CCCCTC, GCCCTCCG, GTGGTTC, ACGCCA, CCCCG	AGGCCA, GGTGCT, TCGCTC, GCCCG, GGGGG, NotAorT
GSE13631	CTGGCC, GGTTGCCC	GGTGCT, ACGCCA

Nine experiments are listed above. There are two other experiments with at least 10 CEL files: 9374, (25 CEL files) and 10251 (10 CEL files). The bulk of the CEL files (more than 2500) come from GSE11611, while both GSE7124 and GSE9687 include over 100 CEL files.

Drosophila — Drosgenome

Experiment	Estimate relatively rarely important	Estimate relatively commonly important
GSE3057	CCGCCTCCC, ATCCGC	
GSE3069	CCGCCTCCC	
GSE3830	CCGCCTCCC	
GSE3842	CCGCCTCCC	GGTTGCCC, GCTCCTCG
GSE3854		ATCCGC, CTGCCT, GGTTGCCC, GCTCCTCG
GSE4174		CTGCCT, GCTCCTCG
GSE4188	CCGCCTCCC, ATCCGC	
GSE4235		GCTCCTCG
GSE6515	CCGCCTCCC, ATCCGC	GCTCCTCG
GSE6542	ATCCGC	
GSE6558	CCGCCTCCC, ATCCGC	
GSE7110	CCGCCTCCC, ATCCGC, GGGG, GCCCG	GTGGTTC, GGTTGCCC,
GSE7159	GGGG	CTGCCT
GSE7655		GGGGG
GSE7873	GGGG	CTGCCT, GCTCCTCG
GSE9425	GGGG	
GSE9889		CTGCCT
GSE10012	ATCCGC	CCCCC, CCCCG
GSE10013		CCCCG
GSE10014	ATCCGC	CCCCC, CCCCG
GSE12477	GGGG	
GSE27376	GGGG	

The table above includes 22 experiments. There are 16 other experiments with at least 10 CEL files: GSE2780 (10 CEL files), GSE2828 (12), GSE3060 (12), GSE3826 (12), GSE3828 (12), GSE3829 (12), GSE3831 (12), GSE3832 (12), GSE6490 (12), GSE6491 (12), GSE6492 (12), GSE6493 (12), GSE8751 (30), GSE9088 (30), GSE9149 (29), and GSE11203 (24).

Drosophila — Drosgenome_2

Experiment	Estimate relatively rarely important	Estimate relatively commonly important
GSE2863	CCCCC	ACGCCA, GGGG
GSE5404	CCCCC	CTGCCT, ACGCCA, GGGGG
GSE5430		ACGCCA
GSE7614	CCCCC	
GSE7763	CCGCCTCCC, ATCCGC, NotAorT, CCCCG	CTGGCT, CTGGCC, GGGG, NotAorT, CCCCC, ACGCCA
GSE8623		CTGCCT
GSE8775	CCCCC	
GSE8892	NotAorT, CCCCC, CCCCG, GCCCG	GGGG, GGGGG, TGGGG
GSE8938		CTGCCT, GGGG
GSE9107	CCCCC, CCCCG	CTGCCT, TCGCTC, ACGCCA, GGGG, GGGGG
GSE9552	CCCCC	ACGCCA
GSE23344	CCGCCTCCC, ATCCGC, NotAorT, CCCCG, GCCCG	GGGG
GSE23802	CCGCCTCCC, ATCCGC, ACGCCA	CCCCC
GSE23880		CTGCCT, GGGGG
GSE24167		GGGGG
GSE24503	CCGCCTCCC, ATCCGC, NotAorT, CCCCC, CCCCG, GCCCG	GGGG, GGGGG, TGGGG
GSE24729	CCGCCTCCC, ATCCGC, GGGG, CCCCC	
GSE24917	CCGCCTCCC, ATCCGC, NotAorT, CCCCG, GCCCG	GGGG
GSE25267	CCGCCTCCC, ATCCGC	
GSE26246	NotAorT	GGGG, TGGGG, CCCCC
GSE26726	GGGG, CCCCC	
GSE27345	GGGG	CTGCCT, ACGCCA
GSE27376		CTGCCT, ACGCCA
GSE27927	CCGCCTCCC, ATCCGC, ACGCCA, TCGCCGCT, NotAorT, CCCCG, GCCCG	GGGG, CCCCC
GSE28728	CCCCG	ACGCCA
GSE29203		CTGCCT, ACGCCA
GSE29815		CTGCCT, TCGCTC, GCTCCTCG, GGGG
GSE30020		TGGGG
GSE30360	CCGCCTCCC, ATCCGC	
GSE31564	GGGG, CCCCC	AGGCCA

Thirty experiments are listed above. There are eight other experiments with at least 10 CEL files: GSE4032 (12 CEL files), GSE8330 (12), GSE24156 (15), GSE24167 (15), GSE24978 (12), GSE26717 (12), GSE27178 (20), and GSE28147 (12). Only GSE7763 contains more than 100 CEL files.

Dependence on position of motif in probe

Mean values of estimates ($\times 100$) arranged as vectors suitable for inclusion in an R program. A selection of these values appear in Figure 3 of the paper after the application of the Friedman smoother (R's *supsmu* function).

nC=c(44,103,144,145,111,98,154,86,146,113,118,45,78,84,59,18,64,26,18,39,82)

nA=c(-3,-37,-13, -8,-39,-7,-40,-31,-213,-3,5, -16,-241, -49,-43,-34,-5,13,-28,-12,-9)

nG=c(26,124,138,124,83,114,112,107,-81,167,146,116,-113,74,56,66, 87,58,56,76,56)

nCCGCCTCCC=c(43,47,54,59,61,72,67,56,57,52,62,40,44,58,59,72,64,67,61,46,36)

nCCCCTC=c(9,23,25,36,38,37,35,26,32,20,28,13,23,23,24,38,36,30,29,23,19)

nCCCGCCCC=c(45,54,50,49,46,47,49,48,41,33,26,41,34,36, 58,52,41,37,34,22,13)

nATCCGC=c(29,30,27,31,28,26,26,37,25,-8,45,47,22,36,33,47,33,23,30,14, 6)

nAGGCCA=c(-9,-20,-24,-19,-23,-27,-23,-21,-22,-16,-22,-23,-19,-23,-21,-24, -22,-21,-15,-19,-14,)

nCTGCCT=c(18,25,26,25,27,24,23,22,20,24, 22,16,22,21, 27,28,24,21,16,12,9)

nCTGGCC=c(-16,-20,-17,-19,-27,-19,-14,-13,-22,-15,-19,-6,-19,-18, -18,-23,-18,-10,-8,-8,-9)

nTCGCCGCT=c(25,34,33,27,23,22,26,30,25,34,12,26,28, 28,27,32,39,26,25,15,9)

nTCGCTC=c(4,25,18,24,14,17,20,9,14,25,-10,-4,-2, 22,5,20,16,13,18,4,15) nGGTGCT=c(-3,-3,1,0,0,1,4,7,4,9,3,6,2,8,15,8,5,4,6,-1,1) nGGTTGCCC=c(3,15,8,12,11,10,8,4,7,6,3,4,8, 10,11,12,11,14,10,7,1)

nGCTCCTCG=c(13,15,15,18,19,23,18,18,16,16,19,17,12,10,14,20,17,19,12,10,9)

nACGCCA=c(22,18,24,16,11,17,11,17,22,24,14,3,2,21,23,19,24,12,7, 10,9)

nTCCCT=c(-4,13,18,20,31,24,30,25,28,30,27,19,14,12,18,26,24,27,25,30,26)

nGGGG=c(88,56,54,46,38,30,28,24,22,18,22,21,24, 21,24,37,26,27,26,24,46)

nGGGGG=c(0,-6,2,8,9,19,6,8,12,3,5,3,14, 14,22,33,21,22,28,28,46)

nTGGGG=c(-62,-34,-27,-20,-19,-14,-15,-12,-6,-15,-14,-8,-28,-9,-12,-14,-9,-11,-10,-13, -13)

nNotAorT=c(-11,-20,-21,-19,-18,-18,-18,-18,-15,-14,-13,-10,-16, -13,-16,-15,-15,-13,-15,-12,-16)

nCCCCC=c(26,56,67,74,81,79,79,88, 72,60,82,42,57,34,72,76,82,68,63,50,36)

nCCCCG=c(-3,8,14,26,27,30,19,29,16,42,22,41, 25,28,31,23,27,39,35,30,37)

nTCCCC=c(-13,14,24,29,26,29,32,34,29,30,27,25,13,25,14,28,33,27,31,31,17)

nRSQ=c(81,88,89,88,88,88,90,91,91,82,79,76,87,87, 83,83,79,77,75,74,76)