Connecting Gene Expression Data from Connectivity Map and In Silico Target Predictions For Small Molecule Mechanism-of-Action Analysis

Aakash Chavan Ravindranath,[a‡] Nolen Perualila-Tan,[b‡] Adetayo Kasim,[c] Georgios Drakakis,[a] Sonia Liggi,[a] Suzanne C Brewerton,[d] Daniel Mason,[a] Michael J Bodkin,[d] David A Evans,[d] Aditya Bhagwat,[e] Willem Talloen,[f] Hinrich W.H. Göhlmann,[f] QSTAR Consortium,[g] Ziv Shkedy,[*b] and Andreas Bender[*a]

† Electronic Supplementary Information (ESI)
See DOI: 10.1039/b000000x/

S1 Introduction

This supplementary appendix contains additional materials that were cited and briefly discussed in the article but were not presented in the manuscript. Section S2.1 provides more information about FARMS as a summarization method and the motivation of using it in the paper. Section S2.2 describes in more details the MLP algorithm for gene-set enrichment analysis and in Sections S3 and S4, the supporting tables and figures that were referred in the paper are presented.

S2 Supplementary Information

S2.1 Microarray Summarization by FARMS

Factor Analysis for Robust Microarray Summarization (FARMS) is a model based approach for summarizing microarray data[1]. The main idea of the FARMS algorithm for expression arrays is to detect a common hidden cause of the observed measurements that cannot arise from noise which is uncorrelated for different measurements. The hidden cause is the true but unobserved mRNA concentration in the tissue sample which leads to a simultaneous decrease or increase in probe intensities measuring this mRNA. The hidden cause is called signal since it indicates the mRNA concentration. The core of the FARMS algorithm is a factor analysis a multivariate technique to detect a common structure in the data of multiple probes that measure the same target.

FARMS is currently the leading method in an international challenge, the Benchmark for Affymetrix GeneChip Expression Measures, in which FARMS outperformed 88 competing methods, if both sensitivity and specificity are simultaneously considered by the area under the receiver operating characteristic curve (AUC). AUC is commonly considered the best suited measurement of the quality of a summarization

[a] *Unilever Centre for Molecular Science Informatics, Department of Chemistry, University of Cambridge, Lensfield Road, Cambridge CB2 1EW, United Kingdom*
[b] *Interuniversity Institute for Biostatistics and statistical Bioinformatics, Universiteit Hasselt, Agoralaan 1, B3590 Diepenbeek, Belgium*
[c] *Wolfson Research Institute for Health and Wellbeing, Durham University, United Kingdom*
[d] *Eli Lilly U.K., Erl Wood Manor, Windlesham, Surrey GU206PH, United Kingdom*
[e] *Open Analytics, 2600, Antwerp, Belgium*
[f] *Janssen Pharmaceutical Companies of Johnson and Johnson, 2340, Beerse, Belgium*
[g] http://qstar-consortium.org

‡ 'These authors contributed equally to this work'

method. The FARMS algorithm is implemented in the function qfarms in the Bioconductor package farms: `http://www.bioconductor.org/packages/release/bioc/html/farms.html`

S2.2 MLP method

For the pathway analysis using MLP, the search for enriched gene sets starts from the list of p-values that quantify the degree of differential expression for each gene probed across the experiment conditions. For the application presented in the paper, we use the Limma p-values obtained from testing differentially expressed genes. For a predefined gene set from GO annotations, it calculates a score that summarizes the significance of all the genes included in that specific set. This score is the mean of the negative logarithm of the p-values:

$$MLP = mean(-log(p - values))$$

Once the MLP statistic is calculated for each gene set in the dataset, a permutation procedure is applied to determine whether or not a particular gene set is significant. If the observed MLP is larger than most of the value of the MLP*(MLP test statistic) for that same gene set obtained across multiple random permutations, then the gene set is declared significant.[2,3] A more detailed description of the method and its output is provided by Raghavan *et al.* (2006), Raghavan *et al.* (2007) and Amaratunga *et al.* (2014).

The MLP algorithm is implemented in the R package MLP. The package is available freely in Bioconductor: `http://www.bioconductor.org/packages/release/bioc/html/MLP.html`

The MLP package provides several plot functions. A GO graph can be used to visualise the top set of significant GO terms according to their structure in ontology with the biggest and least specific GO terms shown at the bottom. An example of which is presented in Figure S4 showing the top set of significant GO terms for the antipsychotic drugs. The darker the color, the more significant is the GO terms. In that sense, the genes that are part of cholesterol biosynthetic are a subset of the genes in sterol biosynthetic. The genes contributing most to the significance of a certain GO terms can be identified and plotted according to the level of significance using a barplot as shown in Figure S5.

S3 Supplementary Tables

Table S1 Overlapping Pathways for MCF7 cell line. Pathway Search involving the top protein targets and genes regulated by distinct compound set. MoA of the compound cluster is comprehended through pathway overlap between the significant genes and predicted protein targets, as shown in the table. The compound clustering based upon similar targets lead to sub-clustering of compounds with similar therapeutic classes. Cluster 1 containing antipsychotic drugs and cluster 7 containing compounds geldanamycin and tanespimycin, were studied in detail. As an example, in cluster 1 our method suggests genes INSIG1 and LDLR and target CYP450, share pathway ”steroid metabolic process” for the listed antipsychotic compounds. This study was also supported by the findings of Polymeropoulos *et al.* [6].

No.	Compounds	Pathway	Target	Genes
1	amitriptyline clozapine thioridazine chlorpromazine trifluoperazine prochlorperazine fluphenazine	Steroid metabolic process	Cytochrome P450 2D6	INSIG1 LDLR
2	verapamil dexverapamil	MAPK signaling pathway	Voltage gated T type calcium channel alpha 1G subunit	DUSP9
3	estradiol alphaestradiol fulvestrant	**no overlapping pathways found**		
4	dexamethasone prednisolone fludrocortisone	**no overlapping pathways found**		
5	nifedipine nitrendipine felodipine	Aging	Induced myeloid leukemia cell differentiation protein Mcl.1	IFIT1
6	15-delta prostaglandin J2 arachidonic acid	Arachidonic acid metabolism	Thromboxane A synthase	AKR1C3
7	Tanespimycin geldanamycin	Response to Unfolded Protein	Heat shock protein HSP 90 alpha	HSP90AB1 HSPA6 HSPA4L DNAJB4 HSPA1B DNAJB1 DNAJA1 HSP90AA1
		Antigen processing and presentation	Heat shock protein HSP 90 alpha	HSP90AB1 HSPA1B HSPA1A HSP90AA1
8	tacrolimus sirolimus	Activation of cysteine-type endopeptidase activity involved in apoptotic process	Proteinase activated receptor 1	MOAP1
		Negative regulation of cell proliferation	Proteinase activated receptor 1	S100A11
		Regulation of actin cytoskeleton	Proteinase activated receptor 1	PFN1
		Prostate cancer	Heat shock protein HSP 90 alpha	GSK3B

Table S2 Overlapping Pathways for PC3 cell line. Pathway Search involving the top protein targets and genes regulated by compound set. MoA of the compound cluster is comprehended through pathway overlap between the significant genes and predicted protein targets as mentioned. The compounds clustering based on similar targets lead to sub-clustering of compounds with similar therapeutic class. For example in cluster 10, our method suggests genes FABP4 and ANGPTL4 and targets PPAR-gamma, PPAR-alpha and acyl CoA desaturase, share pathway ”PPAR signalling” for the listed thiazolidinediones and findings confirmed by KEGG database pathway(hsa03320).

No.	Compounds	Pathways	Target	Genes
1	thioridazine amoxapine cyproheptadine perphenazine	Protein modification process	Muscarinic acetylcholine receptor M3	HERPUD1
2	loperamide haloperidol	**no overlapping pathways found**		
3	bromocriptine lisuride	**no overlapping pathways found**		
4	estradiol alpha-estradiol fulvestrant	**no overlapping pathways found**		
5	prednisone lynestrenol danazol	Neuroactive ligand-receptor interaction	Glucocorticoid receptor	GHR
		Focal adhesion	Protein kinase C alpha	BIRC3
		MAPK signaling pathway	Protein kinase C alpha	IL1R2
6	sulfathiazole sulfaguanidine	Small cell lung cancer	Cyclin dependent kinase2	LAMB3
7	hydrochlorothiazide metolazone	**no overlapping pathways found**		
8	erythromycin oleandomycin	**no overlapping pathways found**		
9	mercaptopurine azathioprine	Cytokine-cytokine receptor interaction	Vascular endothelial growth factor receptor2	IL17RA
		ErbB signaling pathway Wnt signaling pathway Colorectal Cancer Endometrial Cancer	Glycogen synthase kinase.3 beta	MYC
		p53 signaling pathway	Cyclin dependent kinase 1	THBS1
		Focal Adhesion	Glycogen synthase kinase.3 beta Vascular endothelial growth factor receptor2	THBS1
10	rosiglitazone troglitazone	Induction of apoptosis	Peroxisome proliferator.activated receptor gamma	PRKCD
		Positive regulation of transcription from RNA polymerase II promoter	Peroxisome proliferator.activated receptor gamma	TNFRSF1A
		PPAR signaling pathway	Peroxisome proliferator.activated receptor gamma Peroxisome proliferator.activated receptor alpha Acyl.CoA desaturase	FABP4 ANGPTL4
		Adipocytokine signaling pathway	Peroxisome proliferator.activated receptor alpha	TNFRSF1A
11	fisetin genistein	Wnt signaling pathway	Glycogen synthase kinase.3 beta	VANGL1 DKK1
		Axon guidance	Glycogen synthase kinase.3 beta	EPHA2

S4 Supplementary Figures

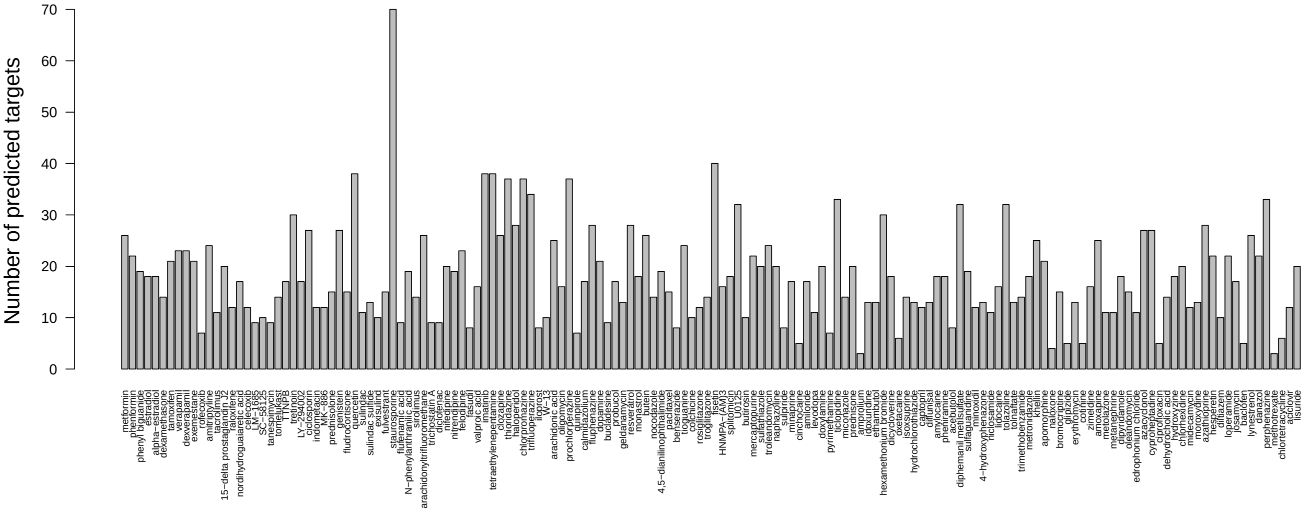

Fig. S1 Barplot of predicted targets per compound The barplot displays the number of predicted targets found for each compound from both cell lines MCF7 and PC3.

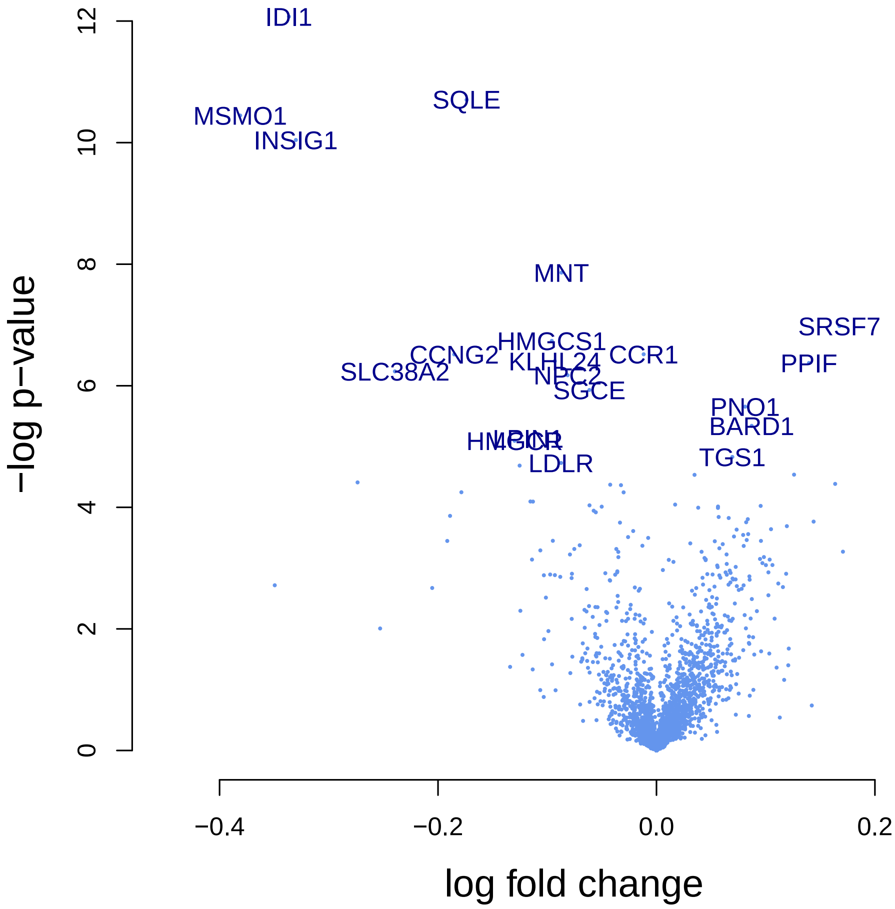

Fig. S2 Volcano plot. -log(P-value) Vs. fold change. Every gene is represented by a dot in the graph. Genes such as IDI1, INSIG1, MSMO1, SQLE and MNT among others labelled in the graph have the smallest P-values (*i.e.* the highest evidence for statistical significance) when testing for differentially expressed genes between the cluster of interest and the other compounds in the compound set. Genes at the left and right side of the graph have the largest effect size (fold-change). The plot displays the genes of high significance for the sub-cluster containing antipsychotic compounds. Studies by Iskar *et al.* has shown INSIG1, LDLR, IDI1 and SQLE are responsible for the "cholesterol metabolic process, which are in accordance with our MLP results shown in (Figure S4)[7]. The genes INSIG1 and LDLR were found to be significant for the antipsychotic compounds MoA where they shared similar pathways with the predicted targets.

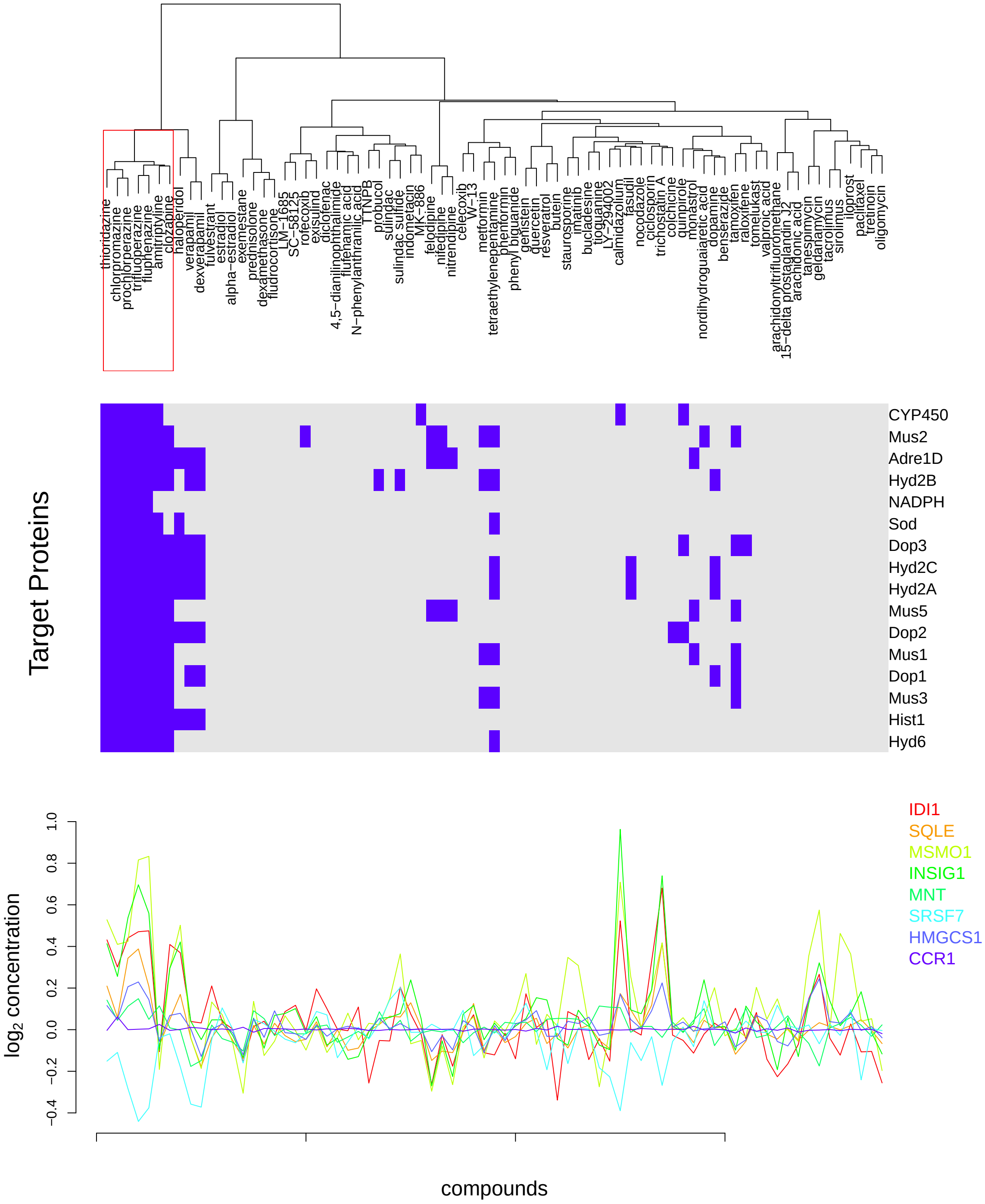

Fig. S3 Genes and Protein targets regulated by antispsychotic compound cluster.(a)Protein-target-similarity-based hierarchical clustering of compounds; (b) heatmap of the proteins target (rows) and compounds(columns) coloured according to activation/inactivation of protein targets; (c) the profile plot of the top 8 differentially expressed genes, with compounds ordered on the x-axis and fold-change in the y-axis. The selected compound sub-cluster contains compounds that are predicted on the targets represented in blue. The neighbouring compound cluster (haloperidol, verapamil and dexverapamil) of calcium channel binders are known to have antipsychotic effects, thus large number of similar targets are predicted. The genes (IDI1, SQLE, MSMO1, INSIG1, MNT, SRSF7, HMGCS1 and CCR1) are perturbed on sub-cluster and also have similar gene perturbation on the neighbouring cluster.

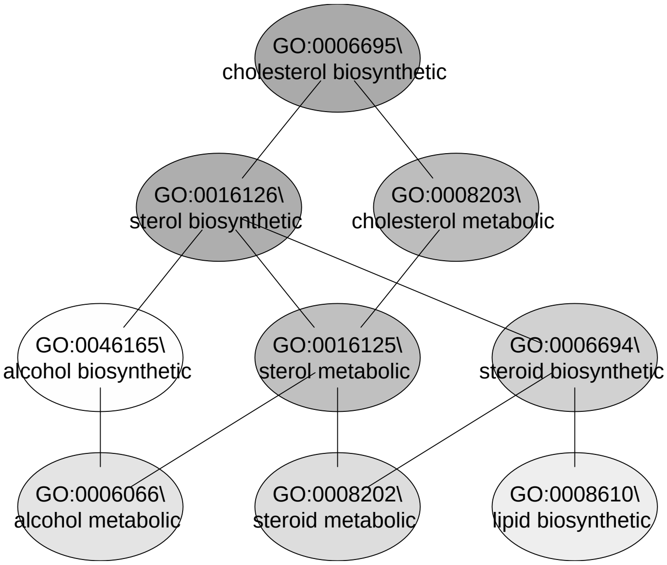

Fig. S4 GO pathways containing the top gene sets with MLP for the antipsychotic drugs. Every ellipse represents a gene set. The color indicates significance; the darker, the more significant. The connectors indicate the parent-child relationship. The antipsychotic drugs are affecting numerous processes related to the cholesterol biosynthetic process. As stated by Polymeropoulos *et al.* activation of antipsychotics by genes associated with lipid homeostasis is not just a common off target effect of these drugs but rather the common central mechanism by which they achieve their antipsychotic activity [6].

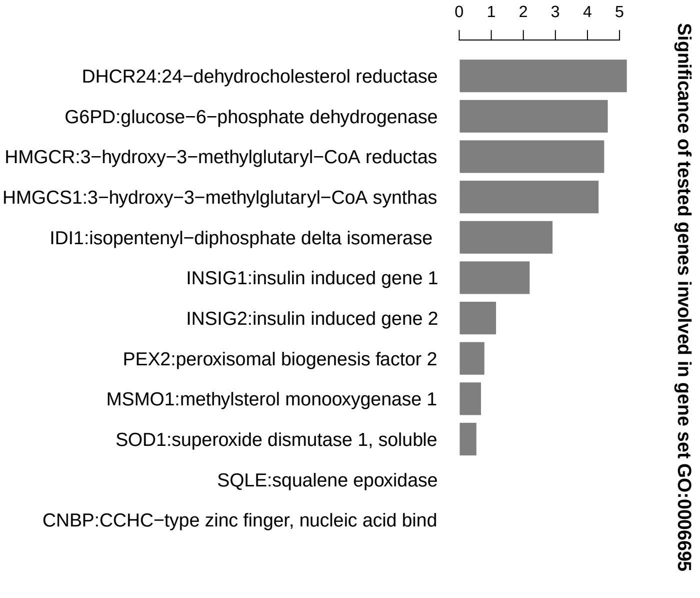

Fig. S5 Significance plot of the top genes contributing to ”cholesterol biosynthetic process”. The plot represents the top 12 genes contributing with the level of significance in the bar for the respective pathway in the MLP analysis for antipsychotic compound cluster. Gene Dhcr24 is predicted to be highly significant in the ”cholesterol biosynthetic process” and is known to code for the protein cholesterol-synthesizing enzyme seladin-1 (in agreement with the study by Crameri *et al.*). [8 9] Other genes such as G6PD were also known to regulate the pathway through other proteins such as SREBP. [10]

References

1 S. Hochreiter, D.-A. Clevert and K. Obermayer, *Bioinformatics*, 2006, **22**, 943–949.

2 D. Lin, Z. Shkedy, D. Yekutieli, D. Amaratunga and L. Bijnens, *Modeling Dose-response Microarray Data in Early Drug Development Experiments Using R, Order Restricted Analysis of Microarray Data*, Springer, 2012.

3 D. Amaratunga, J. Cabrera and Z. Shkedy, *Exploration and Analysis of DNA Microarray and Other High-Dimensional Data, Second Edition.*, Wiley, 2014.

4 N. Raghavan, D. Amaratunga, J. Cabrera, A. Nie, J. Qin and M. McMillian, *Journal of Computational Biology*, 2006.

5 N. Raghavan, A. De Bondt, W. Talloen, D. Moechars, H. W. H. Göhlmann and D. Amaratunga, *Bioinformatics*, 2007, **23**, 3032–3038.

6 M. H. Polymeropoulos, L. Licamele, S. Volpi, K. Mack, S. N. Mitkus, E. D. Carstea, L. Getoor, A. Thompson and C. Lavedan, *Schizophr. Res.*, 2009, **108**, 134–142.

7 M. Iskar, G. Zeller, P. Blattmann, M. Campillos, M. Kuhn, K. H. Kaminska, H. Runz, A.-C. Gavin, R. Pepperkok, V. van Noort and P. Bork, *Mol. Syst. Biol.*, 2013, **9**, 662.

8 A. Crameri, E. Biondi, K. Kuehnle, D. Ltjohann, K. M. Thelen, S. Perga, C. G. Dotti, R. M. Nitsch and M. H. Ledesma, Maria Dolores andMohajeri, *EMBO J.*, 2006, **25**, 432–443.

9 Wechsler, A. and Brafman, A. and Faerman, A. and Björkhem, I. and Feinstein, E., *Science*, 2003, **302**, 2087.

10 J. D. Horton, J. L. Goldstein and M. S. Brown, *J. Clin. Invest.*, 2002, **109**, 1125–1131.