

Next-generation high-density self-assembling functional protein arrays

Niroshan Ramachandran, Jacob V Raphael, Eugenie Hainsworth, Gokhan Demirkan, Manuel G Fuentes, Andreas Rolfs, Yanhui Hu & Joshua LaBaer

We developed a high-density self-assembling protein microarray, based on the nucleic acid programmable protein array (NAPPA) concept, to display thousands of proteins that are produced and captured *in situ* from immobilized cDNA templates. We arrayed up to 1,000 unique human cDNAs and obtained high yields of protein expression and capture with minimal variation and good reproducibility. This method will enable various experimental approaches to study protein function in high throughput.

High-density functional protein arrays allow the functional testing of thousands of proteins simultaneously^{1,2}. A key remaining challenge for producing protein microarrays has been uniting high content (many different proteins) with high density and functionality³. Most approaches rely on expressing and purifying proteins to print on the array surfaces and have succeeded at displaying many different proteins at high density^{1,2}.

Considerable challenges, however, accompany the use of purified protein for printing microarrays. Variable yields of protein result in dynamic ranges in the amount of protein displayed that cover several orders of magnitude, depending on protein size, hydrophobicity and other properties. Array batch-to-batch variation may affect experimental reproducibility and the folding and function of some proteins may also be lost during purification, printing and storage.

To address these concerns, we had previously developed a protein microarray method called nucleic acid programmable protein array (NAPPA), which allows for functional proteins to be synthesized *in situ* directly from printed cDNAs just in time for assay⁴. The proteins are translated using a T7-coupled rabbit reticulocyte lysate *in vitro* transcription-translation (IVTT) system. The expressed proteins are captured locally with an antibody to a C-terminal glutathione S-transferase (GST) tag on each protein (anti-GST). This approach eliminates the need for high-throughput protein isolation and ensures that all proteins are produced fresh (that is, coincident with or minutes before use) for each experiment. Many experiments have confirmed that NAPPA produces functional protein⁵. Other alternate strategies for producing protein

microarrays have also been introduced. The multiple spotting technique, MIST, prints an *Escherichia coli*-based IVTT extract directly on top of a printed PCR template⁶. Another approach uses a variation of ribosome display to immobilize an mRNA-DNA hybrid and express proteins using a cell-free translation mix⁷. In a recent approach called DNA array to protein array, DAPA, proteins are translated on a cDNA array and then diffuse across a cell-free extract-infused membrane to a protein capture surface⁸. Although encouraging, these strategies have only been tested with relatively small numbers of proteins compared with printing purified proteins and have yet to demonstrate the robust ability to produce the high content needed to justify protein microarrays as a routine proteomics tool.

Here we describe a next-generation NAPPA method for making fresh protein *in situ* to produce high-content protein microarrays that begin to address many of these important issues. The key printed substrate for NAPPA is purified DNA, which is simpler to prepare, quantify, print and store than protein. In performing optimization experiments, we observed that high-quality supercoiled DNA provided the best substrate for cell-free protein expression, and that commercial chemistries had insufficient yield and purity for this purpose (data not shown). We therefore investigated the use of a resin derivatized with diamine chemistry, which allowed us to purify high-quality DNA efficiently. DNA bound the positively charged diamines at low pH and eluted when the diamines became neutrally charged under alkaline conditions (**Supplementary Methods** and **Supplementary Protocol** online). Using this method, one technician can process 5,000 samples per week with yields of 18 μ g of supercoiled DNA per 1 ml of culture (5–10-fold greater than commercial systems). The DNA is of sufficient quality for use in mammalian cell transfections (data not shown).

We also developed a new printing chemistry that relies on the surprising (and unexplained) ability of bovine serum albumin (BSA) to dramatically improve DNA-binding efficiency (**Supplementary Methods** and **Supplementary Protocol**). BSA and the capture antibody are coupled to the amine-coated glass surface via an activated ester-terminated homo-bifunctional cross-linker. Using fluorescently labeled DNA, we estimated that 64% of the DNA is captured onto the surface (**Supplementary Fig. 1** online).

To assess protein yield and reproducibility, we printed a test array of cDNAs for 96 genes (**Fig. 1a**), non-expressing plasmid DNA as a negative control and a concentration series of purified recombinant protein (**Supplementary Fig. 2** online). By PicoGreen staining for double-stranded DNA (**Fig. 1a,b**, **Supplementary Methods** and **Supplementary Protocol**), we observed that 97% of the printed samples were detectable (3 s.d. above signal for control features without DNA). Using an antibody to GST (distinct from the one used to capture the nascent protein; **Supplementary Methods**),

Harvard Institute of Proteomics, Harvard Medical School, 320 Charles Street, Cambridge, Massachusetts 02141, USA. Correspondence should be addressed to J.L. (joshua_labaer@hms.harvard.edu).

RECEIVED 7 DECEMBER 2007; ACCEPTED 27 MARCH 2008; PUBLISHED ONLINE 11 MAY 2008; DOI:10.1038/NMETH.1210

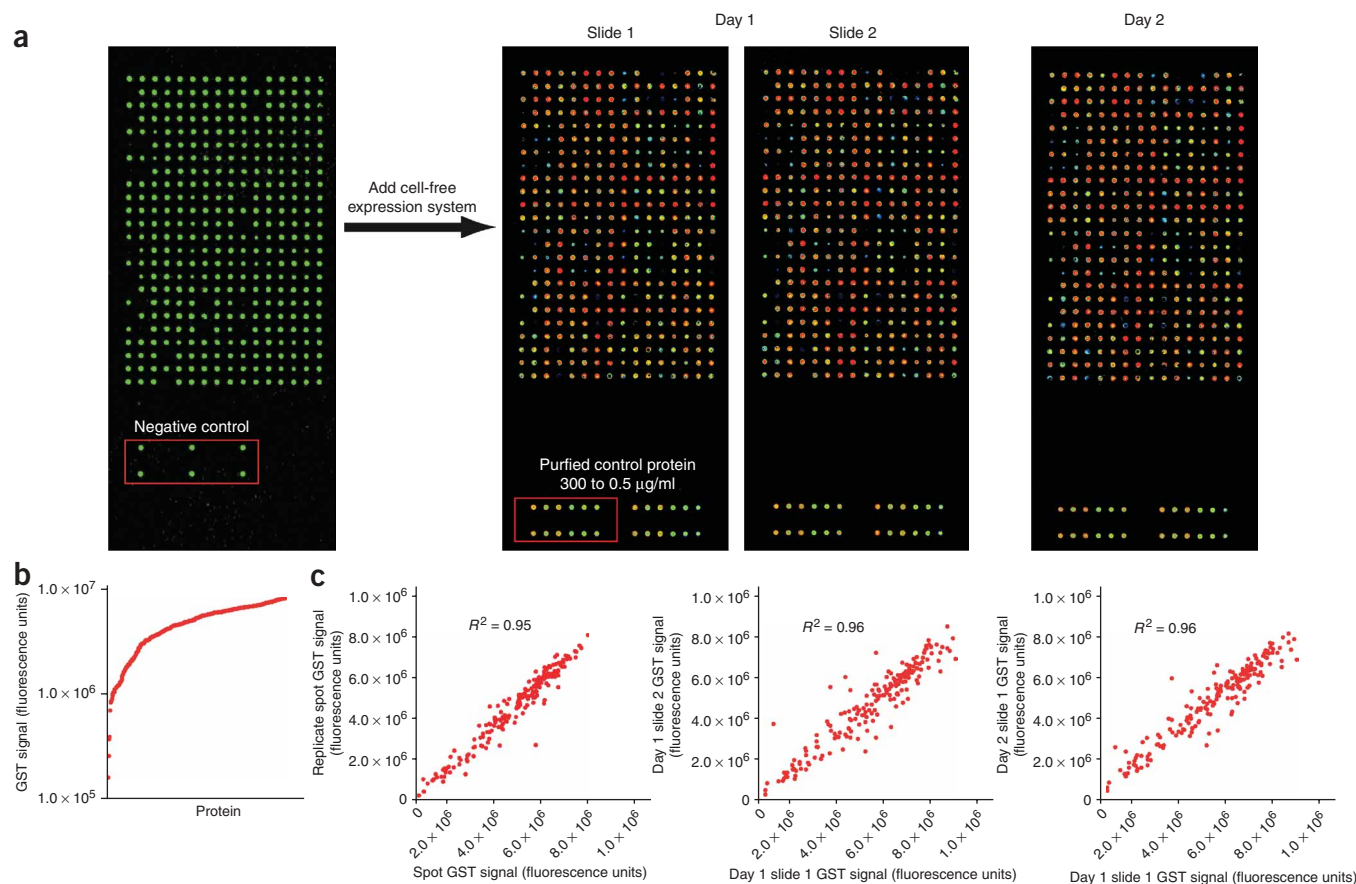


Figure 1 | Test array. (a) A test array representing 96 unique genes was printed and stained with PicoGreen dye to assay DNA binding (left). The negative control for protein expression includes a non-expressing plasmid (vector plasmid). Genes on the test arrays were expressed using T7 transcription-translation-coupled rabbit reticulocyte lysate and the proteins were detected using anti-GST. Feature color represents an artificial color scheme to indicate signal intensity (white > red > yellow > green > blue). Two slides were processed on the same day (middle two panels) and a third slide was processed on a different day (right). The boxed signal at the bottom of the slide includes a titration of different amounts of purified GST (control protein), used to estimate the yield of protein produced and captured. (b) The amount of protein produced in each feature (**Supplementary Table 1**). (c) A comparison of the raw protein signal from arrays processed on the same day and on different days. The correlation of signal between a spot and its replicate on the same array (left), correlation of average signal (average of spot and replicate) between two replicate slides (slide 1 and slide 2) processed on the same day (middle) and correlation of average signal between slide 1 processed on day 1 and a replicate slide processed on a different day (day 2; right). Average correlation of signal was >0.95.

which recognizes the C-terminal GST tag genetically encoded in each coding sequence and thus confirms full-length translation, we detected protein signal for 99% of the 96 printed genes (3 s.d. above the signal from non-expressing plasmid; **Fig. 1a,b**, **Supplementary Methods**, **Supplementary Table 1** and **Supplementary Protocol**). Compared to the printed recombinant purified GST, the average protein yield was 9 fmol per feature (4–13 fmol, 10th percentile to 90th percentile). Slide processing for protein display was uniform and reproducible between replicates within an array ($R^2 = 0.95$) and between duplicate arrays ($R^2 = 0.96$; **Fig. 1c**).

To demonstrate that a variety of proteins can be displayed by this format, we selected 1,000 human genes for which we had isolated plasmids from single colonies, verified the full-length sequence and previously made them readily available through the PlasmID repository^{9,10}. We detected DNA signal for 99% of the samples (coefficient of variation, CV = 18%; **Fig. 2a,b** and **Supplementary Table 2** online). Although we observed a slight variability in protein yield depending on the amount of DNA per feature, 96% of the genes showed readily detectable protein signal. Examining these data by protein class, we observed that kinases and transcription factors

expressed and captured well with success rates of 98% and 96%, respectively (**Fig. 2c**). Moreover, even membrane proteins, which are typically difficult to produce in heterologous systems, showed good signal for 93% of those tested. The range of protein signal was similar for the various protein families. With increasing protein size, there was only a small reduction in the protein display success rate.

To test for zone effects, we printed the same cDNA sample (encoding p53) in 40 features distributed evenly throughout the array and used an antibody to p53 for detection, which demonstrated an average coefficient of variation of only 7% (**Supplementary Fig. 3** online). To assess the level of signal cross-talk potentially caused by protein diffusion to nearby features, we examined all features neighboring the p53 protein. Immediate neighbors to p53 features had signals that were 1.9% of the p53 signal in p53 features (average of 160 spots), whereas background signal was 0.7% (average of signals in 392 spots that were at least 4 spots (2,572 μm) removed from p53). Moreover, the appropriate proteins were displayed as expected, as demonstrated by protein-specific antibody signals, which revealed little variation (CV = 6%) when we tested independently processed samples.

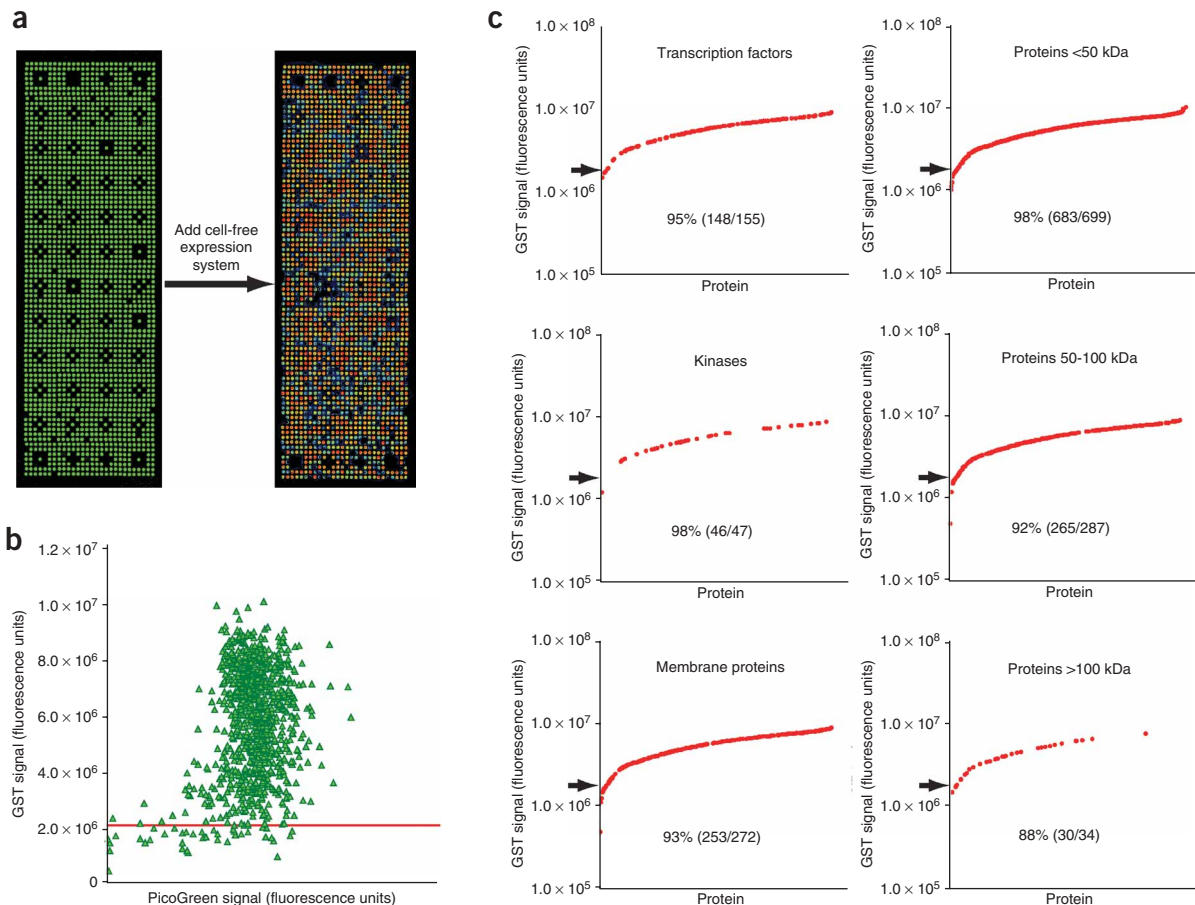


Figure 2 | High-density array. **(a)** A high density array with DNA representing 1,000 unique genes, each printed in duplicate (see **Supplementary Table 2** for gene list). The arrays were stained with PicoGreen to show DNA binding (left) and proteins were detected with anti-GST after addition of cell free expression lysate (right). **(b)** The protein signal at every feature plotted against the amount of DNA captured there. Most of the DNA (98%) and protein signals (92%) were within two fold of their respective means. The background (3 s.d. above the signal from the negative control) is indicated with a red line across the graph. **(c)** The respective success of protein signal was examined by protein class and size. The percentage of proteins with signal above background (3 s.d. above the negative control signal; arrow) is indicated.

To confirm protein function on high-density arrays, we printed an array expressing 647 unique genes in duplicate, including 449 genes that we had not previously tested (**Fig. 3a** and **Supplementary Table 3** online). We tested for binary interactions between several well-characterized interacting pairs including Jun-Fos¹¹ and p53-MDM2 (ref. 12) (**Fig. 3b**). We co-expressed the query protein along with the arrayed proteins by adding the appropriate cDNA clone (without a GST tag) to the cell-free expression lysate⁵. After protein expression and washing, we treated the arrays with protein-specific antibodies to detect the query protein, revealing the positions where it bound. Using Jun, Fos and MDM2 as queries, we detected selective binding to their appropriate interacting partners. There are no simple tests to confirm protein folding, and function must be tested on a protein-by-protein basis. All of the interaction pairs we tested behaved as expected.

The folding of large multidomain mammalian proteins often relies on the presence of chaperones and cofactors. Our IVTT-based method uses mammalian ribosomal machinery and the presence of chaperones, including hsp90, hsc70 and others, which may encourage folding. The role of chaperones in producing properly folded proteins such as kinases, structural proteins, membrane proteins and even viral proteins in rabbit reticulocyte lysate is well documented¹³.

Proteins may occur in various activity states depending on co- or post-translational modifications. Post-translational modifications represent a challenge for all protein microarray formats because proteins produced and purified in heterologous systems may either lack modifications or display unnatural ones. Proteins expressed using the rabbit reticulocyte lysate IVTT system typically lack most post-translational modifications. However, because it is an open system, it is possible to add modifying enzymes or extracts, such as kinases or canine pancreatic microsomal membranes, to test the effect of post-translational modifications¹³. Additionally, some proteins require association with activating partners for function. We have previously shown that multi-protein complexes function in the NAPPA setting⁵.

The ideal method for producing protein microarrays would evince several important properties. First, it must be reliable and reproducible, from sample to sample and array to array. Second, the method should be capable of displaying a broad variety of proteins, insensitive to protein class or size. Third, it should display a high yield of protein per feature while maintaining a narrow range of protein yield from protein to protein. Fourth, the method must be readily executable at large scale and high density. And finally, the method must display functional protein.

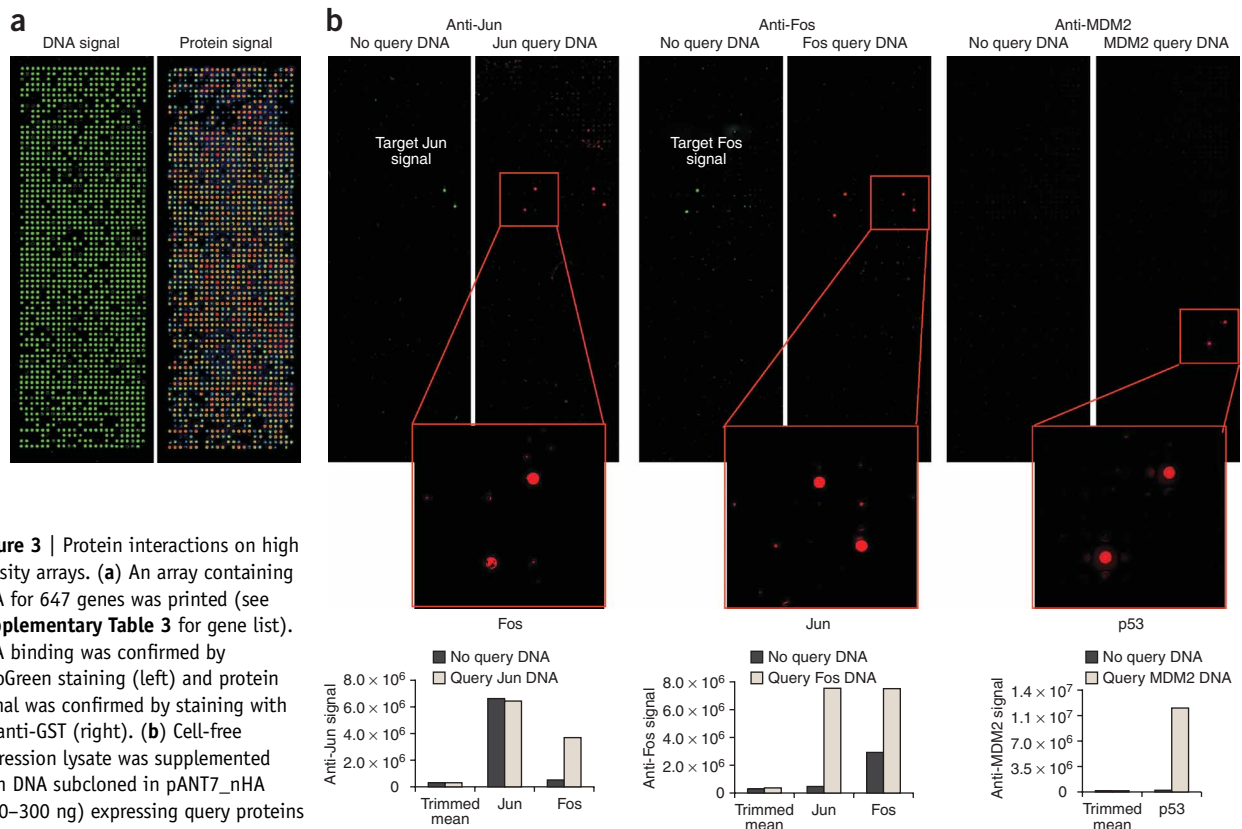


Figure 3 | Protein interactions on high density arrays. (a) An array containing DNA for 647 genes was printed (see **Supplementary Table 3** for gene list). DNA binding was confirmed by

PicoGreen staining (left) and protein signal was confirmed by staining with an anti-GST (right). (b) Cell-free expression lysate was supplemented with DNA subcloned in pANT7_nHA (100–300 ng) expressing query proteins Jun, Fos and MDM2. The binding of the

query to the target proteins on the array was detected using the appropriate protein-specific antibodies. A control array was processed in each case where no query plasmid was added to the cell-free expression lysate (target Jun signal is from spots where gene encoding Jun was printed). The graphs below show the trimmed mean signal (25–75%) for each array and the average of the replicate signals for the target protein.

Our next-generation NAPPA approach routinely produced 9 fmol of protein per feature with ~90% success for a broad variety of proteins of different sizes, including membrane proteins and proteins >100 kDa. To our knowledge, this is the first non-protein printing method to produce over a thousand unique proteins on a microarray surface. Notably, the range of protein signals for nearly all proteins was narrow: 92% of displayed protein yields were within twofold of the mean. This limited variation may be due to the saturation of the capture sites on the array by the expressed target. The method was highly reproducible from array to array and sample to sample (CV = 6%), which compares favorably with that reported for DNA microarray chemistries in which coefficients of variation range from 20 to 40%¹⁴. This is particularly important considering that NAPPA entails not only printing cDNA but also transcription, translation and protein capture. The ability to array proteins at high density will be well suited for testing protein-protein interactions, screening for enzyme substrates and measuring selectivity of small-molecule drug binding.

Note: Supplementary information is available on the Nature Methods website.

ACKNOWLEDGMENTS

We thank J. Williamson and M. Fernandez for their help with the robotics and D. Zhu and R. Boyce for developing the DNA normalization tool. This study was supported by the Early Detection Research Network (US National Cancer Institute

grant 5U01CA117374-02) and the US National Institute of Allergy and Infectious Diseases (contract HHSN2332200400053C).

AUTHOR CONTRIBUTIONS

N.R. designed the experiment, processed slides and wrote the manuscript; J.V.R. and G.D. implemented automation, purified and printed DNA; E.H. analyzed array data; M.G.F. tested surface and printing chemistries; Y.H. performed informatics analysis of gene collection; A.R. cloned genes; and J.L. designed the experiment and wrote the manuscript.

Published online at <http://www.nature.com/naturemethods/>
Reprints and permissions information is available online at <http://npg.nature.com/reprintsandpermissions>

- MacBeath, G. & Schreiber, S. *Science* **289**, 1760–1763 (2000).
- Zhu, H. *et al. Science* **293**, 2101–2105 (2001).
- Braun, P. *et al. Proc. Natl. Acad. Sci. USA* **99**, 2654–2659 (2002).
- Ramachandran, N. & LaBaer, J. *Curr. Opin. Chem. Biol.* **9**, 14–19 (2005).
- Ramachandran, N. *et al. Science* **305**, 86–90 (2004).
- Angenendt, P., Kreutzberger, J., Glokler, J. & Hoheisel, J.D. *Mol. Cell. Proteomics* **5**, 1658–1666 (2006).
- Tao, S.C. & Zhu, H. *Nat. Biotechnol.* **24**, 1253–1254 (2006).
- He, M. *et al. Nat. Methods* **5**, 175–177 (2008).
- Murthy, T. *et al. PLoS ONE* **2**, e577 (2007).
- Rolfs, A. *et al. PLoS ONE* **3**, e1528 (2008).
- Newman, J.R. & Keating, A.E. *Science* **300**, 2097–2101 (2003).
- Boutell, J.M., Hart, D.J., Godber, B.L., Kozlowski, R.Z. & Blackburn, J.M. *Proteomics* **4**, 1950–1958 (2004).
- Rickman, D.S., Herbert, C.J. & Aggerbeck, L.P. *Nucleic Acids Res.* **31**, e109 (2003).
- Arduengo, M., Schenborn, E. & Hurst, R. *Cell Free Protein Expression* (Landes Bioscience, Austin, Texas, USA, 2007).