

GeneAmp 5700 (Applied Biosystems) with a Quantitect SYBR Green RT-PCR kit (Qiagen). Expression levels were normalized on the basis of the amount of *polyubiquitin* transcripts. All values are means and standard deviations for three replicates of three biological repeats.

**Construction of fluorescence-tagged fusions**

cDNA fragments of *CASTOR* and *POLLUX* were amplified by PCR with the primers CAS-F/CAS-R (5'-ACGCGTCGACATGTCCTTGATTCCGGAG-3', 5'-CATGCCATGGATT CCTTTTCAGTAATTAC-3') and POL-F/POL-R (5'-ACGCGTCGACATGATACCACT ACCAGTA-3', 5'-CATGCCATGGAATCGCCTGAAGCAATCAC-3'), respectively. Each fragment was digested with *Sall* and *NcoI* and ligated into the same restriction sites of pUC18-CaMV35S-sGFP (S65T)-nos<sup>30</sup>. For the construction of *AtrecA*-DsRed2 fusion, a fragment encoding the transit peptide of *AtrecA*<sup>19</sup> was amplified from *Arabidopsis thaliana* genomic DNA with the use of the primers *AtrecA*-*Sall*-F/*AtrecA*-LFH-R (5'-ACGCGT CGACATGGATTACACAGCTAGTC-3', 5'-ATCGAATTCAGAACTGATTTTGTG-3'). DsRed2 fragment was amplified from pDsRed2-1 (Clontech) with the use of DsRed2-LFH-F/DsRed2-*NotI*-R (5'-CTGAATTCGATCGCGCATGGCCTCCTCCGAGAA-3', 5'-ATTTCGGCCGCTACAGGAACAGGTGGTG-3') primers. *AtrecA*-LFH-R and DsRed2-LFH-F primers were designed to introduce overlapping nucleotides (underlined) at the 3' and 5' ends of the *AtrecA* and DsRed2 fragments, respectively. Using both fragments as templates, joint PCR was performed with *AtrecA*-*Sall*-F and DsRed2-*NotI*-R primers. The resulting fusion fragment was digested with *Sall* and *NotI* and cloned into the same restriction site of pUC18-CaMV35S-sGFP (S65T)-nos vector.

**Microprojectile bombardment and confocal laser scanning microscopy**

Microprojectile bombardment was performed with a Biolistic PDS-1000/He Particle Delivery System (Bio-Rad). Epidermis of *Allium cepa* scaly bulb and roots of *Pisum sativum* were bombarded with a rupture-disk pressure of 1,100 p.s.i. (~7.6 MPa) at a target distance of 6 cm. At 24–40 h after bombardment, they were analysed with a Bio-Rad Radiance2000 confocal laser scanning microscope. Green fluorescence of GFP and red fluorescence of DsRed2 were excited at 488 nm with an argon laser and collected sequentially with a filter set (HQ530/60 and E570LP). Images of both fluorescences were processed and merged with the Lasersharpp2000 program system (Bio-Rad).

**Computer analysis**

Sequences were analysed by BLAST (<http://www.ncbi.nlm.nih.gov/BLAST/>) and GENSCAN version 1.0 (<http://genes.mit.edu/GENSCAN.html>). Clustal W (<http://www.ebi.ac.uk/clustalw/>) was used for multiple alignment and evolutionary relationships. The target peptide and transmembrane regions were predicted by TargetP version 1.01 (<http://www.cbs.dtu.dk/services/TargetP/>) and TMHMM version 2.0 (<http://www.cbs.dtu.dk/services/TMHMM-2.0/>). For domain and structure analyses, both Pfam (<http://www.sanger.ac.uk/Software/Pfam/>) and FUGUE v.2.0 (<http://www-cryst.bioc.cam.ac.uk/~fugue/>) were applied.

Received 30 September; accepted 29 November 2004; doi:10.1038/nature03237. Published online 22 December 2004.

1. Smith, S. E. & Read, D. J. *Mycorrhizal Symbiosis* (Academic, London, 1997).
2. Oldroyd, G. E. D. Dissecting symbiosis: developments in Nod factor signal transduction. *Ann. Bot.* **87**, 709–718 (2001).
3. Kistner, C. & Parniske, M. Evolution of signal transduction in intracellular symbiosis. *Trends Plant Sci.* **7**, 511–518 (2002).
4. Ehrhardt, D., Wais, R. & Long, S. Calcium spiking in plant root hairs responding to Rhizobium nodulation signals. *Cell* **85**, 673–681 (1996).
5. Truchet, G. et al. Sulphated lipooligosaccharide signals from *Rhizobium meliloti* elicit root nodule organogenesis in alfalfa. *Nature* **351**, 670–673 (1991).
6. Radutoiu, S. et al. Plant recognition of symbiotic bacteria requires two LysM receptor-like kinases. *Nature* **425**, 585–592 (2003).
7. Cárdenas, L. et al. Ion changes in legume root hairs responding to Nod factors. *Plant Physiol.* **123**, 443–452 (2000).
8. Senoo, K. et al. Isolation of two different phenotypes of mycorrhizal mutants in the model legume plant *Lotus japonicus* after EMS-treatment. *Plant Cell Physiol.* **41**, 726–732 (2000).
9. Szczyglowski, K. et al. Nodule organogenesis and symbiotic mutants of the model legume *Lotus japonicus*. *Mol. Plant Microbe Interact.* **11**, 684–697 (1998).
10. Bonfante, P. et al. The *Lotus japonicus* *LjSym4* gene is required for the successful symbiotic infection of root epidermal cells. *Mol. Plant Microbe Interact.* **13**, 1109–1120 (2000).
11. Novero, M. et al. Dual requirement of the *LjSym4* gene for mycorrhizal development in epidermal and cortical cells of *Lotus japonicus* roots. *New Phytol.* **154**, 741–749 (2002).
12. Harris, J. M., Wais, R. & Long, S. R. *Rhizobium*-induced calcium spiking in *Lotus japonicus*. *Mol. Plant Microbe Interact.* **16**, 335–341 (2003).
13. Hayashi, M. et al. Construction of a genetic linkage map of the model legume *Lotus japonicus* using an intraspecific F2 population. *DNA Res.* **8**, 301–310 (2001).
14. Nakamura, Y. et al. Structural analysis of a *Lotus japonicus* genome. II. Sequence features and mapping of sixty-five TAC clones which cover the 6.5-mb regions of the genome. *DNA Res.* **9**, 63–70 (2002).
15. Kawasaki, S. & Murakami, Y. Genome analysis of *Lotus japonicus*. *J. Plant Res.* **113**, 497–506 (2000).
16. Kawaguchi, M. et al. Providing the basis for genomics in *Lotus japonicus*: the accessions Miyakojima and Gifu are appropriate crossing partners for genetic analyses. *Mol. Gen. Genomics* **266**, 157–166 (2001).
17. Stracke, S. et al. A plant receptor-like kinase required for both fungal and bacterial symbiosis. *Nature* **417**, 959–962 (2002).
18. Ane, J. M. et al. *Medicago truncatula* *DMII1* required for bacterial and fungal symbioses in legumes. *Science* **303**, 1364–1367 (2004).
19. Köhler, R. H. et al. Exchange of protein molecules through connections between higher plant plastids. *Science* **276**, 2039–2042 (1997).
20. Shi, J., Blundell, T. L. & Mizuguchi, K. FUGUE: sequence-structure homology recognition using environment-specific substitution tables and structure-dependent gap penalties. *J. Mol. Biol.* **310**, 243–257 (2001).

21. Jiang, Y. et al. Crystal structure and mechanism of a calcium-gated potassium channel. *Nature* **417**, 515–522 (2002).
22. Jiang, Y. et al. Structure of the RCK domain from the *E. coli* K<sup>+</sup> channel and demonstration of its presence in the human BK channel. *Neuron* **29**, 593–601 (2001).
23. Kwok, E. Y. & Hanson, M. R. Plastids and stromules interact with the nucleus and cell membrane in vascular plants. *Plant Cell Rep.* **23**, 188–195 (2004).
24. Kawaguchi, M. et al. Root, root hair, and symbiotic mutants of the model legume *Lotus japonicus*. *Mol. Plant Microbe Interact.* **15**, 17–26 (2002).
25. Perry, J. A. et al. A TILLING reverse genetics tool and a web-accessible collection of mutants of the legume *Lotus japonicus*. *Plant Physiol.* **131**, 866–871 (2003).
26. Schaefer, L. et al. Symbiotic mutants deficient in nodule establishment identified after T-DNA transformation of *Lotus japonicus*. *Mol. Gen. Genet.* **259**, 414–423 (1998).
27. Niwa, S. et al. Responses of a model legume *Lotus japonicus* to lipochitin oligosaccharide nodulation factors purified from *Mesorhizobium loti* JRL501. *Mol. Plant Microbe Interact.* **14**, 848–856 (2001).
28. Broughton, W. J. & Dilworth, M. Y. Control of leghemoglobin synthesis in snake beans. *Biochem. J.* **125**, 1075–1080 (1971).
29. Firmin, J. L. et al. Resistance to nodulation of cv. Afghanistan peas is overcome by nodX, which mediates an O-acetylation of the *Rhizobium leguminosarum* lipo-oligosaccharide nodulation factor. *Mol. Microbiol.* **10**, 351–360 (1993).
30. Isono, K. et al. Leaf-specifically expressed genes for polypeptides destined for chloroplasts with domains of  $\sigma^{70}$  factors of bacterial RNA polymerases in *Arabidopsis thaliana*. *Proc. Natl Acad. Sci. USA* **94**, 14948–14953 (1997).

Supplementary Information accompanies the paper on [www.nature.com/nature](http://www.nature.com/nature).

**Acknowledgements** We thank K. Szczyglowski, J. Webb and J. Stougaard for providing mutant seeds; M. Hayashi for help with mapping; T. Kojima and R. Ohtomo for mycorrhiza analysis; Y. Niwa for providing pUC18-CaMV35S-sGFP (S65T)-nos vector; G. Oldroyd and J. Sun for help with Ca-spiking assays; J. Krüger and B. B. H. Wulff for critical reading of the manuscript; J. Soll for providing the pea root transformation protocol before publication; and M. Durrant for help with modelling the CASTOR pore structure. Part of this work was supported by the fund of Promotion of Basic Research Activities for Innovative Biosciences (BRAIN), and Core Research for Evolutional Science and Technology (CREST), Japan Science and Technology Agency. Research at the Sainsbury Laboratory is funded by the Gatsby Charitable Foundation.

**Competing interests statement** The authors declare that they have no competing financial interests.

**Correspondence** and requests for materials should be addressed to S.K. ([kawasa@nias.affrc.go.jp](mailto:kawasa@nias.affrc.go.jp)). The sequences have been deposited at the DNA Data Bank of Japan with the following accession numbers: LJTO2K14a (AP006732), LJTO2K14b (AP006733), LJTO2K14c (AP006734), LJT45115 (AP006736), LJT20F11 (AP006737), LJT46G19 (AP006735), LJT45B09a (AP006729), LJT45B09b (AP006730) and LJT45B09c (AP006731); genomic sequences (B-129 Gifu) of CASTOR (AB162016), POLLUX (AB162017), and mRNA sequences (B-129 Gifu) of CASTOR (AB162157) and POLLUX (AB162158).

**Interaction network containing conserved and essential protein complexes in *Escherichia coli***

Gareth Butland<sup>1</sup>, José Manuel Peregrín-Alvarez<sup>2</sup>, Joyce Li<sup>1</sup>, Wehong Yang<sup>1</sup>, Xiaochun Yang<sup>1</sup>, Veronica Canadien<sup>3</sup>, Andrei Starostine<sup>1</sup>, Dawn Richards<sup>3</sup>, Bryan Beattie<sup>3</sup>, Nevan Krogan<sup>1</sup>, Michael Davey<sup>1</sup>, John Parkinson<sup>2,4,5</sup>, Jack Greenblatt<sup>1,3,5</sup> & Andrew Emili<sup>1,5</sup>

<sup>1</sup>Banting and Best Department of Medical Research, University of Toronto, 112 College Street, Toronto, Ontario M5G 1L6, Canada  
<sup>2</sup>Hospital for Sick Children, 555 University Avenue, Toronto, Ontario M4K 1X8, Canada  
<sup>3</sup>Affinium Pharmaceuticals, 100 University Avenue, Toronto, Ontario M5J 1V6, Canada  
<sup>4</sup>Department of Biochemistry and <sup>5</sup>Department of Medical Genetics and Microbiology, University of Toronto, Medical Sciences Building, 1 King's College Circle, Toronto, Ontario M5S 1A8, Canada

Proteins often function as components of multi-subunit complexes. Despite its long history as a model organism<sup>1</sup>, no large-scale analysis of protein complexes in *Escherichia coli* has yet been reported. To this end, we have targeted DNA cassettes into the *E. coli* chromosome to create carboxy-terminal, affinity-tagged alleles of 1,000 open reading frames (~23% of the

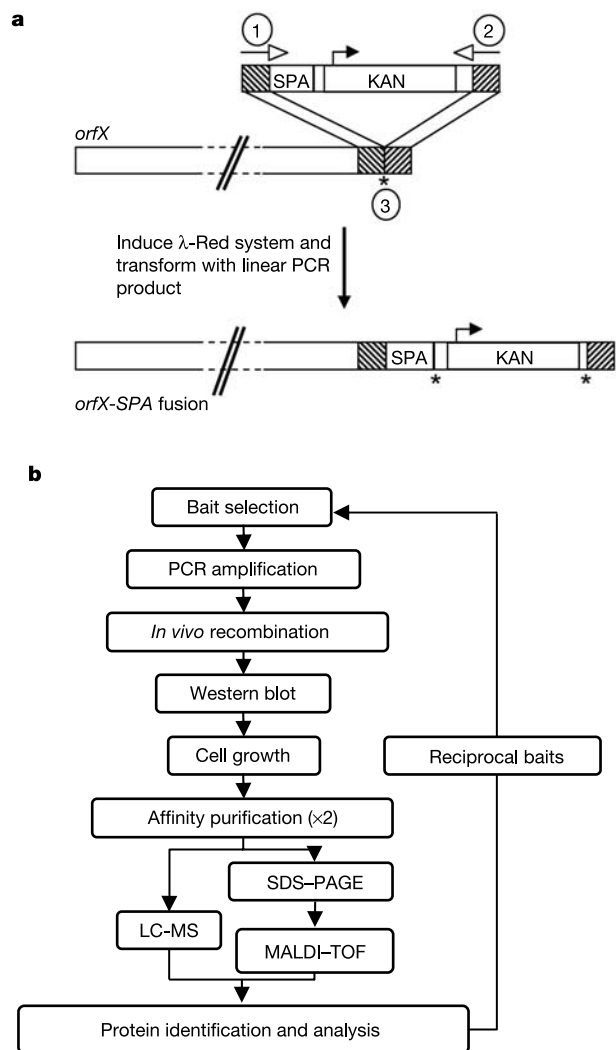
genome). A total of 857 proteins, including 198 of the most highly conserved, soluble non-ribosomal proteins essential in at least one bacterial species, were tagged successfully, whereas 648 could be purified to homogeneity and their interacting protein partners identified by mass spectrometry. An interaction network of protein complexes involved in diverse biological processes was uncovered and validated by sequential rounds of tagging and purification. This network includes many new interactions as well as interactions predicted based solely on genomic inference or limited phenotypic data<sup>2</sup>. This study provides insight into the function of previously uncharacterized bacterial proteins and the overall topology of a microbial interaction network, the core components of which are broadly conserved across Prokaryota.

The yeast-based tandem affinity purification (TAP) procedure for isolating protein complexes makes use of site-specific recombination to introduce a dual tagging cassette into chromosomal loci<sup>3</sup>. *Escherichia coli* does not readily recombine exogenous linear DNA fragments into its chromosome, but expression of the lambda general recombination system ( $\lambda$ -Red) markedly enhances integration<sup>4,5</sup>. We adapted one such system<sup>5</sup> to introduce a DNA cassette, bearing a selectable marker and either the TAP or sequential peptide affinity (SPA) tags<sup>3,6</sup>, into the C termini of open reading frames (ORFs) in the lysogenic *E. coli* strain DY330 (ref. 6), which harbours  $\lambda$ -Red under control of a temperature-sensitive repressor<sup>5</sup> (Fig. 1a). The tagged bait proteins, expressed at endogenous levels, were purified  $\sim 10^6$ -fold to homogeneity from log-phase cultures using two rounds of affinity chromatography<sup>6</sup>. To minimize nucleic-acid-mediated interactions, extracts were pre-treated with nuclease. Polypeptide components of isolated complexes were then identified using two forms of mass spectrometry: peptide mass fingerprinting was performed on all silver-stained polypeptide bands visible by SDS-polyacrylamide gel electrophoresis (PAGE) not seen in parallel control purifications, whereas gel-free shotgun sequencing was used to identify small and lower-abundance proteins. To minimize the false discovery rate (false positives), protein-protein interactions were deemed authentic provided a reciprocal interaction could be confirmed or if the data were reproducible. Figure 1b shows a schematic overview of the methodology.

The effectiveness of the procedure was confirmed in pilot purifications of DNA-dependent RNA polymerase (RNAP). Tagged core subunit  $\beta$  (RpoB) co-purified specifically with essential elongation factors (NusA and NusG), specialized sigma factors involved in promoter recognition ( $\sigma^{32}$  (RpoH),  $\sigma^{38}$  (RpoS),  $\sigma^{54}$  (RpoN),  $\sigma^{70}$  (RpoD)), and with accessory factors  $\omega$  (RpoZ), HepA (RapA) and YacL (15 kDa acidic protein of unknown function not previously known to interact with RNAP) (Fig. 2a; see also Supplementary Table 1). Similarly, NusG co-purified with YacL, HepA, core enzyme and termination factor Rho, whereas  $\omega$  bound  $\sigma^{70}$ , NusA and b1731, another small protein of unknown function. In reciprocal experiments, tagged b1731 co-purified with  $\beta$ ,  $\beta'$  (RpoC),  $\alpha$  (RpoA),  $\sigma^{70}$  and  $\omega$ , but not with Nus factors, HepA or YacL (Fig. 2b), implying an exclusive association with initiating holoenzyme. In contrast, tagged YacL bound  $\omega$ , NusG and HepA together with core enzyme (Fig. 2c; see also Supplementary Table 1), suggesting a role in elongation. On the basis of their specific and reproducible association with RNAP, we suggest that b1731 and YacL be renamed Rap (RNAP-associated protein) B and C, respectively. YacL and b1731 are not required for viability, and homologues are restricted to  $\gamma$ -Proteobacteria (Supplementary Fig. 1a), suggesting a specialized function. Notably, tagged  $\sigma^{54}$  and  $\sigma^{38}$  bound to the sugar transporter ManX, whereas  $\sigma^{32}$  co-purified with quinone oxidoreductase Qor, indicating that protein-protein interactions may regulate alternative sigma factor activities.

Purification of DNA-dependent DNA polymerase indicated that low-abundance complexes are amenable to analysis. A nine-subunit holo-complex, DNA polymerase III\*, which contains a core complex ( $\alpha$  (DnaE),  $\epsilon$  (DnaQ) and  $\theta$  (HolE)) requiring only processivity

factor  $\beta$  (DnaN) for full replicative activity<sup>7</sup>, was readily isolated (Supplementary Table 1). Purification of tagged  $\alpha$ ,  $\epsilon$  and  $\theta$  yielded core complex as well as the clamp loader ( $\gamma$  and  $\tau$  (dnaX),  $\delta$  (HolA),  $\delta'$  (HolB),  $\chi$  (HolC) and  $\psi$  (HolD)), which recruits  $\beta$  onto DNA (Fig. 2d;  $\theta$  detected by gel-free mass spectrometry). In reciprocal experiments (Supplementary Table 1), tagged  $\chi$  and  $\psi$  co-purified with core enzyme and the clamp loader.  $\chi$  also bound PriA (primosomal DNA helicase), b1808 (putative ATP-dependent helicase) and TopA (DNA topoisomerase I), along with three conserved replication factors that act coordinately *in vitro*<sup>8</sup>: Ssb, single-strand DNA-binding protein previously reported to bind  $\chi$ <sup>9</sup>; RecQ, a replicative DNA helicase; and TopB, DNA topoisomerase III (Fig. 2d). In turn, Ssb co-purified with TopB, exonucleases RecJ and SbcB, and helicases PriA, RecG and RecQ (Supplementary Table 1), consistent with multiple roles in chromosome dynamics.



**Figure 1** Systematic identification and validation of protein complexes in *E. coli*. **a**, Gene-specific affinity-tagging cassettes produced by polymerase chain reaction (PCR)<sup>6</sup> using primers (1, 2) homologous to a target translational termination codon (3) were integrated into the *E. coli* chromosome using the  $\lambda$ -Red recombination system<sup>5</sup>. KAN, kanamycin-resistance cassette. Asterisks indicate stop codons. **b**, Flow chart of steps in purifying and validating protein complexes. Interactions were confirmed either by reciprocal tagging and purification, or by repeat analysis. LC-MS, gel-free liquid chromatography-tandem mass spectrometry; MALDI-TOF, gel-based peptide mass fingerprinting using matrix-assisted laser desorption/ionization-time-of-flight mass spectrometry.

TopB co-purified with Ssb and RecQ, suggesting that these factors serve a similar role in coordinating DNA replication in prokaryotes as suggested for their eukaryotic homologues<sup>10</sup>.

We expanded our analysis to cover about one-quarter of the *E. coli* genome, targeting 1,000 ORFs (Supplementary Fig. 2 and Supplementary Table 2), including 248 uncharacterized 'y' genes, 168 putative 'b' genes and 209 of the most broadly conserved genes encoding soluble (non-membrane) non-ribosomal proteins essential in *E. coli* or another bacterium (see Methods). We successfully tagged 857 proteins (86%; confirmed by western blot analysis), including 198 essential and conserved (essential-conserved) proteins, and were able to purify 648 (65%; detected by mass spectrometry), which compares favourably to analogous studies of protein complexes in yeast<sup>3</sup>. A total of 118 of these proteins had no detectable partners, whereas 5,254 putative protein-protein interactions were detected for the other 530 baits (Supplementary Table 1). To eliminate false-positives, we reciprocally tagged and purified a large subset of candidate partners. A validation rate of ~53% was achieved (see Supplementary Information), which compares favourably to studies in yeast<sup>11</sup>, confirming the stringency of the methodology. A total of 716 non-redundant binary interactions, involving 83 essential (excluding the ribosome) and 152 non-essential proteins, have been validated so far (highlighted in Supplementary Table 1). The entire validated data set is shown graphically in Fig. 3a.

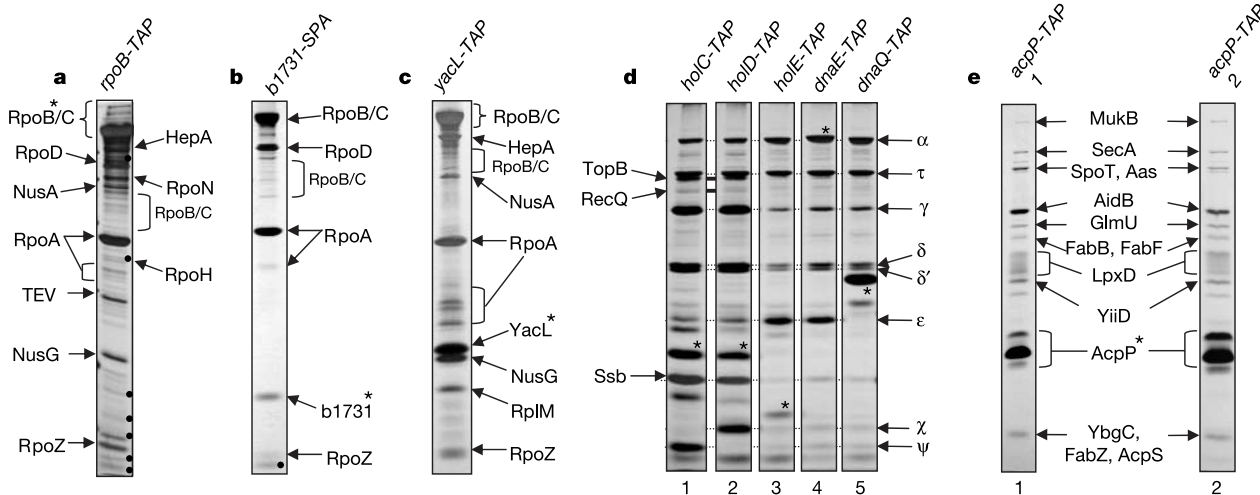
Eighty-five per cent of the validated interactions are new, as they are not described in the Database of Interacting Proteins (DIP)<sup>12</sup>, Biomolecular Interaction Network Database (BIND)<sup>13</sup>, STRING<sup>14</sup>, or Prolinks databases<sup>15</sup>, whereas only ten orthologous interactions (interologs) were reported in a two-hybrid interaction screen in *Helicobacter Pylori*<sup>16</sup> (Supplementary Table 3). The significance of these novel interactions is reinforced by functional annotation (Supplementary Table 2). For instance, acyl carrier protein (ACP), a key carrier of growing fatty acid chains, bound specifically and reproducibly to enzymes linked to biogenesis of fatty acids, phospholipids and lipid A (essential outer-membrane constituent) (Fig. 2e; see also Supplementary Table 1), including two 3-ketoacyl-ACP synthases (FabB, FabF), 3-ketoacyl-ACP reductase (FabG), 3-hydroxyacyl-ACP dehydrase (FabZ), LpxD (essential protein

required for lipid A biogenesis), YbgC (*tol-pal* cluster hydrolase of short-chain acyl-CoA thioesters), AcpS (involved in transfer of 4'-phosphopantethein to ACP), Aas and PlsB (membrane proteins involved in phospholipid acylation), and YiiD (putative acetyltransferase). ACP also co-purified with GlmU (an essential bi-functional enzyme that converts glucosamine-1-phosphate to UDP-GlcNAc (lipid A precursor)), AidB (isovaleryl-CoA dehydrogenase), SecA (pre-protein translocase), as well as MukB and SpoT, as previously reported<sup>17</sup>. We did not detect IscS, which was predicted to interact with ACP in an overexpression study<sup>17</sup>, possibly because transient or low-affinity interactions are missed by our method.

Many other informative complexes were detected (see Supplementary Information). These included notable interactions mediated by the cysteine desulfurase IscS (IscS-FdhD and IscS-YhhP), between two uncharacterized essential proteins (YgjD-YeaZ), and by a sizeable group of uncharacterized proteins with factors involved in ribosome function, RNA processing, and/or RNA binding.

Graph network analysis of the validated data set provided evidence of 'scale-free' behaviour<sup>18</sup>. Most of the proteins had few interacting partners, whereas a subset of 'hubs' formed a far greater number of connections (Fig. 3b). Comparable connectivity was observed for the essential-conserved proteins alone (Supplementary Fig. 3). Scale-free networks are predicted to be robust against random node removal but vulnerable to hub removal, a property that might be expected to be preserved across evolution<sup>18</sup>. Indeed, removal of the 20 most highly connected nodes (>15 interactions; Supplementary Table 4) markedly reduced the network connectivity (see Supplementary Figs 4a and b). Notably, these same hubs were all highly conserved (detected in ≥125 genomes; Supplementary Fig. 4c). Moreover, protein connectivity was proportional (positively correlated) to the number of genomes a homologue was detected in (Supplementary Fig. 4d). Although a previous analysis of bacterial two-hybrid data<sup>19</sup> failed to detect such a dependency, possibly due to a high false-positive rate, our results are in agreement with a more in-depth analysis of the relationship of protein evolutionary rates to the number of interactions in eukaryotes<sup>20</sup>.

To investigate further the conserved nature of the bacterial network, we analysed co-occurrence of BLAST homologues of



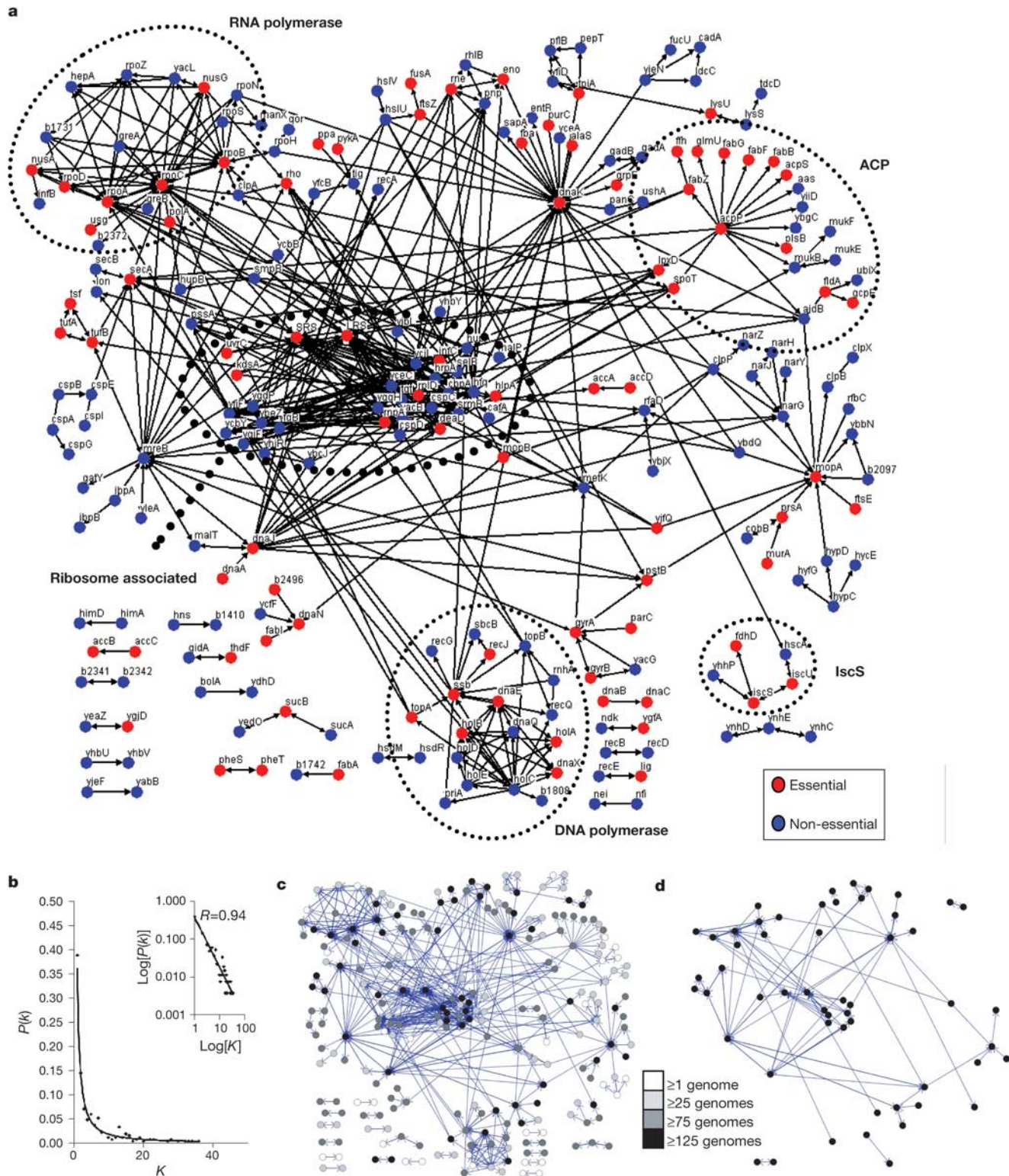
**Figure 2** Analysis of affinity-purified protein complexes. SDS-PAGE silver-stain analysis of the components of affinity-purified complexes from *E. coli*. **a-c**, Purification of TAP-tagged *E. coli* RNAP subunit  $\beta$  (**a**) and two associated proteins: SPA-tagged b1731 (**b**) and TAP-tagged YacL (**c**). **d**, Purification of TAP-tagged HoID ( $\psi$ ) and HoIC ( $\chi$ ) subunits of the processivity clamp loader (lanes 1 and 2), and DNA polymerase III core subunits

HoIE ( $\theta$ ), DnaE ( $\alpha$ ) and DnaQ ( $\epsilon$ ) (lanes 3–5). **e**, Independent purifications of TAP-tagged AcpP (lanes 1 and 2). Only validated interacting proteins are labelled. Black circles indicate bands that failed to yield spectra or validated results; asterisks indicate tagged bait; and brackets indicate degradation products.



each pair of interacting proteins across all three domains of life (Archaea, Prokaryota and Eukaryota)<sup>21</sup> (see Methods). As seen in Supplementary Fig. 5, the interacting proteins were more likely to be co-conserved than control, randomly selected protein pairs,

indicating that the interactions are similarly conserved. Notably, the most highly conserved proteins were highly connected, forming a single interconnected component (Fig. 3c, d). This core set of 154 interactions involving 71 proteins (including the ribosome;



**Figure 3** Network properties of bacterial protein–protein interactions. **a**, Network of validated protein complexes. Interactions are represented as directional edges extending from the tagged protein. Baits without partners are removed for clarity. Red nodes, essential proteins; blue nodes, non-essential proteins; black ovals, complexes discussed in text. **b**, Connectivity distribution of validated interactions ( $K$ ) per protein plotted as a

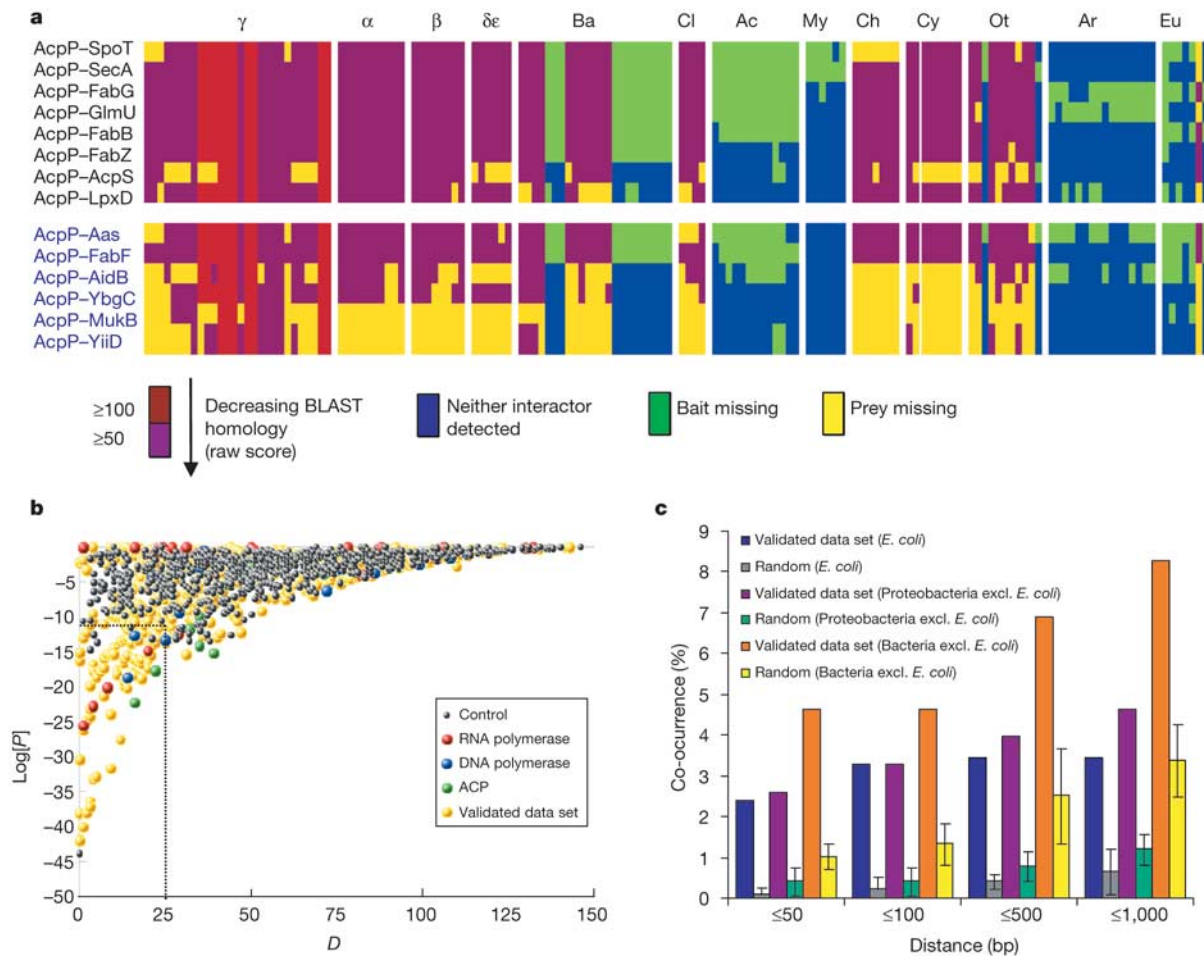
function of frequency,  $P(k)$ . Inset: log-plot power law distribution,  $P(k) \approx k^{-\gamma}$ , where  $\gamma$  is the degree exponent.  $R$ , Pearson's correlation coefficient (see Methods). **c**, Network sequence conservation. Node shading (white-to-black) is scaled according to the number of genomes that pairs of interacting proteins co-occur in. **d**, Network of highly conserved proteins co-occurring in  $\geq 125$  genomes (homologue raw BLAST bit score  $\geq 50$ ).

Supplementary Table 5) potentially fulfils critical roles across all bacteria. Similar results were obtained using clusters of orthologous groups of proteins<sup>22</sup> (COGs; Supplementary Fig. 6a, b).

Co-occurrence of homologues across different genomes (phylogenetic profiles) has previously been used to explore functional links between genes<sup>23,24</sup>. Different approaches for constructing these profiles include the use of orthology assignments (for example, COGs) and sequence homology<sup>25</sup>. Recently, a study<sup>26</sup> introduced a mapping approach based on sequence homology methods to assess the degree of conservation of interologs between species. We adopted a similar approach to examine patterns of interolog conservation within the complexes detected in this study. Intriguingly, AcpP and several of its interacting partners displayed significant divergence in Bacilli, Actinobacteridae, *Mycoplasma* spp. as well as Archaea and eukaryotes (Fig. 4a). The lack of obvious homologues of AcpP in actinobacteria is consistent with the highly diverged nature of the predicted orthologues from this phylum in the COGs classification scheme. Evolutionary divergence was also evident with the DNA and RNA polymerase complexes. Although core RNAP subunits and cofactors *nusA*, *rpoA*, *rpoB*, *rpoC*, *rpoD*, *rpoH* and *rpoS* are found in virtually all eubacteria, other subunits

(such as *hepA*, *rpoZ*, *b1731* and *yaqL*) are restricted to  $\gamma$ -Proteobacteria (Supplementary Fig. 1a). Likewise, the DNA polymerase clamp loader subcomplex (*hola-D*) is similarly restricted (Supplementary Fig. 1b). These data suggest that sequence divergence may lead to functional diversification and the formation of novel modules. By clustering interacting proteins based on their phylogenetic profiles, it may be possible to identify new modules. For instance, the interacting PflB–PepT gene products cluster together (see Supplementary Fig. 2).

Statistical methods such as mutual information, Pearson correlation coefficient, Hamming distance ( $D$ ) and the chance co-occurrence probability distribution ( $P$ ) have been developed to predict functional relationships among genes based on phylogenetic profiles<sup>23</sup>. We applied the latter two metrics to quantify the extent of correlation between different phylogenetic profiles. Using relatively modest cutoff values ( $D < 25$ ,  $P < 10^{-11}$ ), a small but significant subset (~14%) of the interacting proteins showed closely correlated phylogenetic profiles relative to a control set of randomly selected pairs of bacterial proteins (Fig. 4b). Chromosomal proximity has also been used to infer functional linkages between evolutionarily conserved proteins<sup>27</sup>. Investigations into the distribution of relative



**Figure 4** Bioinformatic analyses of interacting protein modules. **a**, Phylogenetic profile of ACP interactions (bait–partner) within 148 genomes. Coloured boxes indicate degree of BLAST homology. Black font, bait–partner both essential-conserved proteins; blue font, only bait is an essential-conserved protein. Phylogenies from the NCBI taxonomy database (see Supplementary Information).  $\gamma$ ,  $\alpha$ ,  $\beta$  and  $\delta\epsilon$  indicate respective proteobacteria. Ba, Bacilli; Cl, Clostridia; Ac, Actinobacteridae; My, *Mycoplasma*; Ch, Chlamydiae; Cy, Cyanobacteria; Ot, unclassified bacteria; Ar, Archaea; Eu, Eukaryota. **b**, Scatter log-plot of Hamming distance ( $D$ ) versus chance co-occurrence probability distribution ( $P$ ) for

interactions associated with RNAP, DNA polymerase, ACP, the complete validated data set or random protein pairs (control). The dotted box indicates interactions (~14%) with significantly correlated profiles. **c**, Bar graph of relative genomic distances (base pairs) between genes encoding interacting proteins. Orange/yellow, orthologues in bacteria (excluding *E. coli*; see Supplementary Fig. 1); purple/green, orthologues in proteobacteria (excluding *E. coli*); blue/grey, *E. coli*. Error bars indicate standard deviation for 20 replicate controls.



chromosomal distances between loci encoding pairs of interacting proteins revealed that only a modest proportion (3.4%) of the interacting proteins were encoded by genes located within 500 base pairs (bp) of each other within the *E. coli* genome (Fig. 4c). A slightly larger fraction (6.9%) of putative orthologues was similarly separated by less than 500 bp in at least one other bacterial genome (Fig. 4c; see also Supplementary Table 6). Importantly, of the 42 protein pairs satisfying this criterion, 18 are encoded within the same operon in *E. coli* (see Supplementary Information), including 5 of the 15 interactions involving AcpP, further validating these data. These results indicate that only a fraction of our experimentally detected bacterial protein interactions could be readily predicted by genome-context methods (see above and Supplementary Table 3).

In summary, a reliable network of functionally diverse protein complexes was elucidated in *E. coli* using an experimental approach that can be readily adapted to other prokaryotes<sup>6</sup>. These data offer an insight into the function of uncharacterized proteins and outline the topological organization of a bacterial interactome. Because only about 30 bacterial proteins are currently targeted by prescription drugs<sup>28</sup>, knowledge of physical interactions mediated by conserved, essential bacterial proteins should facilitate the design of broad-range antimicrobials. These data should also prove valuable for calibrating computational approaches designed to predict functional associations between proteins. Moreover, the tagged strain collection should facilitate biochemical studies using traditional or microarray-based assays. □

## Methods

### Construction of TAP/SPA-tagged *E. coli* strains

*Escherichia coli* strains bearing either TAP- or SPA-tagged alleles were constructed by targeted homologous recombination of DNA cassettes into the *E. coli* strain DY330 as previously reported<sup>6</sup>. Primer sequences are available upon request. The SPA tag was selected for use after initial trials due to the reduced protein degradation observed when compared with TAP-purified protein complexes.

### Large-scale SPA/TAP purification of *E. coli* protein complexes

TAP or SPA purification was performed using 2–41 log-phase cultures as previously described<sup>6</sup> except that Benzonase nuclease (Novagen; 3 U) was incubated with the cleared cell extract on ice for 30 min before purification. Purified complexes were split into aliquots for SDS-PAGE and liquid chromatography-tandem mass spectrometry analysis as detailed in the text.

### Identification of proteins by mass spectrometry

Complex subunits were separated by SDS-PAGE on 12% acrylamide gels with a Whatman V16 vertical gel apparatus run at low current (9 mA) for 16 h. Gels were silver-stained using a standard protocol, except that formaldehyde crosslinking was not performed (details available upon request). Protein bands were excised and analysed as described previously<sup>29</sup>. Gel-free protein sequencing was performed by microcapillary-scale liquid chromatography-electrospray-ion trap tandem mass spectrometry as described<sup>29</sup>. Spectra were searched against an in-house database of predicted *E. coli* protein-coding sequences.

### Selection of target gene products

ORFs were selected for study to obtain broad biological coverage, including highly conserved essential and non-essential proteins, proteins with putative functional assignments, and hypothetical uncharacterized ORFs. Proteins known or predicted to contain trans-membrane helices were avoided owing to technical difficulties associated with purifying membrane proteins. The essential-conserved set of genes was selected using a basic rule set (see Supplementary Information).

### Connectivity distribution

To assess the correlation between connectivity ( $k$ ) and frequency ( $P(k)$ ), and between connectivity and the number of genomes a homologue was detected in, we calculated the Pearson's correlation coefficient  $R$ .

### Phylogenetic analysis of protein complexes

For each *E. coli* sequence, a TBLASTN<sup>30</sup> search was performed against each of the different organism genome data sets. In addition, to avoid complications caused by intronic regions, a protein data set was obtained for each eukaryotic organism considered here and a BLASTP<sup>30</sup> performed. The raw score for the highest sequence match to each data set was extracted and stored in a local database. Phylogenetic interaction profiles were visualized using a java applet developed in house.

### Phylogenetic distribution of proteins and interactions

The phylogenetic distribution of the proteins was obtained as previously described<sup>21</sup> using

as reference the non-redundant protein sequence database SWALL (SwissProt plus TrEMBL; see Supplementary Methods).  $P(h)$  represents the frequency of proteins with a homologue in the corresponding taxonomic groups. We used BLAST<sup>30</sup> with a threshold raw score of 50 and default parameters. We considered a protein interaction to be conserved if both interacting proteins have detectable homologues in any of the 148 complete genomes analysed (see legend to Supplementary Fig. 1). To analyse statistical significance, 30 control set samples of equal size were taken from the *E. coli* genome as previously described<sup>21</sup>. We used a two-tailed  $t$ -test, at a 95% confidence level.

### Phylogenetic profiles

Each gene is represented by a vector representing the pattern of co-occurrence across 148 genomes (with a value of 1 assigned when a homologue is present, and 0 when one is not). The extent of correlation between phylogenetic profiles of pairs of interacting proteins was assessed by computing Hamming distance ( $D$ ) and the chance co-occurrence probability distribution ( $P$ )<sup>21</sup>. If  $N$  is the total number of genomes over which we construct a phylogenetic profile for  $R$  genes,  $X$  and  $Y$  are the number of genomes in which homologues of two genes occur, and  $z$  is the number of genomes in which the genes co-occur, then  $P = w_z \bar{w}_z / W$ , where  $w_z$  is the number of ways to distribute  $z$  co-occurrences over  $N$  genomes,  $\bar{w}_z$  is the number of ways to distribute the rest of the  $X - z$  and  $Y - z$  genes over the rest of the  $N - z$  lineages, and  $W$  is the number of ways to distribute  $X$  and  $Y$  over  $N$  genomes without restriction, and  $D = x + y - 2z$ .

### Genomic distance

Relative distances between genes encoding each pair of interacting proteins were calculated using the chromosome location coordinates of genes in COGs<sup>22</sup>.

Received 6 October; accepted 3 December 2004; doi:10.1038/nature03239.

- Neidhardt, F. (ed.) *Escherichia coli and Salmonella: Cellular and Molecular Biology* (ASM Press, Washington DC, 1996).
- Serres, M. H. *et al.* A functional update of the *Escherichia coli* K-12 genome. *Genome Biol.* **2**, research0035.1–0035.7 (2001).
- Gavin, A. C. *et al.* Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature* **415**, 141–147 (2002).
- Datsenko, K. A. & Wanner, B. L. One-step inactivation of chromosomal genes in *Escherichia coli* K-12 using PCR products. *Proc. Natl Acad. Sci. USA* **97**, 6640–6645 (2000).
- Yu, D. *et al.* An efficient recombination system for chromosome engineering in *Escherichia coli*. *Proc. Natl Acad. Sci. USA* **97**, 5978–5983 (2000).
- Zeghouf, M. *et al.* Sequential Peptide Affinity (SPA) system for the identification of mammalian and bacterial protein complexes. *J. Proteome Res.* **3**, 463–468 (2004).
- Stukenberg, P. T. & O'Donnell, M. Assembly of a chromosomal replication machine: two DNA polymerases, a clamp loader, and sliding clamps in one holoenzyme particle. V. Four different polymerase-clamp complexes on DNA. *J. Biol. Chem.* **270**, 13384–13391 (1995).
- Harmon, F. G., Brockman, J. P. & Kowalczykowski, S. C. RecQ helicase stimulates both DNA catenation and changes in DNA topology by topoisomerase III. *J. Biol. Chem.* **278**, 42668–42678 (2003).
- Witte, G., Urbanek, C. & Curth, U. DNA polymerase III chi subunit ties single-stranded DNA binding protein to the bacterial replication machinery. *Nucleic Acids Res.* **31**, 4434–4440 (2003).
- Nakayama, H. RecQ family helicases: roles as tumor suppressor proteins. *Oncogene* **21**, 9008–9021 (2002).
- von Mering, C. *et al.* Comparative assessment of large-scale data sets of protein–protein interactions. *Nature* **417**, 399–403 (2002).
- Salwinski, L. *et al.* The Database of Interacting Proteins: 2004 update. *Nucleic Acids Res.* **32** (Database issue), D449–D451 (2004).
- Bader, G. D., Betel, D. & Hogue, C. W. BIND: the Biomolecular Interaction Network Database. *Nucleic Acids Res.* **31**, 248–250 (2003).
- von Mering, C. *et al.* STRING: a database of predicted functional associations between proteins. *Nucleic Acids Res.* **31**, 258–261 (2003).
- Bowers, P. *et al.* Prolinks: a database of protein functional linkages derived from coevolution. *Genome Biol.* **5**, R35 (2004).
- Rain, J. C. *et al.* The protein–protein interaction map of *Helicobacter pylori*. *Nature* **409**, 211–215 (2001).
- Gully, D., Moinier, D., Loiseau, L. & Bouveret, E. New partners of acyl carrier protein detected in *Escherichia coli* by tandem affinity purification. *FEBS Lett.* **548**, 90–96 (2003).
- Wuchty, S., Oltvai, Z. N. & Barabasi, A. L. Evolutionary conservation of motif constituents in the yeast protein interaction network. *Nature Genet.* **35**, 176–179 (2003).
- Jordan, I. K., Wolf, Y. I. & Koonin, E. V. No simple dependence between protein evolution rate and the number of protein–protein interactions: only the most prolific interactors tend to evolve slowly. *BMC Evol. Biol.* **3**, 1 (2003).
- Fraser, H. B., Wall, D. P. & Hirsh, A. E. A simple dependence between protein evolution rate and the number of protein–protein interactions. *BMC Evol. Biol.* **3**, 11 (2003).
- Peregrin-Alvarez, J. M., Tsoka, S. & Ouzounis, C. A. The phylogenetic extent of metabolic enzymes and pathways. *Genome Res.* **13**, 422–427 (2003).
- Tatusov, R. L., Koonin, E. V. & Lipman, D. J. A genomic perspective on protein families. *Science* **278**, 631–637 (1997).
- Wu, J., Kasif, S. & DeLisi, C. Identification of functional links between genes using phylogenetic profiles. *Bioinformatics* **19**, 1524–1530 (2003).
- Pellegrini, M., Marcotte, E. M., Thompson, M. J., Eisenberg, D. & Yeates, T. O. Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proc. Natl Acad. Sci. USA* **96**, 4285–4288 (1999).
- Date, S. V. & Marcotte, E. M. Discovery of uncharacterized cellular systems by genome-wide analysis of functional linkages. *Nature Biotechnol.* **21**, 1055–1062 (2003).
- Yu, H. *et al.* Annotation transfer between genomes: protein–protein interologs and protein–DNA regulogs. *Genome Res.* **14**, 1107–1118 (2004).
- Overbeek, R., Fonstein, M., D'Souza, M., Pusch, G. D. & Maltsev, N. The use of gene clusters to infer

- functional coupling. *Proc. Natl Acad. Sci. USA* **96**, 2896–2901 (1999).
28. Haselbeck, R. *et al.* Comprehensive essential gene identification as a platform for novel anti-infective drug discovery. *Curr. Pharm. Des.* **8**, 1155–1172 (2002).
29. Krogan, N. J. *et al.* High-definition macromolecular composition of yeast RNA-processing complexes. *Mol. Cell* **13**, 225–239 (2004).
30. Altschul, S. F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389–3402 (1997).

Supplementary Information accompanies the paper on [www.nature.com/nature](http://www.nature.com/nature).

**Acknowledgements** The authors wish to thank C. J. Ingles and M. Shales for comments on the manuscript. This work was supported by funds from the Ontario Research and Development Challenge Fund and Genome Canada to A.E. and J.G. G.B. was a recipient of a Charles H. Best Post-Doctoral Fellowship. J.M.P.-A. acknowledges support from the Hospital for Sick Children (Toronto, Ontario, Canada) Research Training Centre. Computer analyses were undertaken at the Centre for Computational Biology, Hospital for Sick Children.

**Authors' contributions** Informatics studies were performed and analysed by J.M.P.-A. and J.P. Experimental design and data analysis were coordinated by G.B. Tagging and purification experiments were performed by W.Y., X.Y., J.L. and G.B. V.C., A.S., D.R., B.B., N.J.K. and M.D. performed and assisted with mass spectrometry analysis. The manuscript was jointly drafted by G.B., A.E., J.G., J.M.P.-A. and J.P. The project was conceived and designed by J.G. and was directed by A.E.

**Competing interests statement** The authors declare that they have no competing financial interests.

**Correspondence** and requests for materials should be addressed to A.E. ([andrew.emili@utoronto.ca](mailto:andrew.emili@utoronto.ca)) or J.G. ([jack.greenblatte@utoronto.ca](mailto:jack.greenblatte@utoronto.ca)).

## Nanoarchaeum equitans creates functional tRNAs from separate genes for their 5' - and 3' -halves

Lennart Randau<sup>1,2</sup>, Richard Münch<sup>2</sup>, Michael J. Hohn<sup>1,3</sup>, Dieter Jahn<sup>2</sup> & Dieter Söll<sup>1,4</sup>

<sup>1</sup>Department of Molecular Biophysics and Biochemistry, Yale University, 266 Whitney Avenue, New Haven, Connecticut 06520-8114, USA

<sup>2</sup>Institut für Mikrobiologie, Technical University Braunschweig, P.O. Box 3329, D-38023 Braunschweig, Germany

<sup>3</sup>Lehrstuhl für Mikrobiologie und Archäozentrum, Universität Regensburg, Universitätsstrasse 31, D-93053 Regensburg, Germany

<sup>4</sup>Department of Chemistry, Yale University, 266 Whitney Avenue, New Haven, Connecticut 06520-8114, USA

Analysis of the genome sequence of the small hyperthermophilic archaeal parasite *Nanoarchaeum equitans*<sup>1,2</sup> has not revealed genes encoding the glutamate, histidine, tryptophan and initiator methionine transfer RNA species. Here we develop a computational approach to genome analysis that searches for widely separated genes encoding tRNA halves that, on the basis of structural prediction, could form intact tRNA molecules. A search of the *N. equitans* genome reveals nine genes that encode tRNA halves; together they account for the missing tRNA genes. The tRNA sequences are split after the anticodon-adjacent position 37, the normal location of tRNA introns. The terminal sequences can be accommodated in an intervening sequence that includes a 12–14-nucleotide GC-rich RNA duplex between the end of the 5' tRNA half and the beginning of the 3' tRNA half. Reverse transcriptase polymerase chain reaction and aminoacylation experiments of *N. equitans* tRNA demonstrated maturation to full-size tRNA and acceptor activity of the tRNA<sup>His</sup> and tRNA<sup>Glu</sup> species predicted *in silico*. As the joining mechanism possibly involves tRNA *trans*-splicing, the presence of an intron might have been required for early tRNA synthesis.

The origin of the tRNA molecule is the subject of continuing

discussions and has led to different models postulating that tRNA evolved by duplication or ligation of an RNA hairpin<sup>3,4</sup>. To examine these models further, the investigation of ancient tRNA genes was central. An interesting organism for this task was *Nanoarchaeum equitans*, currently the only characterized member of the kingdom Nanoarchaeota, which roots early in the archaeal lineage, before the emergence of Euryarchaeota and Crenarchaeota<sup>5</sup>. A significant fraction of the small number of *N. equitans* open reading frames consists of 'split genes' that are encoded as fused versions in other archaeal genomes. Our attention was caught by the 'absence' of four tRNA genes encoding the glutamate, histidine, tryptophan and initiator methionine acceptors<sup>6</sup>.

We therefore developed a computational approach to search for tRNA signature sequences in the *N. equitans* genome. Our program, trained by an alignment of 4,000 tRNA gene sequences (taken from ref. 6), identifies sequences comprising the highly conserved T-loop region and defines the adjacent 3'-acceptor stem sequence. The reverse complementary sequence (defining the 5'-acceptor stem sequence) plus a D-stem position weight matrix identifies the corresponding 5' half. The length of the position weight matrices can be adjusted and mismatches in the acceptor stem can be included. Finally, putative tRNA-halves are ligated *in silico* and analysed by COVE<sup>7</sup>. In addition to identifying the set of tRNAs predicted by the tRNAscan-SE program<sup>8</sup>, our algorithm found nine tRNA halves spread throughout the chromosome. Surprisingly, these tRNA halves could be joined *in silico* to form the missing tRNA<sup>His</sup>, tRNA<sup>Met</sup>, tRNA<sup>Trp</sup> and two tRNA<sup>Glu</sup> species (Fig. 1). Further analysis of the tRNA half genes revealed several striking features. First, the location of the sequence separation that generated all nine tRNA half genes is after position 37, one nucleotide downstream of the anticodon and the common location of tRNA introns<sup>9</sup>. Second, a consensus sequence matching the highly conserved archaeal Box A promoter element<sup>10</sup> was found upstream of all 5' tRNA halves. Third, this same consensus sequence (5'-TTTT/ATAAA-3') was located 17–25 base pairs (bp) further upstream of the 3' tRNA halves, resulting in a transcript with a 12–14-bp-long GC-rich leader sequence. Last, it is remarkable that this leading sequence is in all cases the exact reverse complement to a sequence following the corresponding 5' tRNA half.

The existence of three tRNA<sup>Glu</sup> half genes was most exciting. Two 5' tRNA halves were identified that differed solely by one anticodon base (isoacceptors with UUC and CUC anticodon), whereas only one 3' tRNA<sup>Glu</sup> half gene was found. Both 5' tRNA<sup>Glu</sup> half genes were followed by the identical 14-bp sequence that was the exact reverse complement of the single 3' tRNA<sup>Glu</sup> half upstream sequence. All identified split tRNA genes contained the consensus bases of all archaeal elongator tRNAs<sup>6</sup>, namely U8, A14, G15, G18, G19, C32, U33 and the T-loop GTTCA/GAATC (53–61), with the exception of the putative tRNA<sup>Trp</sup> harbouring an unusual GG sequence preceding the anticodon. The identified tRNA<sup>Met</sup> displays the consensus sequences of archaeal initiator tRNAs such as the anticodon stem/loop nucleotides (nt) 29–41 (GGGCU-CAUAACCC) and the R11:Y24 base pair (G11:C24), which is the reverse of the Y11:R24 base pair found in elongator tRNAs including the annotated *N. equitans* tRNA<sup>Met</sup>. Therefore we define the split tRNA<sup>Met</sup> as the missing initiator tRNA. The sequences also reveal characteristic nucleotides in the respective tRNA species needed for recognition by the cognate aminoacyl-tRNA synthetase. For example, the tRNA<sup>His</sup> half genes encode the unique G-1:C73 base pair required for aminoacylation of tRNA<sup>His</sup> by histidyl-tRNA synthetase<sup>11</sup>, and the tRNA<sup>Glu</sup> isoacceptors contain the characteristic D-loop nucleotides 20a and 20b and the deletion of base 47 essential for making the 'augmented D-helix'<sup>12</sup>.

We performed reverse transcriptase polymerase chain reaction (RT-PCR) analysis of *N. equitans* total tRNA to verify the computationally predicted sequence of the newly discovered joined tRNAs. Our sequencing results confirmed the sequences for tRNA<sup>Glu</sup>