Published online February 10, 2004

Nucleic Acids Research, 2004, Vol. 32, No. 3 ***1059–1064***
DOI: 10.1093/nar/gkh254

Protein interaction mapping on a functional shotgun sequence of *Rickettsia sibirica*

Joel A. Malek*, Jamey M. Wierzbowski, Wei Tao, Stephanie A. Bosak, David J. Saranga, Lynn Doucette-Stamm, Douglas R. Smith, Paul J. McEwan and Kevin J. McKernan

Agencourt Bioscience Corporation, 100 Cummings Center Suite 107G, Beverly, MA 01915, USA

Received November 20, 2003; Revised and Accepted January 7, 2004 DDBJ/EMBL/GenBank accession no. AABW01000001

ABSTRACT

Protein interaction maps can reveal novel pathways and functional complexes, allowing 'guilt by association' annotation of uncharacterized proteins. To address the need for large-scale protein interaction analyses, a bacterial two-hybrid system was coupled with a whole genome shotgun sequencing approach for microbial genome analysis. We report the first large-scale proteomics study using this system, integrating *de novo* genome sequencing with functional interaction mapping and annotation in a high-throughput format. We apply the approach by shotgun sequencing and annotating the genome of *Rickettsia sibirica* strain 246, an obligate intracellular human pathogen among the Spotted Fever Group rickettsiae. The bacteria invade endothelial cells and cause lysis after large amounts of progeny have accumulated. Little is known about specific Rickettsial virulence factors and their mode of pathogenicity. Analysis of the combined genomic sequence and protein–protein interaction data for a set of virulence related Type IV secretion system (T4SS) proteins revealed over 250 interactions and will provide insight into the mechanism of Rickettsial pathogenicity.

INTRODUCTION

The utility of protein interaction maps is extensively documented and it is well appreciated that genome annotation could benefit from protein interaction data (1–6). Two-hybrid projects require protein-coding regions be cloned in expression vectors as reagents for genetic screens (7). To obtain these, a shotgun strategy is preferable. This approach rapidly generates multiple overlapping fragments for any region. Use of fragment libraries in two-hybrid screens has been shown to reduce false-negatives (8,9). An additional benefit of this strategy is the ability to localize domains responsible for interactions, in both bait and prey constructs, similar in concept to deletion studies. To link genome sequencing and protein interaction mapping in a pipeline, a bacterial version of the yeast two-hybrid (Y2H) system is well suited. Bacterial two-hybrid (B2H) systems have been developed and proven reliable for several selected gene products (10,11), but to date B2H has not been applied to large scale protein interaction mapping as with the well established Y2H system (4–6).

The B2H system used in this study was developed by Hochschild and colleagues (12–14), and is similar in concept to the standard Y2H system. We selected this version of the B2H system as it allows for random cloning of fragments because proteins are fused C-terminally to binding or activation domains. In addition, the system has been validated in various screening studies and shown to provide reproducible interaction data (13–15). Briefly, a protein of interest (the bait) is fused to λ cI, a DNA binding domain, which binds to a λ operator sequence, OR2, placed upstream of a weak promoter. In addition, a second protein of interest (the prey) is fused to the RNA polymerase (RNAP) α subunit, an activation domain, which is part of the RNAP holoenzyme (Fig. 1a). If the two proteins of interest interact, RNAP is recruited to the weak promoter causing increased transcription of the downstream reporter genes, β-lactamase and β-galactosidase (Fig. 1b). Colonies expressing the reporter genes can then be selected on appropriate media (see Materials and Methods). Since fusion proteins are generated from standard backbone vectors and expressed in *Escherichia coli*, sequencing of inserts to determine interacting proteins is greatly simplified. Utilizing this system, we developed a process termed 'functional shotgun sequencing' in which a shotgun library is constructed in the bait vector, followed by determination of open reading frame (ORF) fragments that are cloned in the correct frame and can be used as bait (Fig. 2). Alternatively, the fragments can be shuttled to the appropriate vector for use as prey. This process allows for the rapid development of a resource of cloned overlapping ORF fragments from which protein interaction screens in the B2H can begin.

MATERIALS AND METHODS

Modification of the bacterial two-hybrid vectors

Original vectors from the system developed by Hochschild and colleagues (10) were modified as follows: pACλcI32 and pBRstar (15) were modified by re-introducing the NotI site followed by a BstXI restriction site, XhoI restriction site plus three frame stop codons. Constructs were renamed pBAIT and pPREY respectively. To verify that vector modifications did not alter functionality, $Gal11^{P}$ and Gal4 fragments (14), were

*To whom correspondence should be addressed. Tel: +1 978 867 2632; Fax: +1 978 867 2601; Email: jamalek@agencourt.com

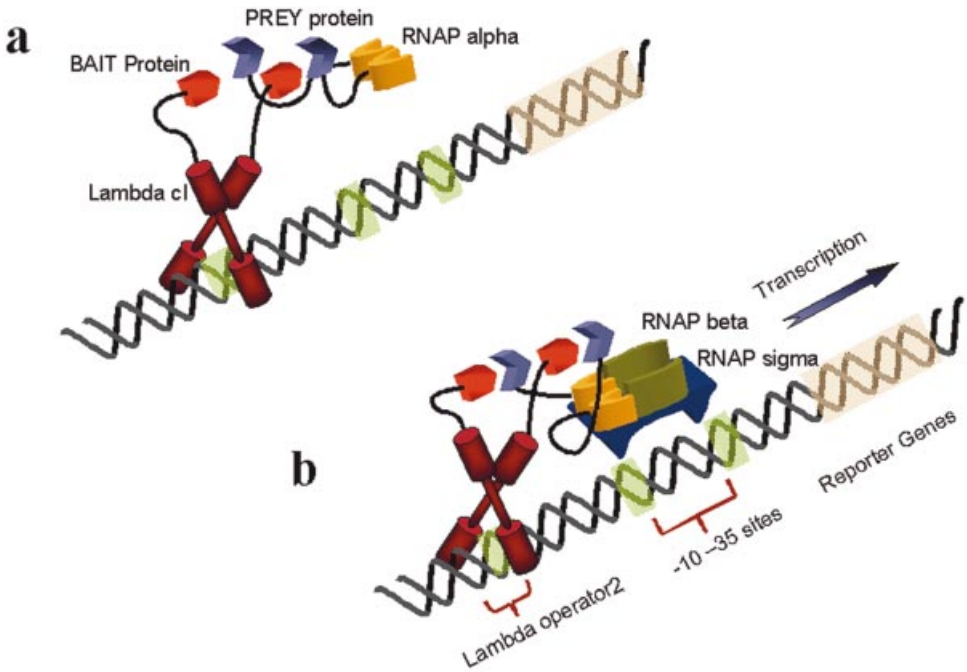

Figure 1. The bacterial two-hybrid system. (**a**) The λ cI fused to the bait protein dimerizes and binds the λ operator. (**b**) A protein interaction between the bait and prey protein recruits the RNAP complex, via the RNAP α subunit, to the weak promoter site directing transcription of the reporter genes.

cloned into pBAIT and pPREY using BstXI adapters, screened and the capability to interact verified.

Functional shotgun sequencing of *Rickettsia sibirica*

Genomic DNA from *Rickettsia sibirica* was randomly sheared using nebulization. Fragments of 750 bp were gel purified, adapted with BstXI adapters and ligated into pBAIT. From this pBAIT library 27 314 sequences were attempted yielding 25 210 successful reads with average assembled length of 621 bp. Sequence coverage of the ~1.25 Mbp genome by pBAIT sequences was 12.5×. Coverage of the genome was checked for significant deviation from the expected mean and no regions of unusual over- or under-representation were found. To improve and verify the assembly, a fosmid library was generated by shearing genomic DNA to average size of 40 kb followed by cloning using fosmid packaging. Paired-end reads from 5044 fosmid clones were attempted, yielding 8322 successful reads. Sequence coverage of the genome by fosmid reads was 3.3×.

Assembly and annotation

Assembly was conducted using the Paracel Genome Assembler™ (Paracel, Pasadena, CA, USA) and ordered by paired-end reads. Protein coding regions were initially determined by GeneMarkS (16) and assigned function based on BLASTP analysis using the GenBank NR database. Sequences from interactions were annotated using the COG database to create protein families (see Supplementary Material, Table S2).

Selection of in-frame fragments and creation of the prey libraries

pBAIT clones containing in-frame fragments of genes were determined by translation of nucleotide sequence oriented by the vector/insert junction. Translated fragments were then searched against the set of determined ORFs of *R.sibirica* for similarity. Clones determined to contain ORF fragments expressed in the correct frame were re-arrayed to fresh plates creating a set of ORF fragments for screening. For the ORF fragment prey library, all pBAITs determined through sequencing to contain in-frame fragments were arrayed, grown to stationary phase and plasmid DNA prepared. Inserts were excised and re-ligated into pPREY sites maintaining directionality. The ligation was transformed, plated and ~2 million colonies were scraped for DNA preparation. For the screening shotgun library, adapted *R.sibirica* DNA from the shotgun sequencing project was mixed at a 1000:1 ratio with a control insert Gal11^{P}, to serve as a downstream positive control, ligated into pPREY vector, transformed and ~6 million colonies were plated. After overnight growth the colonies were scraped and plasmid DNA extracted using standard methods. One hundred clones from both the ORF and shotgun library pPREY library were sequenced to ensure the library was random. For baits, 17 overlapping peptide fragments were found spanning regions of; VirD4 (rsib_orf.311) a.a. 2–591, VirB11 (rsib_orf.312) a.a. 108–334, VirB10 (rsib_orf.313) aa 9–89 and a.a.125–483, VirB9 (rsib_orf.314) a.a. 80–157, VirB8′ (rsib_orf.315) a.a. 83–243, VirB7 (rsib_orf.316) a.a. 39–52, VirB8 (rsib_orf.317) a.a. 3–227.

Screening in the bacterial two-hybrid system

For screening, the Bacteriomatch™ reporter strain (Stratagene, La Jolla, CA, USA) was used. This strain harbors the reporter episome pFWO62SD+bla15 used in reporter strain US3F′3.1. pBAIT DNA from clones containing peptide fragments of interest was prepared in 96-well plates using standard alkaline lysis methods. Each peptide of interest was

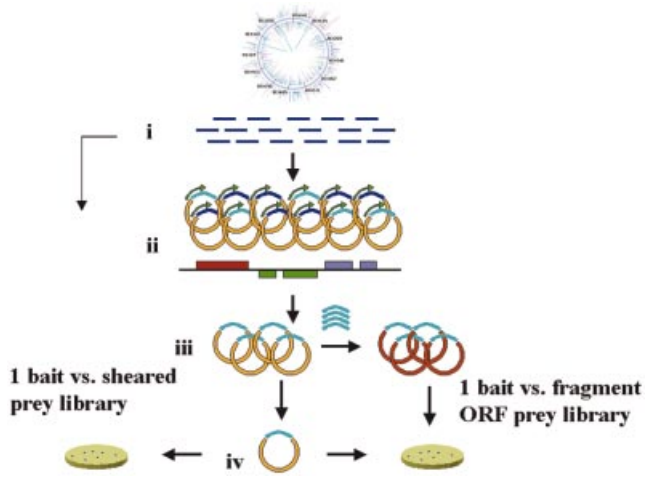

Figure 2. Functional shotgun sequencing pipeline. (i) Genomic DNA is sheared and cloned into bait and prey vectors. (ii) Randomly selected bait clones are sequenced, the data assembled and the genome annotated. (iii) Clones determined to contain fragments of genes expressed in the correct frame are re-arrayed for screening. A copy of the set is pooled, and the inserts transferred to the prey vector creating the fragment ORF prey library. (iv) Baits from proteins of interest are either screened against the previously created sheared genomic prey library, or the shuttled fragment ORF prey library. Sequencing of positive clones directly from selected colonies is conducted with pBAIT or pPREY specific primers.

transformed using 100 µl of Bacteriomatch™ reporter strain cells, 50 ng of pBAIT and 50 ng of either ORF library or shotgun library pPREY DNA. This yielded 650 000 dual transformants on average. Dual transformants were plated on 25 cm^2 plates containing LB agar supplemented with 25 µg/ml IPTG, 300 µg/ml carbenicillin, 2 µg/ml tetracycline, 50 µg/ml kanamycin and 12.5 µg/ml chloramphenicol. Small aliquots were also plated on media lacking carbenicillin to determine total dual transformation numbers. Screening was also conducted on minimal media plates containing the same antibiotics, IPTG amounts, but with lactose as the sole carbon source. This latter method selects for expression of both β-lactamase and β-galactosidase. At this level of dual transformation (650 000), the ORF fragment library was oversampled ~160×. In the case of the shotgun library this was ~40× coverage of the proteome. All colonies, or up to 400 colonies, growing after 16 h were picked for secondary screening on agar plates containing all previous ingredients plus IPTG and Xgal. After overnight incubation, colonies yielding significant blue color when compared with negative controls were picked for sequencing.

Categorization and validation of interactions

Interactions were categorized as follows: observed once, were assigned score 1; more than once were assigned score 2; and more than once by different fragments were assigned score 3. This categorization represents the levels of validation of any given interaction. All screening against libraries was conducted in conjunction with a negative control pBAIT expressing the λ cI alone. Colonies from these screens were sequenced and one false-positive, rsib_orf.1344, was identified. Both plasmids from 24 randomly selected interactors were prepared and re-transformed into the selection strain.

Table 1. Comparison of spotted fever group rickettsiae genomes

	R.sibirica	*R.conorii*
Protein-coding regions	1234	1373
Average protein-coding gene length (bp)	787	746
% coding	77.7	80.8
% G+C	32.9	32.9
Genome size (bp)	1 250 021	1 268 755

Twenty-three of the original 24 clones revealed reconstituted interactions corresponding well with previously determined β-galactosidase activity levels. This suggests that ~4% of interactions may have occurred due to breakthrough of the reporter strain.

RESULTS

The genome of *R.sibirica* 246 was subjected to functional shotgun sequencing, assembly, gene identification and automated annotation. Sequencing reads assembled into one supercontig consisting of seven ordered and oriented contigs. A total of 1234 putative genes were identified having an average coding length of 787 bp (Table 1), comprising 972 024 protein-coding bases or 324 008 amino acids. As expected, the identified *R.sibirica* genes displayed a high degree of sequence conservation with genes of *Rickettsia conorii* whose genome is completely sequenced (17). A total of 3932 sequences, when translated in frame with λ cI, revealed a cloned fragment from a *R.sibirica* ORF. The 3932 in-frame clones spanned 599 602 amino acids or about 1.85× proteome redundancy. At this level of proteome coverage, predicted missing coverage will be $P = e^{-1.85}$ or ~15.7% (18). Thus, ~85% of the proteome, or 1040 proteins, should be represented in the identified clone set. In fact, in-frame fragments from 986 ORFs covering 278 832 unique amino acids were observed, corresponding to 86% of all amino acids being covered at least once. These numbers agree well with predicted coverage. From this set of clones, we were able to select clones spanning regions of interest for screening against either the sheared genomic prey library or the shuttled ORF fragment prey library.

Rigorous testing of the B2H system with proteins known to interact showed that use of minimal media and higher levels of reporter antibiotic in conjunction with IPTG were crucial in obtaining reproducible interactions consistently. We observed that allowing cells to recover greater than 2 h after transformation resulted in increased basal level reporter antibiotic resistance reducing the capacity to select for interactions based solely on the reporter antibiotic resistance. As an initial test of the system with large, dimerizing proteins, we amplified the full-length *E.coli MoeA* gene by PCR from the genome and cloned into pBAIT and pPREY. This protein was selected for study as some criticism of the unmodified B2H system has centered on reported interactions only occurring between small (40–150 aa) moieties which themselves do not dimerize. *MoeA*, a 411 aa, protein has been shown to interact with itself in another B2H system (19). Transformation in the reporter strain and selection revealed a strong interaction in our system. Both pBAIT and pPREY

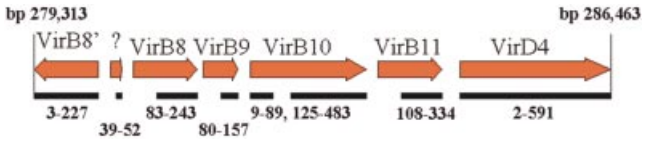

Figure 3. The T4SS region of the genome studied. Genes are represented as arrows, and regions covered by multiple in-frame ORF fragments as the black bars underneath each arrow. Coordinates for each protein are by amino acid position.

containing the *MoeA* gene were transformed with empty pPREY or pBAIT plasmids respectively with no observed interaction verifying the plasmid dependence of the interaction.

The region of the genome including the virulence cluster VirD4-VirB8 was selected for further study. The proteins encoded by genes in this region are of interest because of their apparent role in virulence and their relationship to the Type IV secretion system (T4SS) found in numerous pathogens. In some organisms, T4SS have been shown to be responsible for delivery of effector molecules to cells of the eukaryotic host organism. Studies have been conducted on interactions among subunits of the T4SS complex in several microbes (9,20–22). While interactions among the subunits have been characterized, interactions between the T4SS complex and other proteins, such as secreted effectors, have not been as well characterized. Given the presumed conservation of interactions between ortholog pairs in different species, termed interologs (23), we expected to obtain interactions among the T4SS subunits similar to those identified in other organisms using the Y2H system. Seventeen in-frame ORF fragments spanning portions of the VirD4-VirB8 (Fig. 3) region of the genome were selected for screening as baits against a sheared genomic prey library and a shuttled ORF fragment prey library. Screens against the ORF fragment library identified almost all interactions found using the shotgun library and determination of which approach to use should be set based on the scale of screening. Screens against the ORF fragment library produced fewer false-positives (small non-genic peptides that appear to interact ubiquitously), and required less sequencing. However, the shotgun prey library provided better resolution of the interaction domains (Fig. 4).

Screening yielded 284 distinct interactions between 155 proteins or protein families and the six T4SS subunits screened (Fig. 5). One hundred and sixty-two interactions fell into category 1 (observed once), 48 in category 2 (observed more than once) and 74 in category 3 (observed more than once using different fragments) (as in Fig. 4) (see Supplementary Material, Table S1). Of the proteins involved in the 162 interactions observed once, 84% interacted with at least one other T4SS subunit. Twenty-four proteins lacking any annotation were also found to interact with the T4SS subunits and these interactions have been selected for further validation and study. Forty six percent of the interactions previously reported among T4SS subunits in other organisms using the Y2H were obtained using the B2H system (Table 2). We identified two intra-complex interactions among T4SS subunits not previously detected in studies of other organisms using the Y2H (Table 2). Interactions were found between the T4SS baits and

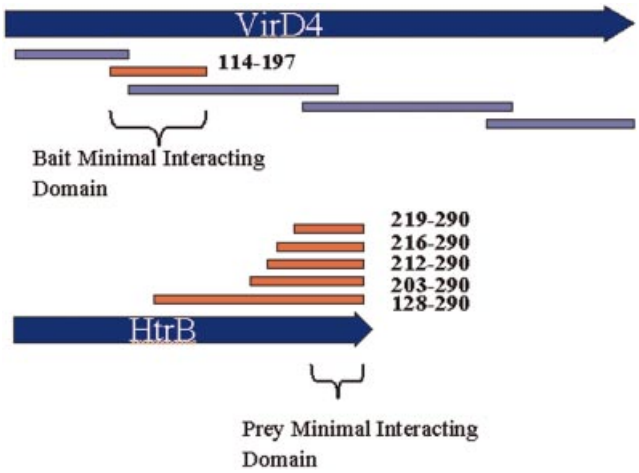

Figure 4. The localized interaction between VirD4 and HtrB. Fragments yielding an interaction are displayed as red bars. Among the various bait fragments screened from VirD4, only one interacted with prey fragments from HtrB. Multiple overlapping fragments were obtained from HtrB localizing the interaction to the C-terminus portion of the protein.

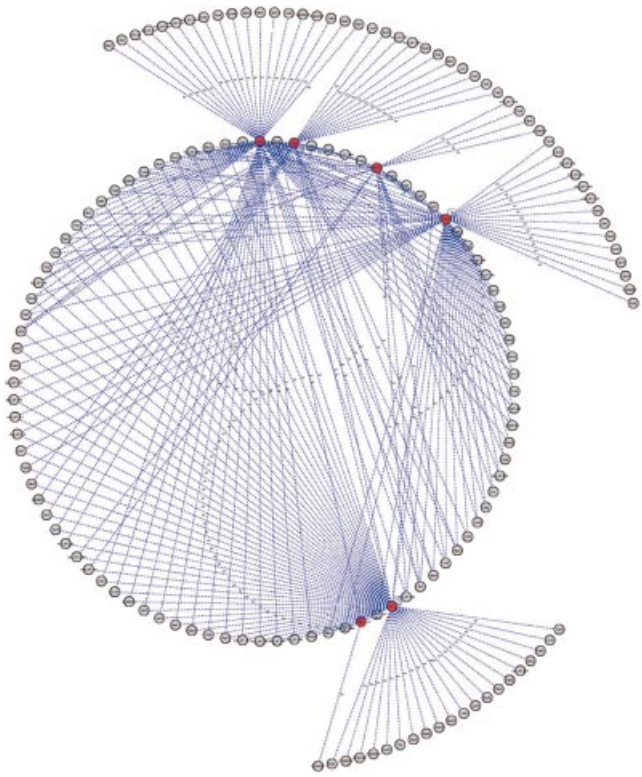

Figure 5. A Map of T4SS protein interactions. Nodes represent proteins while edges represent an interaction (25). Subunits of the T4SS used as baits are highlighted in red. The inner, full circle represents interactions shared among the subunits. The broken outer circle represents interactions distinct to a given subunit. Transported effectors may more likely be found in the inner circle as they would interact with more than one subunit.

lipopolysaccharide-related proteins, hemolysins, protein export proteins, proteases, permeases, outer membrane proteins, ABC transporters, proteins of unknown function and proteins of the T4SS complex among others.

Table 2. Overlap between Y2H studies of *A.tumefaciens* T4SS and this B2H study of *R.sibirica* T4SS

Overlap	Previous studies only	This study only
VirB8-VirB4	VirB8-VirB8	VirB10-VirD4
VirB9-VirB9	VirB8-VirB10	VirD4-VirB4
VirB9-VirB10	VirB9-VirB8	
VirB9-VirB11	VirB10-VirB11	
VirB10-VirB4	VirB11-VirB4	
VirB10-VirB10	VirB11-VirB8	
	VirB11-VirB11	

DISCUSSION

We have presented an approach for integrating genome sequencing and subsequent protein interaction mapping and tested it by sequencing the genome of a human pathogen. Numerous review papers have been written on large-scale two-hybrid studies, analyzing the data, testing the data's validity, and using the data to train *in silico* protein interaction prediction software. The need for further protein interaction information is clear. Among the challenges in generating proteome-wide interaction data are the lack of fully automated processes, the sheer amount of screening necessary to complete one map for one organism, and an incomplete grasp of what constitutes a physiologically important protein interaction.

Use of the B2H system in a large-scale screen required development of the screening protocols. Use of minimal media and IPTG induction were important steps in obtaining reproducible interactions. Our success in observing an interaction with the *E.coli MoeA* protein demonstrates that large proteins or dimers are not absolutely excluded from interacting in this system.

Analysis of some two-hybrid studies has shown a high false-positive rate in the sense that the interaction can be repeated in the system but may not occur in nature. While further validation of interactions observed in this study is necessary, the possible interactions have been narrowed significantly. In this study, 84% of the proteins involved in the 162 interactions that were observed only once, interacted with at least one other T4SS subunit, and 26% of all interactions were observed multiple times with independent fragments of the same protein. These observations lend strength to the potential significance of these interactions. The 46% of interactions overlapping with previous studies seems reasonable given the incomplete proteome coverage of the genome in our screen: regions of T4SS proteins identified as necessary for an interaction in *Agrobacterium tumefaciens* were missing in some of our screens. Furthermore, previous investigations of interologs using the Y2H system only reported between a 16 and 31% recapture rate (23). The three proteins involved in the two intra-T4SS complex interactions not previously observed have been shown to localize to the inner membrane. Taking this into consideration, these observed interactions do not conflict with previous models of the T4SS complex (9) and may indeed represent putative novel intra-complex interactions.

In this study, 24 previously unannotated proteins were found to interact with subunits of the T4SS. While these interactions require further validation, they have assisted in narrowing the search for novel transported effectors. One concern with two-hybrid systems is the existence of 'sticky proteins'; those proteins which interact with numerous proteins in the two-hybrid system but for which the interactions may not be physiologically relevant. Interaction mapping all remaining ORFs in the genome will reveal which proteins in the *R.sibirica* genome are 'sticky', assisting in further validating these interactions. A potential example of a 'sticky' protein was identified in this study: *DnaK*, a molecular chaperone interacted strongly with many of the T4SS subunits. Previous Y2H studies have shown chaperones and heat shock proteins to be candidate 'sticky' proteins.

With respect to virulence, interactions between subunits of the T4SS and *fadB*, *tlyC* and *tlyA* like (rsib_orf.1275) and their related multi-drug efflux proteins *acrAB*, were of interest. *FadB* was previously identified in a screen for genes expressed during intracellular infection. It was shown that *fadB* is activated in *Salmonella typhimurium* specifically during intracellular infection (24). *FadB* is involved in β-oxidation of fatty acids that may help suppress host inflammatory response (24). Observations of and interaction between the T4SS components and *fadB* strengthen the case for a role of this protein in pathogenicity. *TlyC* and *tlyA* are rickettsial hemolysins for which little is known aside from hemolytic activity. Evidence of association with the T4SS should be of interest for further investigation and validation.

The benefits of this approach are clear. Sequencing the genome in B2H vectors, while requiring higher coverage, yields in-frame ORF fragments ready for functional analysis. Use of a B2H system for screening of potential interactions then allows direct entry of plasmids into the well-developed DNA sequencing pipeline. Furthermore, use of bacterial strains offer high dual-transformation efficiencies allowing rapid, deep screening for multiple proteins of interest.

It is our hope that the present data will advance the understanding of mechanisms of virulence through the T4SS. It is our belief that using comparative interaction data will allow deciphering of what interactions are physiologically valid. These types of comparisons can truly begin once a few related genomes have had their protein interactions mapped. Future direction for this approach includes a large-scale comparison with the Y2H and *in vitro* methods to give a better grasp on how data from the various methods relate to each other. Completion of the proteome-wide interaction map is also a priority.

By implementing a large-scale B2H system, we have demonstrated that it is possible to couple whole genome sequencing and protein interaction mapping in a standard sequencing pipeline. With the interest growing in organisms difficult to culture, an approach that yields functional information starting from genomic DNA should be of interest. We hope that with further development, validation and scale-up of this approach, the number of organisms for which interaction maps have been elucidated will rapidly increase and bring this field into the era of cross-species comparative interaction studies. We believe this approach and data will help in development of new drug targets by providing information on genes that are critical to the pathogenicity, maintenance and spread of microbes.

SUPPLEMENTARY MATERIAL

Supplementary Material is available at NAR Online.

ACKNOWLEDGEMENTS

We thank Gregory A. Dasch and Marina E. Eremeeva for their helpful discussions of the manuscript, Lisa M. Campagnoni for help in development of the bacterial two-hybrid system, Erick Gustafson for extensive manuscript review, and the Agencourt sequencing team for production sequencing of the genome. Genomic DNA from *R.sibirica* strain 246 was kindly provided by Gregory A. Dasch and Marina E. Eremeeva of the Centers for Disease Control and University of Maryland respectively.

REFERENCES

1. Eisenberg,D., Marcotte,E.M., Xenarios,I. and Yeates,T.O. (2000) Protein function in the post-genomic era. *Nature*, **405**, 823–826.
2. Oliver,S. (2000) Proteomics: guilt-by-association goes global. *Nature*, **403**, 601–603.
3. Walhout,A.J., Reboul,J., Shtanko,O., Bertin,N., Vaglio,P., Ge,H., Lee,H., Doucette-Stamm,L., Gunsalus,K.C., Schetter,A.J. *et al.* (2000) Protein interaction mapping in *C.elegans* using proteins involved in vulval development. *Science*, **287**, 116–122.
4. Uetz,P., Giot,L., Cagney,G., Mansfield,T.A., Judson,R.S., Knight,J.R., Lockshon,D., Narayan,V., Srinivasan,M., Pochart,P. *et al.* (2000) A comprehensive analysis of protein–protein interactions in *Saccharomyces cerevisiae*. *Nature*, **403**, 623–627.
5. Ito,T., Tashiro,K., Muta,S., Ozawa,R., Chiba,T., Nishizawa,M., Yamamoto,K., Kuhara,S. and Sakaki,Y. (2000) Toward a protein–protein interaction map of the budding yeast: a comprehensive system to examine two-hybrid interactions in all possible combinations between the yeast proteins. *Proc. Natl Acad. Sci. USA*, **97**, 1143–1147.
6. Ito,T., Chiba,T., Ozawa,R., Yoshida,M., Hattori,M. and Sakaki,Y. (2001) A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc. Natl Acad. Sci. USA*, **98**, 4569–4574.
7. Reboul,J., Vaglio,P., Rual,J.F., Lamesch,P., Martinez,M., Armstrong,C.M., Li,S., Jacotot,L., Bertin,N., Janky,R. *et al.* (2003) *C.elegans* ORFeome version 1.1: experimental verification of the genome annotation and resource for proteome-scale protein expression. *Nature Genet.*, **34**, 35–41.
8. Rain,J.C., Selig,L., De Reuse,H., Battaglia,V., Reverdy,C., Simon,S., Lenzen,G., Petel,F., Wojcik,J., Schachter,V. *et al.* (2001) The protein–protein interaction map of *Helicobacter pylori*. *Nature*, **409**, 211–215.
9. Ward,D.V., Draper,O., Zupan,J.R. and Zambryski,P.C. (2002) Peptide linkage mapping of the *Agrobacterium tumefaciens* vir-encoded type IV secretion system reveals protein subassemblies. *Proc. Natl Acad. Sci. USA*, **99**, 11493–11500.
10. Hu,J.C., Kornacker,M.G. and Hochschild,A. (2000) *Escherichia coli* one- and two-hybrid systems for the analysis and identification of protein–protein interactions. *Methods*, **20**, 80–94.
11. Ladant,D. and Karimova,G. (2000) Genetic systems for analyzing protein–protein interactions in bacteria. *Res. Microbiol.*, **151**, 711–720.
12. Dove,S.L., Joung,J.K. and Hochschild,A. (1997) Activation of prokaryotic transcription through arbitrary protein–protein contacts. *Nature*, **386**, 627–630.
13. Dove,S.L. and Hochschild,A. (1998) Conversion of the omega subunit of *Escherichia coli* RNA polymerase into a transcriptional activator or an activation target. *Genes Dev.*, **12**, 745–754.
14. Dove,S.L. and Hochschild,A. (2001) Bacterial two-hybrid analysis of interactions between region 4 of the sigma(70) subunit of RNA polymerase and the transcriptional regulators Rsd from *Escherichia coli* and AlgQ from *Pseudomonas aeruginosa*. *J. Bacteriol.*, **183**, 6413–6421.
15. Shaywitz,A.J., Dove,S.L., Kornhauser,J.M., Hochschild,A. and Greenberg,M.E. (2000) Magnitude of the CREB-dependent transcriptional response is determined by the strength of the interaction between the kinase-inducible domain of CREB and the KIX domain of CREB-binding protein. *Mol. Cell. Biol.,* **20**, 9409–9422.
16. Besemer,J., Lomsadze,A. and Borodovsky,M. (2001) GeneMarkS: a self-training method for prediction of gene starts in microbial genomes. Implications for finding sequence motifs in regulatory regions. *Nucleic Acids Res.*, **29**, 2607–2618.
17. Ogata,H., Audic,S., Renesto-Audiffren,P., Fournier,P.E., Barbe,V., Samson,D., Roux,V., Cossart,P., Weissenbach,J., Claverie,J.M. and Raoult,D. (2001) Mechanisms of evolution in *Rickettsia conorii* and *R. prowazekii*. *Science*, **293**, 2093–2098.
18. Lander,E.S. and Waterman,M.S. (1988) Genomic mapping by fingerprinting random clones: a mathematical analysis. *Genomics*, **2**, 231–239.
19. Magalon,A., Frixon,C., Pommier,J., Giordano,G. and Blasco,F. (2002) *In vivo* interactions between gene products involved in the final stages of molybdenum cofactor biosynthesis in *Escherichia coli*. *J. Biol. Chem.*, **277**, 48199–48204.
20. Ohashi,N., Zhi,N., Lin,Q. and Rikihisa,Y. (2002) Characterization and transcriptional analysis of gene clusters for a type IV secretion machinery in human granulocytic and monocytic ehrlichiosis agents. *Infect. Immun.*, **70**, 2128–2138.
21. Das,A. and Xie,Y.H. (2000) The Agrobacterium T-DNA transport pore proteins VirB8, VirB9 and VirB10 interact with one another. *J. Bacteriol.*, **182**, 758–763.
22. Christie,P.J. (2001) Type IV secretion: intercellular transfer of macromolecules by systems ancestrally related to conjugation machines. *Mol. Microbiol.*, **40**, 294–305.
23. Matthews,L.R., Vaglio,P., Reboul,J., Ge,H., Davis,B.P., Garrels,J., Vincent,S. and Vidal,M. (2001) Identification of potential interaction networks using sequence-based searches for conserved protein–protein interactions or 'interologs'. *Genome Res.*, **11**, 2120–2126.
24. Mahan,M.J., Tobias,J.W., Slauch,J.M., Hanna,P.C., Collier,R.J. and Mekalanos,J.J. (1995) Antibiotic-based selection for bacterial genes that are specifically induced during infection of a host. *Proc. Natl Acad. Sci. USA*, **92**, 669–673.
25. Ideker,T., Ozier,O., Schwikowski,B. and Siegel,A.F. (2002) Discovering regulatory and signaling circuits in molecular interaction networks. *Bioinformatics*, **18**, S233–S240.