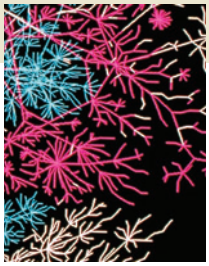

Functional genome annotation through phylogenomic mapping

Balaji S Srinivasan[1,2], Nora B Caberoy[3], Garret Suen[3], Rion G Taylor[3], Radhika Shah[3], Farah Tengra[3], Barry S Goldman[4], Anthony G Garza[3] & Roy D Welch[3]

Accurate determination of functional interactions among proteins at the genome level remains a challenge for genomic research. Here we introduce a genome-scale approach to functional protein annotation—phylogenomic mapping—that requires only sequence data, can be applied equally well to both finished and unfinished genomes, and can be extended beyond single genomes to annotate multiple genomes simultaneously. We have developed and applied it to more than 200 sequenced bacterial genomes. Proteins with similar evolutionary histories were grouped together, placed on a three dimensional map and visualized as a topographical landscape. The resulting phylogenomic maps display thousands of proteins clustered in mountains on the basis of coinheritance, a strong indicator of shared function. In addition to systematic computational validation, we have experimentally confirmed the ability of phylogenomic maps to predict both mutant phenotype and gene function in the delta proteobacterium *Myxococcus xanthus*.

The computational detection of interacting sets of proteins[1] is a topic of considerable interest, and several algorithms for this purpose have been developed in the last five years. These methods include the use of gene fusions as 'Rosetta Stones'[2], the tracking of correlated mutations to infer interactions[3,4], the enumeration of conserved operons[5] and gene neighbors[6], and the calculation of phylogenetic profiles[7]. During the same period a host of approaches for the visualization of multidimensional biological data sets were described, including hierarchical clustering[8], planar graph drawing[9], singular value decomposition[10] and variants of multidimensional scaling[11–13]. The introduction of these algorithms was accompanied by a rapid rise in the number of sequenced genomes, a data set that presents attractive opportunities for computational module detection. To systematically mine this data set we have developed phylogenomic mapping, a way of visually clustering data on the basis of coinheritance, which uses sequenced genomes to provide information about the probable intragenomic interaction partners of a gene.

[1] Department of Developmental Biology, Stanford University School of Medicine, Stanford, California, USA. [2]Department of Electrical Engineering, Stanford University, Stanford, California. [3]Department of Biology, Syracuse University, 130 College Place, BRL Room 702A, Syracuse, New York 13244-1170, USA. [4]Monsanto Corporation, St. Louis, Missouri, USA. Correspondence and requests for materials should be addressed to R.D.W. (rowelch@syr.edu). All requests for *M. xanthus* mutants should be made to A.G.G.

Published online 6 June 2005; doi:10.1038/nbt1098

We began with the premise that pairs of genes whose products are dissimilar, but which are consistently coinherited in the same sets of organisms are likely to be functionally linked[14]. Though the determination of coinheritance becomes more accurate as the number of sequenced genomes increases, the visualization of coinheritance patterns with clustered heatmaps[8] becomes impractical as the number of genomes approaches the number of proteins. Furthermore, although invaluable for the analysis of small sets of proteins[15], heatmaps display too much information in each row to allow global visualization of similarity relationships. Phylogenomic mapping addresses both of these issues by providing a global visualization of coinheritance that can readily scale to incorporate the thousands of genome sequences that will soon be available.

Phylogenomic map construction

A phylogenomic map is a topographical visualization[11] of the similarity structure of a phylogenomic matrix[2] (**Fig. 1**). We generated phylogenomic maps for the 204 bacterial and archaeal genomes sequenced as of December 12, 2004, as well as the newly sequenced delta proteobacterium *Myxococcus xanthus* (preliminary sequence data were obtained from The Institute for Genomic Research website at http://www.tigr.org/). To create each map, the basic local alignment search tool (BLAST)[16] was used to align the open reading frames (ORFs) of a predicted proteome against a local database of proteins from every sequenced genome. As a quality filtering step, we retained only those proteins which registered BLAST bit scores >50 to genes in five or more of the sequenced genomes. Bit scores for these proteins were arranged in an N by k dimensional phylogenomic matrix, where N represents the number of proteins and k represents the number of sequenced genomes. An upper triangular similarity matrix of rank correlations was then computed and the threshold set to retain only the top 50 similarity scores for each row. Finally, a combination of multidimensional scaling and force-directed placement was applied to assign (x,y) coordinates in the plane to each protein[11]. The result is a phylogenomic map where the products of genes with similar evolutionary histories cluster together in mountains, and where local height is proportional to the density of proteins within an area. Further details on map construction and mountain discretization are available in **Supplementary Note 1** online.

A phylogenomic map with labeled mountains for the genome of *M. xanthus* is pictured in **Figure 2a**. Mountain functions (**Table 1**) were assigned by the procedures detailed in 'Computational

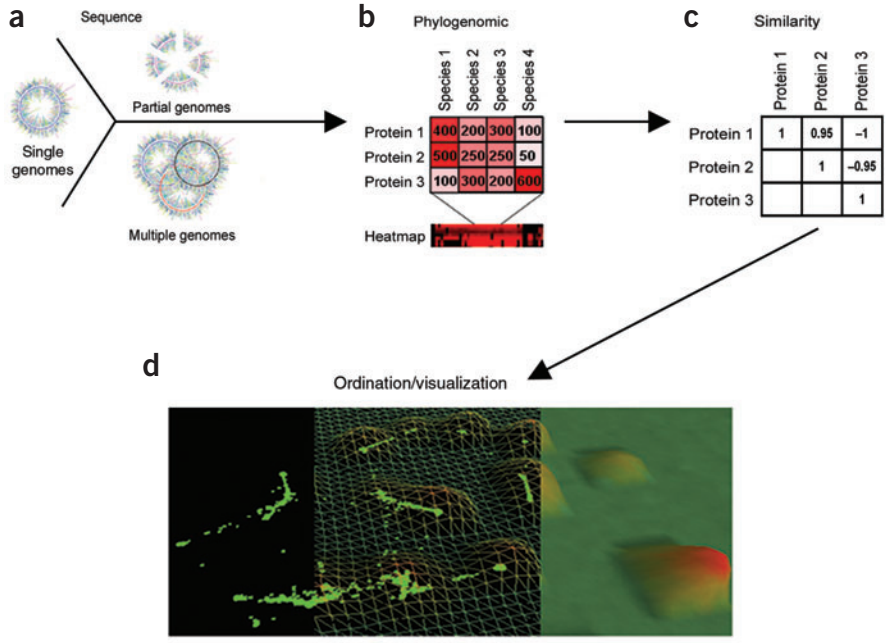

Figure 1 Phylogenomic mapping. (**a**) Sequence data from partial, single or multiple genomes are used to produce protein predictions, which are aligned[16] against a database of proteomes from hundreds of completely sequenced genomes. (**b**) The *N* by *k* phylogenomic matrix. *N* is the number of proteins/rows, *k* is the number of sequenced genomes/columns, and each entry contains bit scores computed from sequence alignment (see text). Below is a heatmap visualization, where array elements with higher bit scores are colored a brighter red. For illustrative purposes, we have zoomed in on a small subsection of the heatmap with $N = 3$ and $k = 4$. (**c**) From the phylogenomic matrix we generate an *N* by *N* similarity matrix using Spearman's rank correlation. This metric is robust with respect to outliers and particularly good for phylogenomic mapping. (**d**) Next, we apply multidimensional scaling and force directed placement to transform the similarity matrix into a two dimensional ordination[11], where each point in the plane represents a protein in the original matrix. We then visualize the data set in three dimensions, where mountain height corresponds to the local density of ordinated proteins and nearby proteins have similar coinheritance patterns.

validation' below. To ensure that our phylogenomic map was an accurate representation of the phylogenetic similarities of the proteins in our data set, we also generated hierarchically clustered heatmaps for each phylogenomic matrix[17]. We observed that tightly correlated profiles in each heatmap were located nearby in each phylogenomic map (**Fig. 2b**,**c**). Our approach can therefore assign proteins from a >200 dimensional profile space to a 2-dimensional map space while still preserving neighborhood fidelity. This built-in scalability will be even more useful when thousands of sequenced genomes are available.

Although the relationship between the mountains of the phylogenomic map and the clusters produced by the associated heatmap is useful in evaluating the internal consistency of our method, the crucial test of the map's utility lies in its ability to predict phenotype. In the following sections, we present both experimental and computational evidence for the accuracy and utility of this mapping method.

Experimental validation in *Myxococcus xanthus*

We chose the complex prokaryote *M. xanthus* as a test case for the experimental validation of the phylogenomic map. *M. xanthus* can exist as both a single-species biofilm and a free-living cell. Though each bacterium is autonomous in metabolism and reproduction, the biofilm is a self-organizing predatory swarm that has many of the characteristics of a multicellular organism. *M. xanthus* is thus a unique model organism in that it combines the genetic tractability of a prokaryote with the behavioral sophistication of a simple eukaryote. In addition to this behavioral complexity, *M. xanthus* is distinguished by its large assortment of predation-enabling antimicrobial agents, a collection that has attracted medical and biotechnological attention[18–20].

The complexity of the *M. xanthus* life cycle is reflected in its 9.45 MB genome, which is the largest prokaryotic genome sequenced to date. Notably, almost 50% of the predicted genes have not been assigned a function. Because of this and its large size, *M. xanthus* is a particularly suitable candidate for phylogenomic map validation. Its assortment of complex and well-characterized genetic subsystems is also important; distinguishing between the genetic subsystems affected by different mutants is made easier when multiple phenotypes can be simultaneously assayed in a given strain.

Mutants in the *M. xanthus* motility apparatus produce pleiotropic phenotypes. *M. xanthus* cells move by 'gliding,' a descriptive term for a complex mechanism that can be divided into two genetically distinct motility systems, referred to as adventurous (A) and social (S) motility[21]. For *M. xanthus* to expand as a swarm or to develop fruiting bodies in response to starvation, cell motility must be coupled to a communication mechanism that coordinates the movement of multiple cells[22]. Genetic mutants that affect swarming motility can thus be subdivided into two categories: those mutants that affect cell movement by one or both motility mechanisms (A or S), and those that interfere with the coordinated motility of the swarm. Both conditions can manifest themselves as a reduced rate of swarm expansion and/or a defect in fruiting body formation in response to starvation (see **Supplementary Note 2** online). The independence of these assays along with their methodological tractability and quantitative output makes them particularly suitable for map validation.

To experimentally validate the map, we posited that disruption of genes whose protein products are within the same predicted functional module would produce similar phenotypes, as each disruption might cause the same network to fail. We should therefore be able to predict the phenotype of mutants of uncharacterized genes based on their phylogenomic proximity to genes of known function. To this end, we used known motility proteins as phylogenomic map 'landmarks' around which more ORFs were selected for disruption. A total of 15 uncharacterized ORFs (**Table 2**) in six mountains (**Fig. 3a**) were chosen in this manner. Each ORF was disrupted using a plasmid-insertion protocol (see **Supplementary Note 2** online).

Whereas some of the disrupted ORFs had putative annotations related to motility, the majority had no apparent connection. Two groups of ORFs are particularly noteworthy: the penicillin-binding and structural proteins in mountains 7 and 33 and the Tol-linked proteins of mountain 25 (**Table 2**). The structural proteins were chosen because of their proximity to proteins known to be involved in pili biogenesis and motility[23,24]. All but one of these mutants demonstrated

Table 1 Computational validaton of the *M. xanthus* phylogenomic map

Mountain	Number of proteins	GO ID	Term	*P* value
1	224	–	–	–
2	212	–	–	–
3	210	GO:0006935	Chemotaxis	1.57E–51
3	210	GO:0007154	Cell communication	6.28E–44
3	210	GO:0045449	Regulation of transcription	1.15E–17
4	207	GO:0003700	Transcription factor activity	0.000287
4	207	GO:0005215	Transporter activity	0.000646
4	207	GO:0051234	Establishment of localization	0.002923
5	207	–	–	–
6	197	GO:0006396	RNA processing	1.27E–10
6	197	GO:0005840	Ribosome	4.44E–08
6	197	GO:0006412	Protein biosynthesis	4.17E–05
7	190	GO:0005215	Transporter activity	1.07E–06
7	190	GO:0016020	Membrane	6.97E–05
8	179	GO:0016053	Organic acid biosynthesis	4.87E–05
9	175	GO:0006468	Protein amino acid phosphorylation	3.98E–142
9	175	GO:0005524	ATP binding	7.97E–49
10	168	GO:0005622	Intracellular	0.000263
10	168	GO:0005737	Cytoplasm	0.003282
10	168	GO:0009059	Macromolecule biosynthesis	0.004557
11	155	–	–	–
12	148	GO:0003995	Acyl-CoA dehydrogenase activity	1.65E–18
12	148	GO:0006118	Electron transport	2.19E–08
13	139	GO:0048037	Cofactor binding	1.23E–19
13	139	GO:0005554	Molecular function unknown	4.96E–13
14	133	GO:0042597	Periplasmic space	0.003207
15	131	GO:0008652	Amino acid biosynthesis	7.78E–07
16	128	–	–	–
17	123	GO:0016787	Hydrolase activity	0.0001
18	118	GO:0005489	Electron transporter activity	1.10E–12
18	118	GO:0003954	NADH dehydrogenase activity	2.75E–12
18	118	GO:0006119	Oxidative phosphorylation	0.000312
19	118	GO:0000160	Two-component signal transduction system (phosphorelay)	1.27E–47
19	118	GO:0009306	Protein secretion	0.003662
19	118	GO:0016043	Cell organization and biogenesis	0.006264
20	117	GO:0008236	Serine-type peptidase activity	1.92E–06
20	117	GO:0016787	Hydrolase activity	2.25E–06
20	117	GO:0005554	Molecular function unknown	0.006173
21	101	GO:0006412	Protein biosynthesis	0.0008
22	95	GO:0000155	Two-component sensor molecule activity	4.65E–52
22	95	GO:0007154	Cell communication	3.26E–31
23	95	–	–	–
24	82	GO:0008653	Lipopolysaccharide metabolism	1.10E–09
24	82	GO:0005515	Protein binding	2.69E–05
25	81	GO:0004222	Metalloendopeptidase activity	1.02E–08
25	81	GO:0009057	Macromolecule catabolism	0.006011
26	71	GO:0004295	Trypsin activity	0.001793
26	71	GO:0016787	Hydrolase activity	0.003435
26	71	GO:0043285	Biopolymer catabolism	0.007552
27	67	–	–	–
28	64	–	–	–
29	63	GO:0004872	Receptor activity	0.000287
29	63	GO:0003964	RNA-directed DNA polymerase activity	0.000298
30	58	–	–	–

Table 1 (continued)

Mountain	Number of proteins	GO ID	Term	*P* value
31	50	–	–	–
32	47	GO:0048037	Cofactor binding	3.11E–31
32	47	GO:0006633	Fatty acid biosynthesis	2.47E–28
32	47	GO:0005554	Molecular function unknown	3.87E–06
33	32	GO:0007155	Cell adhesion	0.002824
34	28	GO:0000156	Two-component response regulator activity	1.04E–07
34	28	GO:0007154	Cell communication	8.38E–06
34	28	GO:0006355	Regulation of transcription, DNA-dependent	4.08E–05
35	27	–	–	–
36	27	–	–	–
37	24	GO:0008914	Leucyltransferase activity	0.000383
37	24	GO:0005509	Calcium ion binding	0.000464
37	24	GO:0007155	Cell adhesion	0.003682
38	18	-	-	-
39	13	GO:0009008	DNA-methyltransferase activity	4.53E–07
39	13	GO:0006310	DNA recombination	0.003222
40	12	GO:0004812	tRNA ligase activity	3.50E–05
40	12	GO:0009451	RNA modification	0.001057
41	10	–	–	–
42	7	–	–	–
43	5	–	–	–

A Gene Ontology[36] file was generated for the *M. xanthus* genome. We used the GO::TermFinder software[33] in conjunction with the phylogenomic map to test whether mountains were statistically enriched for functional groups. Statistically significant associations (see **Supplementary Note 3** online) were found for 27 of the 43 mountains. Mountains which were not assigned annotations are predominantly composed of either phyla-specific proteins or proteins of unknown function. In the former case, the genes within these mountains may be too specialized to have yet received an annotation from the GO consortium. In the latter case, mountains composed largely of conserved hypotheticals (such as 13, 20, and 32) may correspond to novel biological functions.

motility defects, revealing a previously unknown connection between *M. xanthus* structural proteins and its motility system. We hypothesize that these structural proteins play an unappreciated role in *M. xanthus* motility, perhaps by allowing the cell to withstand the mechanical stress of motion[25].

The other group of particular interest are the proteins of mountain 25, selected because of their vicinity to proteins known to be involved in A-motility. It has been suggested that *M. xanthus*' A-motility system involves a 'slime'-powered analog of rocket propulsion[26], wherein the extracellular hydrolysis and expansion of hydrophilic polysaccharides pushes the cell forward on an agar surface. The exact biochemical nature of this slime was previously unknown, as only the hypothetical slime 'nozzles' from the Tol family of proteins have been identified[26]. Our analysis of genes coinherited with these known A-motility genes turned up a group of five Lpx and three Kds proteins in mountain 25, which produce lipopolysaccharides known to be translocated by Tol proteins in *Escherichia coli*[27,28]. Inactivation of two of the genes (*mxan4718* and *mxan1095*) that code for these proteins produced motility defects (**Fig. 3b**). Given that past work has suggested that mutants defective in lipopolysaccharide O-antigen biosynthesis have specific alterations in A-motility [29], we believe that both Lpx and Kds proteins have a different purpose in *M. xanthus* than in *E. coli*. Specifically, we believe they are responsible for generating A-motility polysaccharide slime.

In summary, 12 of 15 ORFs targeted for disruption on the basis of phylogenomic proximity to known motility proteins produced obvious defects in swarm expansion and/or aggregation when inactivated. This success rate of 80% significantly exceeds other methods of finding new genes involved in *M. xanthus* motility. The outcome of the plasmid insertion protocol (see **Supplementary Note 2** online) is similar to the effect of random transposon mutageneis—disruption of an ORF through DNA recombination. Random mutagenesis methods have a much lower success rate; only 1% of 12,000 *magellan-4* insertions inactivated A-motility in an S^- background[30]. Targeted selection of genes based on paralogy to known motility-linked proteins[31] yielded a 10% rate of mutants with motility defects. Finally, given the conservative estimate that 5% of *M. xanthus* proteins are essential for motility[30,32] and that 2% of all mutants are defective for development[32], it is highly unlikely that 12 of 15 disrupted ORFs would produce motility and/or aggregation defects if chosen at random instead of using phylogenomic mapping.

With respect to the prospects of applying this 'landmark-and-disrupt' strategy more generally, we believe that this high level of yield enrichment may be typical for a wide class of bacterial systems. Specifically, if the modular hypothesis advanced by Hartwell *et al.*[1] is correct, many nonessential phenotypes of interest are controlled by a fairly small number of tightly coupled interacting proteins. Although the small size of such groups makes them difficult to initially locate by random mutagenesis, once one member is discovered, their likely interaction partners can be found using phylogenomic mapping. We conclude that phylogenomic mapping is a generally applicable way to speed the experimental investigation of a pathway of interest.

Computational validation

In addition to our experimental validation in *M. xanthus*, we used the descriptive framework produced by the Gene Ontology Consortium (GO)[33] to perform a global computational validation over all 205 phylogenomic maps. The GO framework integrates the knowledge of expert molecular biologists and biochemists in a directed acyclic graph structure that permits automatic inference about the functional

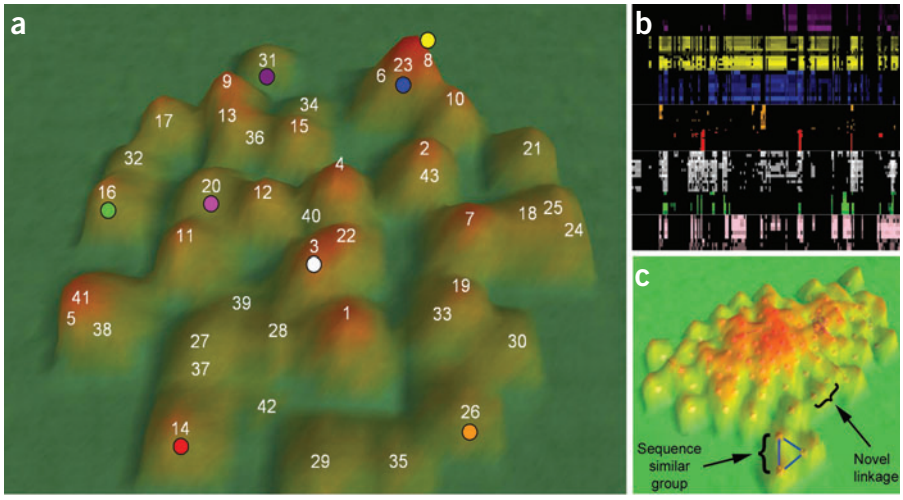

Figure 2 The phylogenomic map of *M. xanthus*. (**a**) The map of the *M. xanthus* genome with labeled mountains (see **Table 1**), at lowest resolution. Each colored dot corresponds to a group of proteins with similar coinheritance patterns. (**b**) Heatmap representations of the colored dots on the map (see also **Fig. 1b**). Importantly, similar heatmap profiles tend to colocate on the phylogenomic map, effecting a dimension reduction from 200+ dimensional profile space to 2-dimensional map space. (**c**) A high resolution view of mountain 23 where the individual proteins which comprise the mountain are clearly visible. Paralogous proteins are connected by blue lines (see **Supplementary Note 1** online). Proteins which are nearby yet unconnected indicate novel linkages that could not have been discerned through sequence alignment alone. We have highlighted a group of three paralogous proteins as well as a pair whose coinheritance could not have been discerned from comparisons of their primary sequence.

significance of gene groups. Whereas GO was originally developed to facilitate systematic analysis of microarray data, it can be applied to confirm the statistical significance of any predicted interaction group, including the mountains produced by phylogenomic mapping.

To apply the GO framework in finding functionally enriched mountains in a phylogenomic map, it is necessary to assign GO terms to the proteins ordinated on the map. When available, we used the manual annotations generated by The Institute for Genomic Research. For all other bacteria, GO files were generated by parsing and integrating data obtained from GenBank, SeqHound[34] and the Gene Ontology Annotation[35] (see **Supplementary Note 3** online for full details).

With these GO annotation files, we were able to apply the cluster scoring algorithm implemented in GO::TermFinder[36] to assess the statistical significance of the mountains generated by all 205 phylogenomic maps. An excerpt of the results for *M. xanthus* and several bacteria of interest can be seen in **Table 3**, and a full table for all 205 bacteria can be found in **Supplementary Table 1** online. We note that all maps had mountains enriched for basic cellular machinery: transcription, translation and DNA replication. In small genomes such as that of *Mycoplasma genitalium*, almost every mountain was implicated in one of these basic physiological processes. For larger genomes, however, mountains enriched for auxiliary functions such as cell-to-cell signaling and antibiotic production became visible. Furthermore, because of the nature of the GO consortium[33], the most globally relevant functions are generally the first ones to be annotated. For this reason, proteins with taxonomically narrow distributions tend to lag in receiving GO annotations, as fewer investigators are working on them. Thus, large genomes tend to have many mountains, which are substantially enriched only for their concentration of phyla-specific hypotheticals. *M. xanthus* is an excellent example; several of the mountains, such as 13, 20 and 32, are enriched for conserved proteins of unknown function (**Table 1**). This division

Table 2 Mutant phenotypes of proteins selected for disruption using the *M. xanthus* phylogenomic map

Mountain	MXAN	Putative annotation	Swarm expansion	Aggregation
7	MXAN6450	Putative 6-aminohexanoate-dimer hydrolase	–	–
7	MXAN2136	Penicillin-binding protein, putative	+	+
7	MXAN7171	Beta lactamase family protein	–	+
17	MXAN3003	MotA/TolQ/ExbB proton channel family protein, putative	+	–
25	MXAN1448	Biopolymer transport protein TolQ	+	–
25	MXAN0346	TolB protein, putative	+	+
25	MXAN1995	Lipid-A-disaccharide synthase	–	–
25	MXAN4718	3-deoxy-8-phosphooctulonate synthase	–	+
25	MXAN1095	TonB system transport protein ExbB/TolQ	–	+
25	MXAN0275	Twitching mobility protein	–	+
33	MXAN5605	Cell division protein FtsW	–	–
33	MXAN5610	Cell division protein FtsI/penicillin-binding protein 2	–	–
33	MXAN2648	Rod shape-determining protein RodA	–	–
36	MXAN2465	TPR domain protein, putative	+	+
36	MXAN1324	TPR domain protein	–	–

We chose ORFs which were likely to give phenotypes of interest, particularly with respect to motility and cellular development. Each mutant was assayed for normal (+) or defective (–) swarm expansion (cell motility) or aggregation (coordinated swarm motility) as compared to *M. xanthus* DK1622 wild type. Swarm expansion was assayed on 0.4% and 1.5% agar. Details of each assay are described in **Supplementary Note 2** online. These results show that 12 out of 15 targeted insertions produce defects in swarm expansion and/or aggregation.

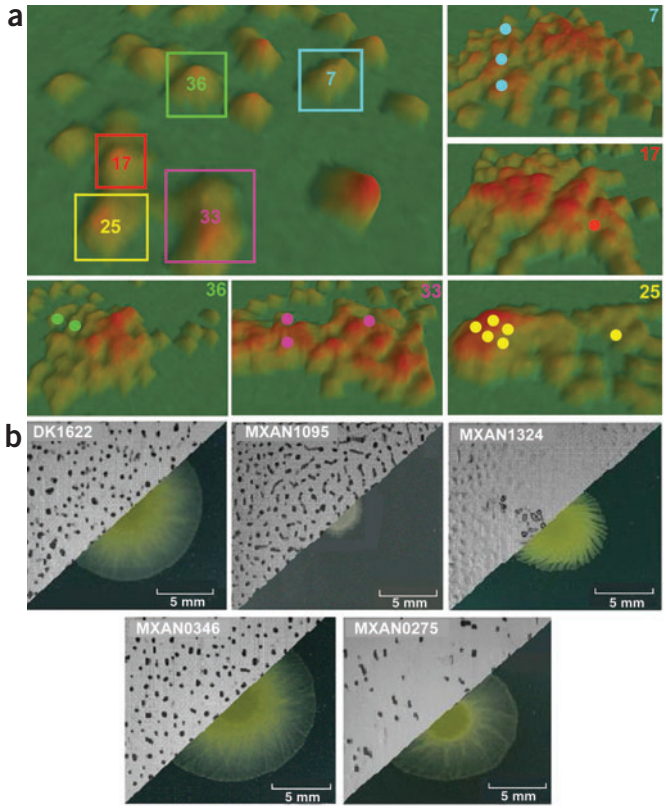

Figure 3 Experimental validation of phylogenomic map predictions in *M. xanthus*. (**a**) Colored boxes denote mountains selected for experimental validation (see text). The specific proteins which were selected for disruption are colored in each medium resolution view. (**b**) Assay images of representative mutant strains displaying aggregation (top) and growth (bottom) phenotypes. From left to right: DK1622 wild type (+/+), MXAN1095 (+/–), MXAN1324 (–/–), MXAN0346 (+/+), and MXAN0275 (–/+).

of hypotheticals into mountains is useful because it provides a classification of unknown proteins that extends beyond the relatively uninformative assignment of 'hypothetical.' In particular, we believe that some of these hypothetical-rich phylogenomic mountains may correspond to novel biological processes.

Functional genome annnotation

We have presented and experimentally tested a method for performing functional genomic annotation directly from sequence data—a method which naturally scales with the rapidly increasing number of sequenced genomes. Phylogenomic mapping relies on the premise that coinheritance implies cofunctionality. It allows us to identify interacting sets of ORFs, predict experimental associations, confirm standing hypotheses and suggest new functions for misannotated and hypothetical sequences.

In addition to its use in single genome annotation, phylogenomic mapping can be generalized to annotate multiple genomes simultaneously. As it is becoming more common to sequence several genomes at the same time, we note that simultaneous ordination of proteins from a coherent set of genomes (e.g., gamma-proteobacteria, members of multi-species biofilms or environmental sequence data[37]) can produce a global overview of the available protein modules, even when only partial genomes are available. Given the commensal nature of many microbial communities[38], we believe that one result of this analysis will be a demonstration of functionally complementary protein modules in multi-species symbiotic communities.

With respect to the question of whether the mountains of the phylogenomic map contain protein modules, we note that not all elements within each mountain are necessarily involved in the same function. Furthermore, the granularity of the map (that is, the number of mountains) can be altered to some extent by changing the parameters of the ordination algorithm. Finally, the quality of module identification is largely dependent upon the quality and breadth of GO annotation coverage, which can vary substantially between species. These caveats notwithstanding, it is interesting to note that the phenotypes of *M. xanthus* motility and development mutants from mountain 33 are all the same, and are thus more similar to each other than they are to mutants from other mountains. This result was a side effect of our initial investigation, which primarily sought to find new motility-linked ORFs using characterized motility genes; we believe this result is not a coincidence. Specifically, we expect that more detailed assays will add additional phenotype data to established 'phenotype vectors' associated with mutants from each mountain, and thereby ascertain whether some mountains represent

Table 3 Global computational validation of phylogenomic mapping

Species	Number of ordinated proteins	Number of ordinated proteins with GO annotation	Fraction of ordinated proteins with GO annotation	Number of mountains	Number of mountains w/ statistically significant GO annotation	Fraction of mountains w/ statistically significant GO annotations	GO annotation source
Bacillus subtilis	3,771	3,066	0.81	32	29	0.91	SeqHound + GBK
Bradyrhizobium japonicum	7,258	4,862	0.67	73	67	0.92	SeqHound + GBK
Caulobacter crescentus	3384	2,807	0.83	32	23	0.72	SeqHound + GBK
Deinococcus radiodurans	2,770	2,250	0.81	29	22	0.76	SeqHound + GBK
Escherichia coli K12	4,038	3,653	0.9	35	33	0.94	SeqHound + GBK
Mycoplasma genitalium	475	459	0.97	7	5	0.71	SeqHound + GBK
Myxococcus xanthus DK1622	4,356	3,213	0.73	43	26	0.6	SeqHound + GBK
Pseudomonas aeruginosa	5,223	4,593	0.88	48	39	0.81	SeqHound + GBK

To demonstrate that phylogenomic mapping is a universally applicable tool, we generated maps and Gene Ontology[36] files for more than 180 bacteria and scored each of them with GO::TermFinder[33] (see text). This table is a selected subset of the results for *M. xanthus* and several other bacteria of interest. The entire list of results for all bacteria scored in this manner can be found as **Supplementary Table 1** online. Mountains were considered enriched for a given GO term if the term was enriched in *P* value terms ($P < 0.01$) and at least 20% of the proteins within the mountain were annotated with the term. In all cases, a large fraction of the mountains on each map are highly enriched for GO annotations.

genetically modular subsystems—as suggested by our GO-based computational validation.

The data sets used to produce the phylogenomic maps are quite distinct from the expression data used to produce other topographical maps[12,13]. Coexpression is like coinheritance insofar as it is a sufficient but not necessary condition to infer functional coupling. There are, however, several differences between phylogenomic and expression data sets that make them complementary but distinct. First, unlike microarray analyses, missing data entries are not a critical issue with phylogenomic analyses. The precision of the input data set is limited only by the accuracy of ORF predictions and the accuracy of base pair sequences.

Second, each new genome that is annotated represents a new column for the phylogenomic matrix, and potentially increases discriminatory power. The utility of a new genome sequence is increased if it improves the 'taxonomic dynamic range' of the species involved in the columns of the matrix, an issue that also arises in the generation of phylogenetic trees[39]. Whereas an additional alpha proteobacterial genome may not make a significant difference in the phylogenomic analysis of a delta proteobacteria, the genome sequence of a closely related delta proteobacterium will make it much easier to separate those sequences that are characteristic of the delta proteobacteria from those that are common to all bacteria. Currently, the accumulation of genome sequences is proceeding more rapidly than the accumulation of expression data for all but the most heavily studied model organisms.

Third, ongoing research in genome evolution and horizontal gene transfer offers the prospect of predicting the ancestry for long segments of known genomes[38,40,41]. Such knowledge would give us a strong hint about the phyletic distributions of the genes in that segment, and hence their phylogenetic profiles. This possibility distinguishes phylogenetic profiles from expression profiles, as the state of cellular simulation is not yet advanced enough to forecast computational prediction of expression data.

Fourth, phylogenomic mapping can be naturally generalized to include any genome-scale information expressible in matrix form. Future work should incorporate other types of nonhomology-based interaction predictors, such as gene neighbor[42], Rosetta Stone[2] and data from microarrays and protein-protein interaction experiments[12]. In particular, the union of phylogenomic and expression data is a natural generalization of recent work on conserved coexpression signatures[43].

To summarize, phylogenomic mapping complements expression analysis as a powerful tool for functional genomics and genome annotation. As demonstrated by our experimental confirmation of predicted mutant phenotypes, map-based mesoscale annotation can greatly reduce the search space for sets of interaction partners of a gene or module of interest. Because phylogenomic mapping requires only sequence data, and can be applied to any genome or partial genome, we believe it can add value to any genome annotation process.

Phylogenomic maps and other information are available on our website (http://myxococcus.syr.edu/phylo).

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

ACKNOWLEDGMENTS

We thank Harley McAdams, Lucy Shapiro, William Nierman and Dale Kaiser for helpful discussions. We thank the Monsanto Corporation and the Institute for Genomics Research for providing access to the genome sequence of *M. xanthus* DK1622. This work was supported in part by National Science Foundation (NSF) Grant MCB-0444154 to A.G.G. B.S.S. was supported by a Department of Defense National Defense Science and Engineering Graduate Fellowship through the Army Research Office. Sequencing of *M. xanthus* DK1622 was accomplished with support from the NSF.

Authors' contributions. B.S.S. developed phylogenomic mapping and performed related computational analyses. *M. xanthus* genome sequence data were provided by B.S.G., the TIGR databases and the GO annotation. The annotation of the *M. xanthus* genome was performed by B.S.G., R.D.W.; further data were obtained from TIGR. N.B.C., R.G.T., R.S., G.S., F.T. and A.G.G. produced experimental data on motility mutants. B.S.S., G.S. and R.D.W. wrote the paper. R.D.W. provided a nurturing environment.

COMPETING INTERESTS STATEMENT

The authors declare that they have no competing financial interests.

Published online at http://www.nature.com/naturebiotechnology/

1. Hartwell, L.H., Hopfield, J.J., Leibler, S. & Murray, A.W. From molecular to modular cell biology. *Nature* **402**, C47–C52 (1999).
2. Marcotte, E.M. *et al.* Detecting protein function and protein-protein interactions from genome sequences. *Science* **285**, 751–753 (1999).
3. Gertz, J. *et al.* Inferring protein interactions from phylogenetic distance matrices. *Bioinformatics* **19**, 2039–2045 (2003).
4. Pazos, F. & Valencia, A. In silico two-hybrid system for the selection of physically interacting protein pairs. *Proteins* **47**, 219–227 (2002).
5. 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).
6. Huynen, M.A., Snel, B., von Mering, C. & Bork, P. Function prediction and protein networks. *Curr. Opin. Cell Biol.* **15**, 191–198 (2003).
7. 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).
8. Eisen, M.B., Spellman, P.T., Brown, P.O. & Botstein, D. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. USA* **95**, 14863–14868 (1998).
9. Barabasi, A.L. & Oltvai, Z.N. Network biology: understanding the cell's functional organization. *Nat. Rev. Genet.* **5**, 101–113 (2004).
10. Alter, O., Brown, P.O. & Botstein, D. Singular value decomposition for genome-wide expression data processing and modeling. *Proc. Natl. Acad. Sci. USA* **97**, 10101–10106 (2000).
11. Davidson, G.S., Wylie, B.N. & Boyack, K. Cluster stability and the use of noise in interpretation of clustering. in Proceedings of the IEEE Symposium on Information Visualization 2001 (INFOVIS'01), 23–30 (IEEE Computer Society, 2001).
12. Werner-Washburne, M. *et al.* Comparative analysis of multiple genome-scale data sets. *Genome Res.* **12**, 1564–1573 (2002).
13. Kim, S.K. *et al.* A gene expression map for *Caenorhabditis elegans*. *Science* **293**, 2087–2092 (2001).
14. Marcotte, E.M., Xenarios, I., van Der Bliek, A.M. & Eisenberg, D. Localizing proteins in the cell from their phylogenetic profiles. *Proc. Natl. Acad. Sci. USA* **97**, 12115–12120 (2000).
15. Enault, F., Suhre, K., Poirot, O., Abergel, C. & Claverie, J.M. Phydbac2: improved inference of gene function using interactive phylogenomic profiling and chromosomal location analysis. *Nucleic Acids Res.* **32**, W336–W339 (2004).
16. 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).
17. de Hoon, M.J., Imoto, S., Nolan, J. & Miyano, S. Open source clustering software. *Bioinformatics* **20**, 1453–1454 (2004).
18. Julien, B. & Shah, S. Heterologous expression of epothilone biosynthetic genes in Myxococcus xanthus. *Antimicrob. Agents Chemother.* **46**, 2772–2778 (2002).
19. Gerth, K. *et al.* The myxalamids, new antibiotics from Myxococcus xanthus (Myxobacterales). I. Production, physico-chemical and biological properties, and mechanism of action. *J. Antibiot. (Tokyo)* **36**, 1150–1156 (1983).
20. Pospiech, A., Cluzel, B., Bietenhader, J. & Schupp, T. A new Myxococcus xanthus gene cluster for the biosynthesis of the antibiotic saframycin Mx1 encoding a peptide synthetase. *Microbiology* **141**, 1793–1803 (1995).
21. Shi, W. & Zusman, D.R. The two motility systems of Myxococcus xanthus show different selective advantages on various surfaces. *Proc. Natl. Acad. Sci. USA* **90**, 3378–3382 (1993).
22. Kaiser, D. & Welch, R. Dynamics of fruiting body morphogenesis. *J. Bacteriol.* **186**, 919–927 (2004).
23. Kaiser, D. Coupling cell movement to multicellular development in myxobacteria. *Nat. Rev. Microbiol.* **1**, 45–54 (2003).
24. Wu, S.S. & Kaiser, D. Genetic and functional evidence that Type IV pili are required for social gliding motility in Myxococcus xanthus. *Mol. Microbiol.* **18**, 547–558 (1995).
25. Lowe, J., van den Ent, F. & Amos, L.A. Molecules of the bacterial cytoskeleton. *Annu. Rev. Biophys. Biomol. Struct.* **33**, 177–198 (2004).
26. Wolgemuth, C., Hoiczyk, E., Kaiser, D. & Oster, G. How myxobacteria glide. *Curr. Biol.* **12**, 369–377 (2002).
27. Raetz, C.R. & Whitfield, C. Lipopolysaccharide endotoxins. *Annu. Rev. Biochem.* **71**, 635–700 (2002).

28. Gaspar, J.A., Thomas, J.A., Marolda, C.L. & Valvano, M.A. Surface expression of O-specific lipopolysaccharide in *Escherichia coli* requires the function of the TolA protein. *Mol. Microbiol.* **38**, 262–275 (2000).
29. Fink, J.M. & Zissler, J.F. Defects in motility and development of Myxococcus xanthus lipopolysaccharide mutants. *J. Bacteriol.* **171**, 2042–2048 (1989).
30. Youderian, P., Burke, N., White, D.J. & Hartzell, P.L. Identification of genes required for adventurous gliding motility in Myxococcus xanthus with the transposable element mariner. *Mol. Microbiol.* **49**, 555–570 (2003).
31. Caberoy, N.B., Welch, R.D., Jakobsen, J.S., Slater, S.C. & Garza, A.G. Global mutational analysis of NtrC-like activators in Myxococcus xanthus: identifying activator mutants defective for motility and fruiting body development. *J. Bacteriol.* **185**, 6083–6094 (2003).
32. Kroos, L., Kuspa, A. & Kaiser, D. Defects in fruiting body development caused by Tn5 lac insertions in Myxococcus xanthus. *J. Bacteriol.* **172**, 484–487 (1990).
33. Harris, M.A. et al. The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res.* **32** Database issue, D258–261 (2004).
34. Michalickova, K. *et al.* SeqHound: biological sequence and structure database as a platform for bioinformatics research. *BMC Bioinformatics* **3**, 32 (2002).
35. Camon, E. *et al.* The Gene Ontology Annotation (GOA) Database: sharing knowledge in Uniprot with Gene Ontology. *Nucleic Acids Res.* **32**, D262–D266 (2004).
36. Boyle, E.I. *et al.* GO:TermFinder – open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes. *Bioinformatics* **20**, 3710–3715 (2004).
37. Venter, J.C. *et al.* Environmental genome shotgun sequencing of the Sargasso Sea. *Science* **304**, 66–74 (2004).
38. McAdams, H.H., Srinivasan, B. & Arkin, A.P. The evolution of genetic regulatory systems in bacteria. *Nat. Rev. Genet.* **5**, 169–178 (2004).
39. Holder, M. & Lewis, P.O. Phylogeny estimation: traditional and Bayesian approaches. *Nat. Rev. Genet.* **4**, 275–284 (2003).
40. Daubin, V., Moran, N.A. & Ochman, H. Phylogenetics and the cohesion of bacterial genomes. *Science* **301**, 829–832 (2003).
41. Florea, L., McClelland, M., Riemer, C., Schwartz, S. & Miller, W. EnteriX 2003: Visualization tools for genome alignments of Enterobacteriaceae. *Nucleic Acids Res.* **31**, 3527–3532 (2003).
42. Galperin, M.Y. & Koonin, E.V. Who's your neighbor? New computational approaches for functional genomics. *Nat. Biotechnol.* **18**, 609–613 (2000).
43. Stuart, J.M., Segal, E., Koller, D. & Kim, S.K. A gene-coexpression network for global discovery of conserved genetic modules. *Science* **302**, 249–255 (2003).