Title

Identification of macromolecular assemblies and determination of their structures

Permalink

https://escholarship.org/uc/item/2v53h0bb

Author

Cimermancic, Peter

Publication Date

2014

Peer reviewed|Thesis/dissertation

Identification of macromolecular assemblies and determination of their structures

by

Peter Cimermancic

DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

Biological and medical informatics

of the

UNIVERSITY OF CALIFORNIA, SAN FRANCISCO

[redacted]

To my dearest wife and son,

without whom it would take another two years

before this thesis was completed.

Acknowledgements

I would like to express my deepest gratitude to my advisor, Andrej Sali. He has been a great mentor, and I have learned from him not only how to do rigorous science, but also how to write grants, give talks, and manage projects and collaborations. I have appreciated his challenging questions, that he tries to motivate people rather than delegate, and that he is invariably positive about other scientists and their work. I am thankful for everything he has done for me, including (*i*) giving me the opportunity to join his lab for a few months during my last college year, (*ii*) helping me with applications for graduate schools, (*iii*) numerous recommendation letters he has had to write, (*iv*) his understanding and support when my son was born, and (*v*) all of the help with our most recent business endeavor. While I am sure I have forgotten to add many things to the list, I will not forget that he did not leave me in the ocean with sharks while we were swimming from one site of the Cabo bay to another.

I have also had the pleasure to learn from and work with many other PIs. In particular, I would like to thank Nevan Krogan for teaching me about systems biology and giving me the opportunity to spend time in his lab as a rotation student and onwards. I would also like to express my appreciation to Michael Fischbach, for being an unlimited resource of knowledge from many fields in biology, for giving me the opportunity to work on very interesting projects in his lab as a rotation student and beyond, and for many discussions we had about new methods of genome mining. I would also like to thank both Nevan and Michael for all the valuable feedback they have given me during my thesis committee meetings. I am also grateful to Rahul Andino, Yifan Cheng, Rahul Deo, Jaime Fraser, Jason Gestwicki, and Jim Wells for the time they have put into our collaborations, advice and support they have given me, for letting me use their lab space and equipment, and for letting their lab members help me with my projects. I would also like to thank to John Gross, my final thesis and qualifying exam committee member, as well as Alan Frankel and Matt Jacobson, my qualifying exam committee members, for invaluable feedback during the exam, and for many meetings and advice leading up to it.

During my time at UCSF, I have benefited from having many collaborators and friends. Starting with my classmates, Erik Jonsson and Charles Greenberg, the “Gentleman scientists”: We have spent long nights together isolating RNA and tweaking hand-made microscopes during our team challenges, only to realize months later that it was all done in vain (I owe them an apology: the most stupid and time consuming ideas, such as the “manual clustering”, were most probably mine). I also want to thank them for all the discussions, help, fun nights, support, and friendship outside the university walls. Many thanks to Aram Avila-Herrera and Zac Apte, classmates and initial roommates, for finding our first apartment in dangerous territory, and for all the great time we have had together, especially the occasional brunches at Crepes on Cole. I also want to thank all other classmates from Biophysics and BMI programs for the (healthy and necessary) peer pressure that has kept me going. My thanks also to Avner Schlessinger for all the scientific feedback, running/jogging on the “catwalk,” and nights in North Beach. Many thanks also to Dave Barkan, Javier Angel Velazquez-Muriel, Jeremy Philips, Daniel Russel, Hao Fan, Barak Raveh, Dina Schneidman, Ben Webb, Ursula Pieper, Elina Tjioe, and Keren Lasker for all the help and advice, for a warm welcome when I joined the lab, and for many fun moments that have followed. Also, many thanks to the rest of the Sali lab (former and current members) for all of their help, feedback I have gotten at group meetings, discussions, great times, and gracious presents at my marriage and for the birth of my son. Special thanks also go to the two Italians in the lab, Riccardo Pellarin and Massimilliano Bonomi (mad Max), for basically being my second mentors. I am also thankful for all the help from our lab coordinator, Hilary Mahon, who makes it all possible: taking care of endless reimbursements, scheduling rooms, visa status, job appointments, and generally keeping the lab functioning.

I also owe my deep gratitude to Kiyoshi Egami, for being a great mentor when I decided to do some experimental biology, and for always unselfishly offering to help—even if it meant sitting in front of a microscope all night long. I also truly appreciate scientific discussions and help from Justin Rettenmaier and Jun Zou. Thanks also to Manon Eckhardt for plating and transfecting human cell cultures, and Zapporah Young for testing our compounds in the cell models so many times. I'm also grateful to the members of labs I have rotated in for all the advice and great times: Hannes Braberg, Natali Gulbache, Stefanie Jaeger, and Jeff Johnson from Krogan lab, and Marnix Medema, Jan Claesen, and Mohamed Abou Donia from Fischbach lab. Special thanks to Hannes Braberg and Assen Roguev for not disclosing my super-powers to everyone at UCSF.

Finishing graduate school would have not been possible without people who helped me occasionally escape science. Thanks to Hannes Braberg, Daniel Russel, Mikael Palner, and Claus Enevoldsen for being great mountain biking buddies and for showing me all the gnarly trails in the Bay area and the Sierras (Skeggs point is still my favorite one, though), and Allison Doak, Avner Schlessinger, Hao Fan, Rocco Varela, and Jens Carlson for the basketball time.

Then there are those with whom I have had a lot of fun just hanging around. I am deeply grateful to Yulia and Erik for being great friends, for all the help, and wine we have consumed; to Charles and Rebecca for the great parties you have thrown; to Erin and Phouc (and Anya) for numerous dinners, a trip to Tahoe, and play dates that were always full of fun; to Dina and Alex for all the generous support; to Hannes and Ayse for a lot of good laughs, hikes, and cherries; to Brittany and Mattfor being such a wonderful neighbors; and to Damjan and Tadeja for good times during our visits back in Slovenia.

Tons of thanks to my old friends from early days in college, Damjan Celic, Domen Zafred, and Andrej Ondracka, I hold them responsible for leaving my home country. If it weren't for them, their influence, and subconscious pressure, I would have most likely stayed in Slovenia—and I am not saying that would be bad, but I would miss this great experience.

Finally, I want to thank my family. To my parents, Branka and Joze, for being such great parents and being so supportive of all my choices. Thanks for driving me to accordion and piano classes, and bike races as a kid, for supporting my choice to move so far away to San Francisco, for helping organizing my wedding, and for being great grandparents. To my sisters, Jozi and Anja, thank you for being wonderful sisters, for thinking of me on so many occasions, and for forgetting all the fights we had as kids. Last but foremost, I would like to thank my wife, Klavdija, for being my best friend, for quitting a great job you had back in Slovenia and moving all the way to San Francisco with me, and for making so many other sacrifices that come with living with a graduate student. I also want to thank her for our wonderful son, Adrian, who is such a joy to be spending time with. I love you both the most!

Contributions

Several chapters of this dissertation describe studies that have been published in the journals cited herein. The chapters do not necessarily represent the final published form. These studies were carried out with the collaborators listed as co-authors for each chapter. The study described in Chapter 2, “Insights into secondary metabolism from a global analysis of prokaryotic biosynthesis”, is in press at *Cell*. In this study, Peter Cimermancic developed an algorithm for identification of biosynthetic gene clusters in prokaryotic genomes, in part carried out data analysis, and wrote the paper with co-authors. The study described in Chapter 3, “Global landscape of HIV-human protein assemblies”, was published in January 2012 in *Nature,* Vol. 481, pp. 365-370. In this study, Peter developed a computational score to identify biologically relevant HIV-human protein assemblies, carried out analyses of the interaction data, and wrote the paper with co-authors. The manuscript of the study described in Chapter 4, “Determining architectures of macromolecular assemblies based on cell colony sizes”, is currently in preparation and will be submitted by the end of 2014. Peter has designed and carried the study, and is now writing a manuscript with co-authors. The study described in Chapter 5, “Expanding the druggable proteome by characterization and prediction of cryptic binding sites”, is submitted to Proceedings of the National Academy of Sciences of the Unites States of America. Peter is one of the two corresponding authors of this study; he has designed and developed a computational method for identification of cryptic sites, managed collaborations, as well as designed the experiment and wrote the manuscript with co-authors.

The work of Peter Cimermancic described herein was done under the supervision of Andrej Sali and meets the requirements for a standard PhD dissertation.

Andrej Šali

Identification of macromolecular assemblies and determination of their structures

Peter Cimermancic

Abstract

To understand the workings of a living cell, we need to identify its molecular components and determine how they associate with each other. To date, studies that identify new macromolecular assemblies have been mainly limited to low-throughput biochemistry assays. Structures of macromolecular assemblies also have been difficult to obtain, and are mostly available for a small subset of individual components or their assemblies amenable to conventional approaches, such as X-ray crystallography and nuclear magnetic resonance spectroscopy. In this dissertation, I describe my contributions to the development of four novel pipelines that utilize new technologies and datasets to facilitate the identification of macromolecular assemblies and the determination of their structures. First, we designed an algorithm to identify genes coding for biosynthetic macromolecular assemblies. Second, we developed a platform to identify previously unknown HIV-human protein assemblies based on affinity purification, mass spectrometry, and computational scoring. Third, to aid the structure determination of macromolecular assemblies that are challenging to isolate and purify, we proposed a new strategy based on *in vivo* measurements of genetic interaction and integrative modeling. Finally, to facilitate rational discovery of small molecule modulators of macromolecular assemblies and their components, we presented a new approach based on computational identification of putative ligand-binding sites that are not detectable in ligand-free structures, due to insufficient structure resolutions or flatness in the absence of a ligand.

Table of Contents

List of Figures

Chapter 2

Chapter 3

Chapter 4

Chapter 5

List of Supplementary Materials

Chapter 2

Chapter 3

Chapter 4

Chapter 5

Chapter 1

Introduction

Introduction

Macromolecular assemblies consist of interacting proteins, nucleic acids, and (in some cases) small molecules. These complexes vary widely in size and play crucial roles in most cellular processes (Alberts et al., 2002). For example, biosynthesis of erythromycin, an antibiotic used to treat a number of bacterial infections, is carried out by a 200 kDa assembly of several enzymes and small-molecule building blocks resembling an assembly line (Walsh and Fischbach, 2010). Another assembly, the nuclear pore complex, regulates macromolecular transport across the nuclear envelope and is composed of ~456 proteins (Alber et al., 2007). Moreover, pathogens (viruses in particular) have very small genomes and are only able to replicate by highjacking and forming macromolecular assemblies with host machinery (Jager et al., 2012). A comprehensive characterization of the functions, structures, and dynamics of biological assemblies is essential for a mechanistic understanding of the cell (and viruses) (Alber et al., 2008). Such characterization helps to elucidate the principles that underlie cellular processes. It can provide a starting point for the modulation of macromolecular assemblies by, for example, small molecules, which is of particular interest when a macromolecular assembly is involved in a disease. Key challenges include: incomplete lists of macromolecular assemblies and their components, difficulty in determination of their structures at sufficient resolution by conventional approaches, and absence of detectable ligand binding pockets that could provide a starting point for structure-based substrate and drug discovery.

TOWARDS THE COMPREHENSIVE LIST OF MACROMOLECULAR ASSEMBLIES

Macromolecules seldom function in isolation. Instead, they associate with small molecules, other macromolecules, or other copies of the same macromolecule to form a functional unit. These interactions are rarely permanent; cell growth, replication, and function are based on dynamic systems of ever-changing macromolecular assemblies. Variations in a macromolecular assembly include changes in the content of their components (one or more components can be added or removed) due to stochastic association and disassociations or changes in the components themselves (*eg*, post-translational modification of proteins). A complete list of macromolecular assemblies is therefore difficult to compile, not only because many interactions are too weak or too transient to be observed by current approaches, but also because some only exist under a specific set of conditions.

A number of assays and techniques have been developed for identification of macromolecular assemblies, most of which focus on cataloguing physical protein-protein interactions, such as the yeast two-hybrid method and direct purification *via* affinity tags (Krogan et al., 2006). In addition, a variety of computational approaches based on sequence and structure homology, gene co-expression, phylogenetic profiles, as well as gene co-localization in the case of prokaryotic genomes have been proposed to predict novel protein-protein interactions (Zhang et al., 2012). The data from protein-protein studies have been non-overlapping to a surprising degree, an observation explained partly by experimental inaccuracy and partly by incompleteness of the single screens (von Mering et al., 2002). Therefore, new approaches are needed to obtain more thorough and accurate identification of macromolecular assemblies. My research has focused on completing the lists of two different types of macromolecular assemblies: those involved in prokaryotic biosynthesis of natural products and those between human and pathogen macromolecules.

Long, multistep linear reaction sequences in synthesis of small-molecule natural products are inefficient for solution phase reactions, whether enzymes or abiotic catalysts are involved. Assemblies of different enzymes and other auxiliary proteins, which are used for biosynthesis by many organisms, act as efficient assembly lines by covalently tethering both the growing chain of a natural product and the building blocks to be incorporated into the chain at each step

(Walsh and Fischbach, 2010). Natural products have played a prominent role in the history of organic chemistry, and continue to be important as drugs, biological probes, and targets of study for synthetic and analytical chemists. The modularity of the biosynthetic assembly lines has long been seen as an opportunity to generate large libraries of new natural products by recombining their constituent domains and modules (Sherman, 2005). Although there have been notable successes, the majority of combinatorially generated assembly lines appear to be non-functional, due in part to the limited number of known assembly lines (Menzella et al., 2007; Menzella et al., 2005; Nguyen et al., 2006).

Several approaches have been proposed to identify new microbial assembly lines, including labor-intensive manual annotation, mass spectrometry-guided genome mining approaches (Kersten et al., 2011), and pipelines whereby candidate genes are prioritized for experimental validation based on *in silico* predictions (Medema et al., 2011). In microbial genomes, genes coding for assembly line proteins usually cluster together (known as biosynthetic gene clusters), facilitating their prediction by computational algorithms that search for clusters of signature biosynthetic genes. While such genome mining of biosynthetic gene clusters has become a key method to accelerate their identification and characterization, the approach is limited by design to identifying biosynthetic gene clusters similar to those in the incomplete and biased datasets of known gene clusters. Therefore, better computational and experimental tools are needed to expand the known space of biosynthetic assembly lines and identify those that potentially make novel chemical compounds. In collaboration with the Fischbach lab at the University of California, San Francisco, and the Breitling and Takano labs at the University of Groningen, The Netherlands, I developed a computational tool based on Hidden Markov Models (HMM) that efficiently and accurately predicts any type of biosynthetic gene clusters. We applied this algorithm to all sequenced microbial genomes in public sequence databases, analyzed the distribution and abundance of the predicted biosynthetic landscape, and subsequently selected two gene clusters with novel architectures for determination and characterization of the structures of the corresponding natural products. This study is described in Chapter 2.

Assemblies of human host and pathogen macromolecules have also been challenging to identify, due to the transient and weak nature of the interactions, cell toxicity, and the ability of a single pathogen macromolecule to interact with many different host macromolecules. Identification of interactions between host and pathogen macromolecules is a key goal to help us understand infections and design molecules for therapeutic intervention. However, the resources for studying such interactions are limited to cumbersome biochemical studies of individual interactions (Jager et al., 2012) or to computational predictions based on protein sequence and structure, domain profiles, or techniques that use machine learning to combine a number of different functional genomic data types (Dyer et al., 2007). The Krogan lab at University of California, San Francisco, carried out the first systematic affinity tagging/purification mass spectrometry (AP-MS) study on any host-pathogen system. The resulting dataset of putative HIV-human protein interactions was noisy and contained many non-specifically binding proteins and contaminants; over 90% of proteins detected in an AP-MS experiment belong to one of these two biologically irrelevant types of protein prey. During my rotation in the Krogan lab, I utilized the existing computational tools for the processing of AP-MS data, such as NSAF, CompPASS, and SAInt (Choi et al., 2011; Sowa et al., 2009). However, the existing tools were inaccurate, so we have built a new one, better suited for identification of AP-MS-derived host-pathogen protein assemblies. This tool, MiST (mass spectrometry interaction statistics), is accurate: we confirmed 97 of 127 AP-MS derived HIV-human protein interactions using co-immunoprecipitation/western blot analysis (76% success rate). Moreover, our approach uncovered a number of previously unknown host-pathogen assemblies, including an assembly of HIV protease and eukaryotic translation intimation factor 3, and HIV accessory factor Vif and a ubiquitin ligase complex. The study is described in Chapter 3.

STRUCTURE DETERMINATION OF MACROMOLECULAR ASSEMBLIES BY AN INTEGRATIVE APPROACH

A comprehensive characterization of the structures and dynamics of biological assemblies is essential for a mechanistic understating of the cell. Even a coarse-grained characterization of the configuration of macromolecular components in an assembly can help to elucidate the principles that underlie cellular processes, and provides a starting point for more detailed structural studies (Alber et al., 2008). However, there is a wide gap between the number of identified macromolecular assemblies and more detailed, structural and mechanistic studies. For example, whereas the number of large macromolecular assemblies in the widely studied yeast cell is estimated to be ~800 on the basis of different high-throughput experiments (Krogan et al., 2006), the number of structures of whole or partial assemblies in the Protein Data Bank (PDB) is less than 200. This gap is even wider for the human proteome, which may have an order of magnitude more assemblies than the yeast cell, with only ~900 partial or whole assembly structures available. Therefore, there may be thousands of biologically relevant macromolecular assemblies and transient interactions whose structures are yet to be characterized.

While it is relatively easy to determine structures of rigid individual components, large and dynamic assemblies usually elude conventional structural efforts. For example, X-ray crystallography is limited by the difficulties of growing suitable crystals and building molecular models into large unit cells, and nuclear magnetic resonance (NMR) spectroscopy is limited by the size of an assembly. Electron microscopy (EM) has recently shown great potential for determining the structures of macromolecular assemblies at near-atomic resolution (Liao et al., 2013), but further progress in technology is needed to provide such reconstructions for a wide range of specimen. These three approaches are also limited by sample preparation; purification and isolation of macromolecular assemblies in sufficient quantities is not always trivial. Some of the alternative structure characterization methods (*eg*, affinity purification, yeast two-hybrid system, chemical cross-linking, small-angle X-ray scattering (SAXS), and fluorescence resonance energy transfer (FRET) spectroscopy) work with lesser quantities and sample purity, but produce lower resolution information. Alternatively, computational macromolecular structure modeling and docking based on homology, shape matching, molecular dynamics simulations, and evolutionary sequence information from large sequence alignments are limited by low accuracy and sparseness of the available data (Alber et al., 2008).

The shortcomings of low-resolution methods that produce sparse information can be minimized by simultaneous consideration of all available information about a given assembly through computation. A number of structures have already been determined by such integrative approaches. For example, the structure of the 26S proteasome was solved by relying on the EM density map of the whole assembly, protein-protein interaction data from high-throughput proteomics experiments, residue-based chemical cross-linking, and comparative protein structure models of the protein components (Lasker et al., 2012). The architecture of NPC, an assembly of ~456 proteins, was also determined by integrative modeling, dependent on the stoichiometry from protein quantification, protein proximities from subcomplex purification, protein positions from immuno-EM, sedimentation analysis that sheds light on protein and subcomplex shapes, and the overall NPC shape from EM (Alber et al., 2007).

Despite these successes, structures and thus mechanistic understanding of many macromolecular assemblies remain elusive because they are difficult to purify and isolate for characterization by conventional high-resolution and coarse-grain approaches. Difficult cases include assemblies of weakly interacting macromolecules, those that are short-lived, and those that are scarce in the cell. Therefore, direct *in vivo* measurements of structural aspects of a wild-type assembly are needed. In collaboration with the Krogan lab, I explored how integrative structure determination can benefit from spatial restraints computed from the correlation between the functional impacts of two point mutations, based on *in vivo* mapping of genetic interactions between point mutations in an assembly of interest and an array of genes from a library of gene deletion mutant alleles (Braberg et al., 2013). As described in Chapter 4, we showed that the data are sufficient to determine the architectures of macromolecular assemblies, with the resulting resolution comparable to that of modeling based on sparse cross-linking datasets.

UTILITY OF STRUCTURES OF MACROMOLECULAR ASSEMBLIES IN THE PROCESS OF DRUG DISCOVERY

In addition to a mechanistic understanding of cellular biology, the goal of structure determination is to provide a starting point for structure-based ligand discovery. A molecular modulator of an assembly can serve as a probe to explore the role of the assembly in a broader biological context (Hermann et al., 2007) or, when an assembly or its individual component is associated with a disease, to potentially develop a drug. While small-molecule ligands can be found without structures by, for example, high-throughput ligand screening of chemical libraries against individual targets and phenotypic assays in cells, tissues, or whole organisms, structure-based approaches provide several advantages. For example, computer-based screening of chemical libraries is less expensive and less time-consuming than high throughput screens, even though in practice both docking and high-throughput screens of two similarly sized compound libraries can yield similar numbers of hits (Doman et al., 2002). Moreover, structure-based docking has some practical advantages, despite inaccurate scoring functions andcrude sampling of conformational states of ligands and targets. For example, docking can reliably screen out compounds that do not fit in a pocket or have grossly incorrect electrostatic properties, thus minimizing the size of the libraries used in experimental screens. Furthermore, docking can be used to screen compounds that are not yet synthesized, and can thus greatly facilitate hit and lead optimization phases of drug discovery.

The step that usually follows structure determination of an assembly is analyzing the structure for binding pockets, that is concave shapes on the surface of a macromolecule into which compounds are docked. A number of algorithms have been developed to localize such pockets, but generally they are only able to identify pockets in ~60% of protein structures (Sheridan et al., 2010). Many proteins, therefore, cannot be subjected to structure-based ligand discovery and are thus considered “undruggable.” In practice, the situation is even worse because the presence of a pocket does not necessarily guarantee a drug-like ligand, especially when a ligand is not known, when the pocket is shallow, hydrophilic, or inaccessible, or because ligand binding to a similar pocket in another protein causes adverse effects.

Fortunately, macromolecular structures are not static, and pockets may be sampled transiently, even when not visible in the determined structures (*ie*, cryptic sites). Moreover, structures may be determined at low resolutions or are inaccurate, and hence insufficient for unambiguous pocket localization, which generally requires accurate atomic structures. In Chapter 5, I describe an analysis and computational prediction of such cryptic sites. I also propose a strategy for how to expand the size of the druggable human proteome based on our algorithm and small molecule chemical tethering. In collaboration with the Fraser and Wells labs at University of California, San Francisco, we demonstrate the potential of our approach by applying it to a disease-associated protein, tyrosine-protein phosphatase non-receptor type 1.

References

Alber, F., Dokudovskaya, S., Veenhoff, L.M., Zhang, W., Kipper, J., Devos, D., Suprapto, A., Karni-Schmidt, O., Williams, R., Chait, B.T.*, et al.* (2007). The molecular architecture of the nuclear pore complex. Nature *450*, 695-701.

Alber, F., Forster, F., Korkin, D., Topf, M., and Sali, A. (2008). Integrating diverse data for structure determination of macromolecular assemblies. Annual review of biochemistry *77*, 443-477.

Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K., and Walter, P. (2002). Molecular Biology of the Cell, Vol 4 (New York: Garland Science).

Braberg, H., Jin, H., Moehle, E.A., Chan, Y.A., Wang, S., Shales, M., Benschop, J.J., Morris, J.H., Qiu, C., Hu, F.*, et al.* (2013). From structure to systems: high-resolution, quantitative genetic analysis of RNA polymerase II. Cell *154*, 775-788.

Choi, H., Larsen, B., Lin, Z.Y., Breitkreutz, A., Mellacheruvu, D., Fermin, D., Qin, Z.S., Tyers, M., Gingras, A.C., and Nesvizhskii, A.I. (2011). SAINT: probabilistic scoring of affinity purification-mass spectrometry data. Nature methods *8*, 70-73.

Doman, T.N., McGovern, S.L., Witherbee, B.J., Kasten, T.P., Kurumbail, R., Stallings, W.C., Connolly, D.T., and Shoichet, B.K. (2002). Molecular docking and high-throughput screening for novel inhibitors of protein tyrosine phosphatase-1B. Journal of medicinal chemistry *45*, 2213-2221.

Dyer, M.D., Murali, T.M., and Sobral, B.W. (2007). Computational prediction of host-pathogen protein-protein interactions. Bioinformatics *23*, i159-166.

Hermann, J.C., Marti-Arbona, R., Fedorov, A.A., Fedorov, E., Almo, S.C., Shoichet, B.K., and Raushel, F.M. (2007). Structure-based activity prediction for an enzyme of unknown function. Nature *448*, 775-779.

Jager, S., Cimermancic, P., Gulbahce, N., Johnson, J.R., McGovern, K.E., Clarke, S.C., Shales, M., Mercenne, G., Pache, L., Li, K.*, et al.* (2012). Global landscape of HIV-human protein complexes. Nature *481*, 365-370.

Kersten, R.D., Yang, Y.L., Xu, Y., Cimermancic, P., Nam, S.J., Fenical, W., Fischbach, M.A., Moore, B.S., and Dorrestein, P.C. (2011). A mass spectrometry-guided genome mining approach for natural product peptidogenomics. Nature chemical biology *7*, 794-802.

Krogan, N.J., Cagney, G., Yu, H., Zhong, G., Guo, X., Ignatchenko, A., Li, J., Pu, S., Datta, N., Tikuisis, A.P.*, et al.* (2006). Global landscape of protein complexes in the yeast Saccharomyces cerevisiae. Nature *440*, 637-643.

Lasker, K., Forster, F., Bohn, S., Walzthoeni, T., Villa, E., Unverdorben, P., Beck, F., Aebersold, R., Sali, A., and Baumeister, W. (2012). Molecular architecture of the 26S proteasome holocomplex determined by an integrative approach. Proceedings of the National Academy of Sciences of the United States of America *109*, 1380-1387.

Liao, M., Cao, E., Julius, D., and Cheng, Y. (2013). Structure of the TRPV1 ion channel determined by electron cryo-microscopy. Nature *504*, 107-112.

Medema, M.H., Blin, K., Cimermancic, P., de Jager, V., Zakrzewski, P., Fischbach, M.A., Weber, T., Takano, E., and Breitling, R. (2011). antiSMASH: rapid identification, annotation and analysis of secondary metabolite biosynthesis gene clusters in bacterial and fungal genome sequences. Nucleic acids research *39*, W339-346.

Menzella, H.G., Carney, J.R., and Santi, D.V. (2007). Rational design and assembly of synthetic trimodular polyketide synthases. Chemistry & biology *14*, 143-151.

Menzella, H.G., Reid, R., Carney, J.R., Chandran, S.S., Reisinger, S.J., Patel, K.G., Hopwood, D.A., and Santi, D.V. (2005). Combinatorial polyketide biosynthesis by de novo design and rearrangement of modular polyketide synthase genes. Nature biotechnology *23*, 1171-1176.

Nguyen, K.T., Ritz, D., Gu, J.Q., Alexander, D., Chu, M., Miao, V., Brian, P., and Baltz, R.H. (2006). Combinatorial biosynthesis of novel antibiotics related to daptomycin. Proceedings of the National Academy of Sciences of the United States of America *103*, 17462-17467.

Sheridan, R.P., Maiorov, V.N., Holloway, M.K., Cornell, W.D., and Gao, Y.D. (2010). Drug-like density: a method of quantifying the "bindability" of a protein target based on a very large set of pockets and drug-like ligands from the Protein Data Bank. Journal of chemical information and modeling *50*, 2029-2040.

Sherman, D.H. (2005). The Lego-ization of polyketide biosynthesis. Nature biotechnology *23*, 1083-1084.

Sowa, M.E., Bennett, E.J., Gygi, S.P., and Harper, J.W. (2009). Defining the human deubiquitinating enzyme interaction landscape. Cell *138*, 389-403.

von Mering, C., Krause, R., Snel, B., Cornell, M., Oliver, S.G., Fields, S., and Bork, P. (2002). Comparative assessment of large-scale data sets of protein-protein interactions. Nature *417*, 399-403.

Walsh, C.T., and Fischbach, M.A. (2010). Natural products version 2.0: connecting genes to molecules. Journal of the American Chemical Society *132*, 2469-2493.

Zhang, Q.C., Petrey, D., Deng, L., Qiang, L., Shi, Y., Thu, C.A., Bisikirska, B., Lefebvre, C., Accili, D., Hunter, T.*, et al.* (2012). Structure-based prediction of protein-protein interactions on a genome-wide scale. Nature *490*, 556-560.

Chapter 2

Insights into secondary metabolism from a global analysis of prokaryotic biosynthetic gene clusters

Insights into secondary metabolism from a global analysis of prokaryotic biosynthetic gene clusters

Peter Cimermancic[1*], Marnix H. Medema[2,3*#], Jan Claesen[1*], Kenji Kurita[4], Laura C. Wieland Brown[5], Konstantinos Mavrommatis[6], Amrita Pati[6], Paul A. Godfrey[7], Michael Koehrsen[7], Jon Clardy[8], Bruce W. Birren[7], Eriko Takano[2,9], Andrej Sali[1,10], Roger G. Linington[4], Michael A. Fischbach[1]

[1]Department of Bioengineering and Therapeutic Sciences and the California Institute for Quantitative Biosciences, University of California, San Francisco, San Francisco, CA 94158, USA

[2]Department of Microbial Physiology and [3]Groningen Bioinformatics Centre, Groningen Biomolecular Sciences and Biotechnology Institute, University of Groningen, Nijenborgh 7, 9747AG Groningen, The Netherlands

[4]Department of Chemistry and Biochemistry, University of California, Santa Cruz, Santa Cruz, CA 95064, USA

[5]Department of Chemistry, Indiana University, Bloomington, IN 47405, USA

[6]US Department of Energy, Joint Genome Institute, Walnut Creek, CA 94598, USA

[7]The Broad Institute, Cambridge, MA 02142, USA

[8]Department of Biological Chemistry and Molecular Pharmacology, Harvard Medical School, Boston, MA 02115, USA

[9]Manchester Institute of Biotechnology, Faculty of Life Sciences, University of Manchester, Manchester, United Kingdom

[10]Department of Pharmaceutical Chemistry, University of California, San Francisco, San Francisco, CA 94158, USA

Contact: fischbach@fischbachgroup.org

*Denotes equal contribution

[#]Present address: Max Planck Institute for Marine Microbiology, Celsiusstraße 1, 28359 Bremen, Germany

Summary

Although biosynthetic gene clusters (BGCs) have been discovered for hundreds of bacterial metabolites, our knowledge of their diversity remains limited. Here, we used a novel algorithm to systematically identify BGCs in the extensive extant microbial sequencing data. Network analysis of the predicted BGCs revealed large gene cluster families, the vast majority uncharacterized. We experimentally characterized the most prominent family, consisting of two subfamilies of hundreds of BGCs distributed throughout the Proteobacteria; their products are aryl polyenes, lipids with an aryl head group conjugated to a polyene tail. We identified a distant relationship to a third subfamily of aryl polyene BGCs, and together the three subfamilies represent the largest known family of biosynthetic gene clusters, with more than 1,000 members. Although these clusters are widely divergent in sequence, their small molecule products are remarkably conserved, indicating for the first time the important roles these compounds play in Gram-negative cell biology.

Introduction

Microbial natural products are widely used in human and veterinary medicine, agriculture, and manufacturing, and are known to mediate a variety of microbe-host and microbe-microbe interactions. Connecting these natural products to the genes that encode them is revolutionizing their study, enabling genome sequence data to guide the discovery of new molecules (Bergmann et al., 2007; Challis, 2008; Franke et al., 2012; Freeman et al., 2012; Kersten et al., 2011; Laureti et al., 2011; Lautru et al., 2005; Letzel et al., 2012; Nguyen et al., 2008; Oliynyk et al., 2007; Schneiker et al., 2007; Walsh and Fischbach, 2010; Winter et al., 2011). The thousands of prokaryotic genomes in sequence databases provide an opportunity to generalize this approach through the identification of biosynthetic gene clusters (BGCs): sets of physically clustered genes that encode the biosynthetic enzymes for a natural product pathway.

Besides core biosynthetic enzymes, many BGCs also harbor enzymes to synthesize specialized monomers for a pathway. For example, the erythromycin gene cluster encodes a set of enzymes for biosynthesis of two deoxysugars, d-desosamine and l-mycarose, that are appended to the polyketide aglycone (Oliynyk et al., 2007; Staunton and Weissman, 2001), while BGCs for glycopeptide antibiotics contain enzymes to synthesize the nonproteinogenic amino acids β-hydroxytyrosine, 4-hydroxyphenylglycine, and 3,5-dihydroxyphenylglycine that their core nonribosomal peptide synthetases use in the assembly of their peptidic scaffolds (Kahne et al., 2005; Pelzer et al., 1999). In many cases, transporters, regulatory elements, and genes that mediate host resistance are also contained within the BGC (Walsh and Fischbach, 2010). Although some BGCs are so well understood that the biosynthesis of their small molecule product has been reconstituted in heterologous hosts (Pfeifer et al., 2001) or in vitro using purified enzymes (Lowry et al., 2013; Sattely et al., 2008), little is known about the vast majority of BGCs, even those that have been connected to a small molecule product.

Here, we report the results of a systematic effort to identify and categorize BGCs in 1,154 sequenced genomes spanning the prokaryotic tree of life. We envisioned that the resulting ‘global map’ of biosynthesis would enable BGCs to be systematically selected for characterization by searching for, e.g., biosynthetic novelty, presence in undermined taxa, or patterns of phylogenetic distribution that indicate functional importance. Surprisingly, the map revealed large and very widely distributed BGC families of unknown function. We experimentally characterized the most prominent of these families, leading to the unexpected finding that gene clusters responsible for producing aryl polyene carboxylic acids constitute the largest BGC family in the sequence databases.

Results and discussion

THE CLUSTERFINDER ALGORITHM DETECTS BGCS OF BOTH KNOWN AND UNKNOWN CLASSES

Several algorithms have been developed for the automated prediction of BGCs in microbial genomes (Khaldi et al., 2010; Li et al., 2009; Medema et al., 2011; Starcevic et al., 2008; Weber et al., 2009), but each of these tools is limited to the detection of one or more well-characterized gene cluster classes. As a more general solution to the gene cluster identification problem, we developed a hidden Markov model-based probabilistic algorithm, ClusterFinder, that aims to identify gene clusters of both known and unknown classes. ClusterFinder is based on a training set of 732 BGCs with known small molecule products that we compiled and manually curated (**SI Table I**). To scan a genome for BGCs, it converts a nucleotide sequence into a string of contiguous Pfam domains and assigns each domain a probability of being part of a gene cluster, based on the frequencies at which these domains occur in the BGC and non-BGC training sets, and the identities of neighboring domains (**Figure 1a**, **Experimental Procedures**). Since ClusterFinder is based solely on Pfam domain frequencies, and Nature uses distinct assemblages of the same enzyme superfamilies to construct unrelated natural product classes, ClusterFinder exhibits relatively little training set bias and is capable of identifying new classes of gene clusters effectively (See **Experimental Procedures** for a detailed description of how we validated ClusterFinder).

A GLOBAL PHYLOGENOMIC ANALYSIS OF BGCS PROVIDES A QUANTITATIVE PERSPECTIVE ON BACTERIAL SECONDARY METABOLITE BIOSYNTHESIS

Our method predicted a total of 33,351 putative BCGs (with an estimated false-positive rate of 5%) in 1,154 genomes of organisms throughout the prokaryotic tree of life (**Figure 1c-d, SI Text 1**), which we subjected to an extensive phylogenomic analysis (**SI Text 2-3, SI Figures**

1, 2, 3, SI Tables I-II). We divided the predicted BGCs into two categories – high-confidence (10,724; used in all subsequent analyses) and low-confidence (22,627) – based on assignment to one of ~20 well-validated BGC classes or on manual inspection for clusters that could not be assigned to any known class. Within the high-confidence set, 7,377 of the predicted gene clusters (69%) were not detected by antiSMASH (Blin et al., 2013; Medema et al., 2011); the difference is due primarily to the fact that antiSMASH does not detect certain BGC classes (including many oligosaccharides), highlighting the need for a tool that identifies BGCs independent of class (**Figure 1b**).

Strikingly, 40% of all predicted BGCs encode saccharides, more than twice the size of the next largest class. Notably, only 13% of previously reported BGCs encode the biosynthesis of saccharides (**SI Text 4**). 93% of species harbor saccharide gene clusters, and in 33% of species, more than half of the predicted gene clusters encode saccharides. Cell-associated saccharides such as lipopolysaccharides (Park et al., 2009), capsular polysaccharides (Kadioglu et al., 2008), and polysaccharide A (Mazmanian et al., 2005; Mazmanian et al., 2008) are known to play key roles in microbe-host and microbe-microbe interactions, while diffusible saccharides have a range of biological activities, most notably antibacterial (Flatt and Mahmud, 2007; Weitnauer et al., 2001). The functions of many of the putative saccharide BGCs are still a mystery: 32%, including BGCs from entirely unexplored genera, are not closely related to any known gene cluster (**Figure S1e**). Saccharide BGC repertoires are also surprisingly diverse: only 37% occur in the genomes of two species chosen at random from the same genus (compared to 43% for polyketides, 60% for terpenoids and 74% for fatty acids, **Figure S1f**). The abundance of novel oligosaccharide BGC families raises the possibility that more clinically relevant saccharides such as the antidiabetic drug acarbose and the antibiotics gentamicin and avilamycin will be discovered (Kersten et al., 2013). Another BGC class of unexpectedly large size is the ribosomally synthesized and posttranslationally modified peptides (RiPPs (Arnison et al., 2013)). Notably, RiPP BGCs are as prevalent in our data set as those encoding nonribosomal peptides (**Figure 1b**).

A BGC DISTANCE NETWORK REVEALS UNEXPLORED REGIONS OF THE BIOSYNTHETIC UNIVERSE

We next sought to study the relationships among BGCs systematically, with the ultimate goal of creating a global BGC map that could be searched systematically to identify clusters of biosynthetic or taxonomic interest. We adapted a measure of the evolutionary distance between multi-domain proteins (Lin et al., 2006) to calculate an all-by-all distance matrix for the 10,724 BGCs in our high confidence set along with the 732 members of our training set. Using MCL clustering to identify groups of related nodes, we define 905 BGC families with distinct core genetic components. The resulting BGC distance network (**Figure 2, SI Text 5, Figure S4**) revealed an unexpected finding: the presence of large cliques that represent very widely distributed BGC families without any experimentally characterized members.

While most known families of secondary metabolites are unique to a small set of organisms, a few are taxonomically widespread. These include the O-antigens, capsular polysaccharides, carotenoids and NRPS-independent siderophores, which can all be clearly distinguished as prominent cliques within our distance network. From a fundamental microbiological perspective, these are among the most important families of molecules produced by microbes and, as such, they have been very intensively studied (Challis, 2005; Rehm, 2010; Samuel and Reeves, 2003; Walter and Strack, 2011). Although we had anticipated finding small gene cluster families of unknown function, we were surprised to discover families harboring hundreds of uncharacterized clusters, distributed widely throughout entire bacterial phyla.

We selected the most prominent of these families for experimental characterization: a set of 811 BGCs, distributed between two subfamilies (hereafter, subfamily 1 and 2), that were not detected by any of the existing BGC identification tools (e.g., antiSMASH), likely because the ketosynthase and adenylation domains they harbor are from uncharacterized, evolutionarily distant clades. BGCs in this family are ~20 kb in size and harbor a core set of genes that include adenylation, ketosynthase, acyl/glycosyltransferase, ketoreductase, dehydratase, thiolation, and thioesterase domains, as well as an outer membrane lipoprotein carrier protein and an MMPL family transporter (**Figure 3a, Figure S5**). These clusters are found in a wide variety of Gammaproteobacteria (*Acinetobacter, Aggregatibacter*, *Escherichia*, *Klebsiella*, *Pantoea, Pseudoalteromonas, Pseudomonas*, *Serratia*, *Shewanella*, *Vibrio*, and *Yersinia*), as well as a broader set of Beta- (*Burkholderia*, *Neisseria*) and Epsilonproteobacteria (*Campylobacter*) (**Figure 3a**).

THE UNEXPLORED BGC FAMILY ENCODES THE BIOSYNTHESIS OF ARYL POLYENE CARBOXYLIC ACIDS

We set out to identify the small molecule product of two clusters in the family, one each from subfamilies 1 and 2. We used circular polymerase extension cloning (CPEC) (Quan and Tian, 2009) to amplify and assemble the 18 gene, 15.5 kb cluster from *E. coli* CFT073 (c1186-c1204), and we transferred a plasmid harboring the cluster into *E. coli* Top10. The transformants exhibited a strong yellow pigmentation that was absent in the empty vector control strain and not observed in the native host strain (**Figure 3c**), but the pigment did not appear to diffuse into liquid or solid culture medium. We liberated the pigment from an organic extract of the cell mass by mild base hydrolysis and purified it by HPLC. Comparative HPLC analysis of extracts from the cluster+ and cluster- strains revealed the presence of a compound unique to the cluster+ strain with an absorption maximum of 425 nm, consistent with a yellow chromophore (**Figure**

S7e). Purification of milligram quantities of the compound for structural characterization required the development of an isolation procedure that rigorously excluded exposure to light. A combination of 1D- and 2D-NMR experiments and high-resolution MS on the purified compound revealed that it was an aryl polyene (APE) carboxylic acid consisting of a 4-hydroxy-3-methylphenyl head group conjugated to a hexaenoic acid (**Figure 3b, Figure S7c, e-f**).

To study the 20 gene, 18.9 kb cluster from *Vibrio fischeri* ES114 (VF0841-VF0860), we first deleted the cluster from its native producer. The yellow pigmentation that is observed in wild type *V. fischeri* under normal laboratory growth conditions was absent in the *V. fischeri* knockout strain (**Figure 3c**). We then proceeded to amplify, assemble, and introduce the *V. fischeri* cluster into *E. coli* Top10, but the native cluster failed to confer yellow pigmentation on its heterologous host. We then constructed a modified variant of the cluster in which the *ermE** promoter was inserted upstream of the operon starting with VF0844. Introduction of this construct into *E. coli* resulted in a yellow-pigmented strain that produced a new compound with an absorption maximum at 425 nm (**Figure 3c**). Purification of the *V. fischeri* compound and analysis by a combination of 1D- and 2D-NMR experiments and high-resolution MS revealed a structure with a similar scaffold to the *E. coli* APE but a 4-hydroxy-3,5-dimethylphenyl head group (**Figure 3b and Figure S7d-f**). Taken together, these data suggest that the cluster representatives from this family encode APE carboxylic acids.

THE ARYL POLYENE BGCS ARE THE LARGEST FAMILY IN THE SEQUENCE DATABASES

To our surprise, the *E. coli* and *V. fischeri* APEs are similar in structure to flexirubin (Fuchs et al., 2013; McBride et al., 2009), a pigment that was previously isolated from the CFB group bacterium *Flexibacter elegans*, and xanthomonadin (Goel et al., 2002), the compound that gives *Xanthomonas spp.* their characteristic yellow color. The biosynthetic genes for flexirubin and xanthomonadin are known (Fuchs et al., 2013; Goel et al., 2002; McBride et al., 2009); both are part of a smaller, distinct subfamily in the ClusterFinder results set (subfamily 3 in **Figure 3a**), but little else is known about the genes in either cluster. Intriguingly, although the clusters in subfamily 3 share similar Pfam domain content to those in subfamilies 1 and 2, the percent identities of their constituent proteins are very low (<20% for some amino acid sequences, see **Figure S6a**). When we turned to a more sensitive approach in which we used MultiGeneBlast (Medema et al., 2013) to look for sequence similarity at the level of the entire gene cluster, we observed distant but recognizable homology between multiple gene pairs from BGCs from subfamily 3 and subfamilies 1 and 2, indicating that the APE clusters might share a common ancestor. Indeed, when we performed a maximum-likelihood phylogenetic analysis of the ketosynthase and adenylation enzyme superfamilies based on structure-guided multiple sequence alignments (**SI Text 6, Figure S6b-d**), we found that the APE KS and A enzymes cluster together in separate uncharacterized clades that are only distantly related to all other known members of these enzyme superfamilies. Based on this evidence, we conclude that the three subfamilies together comprise a single BGC family of >1000 gene clusters (**Figure 3a**). Notably, the APE family is, to our knowledge, the largest family of gene clusters in the database, even exceeding the size of the well-known carotenoids (870 clusters, as detected using the same methods, see **SI Table III**).

The lack of homology even between the xanthomonadin and flexirubin biosynthetic genes (both in subfamily 3) is so profound that these pigments have never been connected in the literature: indeed, both previously discovered APEs have been proposed as chemosystematic markers of a genus (*Flexibacter* and *Xanthomonas*) because of their "limited distribution among bacteria" (Fautz and Reichenbach, 1979; Jenkins and Starr, 1982; Reichenbach et al., 1980; Starr et al., 1977; Wang et al., 2013). Our results, however, show that APE family BGCs are widely distributed throughout the Gram-negative bacterial tree of life (**Figure 4, Figure S3**). Notably, their pattern of phylogenetic distribution is markedly discontinuous: clusters are present in some strains but not others of most genera (36.4% of the complete genomes in a typical genus harbor the cluster, but note the high standard deviation of 37.9%). The most parsimonious explanations for this distribution pattern are frequent gene cluster loss from the descendants of a cluster-harboring ancestor, or frequent horizontal transfer among the descendants of a cluster-negative ancestor. Two lines of evidence support the possibility of frequent horizontal transfer: The family 1 cluster from *E. coli* O157:H7 is located on an O-island (Dong and Schellhorn, 2009), and the family 2 cluster from *Acinetobacter sp.* ADP1 resides on an element that has been identified as horizontally transferred (Barbe et al., 2004). Their broad distribution, and the fact that such widely divergent gene clusters have small molecule products that are so similar in structure, suggests the possibility that aryl polyenes play an important role in Gram-negative cell biology.

ARYL POLYENES MIGHT FUNCTION AS PROTECTIVE AGENTS AGAINST OXIDATIVE STRESS

Xanthomonadin has been proposed to play a role in protection from photodamage by visible light (Poplawsky et al., 2000; Rajagopal et al., 1997), an effect that is thought to be due to its ability to quench the reactive oxygen species (ROS) that are generated when the photosensitizer used in these studies, toluidine blue, is exposed to visible light (Poplawsky et al., 2000). Additionally, xanthomonadin has been shown to protect cellular lipids from peroxidation *in vitro* (Rajagopal et al., 1997) and xanthomonadin mutants show reduced epiphytic survival under conditions of natural light exposure (Poplawsky et al., 2000).

Similarly, we hypothesize that other APEs play a role in protecting bacterial cells from exogenous oxidative stress. Membrane-bound APEs could reduce the concentration of free radicals that would otherwise cause damage to other cellular lipids, proteins, or nucleic acids. Notably, many bacteria that harbor APE BGCs are either commensals or pathogens of a eukaryotic host; consequently, they are likely to encounter oxidative stress from immune cells during colonization or infection.

A role for APEs in protecting Gram-negative bacteria against oxidative stress would make them analogous to the chemically similar but biosynthetically distinct Gram-positive carotenoids, whose antioxidant activity is well established. An important example is staphyloxanthin, a membrane-bound carotenoid virulence factor that is responsible for the characteristic yellow pigmentation of *S. aureus* and proposed to protect *S. aureus* from immune-mediated oxidative stress. A *S. aureus* mutant defective in the first committed step of staphyloxantin biosynthesis exhibits higher susceptibility to various reactive oxygen species and in a neutrophil killing assay (Clauditz et al., 2006; Liu et al., 2005). This mutant was also attenuated in murine models for subcutaneous abscess (Liu et al., 2005) and systemic infection (Liu et al., 2008). Experiments to test whether APE-deficient mutants of Gram-negative bacteria harbor colonization or pathogenesis defects will be an important step in testing this model and gaining insight into why APE gene clusters are so widely distributed throughout the Gram-negative tree of life.

USING SYSTEMATIC SEARCHES TO PRIORITIZE BGCS FOR EXPERIMENTAL CHARACTERIZATION

BGCs are commonly selected for characterization on the basis of chemical or enzymatic novelty. Following the example of the APE family, we anticipate that our global BGC map will enable gene clusters to be selected in a new way that is based on a criterion biologists have long used to prioritize genes: what are the most widely distributed gene clusters of unknown function? Various other prioritization criteria could be used to select BGCs of interest (Frasch et al., 2013). For example, one could select BGCs likely to encode new chemical scaffolds by searching for clusters that do not harbor conventional monomer-coupling enzymes.

Many gene cluster families still await characterization: even with conservative assumptions, we estimate the total number of bacterial BGC families (such as those encoding carotenoids or calcium-dependent lipopeptides) present in the biosphere to be ~6,000 **(Figure S1g)**, less than half of which are identified in our current set of genomes (~2,400). Importantly, each of these 6,000 families will likely contain a range of molecules with distinct biological activities. As developments in single-cell genomics and metagenomics are opening up the exploration of a vast microbial dark matter, this number may grow even further: just in the 201 single-cell genomes of uncultivated organisms recently obtained by the JGI (Rinke et al., 2013), our method identified 947 candidate BGCs, of which 655 fall outside all known BGC classes (**Figure S1h**). Even among cultivated organisms, there are still many underexplored taxa (Letzel et al., 2012) (**SI Text 2**). For the foreseeable future, the number of gene clusters encoding molecules with distinct scaffolds will continue to rise as new genomes are sequenced, and computational approaches to systematically study their relationships will be of great value in prioritizing them for experimental characterization.

Experimental methods

GENOME SEQUENCES

A set of 1154 complete genome sequences was obtained from JGI-IMG (Markowitz et al., 2012), version 3.2 (08/17/2010).

CLUSTERFINDER ALGORITHM AND TRAINING DATA

The ClusterFinder prediction algorithm for BGC identification is a two-state Hidden Markov Model (HMM), with one hidden state corresponding to biosynthetic gene clusters (BGC state) and a second hidden state corresponding to the rest of the genome (non-BCG state). The training set for the BGC state was gathered using a comprehensive search of the scientific literature, which yielded 732 clusters. From these, 55 redundant BGCs were filtered out by selecting one random member from each biosynthetic gene cluster family, with a cluster family defined as a connected component in the >0.7 similarity network (see below). Thus, the final BGC state training set consisted of 677 experimentally characterized gene clusters. For the non-BGC state, non-BGC regions were collected from 100 randomly selected genomes, defined as those regions without significant sequence similarity to the BGC state training set sequences (Pfam domain similarities with E-value > 1e-10). ClusterFinder source code is available from the GitHub repository (https://github.com/petercim/ClusterFinder).

CLUSTERFINDER VALIDATION

The algorithm was validated in three ways. First, its output was compared to 10 bacterial genomes manually annotated for BGCs (leading to an area under the ROC curve of 0.84) (**Figure S7a**). Second, its performance was assessed on 74 experimentally characterized BGCs outside the training set (**Figure S7b**). Out of these, 70 (95%) were detected successfully. When tested alongside antiSMASH (Medema et al., 2011) on the genomes of *Pseudomonas* *fluorescens* Pf-5, *Streptomyces griseus* IFO13350 and *Salinispora tropica* CNB-440 (**SI Table IV**), antiSMASH detected 62 out of 65 (95%) manually annotated secondary metabolite gene clusters, while ClusterFinder detected 59 of these (91%). However, ClusterFinder identified 43 (66%) unannotated gene clusters that appeared likely to synthesize small molecule metabolites on manual inspection, whereas antiSMASH detected only five (8%). This highlights the strength of ClusterFinder in detecting gene clusters irrespective of whether they belong to known or *a priori* specified classes. Among the additional gene clusters detected by ClusterFinder are known gene clusters encoding the biosynthesis of, e.g., alginate and lipopolysaccharides, as well as an uncharacterized cluster that was previously predicted to encode a novel type of secondary metabolite (Hassan et al., 2010).

TYPE CLASSIFICATION OF BGCS

ClusterFinder-detected biosynthetic gene clusters were classified by antiSMASH (Medema et al., 2011) to determine their subtypes (e.g., type I polyketide, nonribosomal peptide, terpenoid). The native antiSMASH types were supplemented by a list of profile HMMs for protein domains characteristic of saccharide gene clusters (**SI Table V**), as well as by fatty acid gene clusters, which could be assigned based on the HMMs that antiSMASH uses in polyketide synthase annotation. Gene clusters lacking protein domains characteristic of gene cluster classes included in antiSMASH were binned in a separate class.

BGC DISTANCE METRIC AND SIMILARITY NETWORK

BGC similarity networks were calculated using a modified version of the distance metric from Lin and coworkers (Lin et al., 2006) for multi-domain proteins. The modified version consists of two different indices: the Jaccard index (which measures the similarity in Pfam domain sets from two BGCs) and the domain duplication index, with weights of 0.36, and 0.64, respectively. The Goodman-Kruskal γ index, which was included in the original similarity metric with a low weight of 0.01, was omitted, since the conservation of the order between two sets of domains does not appear to have an important effect on the structure of the small molecule product, except in the case of NRPS and PKS gene clusters (Fischbach et al., 2008). BGC families were calculated with a Lin similarity threshold of 0.5 and MCL clustering with I = 2.0. The similarity network was obtained using the same Lin similarity threshold and visualized using Cytoscape (Smoot et al., 2011).

BIOINFORMATIC ANALYSIS OF APE GENE CLUSTERS

Expansion of the APE BGC family was performed using manual parsing of MultiGeneBlast (Medema et al., 2013) architecture search results (with the *E. coli*, *V. fischeri*, *X. campestris* and *F. johnsonii* APE gene clusters as query) against GenBank version 197 (08/2013), with a 20% sequence identity cut-off and 2000 blastp hits mapped per query sequence. APE Clusters of Orthologous Groups (COGs) were obtained using OrthoMCL (Li et al., 2003) (MCL I = 1.5, sequence identity cutoff 20%), and were used to construct a cladogram with hierarchical clustering using the Lin modified distance metric. Structure-guided multiple sequence alignments of APE A and KS domains were performed using PROMALS3D (Pei et al., 2008), and phylogenetic trees were inferred with MEGA5 (Tamura et al., 2011) using the Maximum Likelihood method.

CONSTRUCTION OF THE V. FISCHERI ES114 APE-CLUSTER DELETION MUTANT

Oligonucleotide primers, plasmids and bacterial strains used and generated in this study are summarized in **SI Tables VI-VIII**. A deletion construct was generated by fusing the ~1 kb up- and downstream regions of the *V. fischeri* cluster into a counterselectable suicide plasmid backbone using circular polymerase extension cloning (CPEC; (Quan and Tian, 2011)). This construct was introduced into *V. fischeri* ES114 by tri-parental mating and integrants were identified by selection for kanamycin resistance. Second homologous recombination events were enriched by non-selective growth, followed by induction of the counterselectable marker to identify cells that had lost the integrated plasmid backbone. Successful deletion mutants were separated from revertants and verified by colony PCR and sequencing.

HETEROLOGOUS EXPRESSION OF APE GENE CLUSTERS

The *E. coli* CFT073 and *V. fischeri* ES114 APE clusters were amplified by PCR in three parts from genomic DNA and assembled into the SuperCos I vector backbone using either the CPEC (Quan and Tian, 2011) or Gibson (Gibson et al., 2009) method. The *V. fischeri* APE cluster was further modified by introducing an apramycin-resistant cassette containing the *ermE** promoter upstream of the operon starting with VF0844 using PCR targeting (Gust et al., 2004). Correct insertion of *ermE**p was verified by sequencing. The heterologous expression constructs for the *E. coli* CFT073 and *V. fischeri* APE clusters were introduced into chemically competent *E. coli* Top10 yielding strains JC087 and JC090, respectively.

APE COMPOUND PURIFICATION

For large-scale isolation and purification of APE_{EC} and APE_{VF}, all steps were performed in a way that avoided exposure to light. Cells were harvested from 32 L of *E. coli* JC087 and 80 L of *V. fischeri* ES114 liquid cultures, respectively. Following lyophilization, the cell material was extracted four times with 1:2 methanol/dichloromethane and the extracts were concentrated, resuspended in 1:2 methanol/dichloromethane and subjected to mild saponification with 0.5 M potassium hydroxide for 1 hour. The mixture was neutralized and the organic layer was collected, washed, dried, and resuspended in acetone for further purification by a two-step RP-

HPLC method. For both extracts, the peaks with absorbance at 441 nm were collected, dried under vacuum and stored at -20 °C in an amber vial prior to structural analysis.

APE STRUCTURAL CHARACTERIZATION

Purified APE methyl esters were analyzed by a combination of high-resolution uPLC-ESI-TOF mass spectrometry and 1D and 2D-NMR experiments, enabling the determination of their molecular formula: $C_{21}H_{22}O_3$ for APE_{EC} ($[\text{M-H}]^-$ adduct at 321.1496 *m/z* (Δppm = -0.310)) and $C_{22}H_{24}O_3$ for APE_{VF} ($[\text{M-H}]^-$ adduct at 335.1652 *m/z* (Δppm = 0.0)). Analysis of the ^{1}H-NMR, COSY, HSQC, HMBC, ROESY and TOCSY spectra of APE_{VF} in D_6 DMSO and APE_{EC} in D_6 acetone enabled the determination of their solution structure (**Figure 3b**). This procedure is described in detail in the **SI Text** section and shown in **Figure S7c-d**.

Author contributions

P.C., M.H.M., J. Claesen, K.K., R.G.L. and M.A.F. designed the research, analyzed the data and wrote the paper, with substantial input from E.T., J. Clardy, and A.S. P.C. and M.H.M. performed the computational research. J. Claesen and K.K. performed the experimental research. K.M. and A.P. provided input data and data integration into the JGI-IMG database. L.C.W.B., P.A.G., M.K., B.W.B., and M.A.F. designed an earlier version of the gene cluster identification algorithm that served as a model for the current version.

Acknowledgments

We thank members of the Fischbach lab for helpful discussions, and an anonymous reviewer for constructive feedback on the manuscript. We thank Edward Ruby (University of Wisconsin) for providing us with *V. fischeri* ES114, Didier Mazel (Institut Pasteur) for plasmid pSW8197, and Mervyn Bibb (John Innes Centre) for plasmids pIJ773, pIJ790 and pIJ10257. This work was supported by an HHMI Predoctoral Fellowship (PC), a Boehringer Ingelheim Fonds travel grant (MHM), Grant 10463 from the GenBiotics programme of the Dutch Technology Foundation STW to ET (MHM), an NWO-Vidi fellowship (RB), NIH grant TW006634 (RGL), the James and Eleanor Delfino Charitable Trust (KK), a Medical Research Program Grant from the W.M. Keck Foundation (MAF), a Fellowship for Science and Engineering from the David and Lucile Packard Foundation (MAF), DARPA award HR0011-12-C-0067 (MAF), and NIH grants OD007290, AI101018, AI101722 and GM081879 (MAF). This project has been funded in part with Federal funds from the National Institute of Allergy and Infectious Diseases, National Institutes of Health, Department of Health and Human Services, under Contract No.: HHSN272200900018C. M.A.F. is on the scientific advisory board of Warp Drive Bio.

Figures

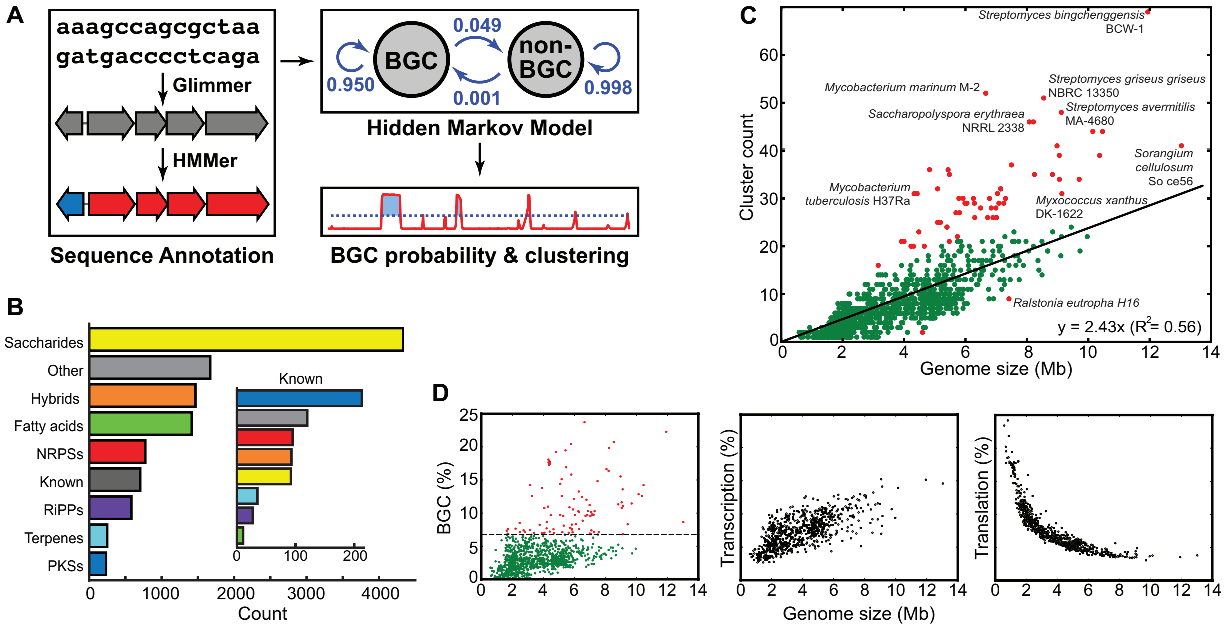


Figure 1. ClusterFinder flowchart and distribution of BGC classes and counts. a, Flowchart of the four-step BGC prediction pipeline: (*i*) annotation of a genome sequence and compression to a string of Pfam domains, (*ii*) calculation of posterior probabilities of a BGC hidden state, (*iii*) clustering of genes that contain Pfam domain(s) with posterior probabilities of BGC hidden state above the threshold, and (*iv*) annotation of the predicted BGCs using an expanded version of the antiSMASH algorithm. **b,** Distribution of BGC classes for known (inset) and predicted BGCs. "Other" gene clusters include gene clusters from other known classes as well as a manually curated set of 1,024 putative gene clusters that fall outside known biosynthetic classes. Unexpectedly, 40% of all predicted BGCs encode saccharides, more than twice the size of the next largest class. **c,** Number of predicted BGCs by genome size. Most bacterial species follow a linear trend (the equation in the bottom-right corner); outliers (defined as having residuals >8) are colored red. **d,** The proportions of bacterial genomes devoted to secondary metabolite biosynthesis (left panel; 6.7% of species that devote >7.5% of their

genome to biosynthesis are marked red), transcription (middle panel), and translation (right panel).

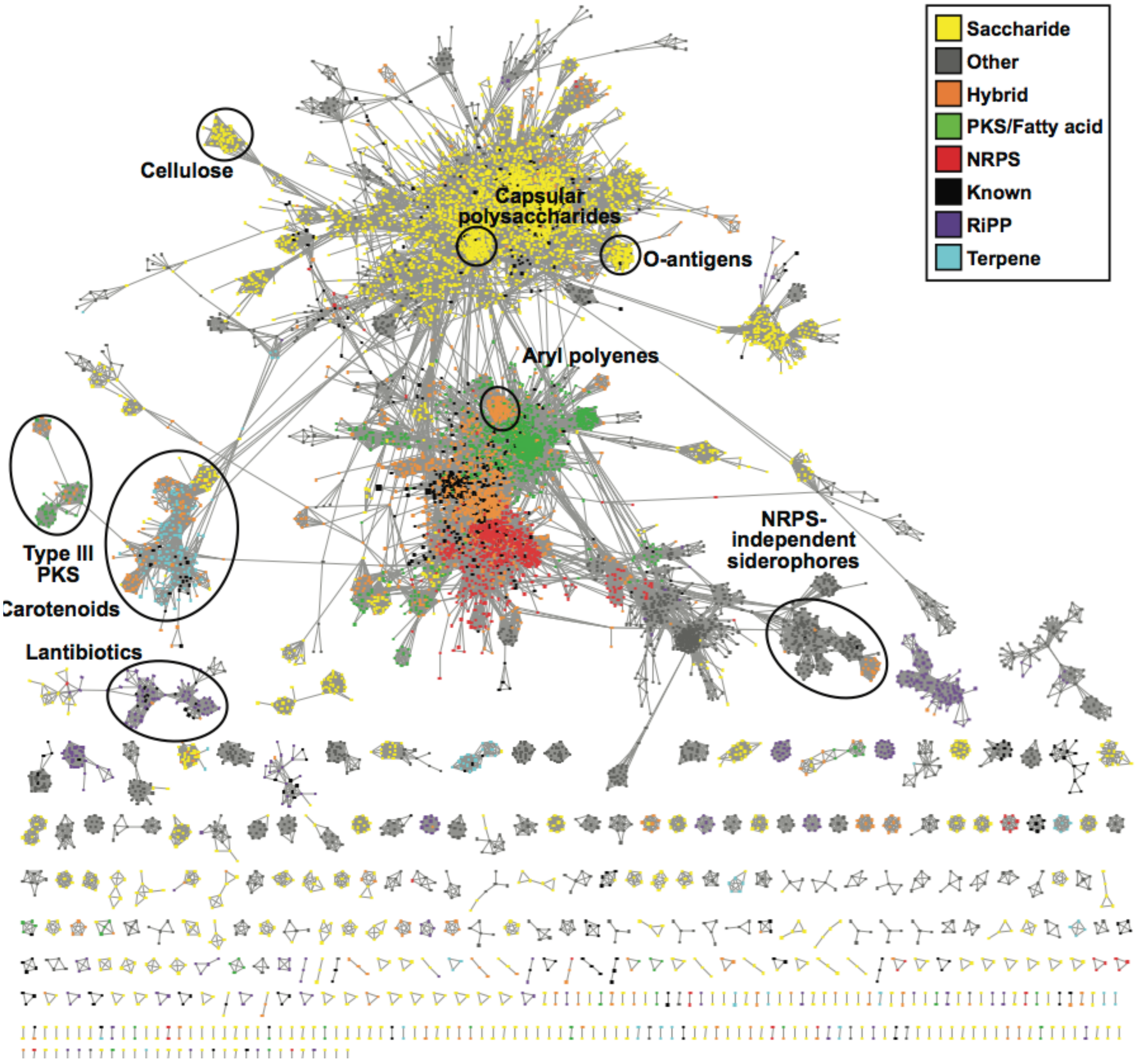


Figure 2. A systematic analysis of bacterial BGCs. Similarity network of known and putative BGCs, with the BGC similarity metric threshold at 0.5. The topology of the network is robust to changes in the distance threshold, as described in the **Experimental Procedures**. One connected component harbors most of the gene clusters (72%), and is largely composed of two linked subgraphs: one dominated by oligosaccharides and the other a mixture of nonribosomal peptides (NRPs) and polyketides/lipids, indicating that BGCs from these classes share a significant number of gene families with one another. Smaller BGC families with more unique

compositions are represented at the bottom of the figure; only 812 BGCs (7.6%) do not have any connections with other BGCs at the chosen cutoff. A selection of node clusters within the network has been highlighted to show how gene cluster families form cliques within the network. The highlighted groups include widely distributed gene cluster families for O-antigens, capsular polysaccharides, carotenoids, and NRPS-independent siderophores, along with one of the lantibiotic BGC families and an unknown family of BGCs with type III polyketide synthases. The aryl polyene family that we characterized further in this study is shown in the middle of the network.

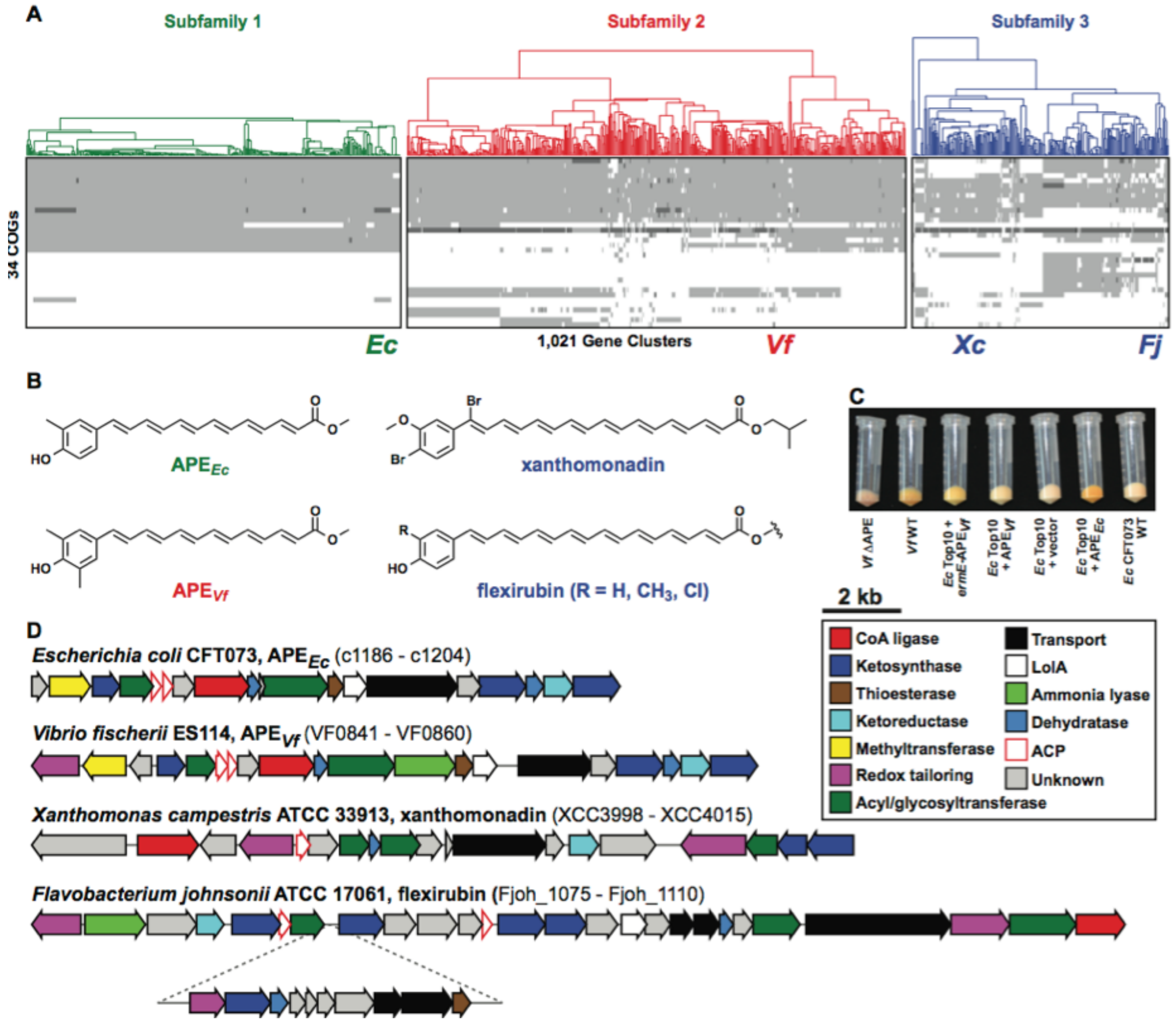


Figure 3. APE gene clusters comprise the largest known BGC family. a, Heat map and dendrogram of all 1,021 detected APE family gene clusters, based on Clusters of Orthologous Groups generated by OrthoMCL (Li et al., 2003) using our adapted version of the Lin distance metric (Lin et al., 2006) that includes sequence similarity. Light grey indicates the presence of one gene from a COG, whereas darker grey tones indicate the presence of two or three genes from a COG. The two BGC subfamilies that functioned as the starting point of our analysis (subfamilies 1 and 2) are shown in green and red, respectively, while the smaller BGC subfamily that includes the xanthomonadin and flexirubin gene clusters (subfamily 3) is shown in blue. The

positions of the two experimentally targeted gene clusters (*Ec* for *Escherichia coli* CFT073 and *Vf* for *Vibrio fischeri* ES114) as well as the *Xanthomonas campestris* ATCC 33913 xanthomonadin (*Xc*) and *Flavobacterium johnsonii* ATCC 17061 flexirubin (*Fj*) gene clusters are indicated below the heat map. See **Figure S5** for a version with more detailed annotations. **b,** Chemical structures obtained for the APE compounds from *E. coli* and *V. fischeri*, and the previously determined chemical structures of xanthomonadin and flexirubin. Note the difference in polyene acyl chain length as well as the distinct tailoring patterns on the aryl head groups. **c**, Bacterial pellets from strains harboring APE gene clusters showing the pigmentation conferred by aryl polyenes. **d,** Genetic architecture of the four characterized aryl polyene gene clusters. The inset in the *Flavobacterium johnsonii* flexirubin gene cluster is a sub-cluster putatively involved in the biosynthesis of dialkylresorcinol (Fuchs et al., 2013), which is acylated to an APE to form flexirubin. See **SI Data File I** for schematics of all 1,021 APE gene clusters from panel **A**.

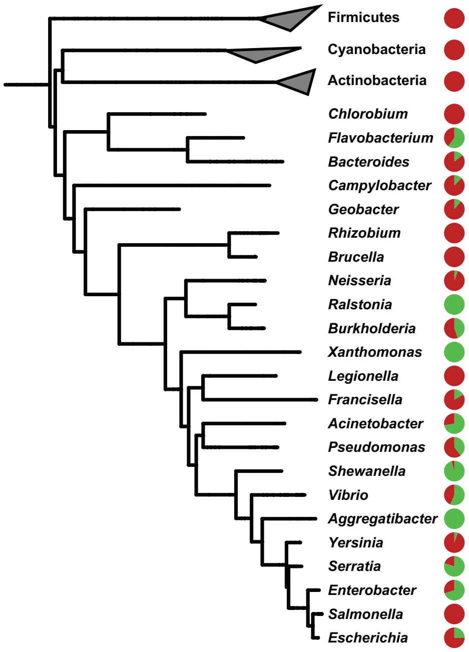


Figure 4. APE gene clusters are widely but discontinuously distributed among Gram-negative bacteria. Presence/absence pattern of APE gene clusters across all complete genomes from selected bacterial genera, mapped onto the PhyloPhLan high-resolution phylogenetic tree (Segata et al., 2013). For each genus, the pie chart represents the percentage of sequenced genomes in which APE gene clusters are present (green) or absent (red). BGCs from the APE family occur throughout all subphyla of the Proteobacteria, as well as in a range of genera from the CFB group. The discontinuous presence/absence pattern suggests that gene cluster gain and/or loss has frequently occurred during evolution. A presence/absence mapping on all the genomes from our initial JGI dataset is provided in **Figure S3**.

Supplementary figures

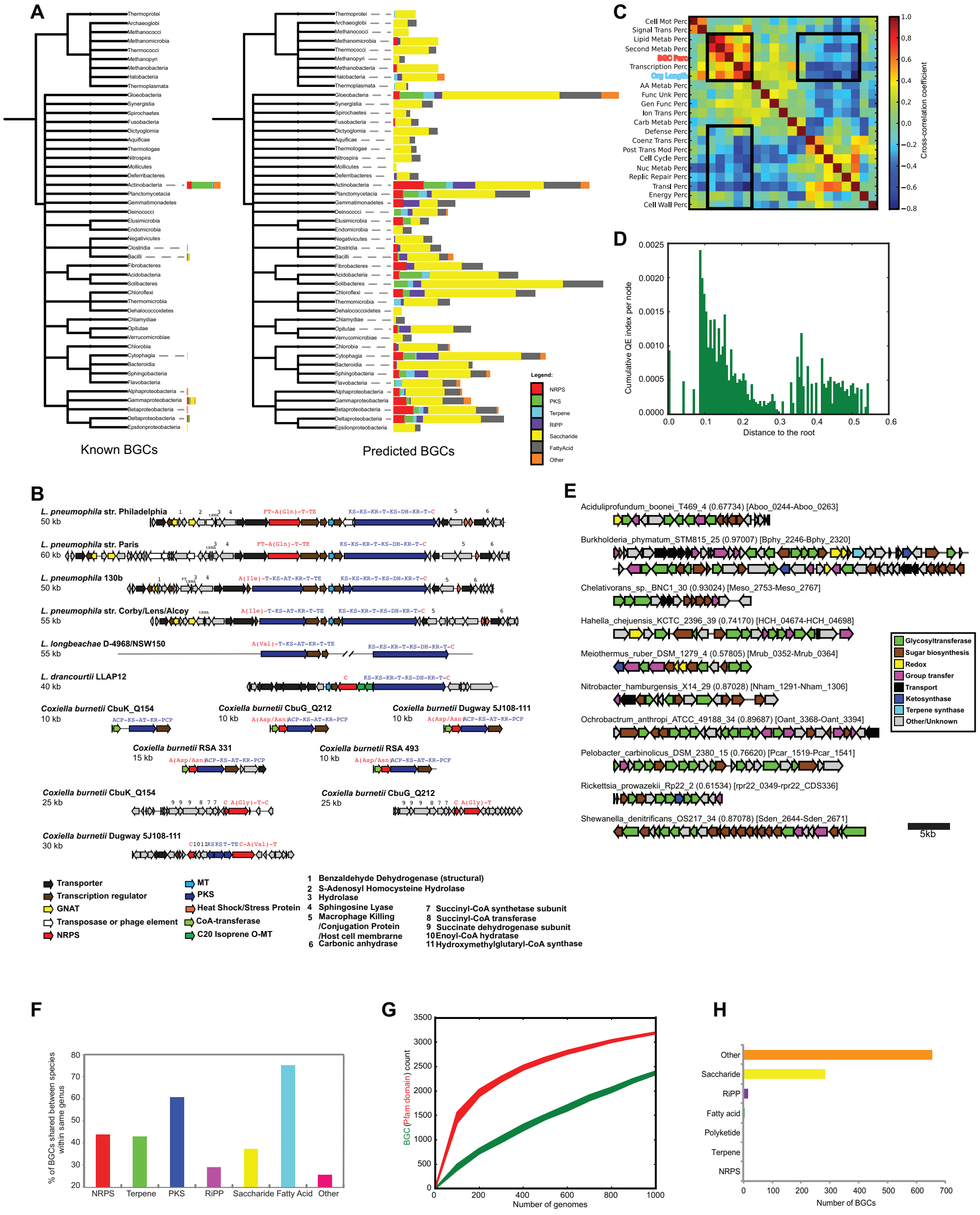


Figure S1. **Global phylogenomic analysis of prokaryotic BGCs, Related to Figure 1. a,** The prokaryotic tree of life is mostly unexplored for BGCs. The phylogenetic tree of bacterial and archaeal classes (as stored in NCBI Taxonomy) shows the distribution of known (left) and predicted BGCs (right). A strong historical bias can be observed: some bacterial classes (such as Actinobacteria) have been heavily studied, whereas other classes with (on average) similarly large numbers of BGCs have been largely neglected. The two graphs are not scaled equally; the left bar plot shows the total number of known BGCs per class, whereas the bar plot on the right displays the average number of predicted BGCs per strain within a class. **b**, Examples of notable PKS and NRPS biosynthetic gene clusters detected in the genomes of the obligate intracellular pathogens *Legionella* and *Coxiella*. Letters above the PKS and NRPS genes signify domain structure, with adenylation domain substrates as predicted by NRPSPredictor2 (Röttig et al., 2011) in brackets. **c**, Cross-correlation matrix of COG protein functions in bacterial genomes. Although we focused on analyzing the association between the number of BGCs (or percentage of the genomes they occupy) and genome lengths (**Figure 1c**), we also investigated whether there are any other COG functions that correlate with genome length. Primary and secondary metabolism, as well as transcription regulation, are linked to genome length, suggesting that genomes become longer by incorporation of biosynthetic and regulatory genes. In contrast, COG functions such as translation, cell cycle regulation, RNA replication and repair, nucleotide metabolism and transport, post-translational modification, protein turnover, and chaperone functions do not seem to be linked to genome length. **d**, Histogram of cumulative quantitative entropy (QE) index with respect to the distance from the root of the phylogenetic tree. A decreasing trend in this histogram suggests decreasing diversification rates on a global evolutionary time-scale. However, a presence of nodes of high diversity closer to the leaves points to recent evolution of BGC repertoires. Each bar plots a sum of QE indices of all nodes within a given bar's limits with respect to the root of the phylogenetic tree. **e**, Examples of previously unknown saccharide gene clusters. The saccharide gene clusters are from unexplored or underexplored genera. Colors represent functions of the genes, as indicated in the figure legend. **f**, Type diversity of BGCs within the same taxonomic genera. The bar graph shows the percentage of gene clusters per class that is shared between two genomes randomly sampled from the same genus. While fatty acid biosynthesis gene clusters are often similar in species of the same genus, RiPP and saccharide BGC repertoires are often radically different between species within the same genus. **g**, Rarefaction analysis of numbers of BGC families (red) and Pfam families (green). BGC families (or “BGC clusters”) were calculated from the BGC similarity network with a similarity threshold of 0.5 and MCL clustering with I = 2.0. For a given number of genomes, a random sample of organisms was selected 20 times (the thickness of the lines denote 68% confidence intervals based on these 20 bootstraps). **h**, Identification and classification of BGCs in 201 single-cell genomes from uncultivated organisms. Functional classification of the 947 BGCs identified in the set of 201 single-cell genomes from JGI (Rinke et al., 2013), using the same antiSMASH-based classification scheme used for the dataset of full genomes from JGI. Besides a significant number of saccharide-encoding gene clusters, the vast majority of putative BGCs falls outside known biosynthetic classes.

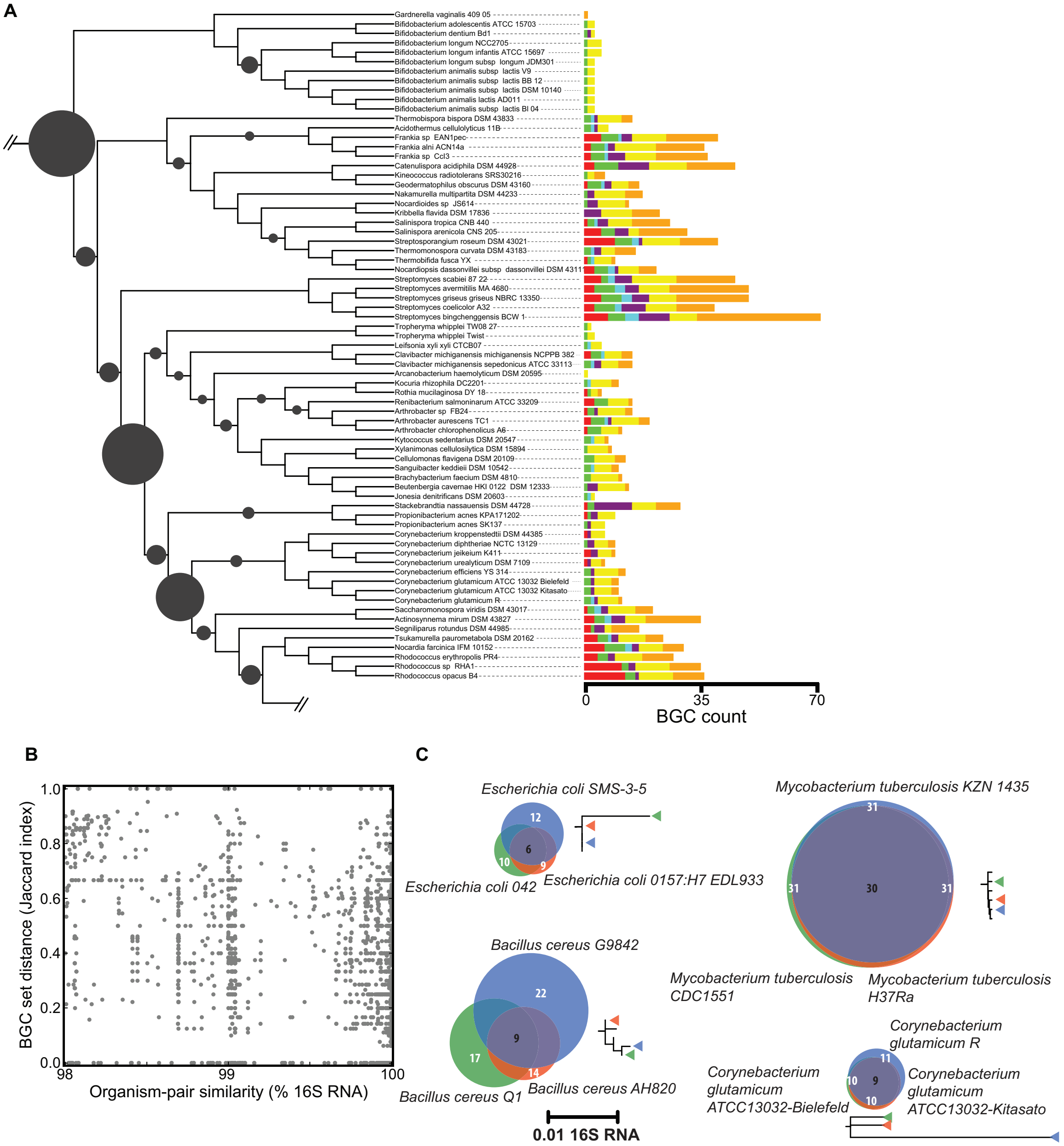


Figure S2. Diversity of BGCs is independent from the phylogeny, Related to Figure 1. a, The decomposition of BGC diversity among species of the phylum Actinobacteria. The diversity of each node in the phylogenetic tree is measured by the quadratic entropy index, and represented by the size of the circle (larger circle defines higher degree of diversity). Color bars at the leaf tips represent number of BGCs per species, with different colors denoting different BGC types (colors as in **Figure 1b**). Hybrid gene clusters (orange) are unusually prominent in

Actinobacteria (~50%). For the entire phylogenetic tree, see **Figure S3**. **b**, The scatter plot shows no correlation between phylogenetic and BGC content distance for a given organism pair. **c,** The Venn diagrams show the number of BGCs shared among three different sets of closely related species. The phylogenetic tree sections to the right of the Venn diagrams are shown using the same scale.

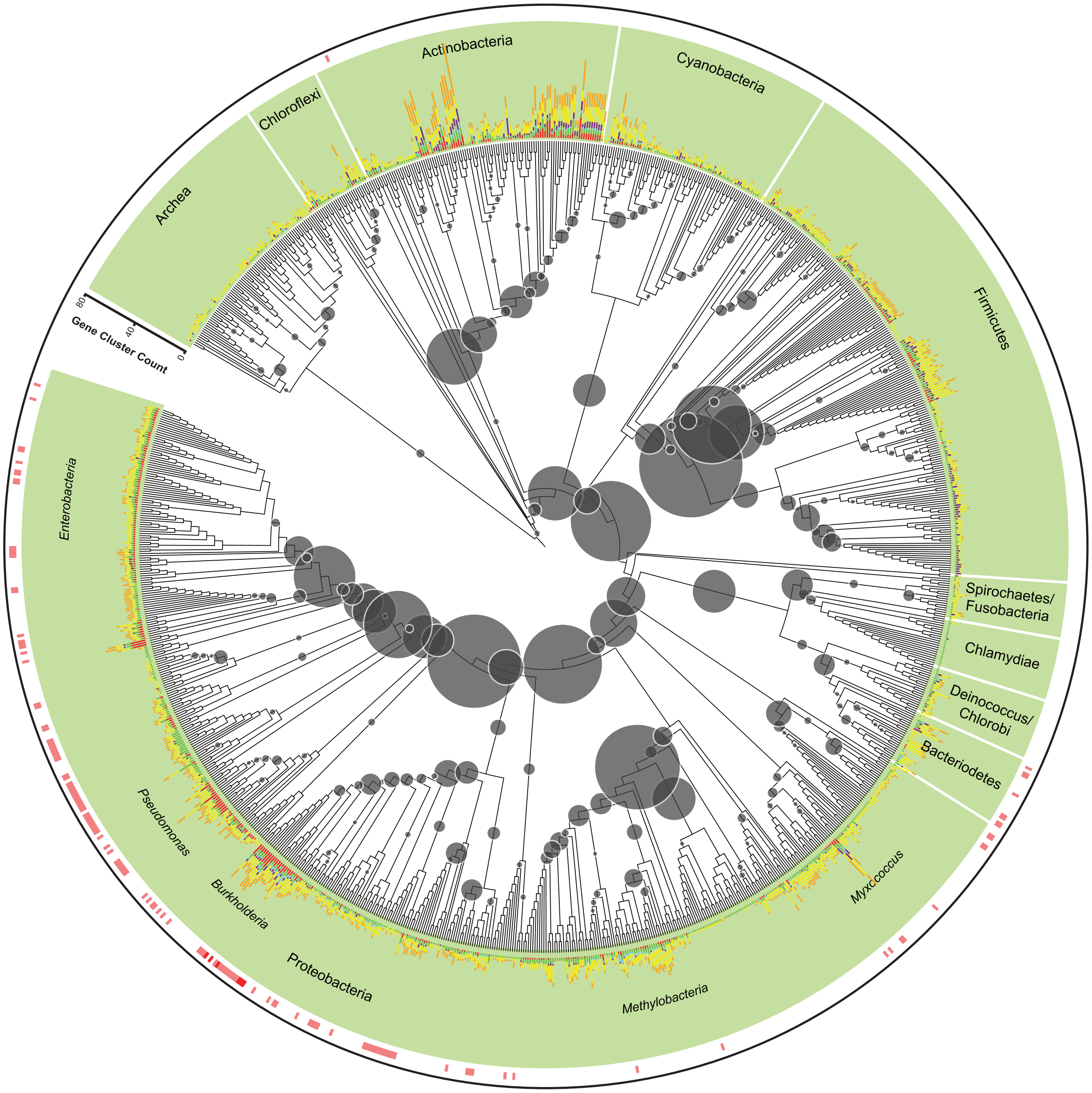
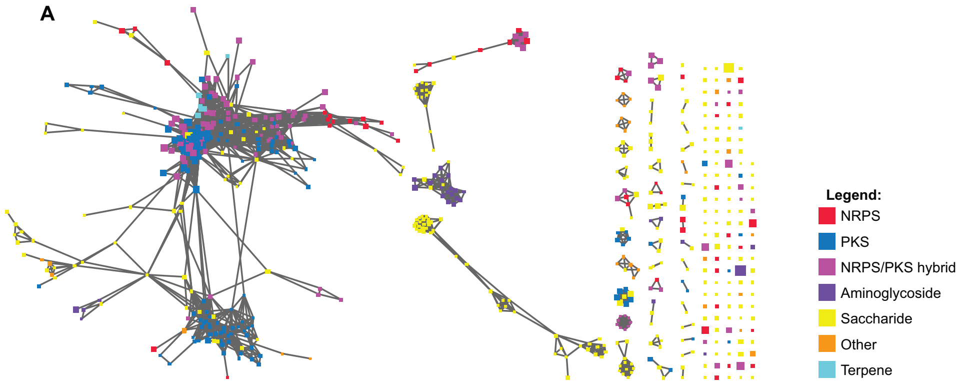


Figure S3. Decomposition of BGC diversity among all sequenced prokaryotic genomes, Related to Figure 1. The diversity of each node in the phylogenetic tree is measured by the quadratic entropy index, and represented by the size of the circle (larger circle defines higher degree of diversity). Color bars at the leaf tips represent the number of BGCs per species, with different colors denoting different BGC types (colors as in **Figure 1b**). The outer ring shows the absence/presence of APE gene clusters in our initial set of 1154 genomes obtained from JGI-IMG. The discontinuous pattern of APE gene cluster conservation suggests frequent horizontal

gene transfer and/or gene cluster loss. Pink indicates the presence of one APE gene cluster in a genome, red indicates the presence of two gene clusters in a genome. Several genomes from *Burkholderia* and *Ralstonia* have two different APE gene clusters located on two different chromosomes. The tree was generated using iTOL (Letunic and Bork, 2007).



B

Statistics for the graph with >0.6 threshold

	#nodes	#edges	γ	L	C	L_{random}	C_{random}	K(k)	p-value $_{K(k)}$
ALL	7,391	136,483	1.66 ± 0.07	1.11	0.69	2.83	0.005	-0.012 ± 0.009	0.45
PKS	1,344	94,357	1.03 ± 0.07	1.11	0.78	1.90	0.100	-0.017 ± 0.010	0.39
Terpene	137	417	0.70 ± 0.49	1.12	0.54	2.90	0.050	0.071 ± 0.061	0.54
Saccharide	2,588	18,896	1.78 ± 0.21	1.1	0.66	3.21	0.006	-0.035 ± 0.032	0.39
RP	290	1,414	0.83 ± 0.29	1.11	0.72	2.73	0.038	-0.131 ± 0.041	0.24
Siderophore	200	1,213	0.47 ± 0.36	1.43	0.79	2.39	0.060	-0.147 ± 0.054	0.19
Hybrid	694	5,525	0.92 ± 0.19	1.13	0.73	2.66	0.023	-0.018 ± 0.016	0.49
NRPS	524	4,431	1.43 ± 0.21	1.16	0.7	2.53	0.030	0.016 ± 0.048	0.74

C

Statistics for the graph with >0.8 threshold

	#nodes	#edges	γ	L	C	L_{random}	C_{random}	K(k)	p-value $_{K(k)}$
ALL	5,152	34,976	2.16 ±0.15	1.07	0.67	3.57	0.0026	-0.028 ± 0.025	0.49
PKS	1,151	18,836	1.52 ±0.14	1.14	0.79	2.36	0.029	-0.005 ± 0.300	0.86
Terpene	97	146	0.03 ±0.99	1.10	0.44	NaN	NaN	0.370 ± 0.460	0.73
Saccharide	1,776	8,199	1.50 ±0.30	1.06	0.64	3.62	0.0056	-0.057 ± 0.030	0.36
RP	221	537	0.73 ±0.71	1.07	0.67	2.29	0.03	-0.044 ± 0.066	0.52
Siderophore	159	609	0.55 ±0.60	1.10	0.67	2.7	0.04	0.120 ± 0.101	0.53
Hybrid	489	1,661	1.03 ±0.41	1.06	0.73	3.45	0.017	-0.130 ± 0.041	0.0034
NRPS	369	2,572	1.07 ±0.30	1.06	0.72	2.53	0.037	0.048 ± 0.240	0.27

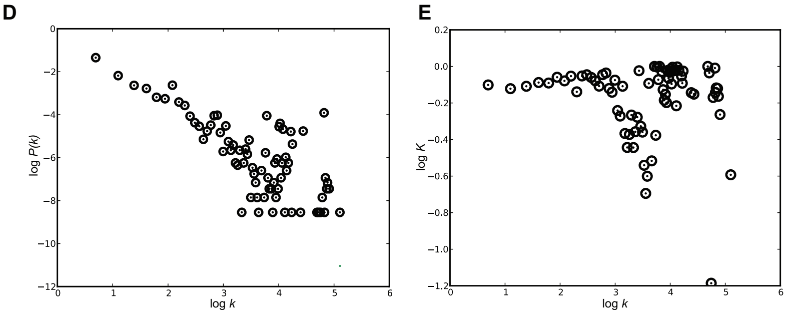


Figure S4. BGC similarity networks, Related to Figure 2. a, Similarity network of known BGCs. The similarities between the BGCs were calculated by taking into account the architecture as well as the sequence similarity features of our distance metric (see Methods for details). This analysis shows that the gene cluster distance metric functions well in separating known families of BGCs, while maintaining links representing known genetic similarities between classes like aminoglycosides and saccharides. Cytoscape(Smoot et al., 2011) was

used to visualize the network. **b**, Analysis of the global BGC similarity network. Network (or graph) topology can be indicative of the relationships among its constituent nodes (here, BGCs). Tables **b** and **c** show different topology parameters for graphs with BGC similarity cutoffs of 0.6 and 0.8, respectively; #nodes indicates the number of nodes in the graph; #edges indicates the number of edges in the graph; gamma equals the exponent of the node degree frequency diagram (the steepness of the linear fit in **d**); L is the average shortest path between any two nodes; C is the average clustering coefficient, Lrand is the average shortest path between any two nodes in the randomized graphs; Crand is the average clustering coefficient in the randomized graphs; and K(k) is coefficient of the linear fit in **e**. The values of the parameters were calculated for all nodes in the graph, as well as for subgraphs of nodes corresponding to individual classes of BGCs. Parameters were calculated using the NetworkX library.

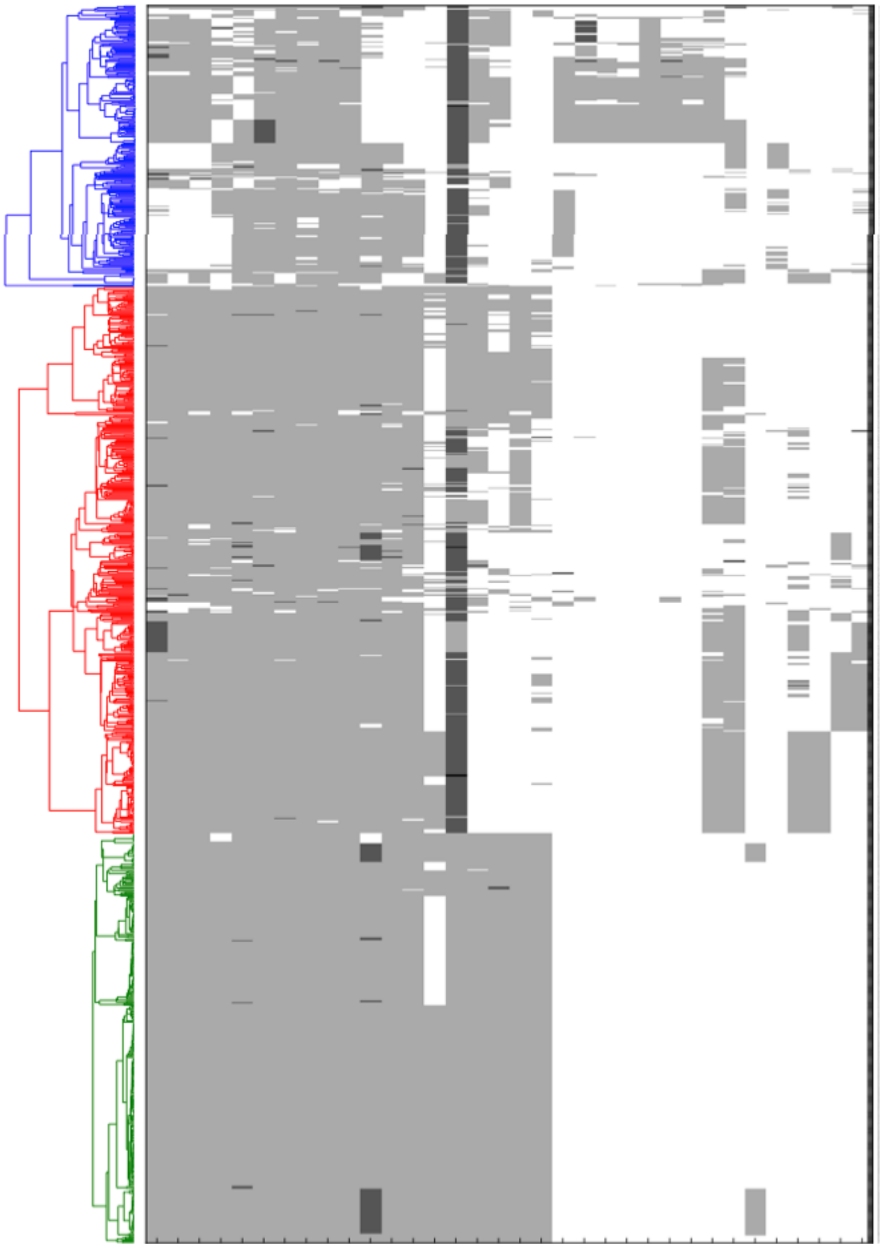


Figure S5. Full annotated APE superfamily clustered heat map including COG annotations, Related to Figure 3. Full version of the clustered heat map shown in Figure 3a. In this version, the COG annotations are shown at the bottom, and the accession number and source strain are shown on the right.

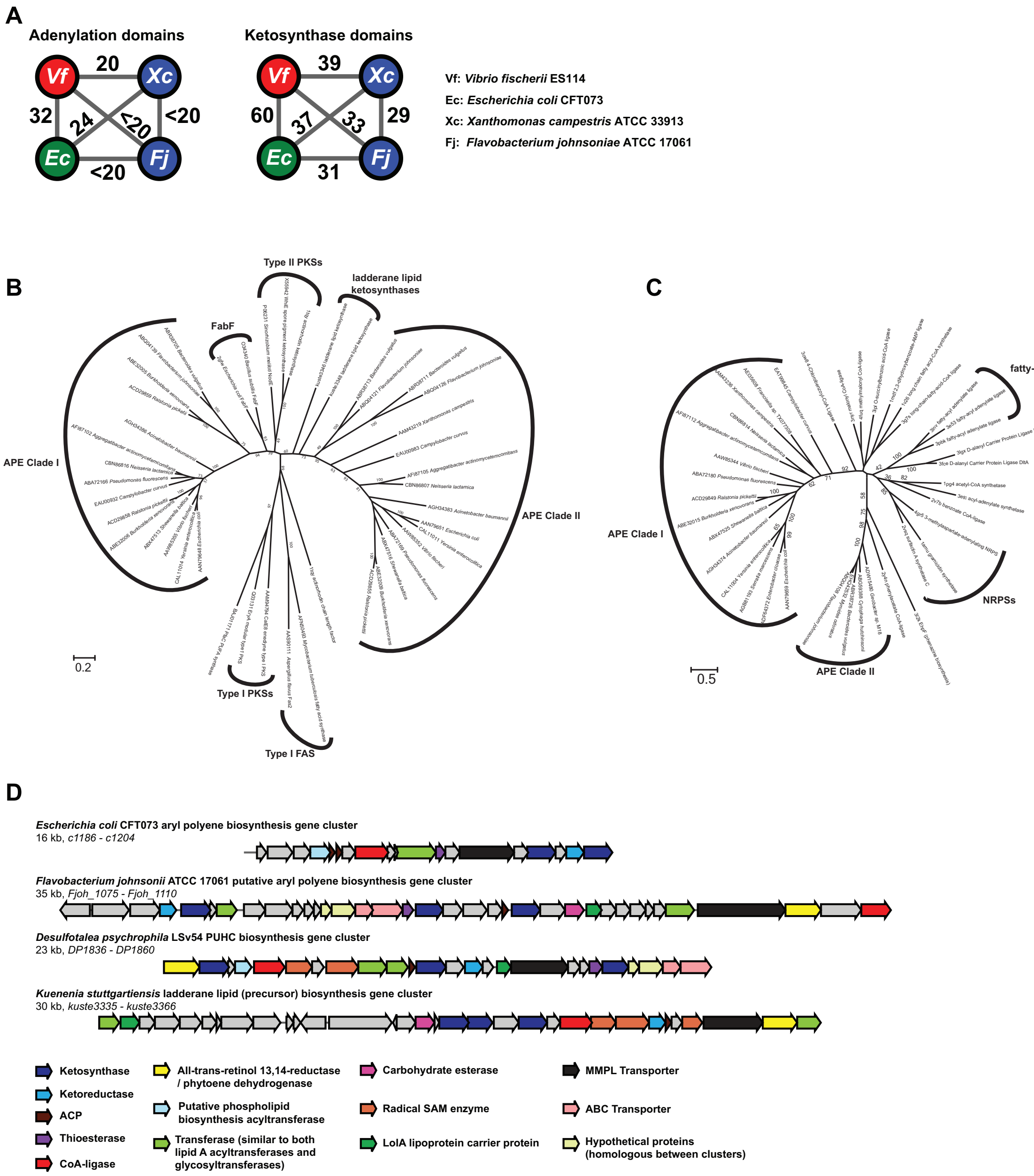

A
Adenylation domains
Vf
Xc
Ec
Fj
20
32
24
<20
<20
<20
Ketosynthase domains
Vf
Xc
Ec
Fj
39
60
37
33
29
31
Vf: Vibrio fischerii ES114
Ec: Escherichia coli CFT073
Xc: Xanthomonas campestris ATCC 33913
Fj: Flavobacterium johnsoniae ATCC 17061
B
Type II PKSs
ladderane lipid ketosynthases
FabF
APE Clade I
APE Clade II
Type I PKSs
Type I FAS
0.2
C
fatty-acyl adenylate ligases
APE Clade I
NRPSs
APE Clade II
0.5
D
Escherichia coli CFT073 aryl polyene biosynthesis gene cluster
16 kb, c1186 - c1204
Flavobacterium johnsonii ATCC 17061 putative aryl polyene biosynthesis gene cluster
35 kb, Fjoh_1075 - Fjoh_1110
Desulfotalea psychrophila LSv54 PUHC biosynthesis gene cluster
23 kb, DP1836 - DP1860
Kuenenia stuttgartiensis ladderane lipid (precursor) biosynthesis gene cluster
30 kb, kuste3335 - kuste3366
Ketosynthase
Ketoreductase
ACP
Thioesterase
CoA-ligase
All-trans-retinol 13,14-reductase / phytoene dehydrogenase
Putative phospholipid biosynthesis acyltransferase
Transferase (similar to both lipid A acyltransferases and glycosyltransferases)
Carbohydrate esterase
Radical SAM enzyme
LolA lipoprotein carrier protein
MMPL Transporter
ABC Transporter
Hypothetical proteins (homologous between clusters)

Figure S6. Phylogenetic analysis of APE gene clusters and key biosynthetic enzymes, Related to Figure 4. a, Pairwise sequence identities of the ketosynthase and adenylation domains in the four characterized gene clusters. The numbers in the graph represent the average percentage identity between the amino acid sequences of the pairs of most closely related adenylation / ketosynthase enzymes in the four gene clusters, as inferred from the structure-guided sequence sequence alignment. Three pairs of adenylation enzymes whose

amino acid sequences are only 12% identical are shown as <20% identical, to account for the inexactness of sequence identity calculations for such distant relationships. **b**, Phylogenetic tree of APE ketosynthase domains with other ketosynthases. The maximum likelihood phylogenetic tree, based on a structure-guided multiple sequence alignment using PROMALS3D (Pei et al., 2008), shows that the ketosynthases from representative APE gene clusters belong to two evolutionary clades. One clade is most closely related to FabF proteins from *Escherichia coli* and *Bacillus subtilis*, while the other clade is most closely related to ketosynthases putatively involved in ladderane lipid biosynthesis in the anammox bacterium *Kuenenia stuttgartiensis*. The gene clusters from *Bacteroides* and *Flavobacterium* contain a duplicate of the ketosynthase from the latter clade, while the xanthomonadin gene cluster from *Xanthomonas campestris* contains no ketosynthase from the first clade. **c**, Phylogenetic tree of APE adenylation domains with other adenylation enzymes. The maximum likelihood phylogenetic tree, based on a structure-guided multiple sequence alignment using PROMALS3D (Pei et al., 2008), shows that the adenylation enzymes involved in APE biosynthesis cluster in two uncharacterized clades within the ANL superfamily that includes Acyl-CoA synthetases, NRPS adenylation domains, and Luciferase enzymes. Most closely related are two adenylation enzymes that are involved in the ligation of two different aryl group-containing compounds, suggesting that convergent evolution may have lead to the independent evolution of two mechanisms to attach an aryl group to the polyene that is synthesized by the same clades of ketosynthases. **d**, Comparison of APE gene clusters with related BGCs. Alignment of the two APE superfamily gene clusters from *Escherichia coli* CFT073 and *Flavobacterium johnsonii* ATCC 17061, the putative ladderane lipid biosynthesis gene cluster from *Kuenenia stuttgartiensis* and the polyunsaturated hydrocarbon biosynthesis gene cluster from *Desulfotalea psychrophila* LSv54. Colors signify homologous genes based on a MultiGeneBlast comparison with the blastp algorithm.

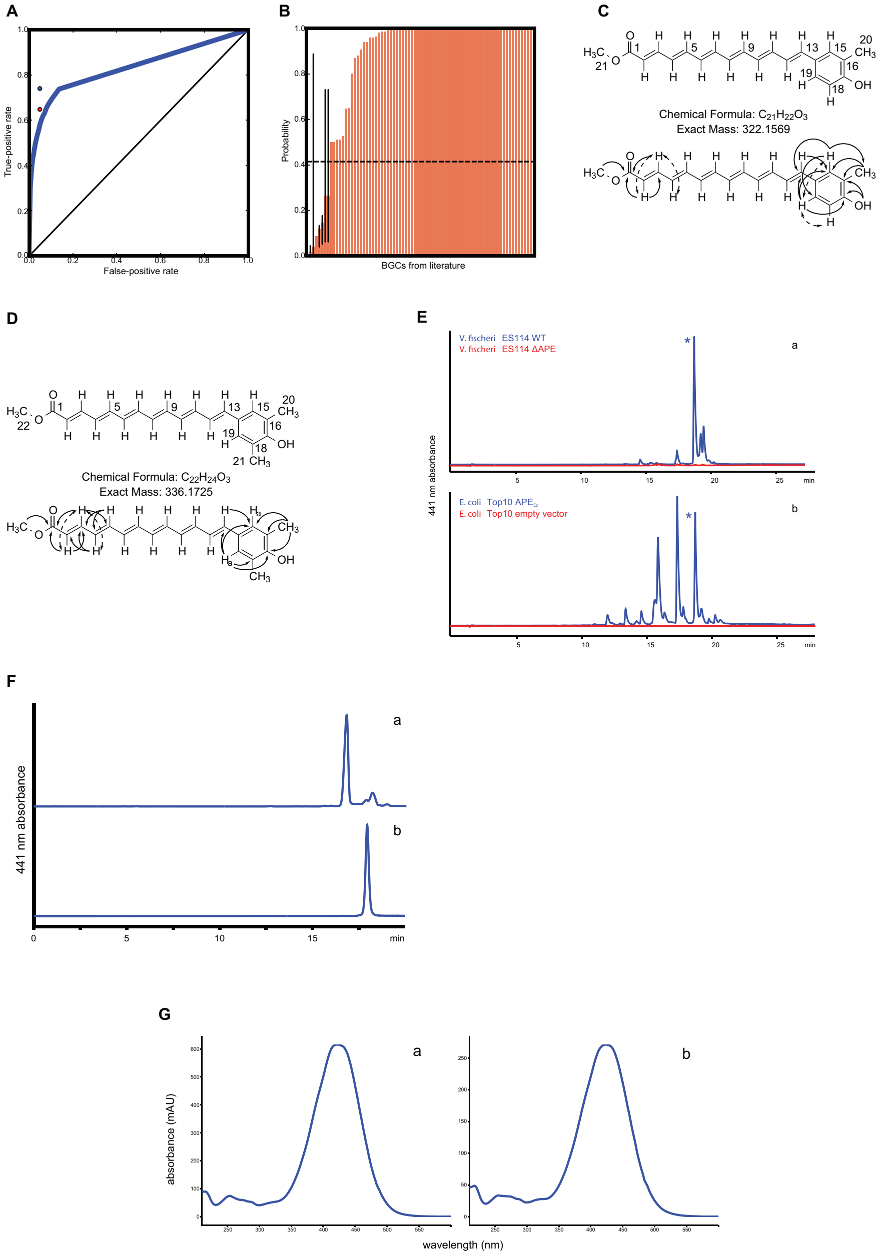


Figure S7. Evaluation of the ClusterFinder algorithm and APE structural characterization, Related to Figure 3. A, The performance of the ClusterFinder algorithm was evaluated by calculating the ROC and AUC using 10 manually annotated genomes (SI Table VII) that were not used in the training of the algorithm. We obtained an AUC of 0.84, which is significantly better than the AUC of a random prediction (AUC of 0.5). The predictions were assessed on protein domain basis; for example, at each probability threshold, a given protein domain was

assigned to the true-positive class if the probability of being in a BGC was higher than the threshold, and if it was manually annotated as being part of a BGC. **B,** We assessed the true-positive rate on a set of 74 BGCs from the literature (SI Table VIII). Only 7 BGCs (9.5%) did not pass our probability threshold of 0.4. **C**, Structure of APE$_{Ec}$ with COSY (dashed lines) and HMBC (solid lines) correlations. **D**, Structure of APE$_{Vf}$ with COSY (dashed lines) and HMBC (solid lines) correlations. **E**, HPLC traces for crude APE extracts. **a)** Overlay of traces for *V. fischeri* ES114 wild type (blue) and the *V. fischeri* ES114 Δ*ape* deletion strain (red). **b)** Overlay of traces for *E. coli* Top10 expressing the CFT073 cluster (blue) and the *E. coli* Top10 control strain containing the empty vector (red). HPLC conditions: gradient of acetonitrile in 0.02% formic acid water: 0% to 30% organic phase in 2 min, 30% to 90% organic phase from 2 min to 22 min, followed by a hold at 90% for 3 minutes and a 3 min wash at 100% organic phase. Detection was at λ = 441 nm. The peak purified and subjected to structural analysis is denoted with an asterisk. **F**, Second RP-HPLC purification for APE$_{Ec}$ **(a)** and APE$_{VF}$ **(b)**. **G**, UV spectrum for APE$_{Ec}$ **(a)** and APE$_{VF}$ **(b)**.

Supplementary text and tables

The manuscript is currently in press, and the material will become available online upon its publication at http://www.cell.com.

References

Arnison, P.G., Bibb, M.J., Bierbaum, G., Bowers, A.A., Bugni, T.S., Bulaj, G., Camarero, J.A., Campopiano, D.J., Challis, G.L., Clardy, J.*, et al.* (2013). Ribosomally synthesized and post-translationally modified peptide natural products: overview and recommendations for a universal nomenclature. Nat Prod Rep *30*, 108-160.

Barbe, V., Vallenet, D., Fonknechten, N., Kreimeyer, A., Oztas, S., Labarre, L., Cruveiller, S., Robert, C., Duprat, S., Wincker, P.*, et al.* (2004). Unique features revealed by the genome sequence of *Acinetobacter* sp. ADP1, a versatile and naturally transformation competent bacterium. Nucleic Acids Res *32*, 5766-5779.

Bergmann, S., Schumann, J., Scherlach, K., Lange, C., Brakhage, A.A., and Hertweck, C. (2007). Genomics-driven discovery of PKS-NRPS hybrid metabolites from *Aspergillus nidulans*. Nat Chem Biol *3*, 213-217.

Blin, K., Medema, M.H., Kazempour, D., Fischbach, M.A., Breitling, R., Takano, E., and Weber, T. (2013). antiSMASH 2.0--a versatile platform for genome mining of secondary metabolite producers. Nucleic Acids Res *41*, W204-212.

Challis, G.L. (2005). A widely distributed bacterial pathway for siderophore biosynthesis independent of nonribosomal peptide synthetases. Chembiochem *6*, 601-611.

Challis, G.L. (2008). Genome mining for novel natural product discovery. J Med Chem *51*, 2618-2628.

Clauditz, A., Resch, A., Wieland, K.P., Peschel, A., and Gotz, F. (2006). Staphyloxanthin plays a role in the fitness of *Staphylococcus aureus* and its ability to cope with oxidative stress. Infect Immun *74*, 4950-4953.

Dong, T., and Schellhorn, H.E. (2009). Global effect of RpoS on gene expression in pathogenic *Escherichia coli* O157:H7 strain EDL933. BMC Genomics *10*, 349.

Fautz, E., and Reichenbach, H. (1979). Biosynthesis of flexirubin: Incorporation of precursors by the bacterium *Flexibacter elegans*. Phytochemistry *18*, 957-959.

Fischbach, M.A., Walsh, C.T., and Clardy, J. (2008). The evolution of gene collectives: How natural selection drives chemical innovation. Proc Natl Acad Sci USA *105*, 4601-4608.

Flatt, P.M., and Mahmud, T. (2007). Biosynthesis of aminocyclitol-aminoglycoside antibiotics and related compounds. Nat Prod Rep *24*, 358-392.

Franke, J., Ishida, K., and Hertweck, C. (2012). Genomics-driven discovery of burkholderic acid, a noncanonical, cryptic polyketide from human pathogenic *Burkholderia* species. Angew Chem Int Ed Engl *51*, 11611-11615.

Frasch, H.J., Medema, M.H., Takano, E., and Breitling, R. (2013). Design-based re-engineering of biosynthetic gene clusters: plug-and-play in practice. Curr Opin Biotechnol *24*, 1144-1150.

Freeman, M.F., Gurgui, C., Helf, M.J., Morinaka, B.I., Uria, A.R., Oldham, N.J., Sahl, H.G., Matsunaga, S., and Piel, J. (2012). Metagenome mining reveals polytheonamides as posttranslationally modified ribosomal peptides. Science *338*, 387-390.

Fuchs, S.W., Bozhuyuk, K.A., Kresovic, D., Grundmann, F., Dill, V., Brachmann, A.O., Waterfield, N.R., and Bode, H.B. (2013). Formation of 1,3-cyclohexanediones and resorcinols catalyzed by a widely occurring ketosynthase. Angew Chem Int Ed Engl *52*, 4108-4112.

Gibson, D.G., Young, L., Chuang, R.Y., Venter, J.C., Hutchison, C.A., 3rd, and Smith, H.O. (2009). Enzymatic assembly of DNA molecules up to several hundred kilobases. Nat Methods *6*, 343-345.

Goel, A.K., Rajagopal, L., Nagesh, N., and Sonti, R.V. (2002). Genetic locus encoding functions involved in biosynthesis and outer membrane localization of xanthomonadin in *Xanthomonas oryzae* pv. *oryzae*. J Bacteriol *184*, 3539-3548.

Gust, B., Chandra, G., Jakimowicz, D., Yuqing, T., Bruton, C.J., and Chater, K.F. (2004). Lambda red-mediated genetic manipulation of antibiotic-producing *Streptomyces*. Adv Appl Microbiol *54*, 107-128.

Hassan, K.A., Johnson, A., Shaffer, B.T., Ren, Q., Kidarsa, T.A., Elbourne, L.D., Hartney, S., Duboy, R., Goebel, N.C., Zabriskie, T.M*., et al.* (2010). Inactivation of the GacA response regulator in *Pseudomonas fluorescens* Pf-5 has far-reaching transcriptomic consequences. Environ Microbiol *12*, 899-915.

Jenkins, C.L., and Starr, M.P. (1982). The pigment of Xanthomonas populi is a nonbrominated aryl-heptaene belonging to xanthomonadin pigment group 11. Curr Microbiol *7*, 195-198.

Kadioglu, A., Weiser, J.N., Paton, J.C., and Andrew, P.W. (2008). The role of *Streptococcus pneumoniae* virulence factors in host respiratory colonization and disease. Nat Rev Microbiol *6*, 288-301.

Kahne, D., Leimkuhler, C., Lu, W., and Walsh, C. (2005). Glycopeptide and lipoglycopeptide antibiotics. Chem Rev *105*, 425-448.

Kersten, R.D., Yang, Y.L., Xu, Y., Cimermancic, P., Nam, S.J., Fenical, W., Fischbach, M.A., Moore, B.S., and Dorrestein, P.C. (2011). A mass spectrometry-guided genome mining approach for natural product peptidogenomics. Nat Chem Biol *7*, 794-802.

Kersten, R.D., Ziemert, N., Gonzalez, D.J., Duggan, B.M., Nizet, V., Dorrestein, P.C., and Moore, B.S. (2013). Glycogenomics as a mass spectrometry-guided genome-mining method for microbial glycosylated molecules. Proc Natl Acad Sci USA *110*, E4407-E4416.

Khaldi, N., Seifuddin, F.T., Turner, G., Haft, D., Nierman, W.C., Wolfe, K.H., and Fedorova, N.D. (2010). SMURF: Genomic mapping of fungal secondary metabolite clusters. Fungal Genet Biol *47*, 736-741.

Laureti, L., Song, L., Huang, S., Corre, C., Leblond, P., Challis, G.L., and Aigle, B. (2011). Identification of a bioactive 51-membered macrolide complex by activation of a silent polyketide synthase in *Streptomyces ambofaciens*. Proc Natl Acad Sci USA *108*, 6258-6263.

Lautru, S., Deeth, R.J., Bailey, L.M., and Challis, G.L. (2005). Discovery of a new peptide natural product by *Streptomyces coelicolor* genome mining. Nat Chem Biol *1*, 265-269.

Letzel, A.C., Pidot, S.J., and Hertweck, C. (2012). A genomic approach to the cryptic secondary metabolome of the anaerobic world. Nat Prod Rep *30*, 392-428.

Li, L., Stoeckert, C.J., Jr., and Roos, D.S. (2003). OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res *13*, 2178-2189.

Li, M.H., Ung, P.M., Zajkowski, J., Garneau-Tsodikova, S., and Sherman, D.H. (2009). Automated genome mining for natural products. BMC Bioinformatics *10*, 185.

Lin, K., Zhu, L., and Zhang, D.Y. (2006). An initial strategy for comparing proteins at the domain architecture level. Bioinformatics *22*, 2081-2086.

Liu, C.I., Liu, G.Y., Song, Y., Yin, F., Hensler, M.E., Jeng, W.Y., Nizet, V., Wang, A.H., and Oldfield, E. (2008). A cholesterol biosynthesis inhibitor blocks *Staphylococcus aureus* virulence. Science *319*, 1391-1394.

Liu, G.Y., Essex, A., Buchanan, J.T., Datta, V., Hoffman, H.M., Bastian, J.F., Fierer, J., and Nizet, V. (2005). *Staphylococcus aureus* golden pigment impairs neutrophil killing and promotes virulence through its antioxidant activity. J Exp Med *202*, 209-215.

Lowry, B., Robbins, T., Weng, C.H., O'Brien, R.V., Cane, D.E., and Khosla, C. (2013). In vitro reconstitution and analysis of the 6-deoxyerythronolide B synthase. J Am Chem Soc *135*, 16809-16812.

Markowitz, V.M., Chen, I.M., Palaniappan, K., Chu, K., Szeto, E., Grechkin, Y., Ratner, A., Jacob, B., Huang, J., Williams, P*., et al.* (2012). IMG: the Integrated Microbial Genomes database and comparative analysis system. Nucleic Acids Res *40*, D115-122.

Mazmanian, S.K., Liu, C.H., Tzianabos, A.O., and Kasper, D.L. (2005). An immunomodulatory molecule of symbiotic bacteria directs maturation of the host immune system. Cell *122*, 107-118.

Mazmanian, S.K., Round, J.L., and Kasper, D.L. (2008). A microbial symbiosis factor prevents intestinal inflammatory disease. Nature *453*, 620-625.

McBride, M.J., Xie, G., Martens, E.C., Lapidus, A., Henrissat, B., Rhodes, R.G., Goltsman, E., Wang, W., Xu, J., Hunnicutt, D.W*., et al.* (2009). Novel features of the polysaccharide-digesting gliding bacterium *Flavobacterium johnsoniae* as revealed by genome sequence analysis. Appl Environ Microbiol *75*, 6864-6875.

Medema, M.H., Blin, K., Cimermancic, P., de Jager, V., Zakrzewski, P., Fischbach, M.A., Weber, T., Takano, E., and Breitling, R. (2011). antiSMASH: rapid identification, annotation and analysis of secondary metabolite biosynthesis gene clusters in bacterial and fungal genome sequences. Nucleic Acids Res *39*, W339-346.

Medema, M.H., Takano, E., and Breitling, R. (2013). Detecting sequence homology at the gene cluster level with MultiGeneBlast. Mol Biol Evol *30*, 1218-1223.

Nguyen, T., Ishida, K., Jenke-Kodama, H., Dittmann, E., Gurgui, C., Hochmuth, T., Taudien, S., Platzer, M., Hertweck, C., and Piel, J. (2008). Exploiting the mosaic structure of trans-acyltransferase polyketide synthases for natural product discovery and pathway dissection. Nat Biotechnol *26*, 225-233.

Oliynyk, M., Samborskyy, M., Lester, J.B., Mironenko, T., Scott, N., Dickens, S., Haydock, S.F., and Leadlay, P.F. (2007). Complete genome sequence of the erythromycin-producing bacterium *Saccharopolyspora erythraea* NRRL23338. Nat Biotechnol *25*, 447-453.

Park, B.S., Song, D.H., Kim, H.M., Choi, B.S., Lee, H., and Lee, J.O. (2009). The structural basis of lipopolysaccharide recognition by the TLR4-MD-2 complex. Nature *458*, 1191-1195.

Pei, J., Tang, M., and Grishin, N.V. (2008). PROMALS3D web server for accurate multiple protein sequence and structure alignments. Nucleic Acids Res *36*, W30-34.

Pelzer, S., Sussmuth, R., Heckmann, D., Recktenwald, J., Huber, P., Jung, G., and Wohlleben, W. (1999). Identification and analysis of the balhimycin biosynthetic gene cluster and its use for manipulating glycopeptide biosynthesis in *Amycolatopsis mediterranei* DSM5908. Antimicrob Agents Chemother *43*, 1565-1573.

Pfeifer, B.A., Admiraal, S.J., Gramajo, H., Cane, D.E., and Khosla, C. (2001). Biosynthesis of complex polyketides in a metabolically engineered strain of *E. coli*. Science *291*, 1790-1792.

Poplawsky, A.R., Urban, S.C., and Chun, W. (2000). Biological role of xanthomonadin pigments in *Xanthomonas campestris* pv. *campestris*. Appl Environ Microbiol *66*, 5123-5127.

Quan, J., and Tian, J. (2009). Circular polymerase extension cloning of complex gene libraries and pathways. PLoS One *4*, e6441.

Quan, J., and Tian, J. (2011). Circular polymerase extension cloning for high-throughput cloning of complex and combinatorial DNA libraries. Nat Protoc *6*, 242-251.

Rajagopal, L., Sundari, C.S., Balasubramanian, D., and Sonti, R.V. (1997). The bacterial pigment xanthomonadin offers protection against photodamage. FEBS Lett *415*, 125-128.

Rehm, B.H. (2010). Bacterial polymers: biosynthesis, modifications and applications. Nat Rev Microbiol *8*, 578-592.

Reichenbach, H., Kohl, W., Bottger-Vetter, A., and Achenbach, H. (1980). Flexirubin-type pigments in *Flavobacterium*. Archives of Microbiology *126*, 291-293.

Rinke, C., Schwientek, P., Sczyrba, A., Ivanova, N.N., Anderson, I.J., Cheng, J.F., Darling, A., Malfatti, S., Swan, B.K., Gies, E.A.*, et al.* (2013). Insights into the phylogeny and coding potential of microbial dark matter. Nature *499*, 431-437.

Samuel, G., and Reeves, P. (2003). Biosynthesis of O-antigens: genes and pathways involved in nucleotide sugar precursor synthesis and O-antigen assembly. Carbohydrate research *338*, 2503-2519.

Sattely, E.S., Fischbach, M.A., and Walsh, C.T. (2008). Total biosynthesis: in vitro reconstitution of polyketide and nonribosomal peptide pathways. Nat Prod Rep *25*, 757-793.

Schneiker, S., Perlova, O., Kaiser, O., Gerth, K., Alici, A., Altmeyer, M.O., Bartels, D., Bekel, T., Beyer, S., Bode, E.*, et al.* (2007). Complete genome sequence of the myxobacterium *Sorangium cellulosum*. Nat Biotechnol *25*, 1281-1289.

Segata, N., Bornigen, D., Morgan, X.C., and Huttenhower, C. (2013). PhyloPhlAn is a new method for improved phylogenetic and taxonomic placement of microbes. Nat Commun *4*, 2304.

Smoot, M.E., Ono, K., Ruscheinski, J., Wang, P.L., and Ideker, T. (2011). Cytoscape 2.8: new features for data integration and network visualization. Bioinformatics *27*, 431-432.

Starcevic, A., Zucko, J., Simunkovic, J., Long, P.F., Cullum, J., and Hranueli, D. (2008). ClustScan: an integrated program package for the semi-automatic annotation of modular biosynthetic gene clusters and in silico prediction of novel chemical structures. Nucleic Acids Res *36*, 6882-6892.

Starr, M.P., Jenkins, C.L., Bussey, L.B., and Andrewes, A.G. (1977). Chemotaxonomic significance of the xanthomonadins, novel brominated aryl-polyene pigments produced by bacteria of the genus *Xanthomonas*. Arch Microbiol *113*, 1-9.

Staunton, J., and Weissman, K.J. (2001). Polyketide biosynthesis: a millennium review. Nat Prod Rep *18*, 380-416.

Tamura, K., Peterson, D., Peterson, N., Stecher, G., Nei, M., and Kumar, S. (2011). MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Mol Biol Evol *28*, 2731-2739.

Walsh, C.T., and Fischbach, M.A. (2010). Natural products version 2.0: connecting genes to molecules. J Am Chem Soc *132*, 2469-2493.

Walter, M.H., and Strack, D. (2011). Carotenoids and their cleavage products: biosynthesis and functions. Nat Prod Rep *28*, 663-692.

Wang, Y., Qian, G., Li, Y., Wang, Y., Wang, Y., Wright, S., Li, Y., Shen, Y., Liu, F., and Du, L. (2013). Biosynthetic mechanism for sunscreens of the biocontrol agent *Lysobacter enzymogenes*. PLoS One *8*, e66633.

Weber, T., Rausch, C., Lopez, P., Hoof, I., Gaykova, V., Huson, D.H., and Wohlleben, W. (2009). CLUSEAN: a computer-based framework for the automated analysis of bacterial secondary metabolite biosynthetic gene clusters. J Biotechnol *140*, 13-17.

Weitnauer, G., Muhlenweg, A., Trefzer, A., Hoffmeister, D., Sussmuth, R.D., Jung, G., Welzel, K., Vente, A., Girreser, U., and Bechthold, A. (2001). Biosynthesis of the orthosomycin antibiotic avilamycin A: deductions from the molecular analysis of the *avi* biosynthetic gene cluster of *Streptomyces viridochromogenes* Tu57 and production of new antibiotics. Chem Biol *8*, 569-581.

Winter, J.M., Behnken, S., and Hertweck, C. (2011). Genomics-inspired discovery of natural products. Curr Opin Chem Biol *15*, 22-31.

Chapter 3

Global landscape of HIV-human protein complexes

Global Landscape of HIV-Human Protein Complexes

Stefanie Jäger[1,2], Peter Cimermancic[2,3], Natali Gulbahce[1,2], Jeff Johnson[1,2], Kathryn McGovern[1,2], Starlynn C. Clarke[4], Michael Shales[1,2], Gaelle Mercenne[5], Kathy Li[1,2,4], Hilda Barry[1,2,4], Gwendolyn M. Jang[1,2,6], Eyal Akiva[2,3], Lars Pache[7], John Marlett[8], Shoshannah L. Roth[9], Melanie Stephens[9], Ivan D'Orso[6], Jason Fernandes[6], Marie Fahey[1,2], Cathal Mahon[1,2,4], Anthony J. O'Donoghue[4], Aleksandar Todorovic[10], John H. Morris[4], David Maltby[4], Tom Alber[11], Gerard Cagney[12], Fredric D. Bushman[9], John A. Young[8], Sumit K. Chanda[7], Wesley I. Sundquist[5], Tanja Kortemme[2,3,13], Ryan Hernandez[2,3,13], Charles S. Craik[2,4], Alma Burlingame[4], Andrej Sali[2,3,13], Alan D. Frankel[2,6,13], Nevan J. Krogan[1,2,13,14*]

1 Department of Cellular and Molecular Pharmacology, University of California, San Francisco, California, USA;

2 California Institute for Quantitative Biosciences, QB3, San Francisco, California, USA;

3 Department of Bioengineering and Therapeutic Sciences, University of California, San Francisco, California, USA;

4 Department of Pharmaceutical Chemistry, University of California, San Francisco, California, USA;

5 Department of Biochemistry, University of Utah, Salt Lake City, UT;

6 Department of Biochemistry and Biophysics, University of California, San Francisco, California, USA;

7 Sanford-Burnham Medical Research Institute, La Jolla, CA;

8 The Salk Institute for Biological Studies, La Jolla, CA;

9 Department of Microbiology, University of Pennsylvania, Philadelphia, PA;

10 Department of Chemistry, University of California-Berkeley, Berkeley, CA;

11 Department of Molecular and Cell Biology, University of California-Berkeley, Berkeley, CA;

12 Conway Institute, University College Dublin, Belfield, Dublin, Ireland;

13 HPC (Host Pathogen Circuitry) Group;

14 J. David Gladstone Institute of Cardiovascular Disease, San Francisco, CA

*To whom correspondence should be addressed: krogan@cmp.ucsf.edu

Summary

HIV has a small genome and therefore relies heavily on the host cellular machinery to replicate. In this study, we used an affinity tagging/purification mass spectrometry (AP-MS) approach to determine the interactions of all 18 HIV-1 proteins and polyproteins with host proteins in two different human cell lines (HEK293 and Jurkat). Using a novel quantitative scoring system, termed MiST, we identified 497 high-confidence HIV-human protein-protein interactions (PPIs) involving 435 individual human proteins, with ~40% of them being identified in both cell types. We found that the host proteins hijacked by HIV are highly conserved across primates, especially those found interacting in both cell types. We uncovered a number of host complexes targeted by viral proteins including the discovery that HIV protease cleaves a component of the eIF3 translational initiation complex, eIF3d, a protein that inhibits HIV replication. This dataset facilitates a more comprehensive and detailed understanding of how the host machinery is manipulated during the course of HIV infection.

Introduction

As described in Chapter 1, the map of the physical interactions between proteins within a particular system is necessary for studying the molecular mechanisms that underlie the system. Analysis of protein-protein interactions (PPIs) has been successfully accomplished in different organisms using a variety of technologies, including mass spectrometry approaches (Gavin et al., 2006; Ho et al., 2002; Krogan et al., 2006; Sowa et al., 2009) and those designed to detect pair-wise physical interactions, including two-hybrid yeast system (Stelzl et al., 2005; Yu et al., 2008) and protein-fragment complementation assays (Tarassov et al., 2008). Although two-hybrid methodologies have been used to systematically study host-pathogen interactions, including studies targeting hepatitis C virus (HCV) (de Chassey et al., 2008), Epstein-Barr virus (EBV) (Calderwood et al., 2007) and influenza virus (H1N1) (Shapira et al., 2009), to date, a systematic AP-MS study has not yet been carried out on any host-pathogen system. Here, we have targeted HIV-1 for such an analysis, uncovering a plethora of host proteins, complexes and pathways that are hijacked by the virus during the course of infection.

Results and discussion

AP-MS PLATFORM FOR ANALYZING HIV-HUMAN PROTEIN COMPLEXES

We aimed to systematically and quantitatively identify host proteins associated with HIV1 proteins using an AP-MS approach (Jäger et al., 2010). To this end, we cloned each gene corresponding to all 18 HIV-1 proteins and polyproteins, including the accessory factors (Vif, Vpu, Vpr and Nef), Tat, Rev, the polyproteins (Gag, Pol and Gp160) and the corresponding processed products (MA, CA, NC and p6; PR, RT and IN; and Gp120 and Gp41, respectively) (**Supplementary Figure 1**). The majority of these factors were codon optimized to express all proteins optimally at comparable levels (**Supplementary Table 1**). Each clone was transiently transfected into HEK293 cells and was also used to generate stably expressed, tetracycline inducible versions in Jurkat cells (**Figure 1a**). Western blot analysis confirmed each factor was expressed in both HEK293 and Jurkat cell lines (**Supplementary Figure 2**). Following multiple purifications of each factor from both cell lines, the material on the anti-Flag or StrepTactin beads, as well as the eluted material, was analyzed by mass spectrometry (**Figure 1a; Supplementary Table 2**). Finally, an aliquot of each purification was subjected to SDS-PAGE, stained (**Supplementary Figure 3**) and subjected to analysis by mass spectrometry.

We identified a range of co-purifying host proteins for each HIV factor that were reproducible regardless of the protocol used (**Supplementary Figures 4,5,7**; **Supplementary Data 1**). Several scoring systems can quantify protein-protein interactions from AP-MS proteomic datasets, including PE (Collins et al., 2007), COMPASS (Sowa et al., 2009) and SAINT (Choi et al., 2010). For this dataset, however, we devised a new scoring system particularly suited for identifying AP-MS derived host-pathogen PPIs, termed MiST (Mass Spectrometry Interaction Statistics) (available as a web-server at http://salilab.org/mist). The MiST score a weighted sum of three measures: 1) protein abundance measured by peak intensities from the mass spectrum (abundance); 2) invariability of abundance over replicated experiments (reproducibility); and 3) uniqueness of an observed host-pathogen interaction across all viral purifications (specificity) (**Figure 1b; Supplementary Methods**). These three metrics are summed through a principal component analysis into an optimal composite score (**Figure 1c**, **Supplementary Data 2**). By using a benchmark of well-characterized HIV-human PPIs (**Supplementary Table 3**), analysis of the MiST scoring system revealed superior performance on our dataset when compared to CompPASS and SAInt (**Supplementary Figure 6**) (and comparable performance using other datasets (**Supplementary Figure 8**)) and allowed us to define a MiST cut-off of 0.75, which corresponds to ~4% of all detected interactions. To estimate how many interactions would exceed this threshold by chance, we randomly shuffled our dataset 1000-times. A random MiST score of 0.75 or greater was assigned to an interaction 10-times less frequently than seen among the MiST scores for the real data, and the probability of an interaction assignment with the random MiST score higher than 0.75 was $2.5 \cdot 10^{-4}$ (**Figure 1d**).

At the MiST threshold >0.75, the number of host proteins we found associated with each HIV protein ranged from 63 (Gp160) to 0 (CA and p6) (**Figure 1e**). In total, we observed 497 different HIV-human PPIs (347 and 348 identified from HEK293 and Jurkat cells, respectively) (**Supplementary Data 3**). 196 interactions (~40%) were detected in both cell types; 150 and 151 were specific to the HEK293 and Jurkat cells, respectively (**Figure 1e**). Only some of these specificities could be explained by differential gene expression in the two cell lines (**Supplementary Figure 9**). Using antibodies against 26 of the human proteins, and affinity tagged versions of an additional 101, we could confirm 100/130 AP-MS derived HIV-human PPIs via co-immunoprecipitation/Western blot analysis (77% success rate) (**Supplementary Figures 10, 11**), suggesting we have derived a high quality physical interaction dataset.

We next analyzed the sets of host proteins associated with each HIV protein with respect to functional categories, uncovering many expected connections. These include an enrichment of host factors involved in transcription physically linked to the HIV transcription factor Tat (Frankel and Young, 1998); and Vpu, Vpr and Vif, HIV accessory factors that hijack ubiquitin ligases, associating with host machinery implicated in the regulation of ubiqutination (Malim and Emerman, 2008) (**Figure 1f**) (**Supplementary Data 4**). Notably, we found an enrichment of host factors involved in translation not only associated with the mRNA export factor Rev, but also Pol and PR. Similar notable trends are present when one considers domain types instead of whole proteins (**Figure 1g**; **Supplementary Table 4**). For example, host proteins interacting with IN are enriched for 14-3-3 domains, which generally bind phosphorylated regions of proteins(Yaffe et al., 1997); and proteins containing β-propellers have a higher propensity for binding to Vpr (for additional domain enrichment analysis, see **Supplementary Figure 12**). These domain analyses could facilitate future structural modeling of HIV-human PPIs.

COMPARISON TO OTHER HIV-HOST RELATED DATASETS

Next, we compared the data derived during the course of our work to other HIV-related datasets, including previously published HIV-human PPIs and host factors implicated in HIV function from genome-wide RNAi screens. For example, the VirusMint database(Chatr-aryamontri et al., 2009) contains 587 HIV-human literature-curated PPIs (**Supplementary Data 5**), which are mostly derived from small-scale, targeted studies. While the overlap between the 497 interactions identified in this work and VirusMint is statistically significant (*p*-value = 8 10^{-8}), it corresponds to only 19 PPIs (**Figure 2a**; **Supplementary Table 5**). However, a greater overlap exists, one that remains statistically significant, when interactions below the MiST threshold of 0.75 are considered using a sliding cut-off (e.g., at a MiST score of 0.2, there exists an overlap of 67, *p*-value = 1 10^{-3}) (**Figure 2c**, red lines; **Supplementary Data 6**). This overlap argues that we have indeed identified many interactions that have been previously reported.

However, it is likely that the higher scoring interactions identified here have a greater chance of being biologically relevant with respect to HIV function compared to many of those in VirusMint.

Recently, four RNAi screens identified host factors that have an adverse effect on HIV-1 replication when knocked down. In total, 1071 human genes were identified in these four studies (**Supplementary Data 7**), 55 of which are overlapping with our 435 proteins (*p*-value = 2.7 10^{-10}) (**Figure 2b**; **Supplementary Table 6**). Again, this overlap increases (as does its statistical significance) as we consider proteins participating in HIV-human PPIs with MiST scores below 0.75 (**Figure 2c**, blue lines; **Supplementary Data 8**).

To identify the evolutionary forces operating on host proteins interacting with HIV-1, we performed a comparative genomics analysis of divergence patterns between human and rhesus macaque. The proteins identified in both HEK293 and Jurkat cell lines exhibit stronger signatures of evolutionary restraint than those identified exclusively in one cell line or in VirusMint (**Figure 2d**). Points in the lower right quadrant of **Figure 2d** show characteristic signatures of strong purifying selection, while the upper right quadrant shows signatures more consistent with neutral evolution. This observation suggests that the PPIs identified in our study, especially the ones identified in both cell types, are more physiologically relevant to mammalian evolution than those identified in VirusMint.

NETWORK REPRESENTATION OF HIV-HUMAN PPI SET

We next plotted the 497 HIV-human interactions identified in this study in a network representation containing nodes corresponding to 16 HIV (yellow) (note we did not detect any high-scoring interactions for CA or p6) and 435 human factors that were derived from the HEK293 cells (blue), Jurkat cells (red) or both (**Figure 3**). We also introduced 289 interactions between human proteins (black edges) derived from several databases (**Supplementary Data**

9), including CORUM (Ruepp et al., 2009) and BIOGRID (Stark et al., 2006). These human-human interactions helped to identify many host complexes, including several that have been previously characterized. For example, we identified components of the cullin-containing ubiqutin ligases associated with the accessory factors that are known to hijack them: Vif (Cul5); Vpu (Cul1) and Vpr (Cul4a) (Malim and Emerman, 2008). We previously identified several new components of the Tat/P-TEFb transcription complex that was derived from this analysis(He et al., 2010), as well as LARP7, a component of an inhibitory snRNP that controls Tat/P-TEFb activity(D'Orso and Frankel, 2010). Other notable connections include associations between: (i) the tRNA synthetase complex with MA, (ii) Vif with the histone deacetylation complex, HDAC3/NCOR1 and (iii) the splicing complex, SMN and Dynein with Vpr. Further discussion of the HIV-human interactions is in **Supplementary Information**. Ultimately, all data will be able to be searched and compared to other HIV-related datasets using the web-based software, GPS-Prot (www.gpsprot.org) (Fahey et al., 2011).

EIF3 IS TARGETED BY HIV PROTEASE

Interestingly, we found Pol and PR associated with the translational initiation complex, eIF3, which binds the 40S ribosomal subunit and serves as a scaffold for the assembly of other translation factors(Hinnebusch, 2006). eIF3 is a 13 subunit complex (eIF3a-m) and we detected 12 of the subunits bound to Pol and/or PR, except eIF3j, which is only loosely associated with the complex(Hinnebusch, 2006) (**Figure 4a**). Interestingly, even though PR is the smallest of the three processed proteins of Pol, we find it associated with the greatest number of host factors (**Figure 4a**). To determine whether or not components of the translation complex are substrates for PR, purified human eIF3 was incubated with active PR, resulting in the removal of a 70 kDa band and appearance of a ~60 kDa protein product (**Figure 4b**). Analysis of the cleaved product by N-terminal sequencing revealed a cleavage of eIF3d between Met114 and

Leu115, which corresponds to the consensus sequence for HIV-1 protease(Schilling and Overall, 2008), and falls with the RNA binding domain (RRM) with eIF3d (Asano et al., 1997) (**Figure 4b**). To validate this result *in vivo*, Flag-tagged versions of 10 eIF3 subunits were independently co-transfected, each with a small amount of active HIV-1 PR into HEK293 cells, and the cell lysates were analyzed by Western blotting (**Figure 4c**). Consistent with the *in vitro* data, only eIF3d was found to be cleaved, resulting in the expected size of the C-terminal fragment of ~60 kDa. Purification of tagged versions of the N- and C-terminal ends of cleaved eIF3d revealed that only the N-terminus of 114 amino acid residues associates with the eIF3 complex (**Supplementary Table 7**). The cleavage occurred with a similar efficiency as the processing of the natural PR substrate Gag (**Figure 4d**), while two cellular proteins previously described to be cleaved by HIV PR, PAPBC1 (Alvarez et al., 2006) and Bcl2 (Strack et al., 1996), were cleaved only at higher PR concentrations or not at all, respectively.

Next, 4-6 siRNAs against different eIF3 subunits were used in HIV infectivity assays (**Figure 4e, f**; **Supplementary Table 8**). Using an HIV-VSVg construct, which only allows for a single round of replication, knockdown of eIF3d, but not other subunits, provided an increase of infectivity (**Figure 4e**), suggesting that this factor is acting at early stages of infection. Interestingly, using the virus NL4.3, which allows for multiple rounds of infection, knock-down of eIF3d, as well as e, and f enhanced NL4.3 infectivity 3 to 5 fold, whereas inhibition of subunits c, g, and i had no promoting effect (**Figure 4f**). Consistent with these results, a previous overexpression screen for factors that restrict HIV-1 replication identified eIF3f as the most potent inhibitory clone (Valente et al., 2009). The knockdown efficiency was >90% for all siRNAs and their impact on cell viability was similar, ruling out an unspecific effect on translation (**Supplementary Figure 14**). Furthermore, we found that knockdown of eIF3d results in an increase of reverse transcription activity using assays monitoring both early and late RT (**Figure 4g**; **Supplementary Figure 15**). This suggests that eIF3 does, in fact, play a role in the early stages of infection, perhaps by binding to the viral RNA through the RRM domain in eIF3d, and thus inhibiting RT, an effect that is overcome by PR cleavage of eIF3d (**Supplementary Figure 16**). Further work will be required to fully understand the advantage the virus seemingly achieves *via* remodeling eIF3.

This work represents the first systematic AP-MS study aimed at characterizing host-pathogen interactions. We identified several previously described human complexes hijacked by HIV proteins, but most of the interactions have not been previously described. We further explored the biological significance of two such identified HIV-human interactions: (i) HIV protease targeting a component of eIF3 that is inhibitory to HIV replication and (ii) CBF (PEBB), a new component of the Vif/Cul5 ubiquitin ligase complex required for APOBEC3G stability and HIV infectivity (Jäger et al., accompanying manuscript). Previous work using this AP-MS pipeline allowed for the identification of four new components of the P-TEFb complex required for Tat activation(He et al., 2010). Further work will be required to determine if, how and at what stage of infection the remaining host factors impinge on HIV function. Ultimately, work should be carried out to characterize the interactions in the context of infection and a more targeted, quantitative genetic, proteomic, and structural analysis of the set of host factors identified in this study will provide a more accurate view of how the host machinery is being re-wired during the course of HIV infection. Finally, analysis of the host factors co-opted by different viruses using the same proteomic pipeline described here will allow for the identification of protein complexes routinely targeted by different pathogens, in turn leading to better therapeutic targets for future studies.

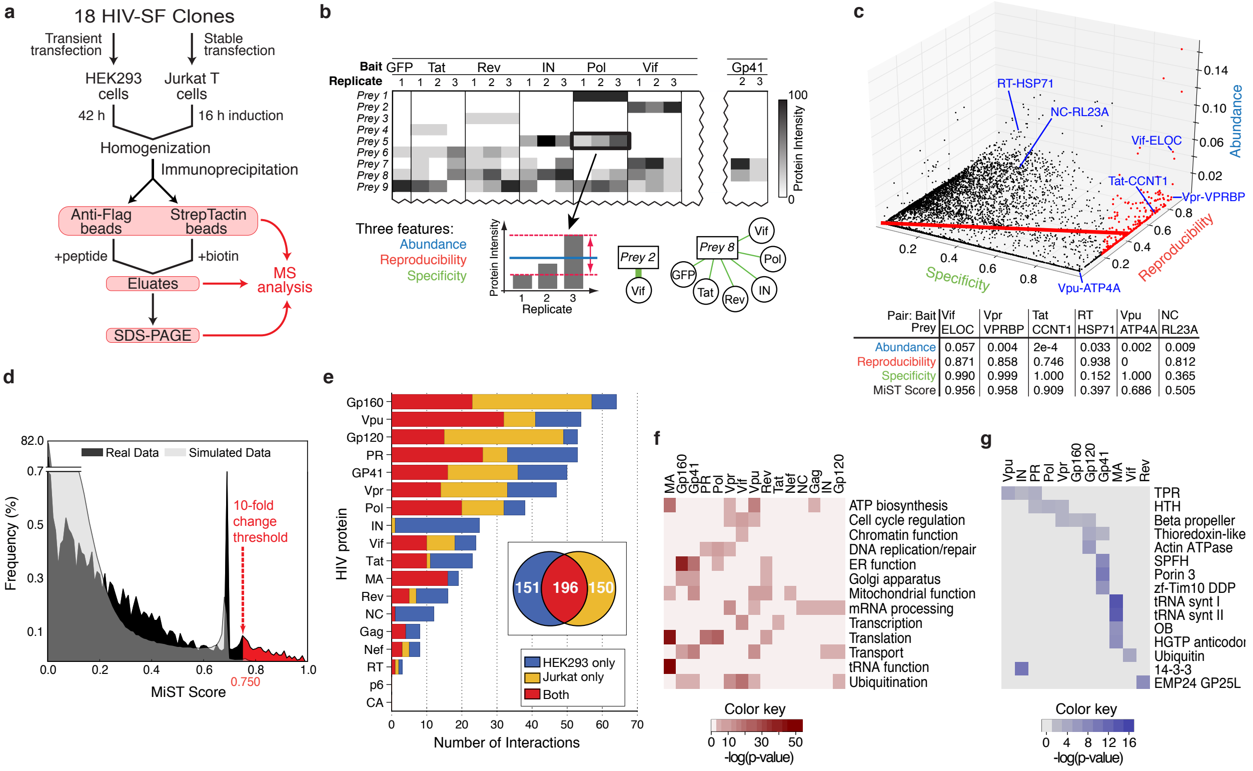
Methods Summary

Details on experimental assays, plasmid constructs, sequences, cell lines, antibodies ,and computational analysis are provided online at http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3310911/. Briefly, affinity tagging and purification was carried out as previously described(Jäger et al., 2010) and the protein samples were analyzed on a Thermo Scientific LTQ Orbitrap XL mass spectrometer. For the evolutionary analysis, genome-wide alignments to rhesus macaque were downloaded from the UCSC genome browser (http://genome.ucsc.edu/) and evolutionary rates for each group of genes considered were measured using the synonymous and non-synonymous rates of evolution. For the *in vitro* protease assay, MBP-tagged PR was expressed in BL21 (Gold) DE3 cells in the presence of 100 µM Saquinavir and purified on a MBP-TRAP column. Purified eIF3 was obtained from Jamie Cate, UC Berkeley. For the infection assays, HeLa P4.R5 cells were transfected with siRNAs and after 48 h infected with pNL4-3 or a pNL4-3 derived VSV-G pseudotyped reporter virus. Infection levels were determined by luminescence readout.

Acknowledgments

We thank Alex Choi, Zanub Rizvi, and Eunju Kwon for cloning of human ORFs. We also thank John Gross, Raul Andino, Reuben Harris, Matt Daugherty and members of the Krogan lab for stimulating discussion. This research was funded by grants from the National Institutes of Health (P50 GM082250 to NJK, ADF and CSC; P01 AI090935 to NJK, SKC, JAY and FDR; and P41RR001614 to AB, NIGMS U54 RR022220 to AS, ASP01GM073732-05, NIGMS to AT, CHRP-ID08-TBI-063 to SKC and the Nomis Foundation to JAY). NJK is a Searle Scholar and a Keck Young Investigator.

Figures



Pair: Bait Prey	Vif ELOC	Vpr VPRBP	Tat CCNT1	RT HSP71	Vpu ATP4A	NC RL23A
Abundance	0.057	0.004	2e-4	0.033	0.002	0.009
Reproducibility	0.871	0.858	0.746	0.938	0	0.812
Specificity	0.990	0.999	1.000	0.152	1.000	0.365
MiST Score	0.956	0.958	0.909	0.397	0.686	0.505

Figure 1: Affinity purification of HIV-1 proteins, analysis and scoring of MS data. **a**, Flowchart of the proteomic AP-MS used to define the HIV-host interactome. **b**, Data from AP-MS experiments are organized in an interaction table with cells representing amount of prey protein purified (e.g. spectral counts or peptide intensities). Three features are used to describe bait-prey relationships: 1) abundance (blue); 2) reproducibility (red) (the invariability of bait-prey pair quantities); and 3) specificity (green) (a measure of how selective is the observed prey for a given bait when compared with other baits). **c**, All bait-prey pairs are mapped into the three feature space (abundance, reproducibility and specificity). The MiST score is defined as a projection on the first principal component (red line). All interactions, represented as nodes, above the defined threshold (0.75) are red. This procedure separates the more likely biologically relevant bait-prey pairs (*e*.g. Vif-ELOC, Vpr-VPRBP, and Tat-CCNT1) from the interactions that

are likely less biologically relevant due to low reproducibility (Vpu-ATP4A) or specificity (RT-HSP71 and NC-RI23A). For a complete list of the three component scores for all pair-wise interactions, see **Supplementary Data 2**. **d**, The histogram of MiST scores (Real Data) is compared to a randomized set of scores obtained from randomly shuffling the bait-prey table (Simulated Data). The MiST score threshold (0.75) was defined using a benchmark (**Supplementary Table 3**) where the predictions are enriched for true, biologically relevant interactions by at least a factor of 10 compared to random predictions (as well as through ROC and recall plots (**Supplementary Figure 6**)). The large peak at ~0.7 corresponds to interactions that were specific to a single HIV protein but were not reproducible. **e**, Bar graph of the number of host proteins we found interacting with each HIV factor (MiST score > 0.75). The cell type in which the interaction was found is represented in blue (HEK293 only), yellow (Jurkat only) or red (both). In total, there are 497 HIV-host interactions involving 435 individual human proteins. **f**, A heatmap representing enriched biological functions of the human proteins identified as interacting with HIV proteins. The biological functions represent manually collapsed Gene Ontology terms obtained *via* DAVID (Methods). **g**, A heatmap representing enriched domains present in the human proteins identified as interacting with HIV proteins. The domain titles represent clan/domain/family names in the Pfam database (Supplementary Methods). Coordinates are colored according to corresponding statistical significance (-log of *p*-value) of the enriched biological functions (**f**) and domain over-representation (**g**).

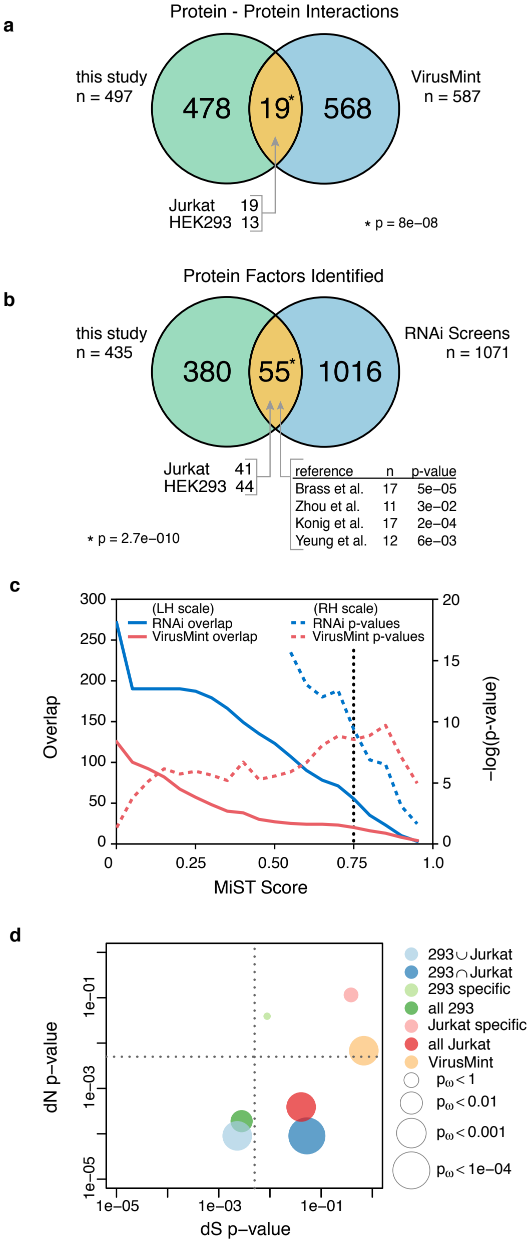


Figure 2: Comparison of PPI data with other HIV datasets. a, Using a MiST score cut-off of 0.75, we identified 497 HIV-human PPIs from both cell types, 19 of which overlap with the 587 PPIs reported in VirusMint (*p*-value = $2.97\ \ 10^{-7}$) (**Supplementary Table 5**). **b**, Out of the possible 1071 human factors identified in 4 HIV-dependency RNAi screens (Brass et al., 2008; König et al., 2008; Yeung et al., 2009; Zhou et al., 2008a), there is an overlap of 55 with the 435 individual host factors identified in our proteomic screen (*p*-value = $2.7\ \ 10^{-10}$) (**Supplementary

Table 6). **c**, Overlapping number of interactions with VirusMint (solid red line) and proteins with RNAi screens (solid blue line) as a function of the MiST cutoff. The *p*-values of the overlap are represented as dashed lines using the same colors (**Supplementary Data 6, 8**). **d**, Comparative genomics analysis of divergence patterns between human and rhesus macaque reveals strong evolutionary constraint. The x- and y-axes represent *p*-values for the synonymous (d*S*) and non-synonymous (d*N*) rates of evolution, respectively, based on 10,000 bootstrap simulations of the human genome (controlling for expression patterns across human tissues), and the size of the circle indicates the significance of the evolutionary parameter ω = *d*N/*d*S (Supplementary Methods). Horizontal and vertical dotted lines are drawn at 0.5% to indicate the Bonferroni significance threshold for each axis. For the VirusMint data, the significance of ω is primarily driven by higher rates of synonymous evolution.

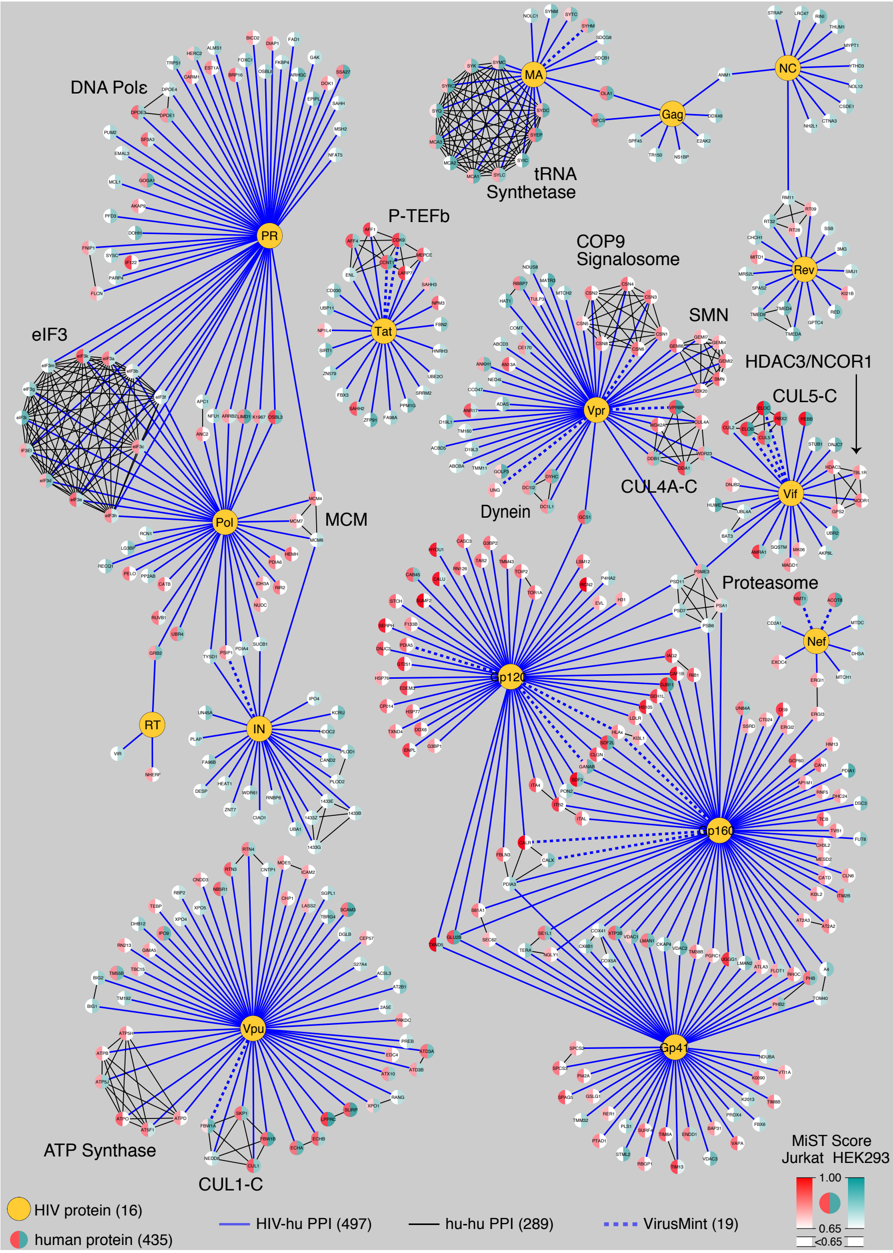
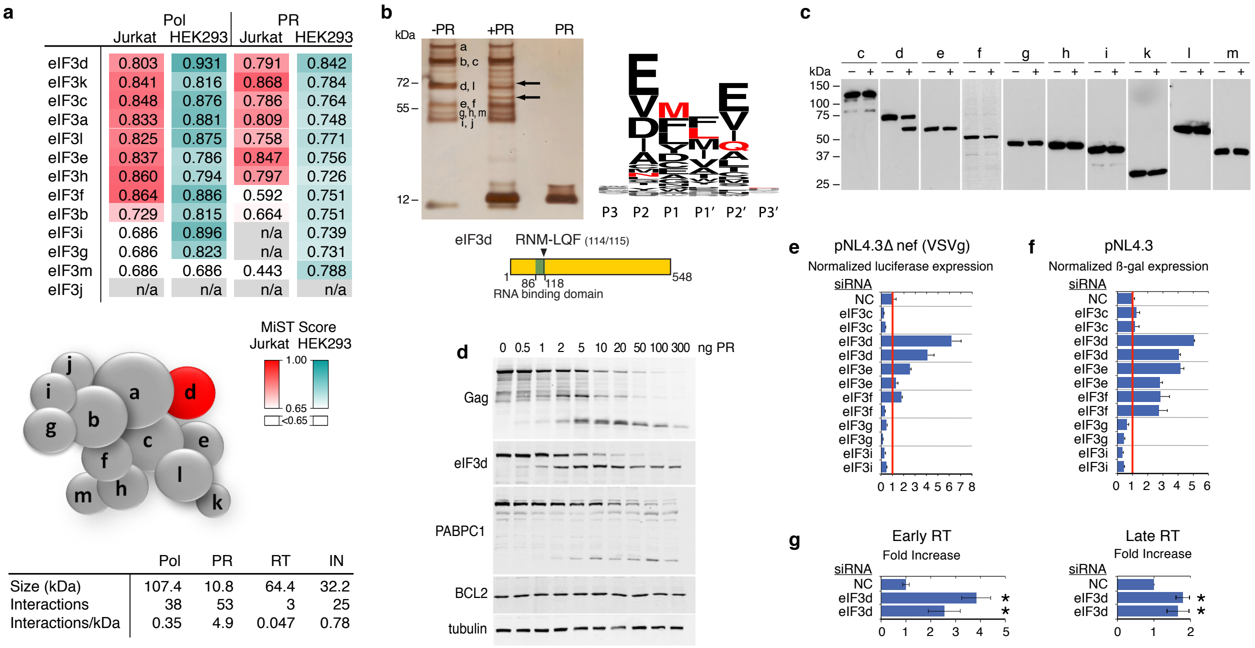


Figure 3: Network representation of the HIV-human PPIs. In total, 497 HIV-human interactions are represented between 16 HIV proteins and 435 human factors. The node is split into two colors and the intensity of the color corresponds to the MiST score from interactions derived from HEK293 (blue) and Jurkat (red) cells. Black edges correspond to interactions between host factors (289) that were obtained from publicly available databases, such as CORUM, HPRD and BIOGRID; several complexes are labeled. Dashed edges correspond to

interactions also found in VirusMint. Ultimately, all data will be able to be searched and compared to other HIV-related datasets using the web-based software, GPS-Prot (www.gpsprot.org) (Fahey et al., 2011).



	Pol Jurkat	Pol HEK293	PR Jurkat	PR HEK293
eIF3d	0.803	0.931	0.791	0.842
eIF3k	0.841	0.816	0.868	0.784
eIF3c	0.848	0.876	0.786	0.764
eIF3a	0.833	0.881	0.809	0.748
eIF3l	0.825	0.875	0.758	0.771
eIF3e	0.837	0.786	0.847	0.756
eIF3h	0.860	0.794	0.797	0.726
eIF3f	0.864	0.886	0.592	0.751
eIF3b	0.729	0.815	0.664	0.751
eIF3i	0.686	0.896	n/a	0.739
eIF3g	0.686	0.823	n/a	0.731
eIF3m	0.686	0.686	0.443	0.788
eIF3j	n/a	n/a	n/a	n/a

	Pol	PR	RT	IN
Size (kDa)	107.4	10.8	64.4	32.2
Interactions	38	53	3	25
Interactions/kDa	0.35	4.9	0.047	0.78

Figure 4: eIF3d is cleaved by HIV-1 PR and inhibits infection. a, MiST scores for eIF3 subunits associated with PR and Pol in HEK293 and Jurkat cells (left). Sizes of the proteins and number of significant interactions (MiST >0.75) detected for Pol and its subunits (right). Modular representation of the eIF3 complex. Subunit positions are based on prior studies(Cai et al., 2010; Zhou et al., 2008b) that indicate subunit d is exposed on the complex surface(Cai et al., 2010; Siridechadilok et al., 2005). **b,** Silver stain of purified eIF3 complex incubated with recombinant HIV-1 PR, demonstrating the decrease of a single band and appearance of an approximately 60 kDa product band. The product band was excised and sequenced, revealing the cleavage location between Met 114 and Leu 115 of eIF3d. Weblogo representing the HIV-1 protease consensus cleavage sequence from P3 to P3'(Schilling and Overall, 2008). The residues corresponding to the eIF3d cleavage site (red) is located within the RNA binding domain(Asano et al., 1997). **c,** Flag western blot of HEK293 cell lysate expressing Flag-tagged eIF3 subunits in the absence (-) or presence (+) of active PR. **d**, HEK293 cells were co-transfected with Gag, or Flag-tagged eIF3d, PABPC1, Bcl2 and increasing amounts of PR. Cell

lysates were probed against CA (upper panel), Flag (middle panel), or tubulin as control (lower panel). **e, f,** Hela-derived P4/R5 MAGI cells were transfected with siRNAs targeting individual subunits of the eIF3 complex (Supplementary Table 7) and subsequently infected with either wild-type pNL4-3 (**f**) or a pNL4-3-derived, VSVg pseudotyped, single-cycle virus (HIV-VSVg) (**e**). The results of two distinct siRNAs targeting non-overlapping sequences are shown for each gene. Values represent average and standard deviation of triplicates, normalized to a negative control (**Supplementary Table 9**). **g,** Early (left) and late (right) HIV-1 DNA levels measured by Q-PCR amplification in cells transfected with two independent eIF3D siRNAs or with control siRNAs. Samples were normalized by input DNA amount or by cellular gene (PBGD) copy number (**Supplementary Tables 10 and 11**). * indicates $p<0.05$ (Kruskal-Wallis test with Dunn's correction for multiple comparisons).

Supplementary text, tables, and figures

Available online at http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3310911/.

References

Alvarez, E., Castello, A., Menendez-Arias, L., and Carrasco, L. (2006). HIV protease cleaves poly(A)-binding protein. Biochem J *396*, 219-226.

Asano, K., Vornlocher, H.P., Richter-Cook, N.J., Merrick, W.C., Hinnebusch, A.G., and Hershey, J.W. (1997). Structure of cDNAs encoding human eukaryotic initiation factor 3 subunits. Possible roles in RNA binding and macromolecular assembly. J Biol Chem *272*, 27042-27052.

Brass, A.L., Dykxhoorn, D.M., Benita, Y., Yan, N., Engelman, A., Xavier, R.J., Lieberman, J., and Elledge, S.J. (2008). Identification of host proteins required for HIV infection through a functional genomic screen. Science *319*, 921-926.

Cai, Q., Todorovic, A., Andaya, A., Gao, J., Leary, J.A., and Cate, J.H. (2010). Distinct regions of human eIF3 are sufficient for binding to the HCV IRES and the 40S ribosomal subunit. J Mol Biol *403*, 185-196.

Calderwood, M.A., Venkatesan, K., Xing, L., Chase, M.R., Vazquez, A., Holthaus, A.M., Ewence, A.E., Li, N., Hirozane-Kishikawa, T., Hill, D.E*., et al.* (2007). Epstein-Barr virus and virus human protein interaction maps. Proc Natl Acad Sci U S A *104*, 7606-7611.

Chatr-aryamontri, A., Ceol, A., Peluso, D., Nardozza, A., Panni, S., Sacco, F., Tinti, M., Smolyar, A., Castagnoli, L., Vidal, M*., et al.* (2009). VirusMINT: a viral protein interaction database. Nucleic Acids Res *37*, D669-673.

Choi, H., Larsen, B., Lin, Z.Y., Breitkreutz, A., Mellacheruvu, D., Fermin, D., Qin, Z.S., Tyers, M., Gingras, A.C., and Nesvizhskii, A.I. (2010). SAINT: probabilistic scoring of affinity purification-mass spectrometry data. Nat Methods *8*, 70-73.

Collins, S.R., Kemmeren, P., Zhao, X.C., Greenblatt, J.F., Spencer, F., Holstege, F.C., Weissman, J.S., and Krogan, N.J. (2007). Toward a comprehensive atlas of the physical interactome of Saccharomyces cerevisiae. Mol Cell Proteomics *6*, 439-450.

D'Orso, I., and Frankel, A.D. (2010). RNA-mediated displacement of an inhibitory snRNP complex activates transcription elongation. Nat Struct Mol Biol *17*, 815-821.

de Chassey, B., Navratil, V., Tafforeau, L., Hiet, M.S., Aublin-Gex, A., Agaugue, S., Meiffren, G., Pradezynski, F., Faria, B.F., Chantier, T.*, et al.* (2008). Hepatitis C virus infection protein network. Mol Syst Biol *4*, 230.

Fahey, M.E., Bennett, M.J., Mahon, C., Jäger, S., Pache, L., Kumar, D., Shapiro, A., Rao, K., Chanda, S.K., Craik, C.S.*, et al.* (2011). GPS-Prot: a web-based visualization platform for integrating host-pathogen interaction data. BMC Bioinformatics *12*, 298.

Frankel, A.D., and Young, J.A. (1998). HIV-1: fifteen proteins and an RNA. Annu Rev Biochem *67*, 1-25.

Gavin, A.C., Aloy, P., Grandi, P., Krause, R., Boesche, M., Marzioch, M., Rau, C., Jensen, L.J., Bastuck, S., Dumpelfeld, B.*, et al.* (2006). Proteome survey reveals modularity of the yeast cell machinery. Nature *440*, 631-636.

He, N., Liu, M., Hsu, J., Xue, Y., Chou, S., Burlingame, A., Krogan, N.J., Alber, T., and Zhou, Q. (2010). HIV-1 Tat and host AFF4 recruit two transcription elongation factors into a bifunctional complex for coordinated activation of HIV-1 transcription. Mol Cell *38*, 428-438.

Hinnebusch, A.G. (2006). eIF3: a versatile scaffold for translation initiation complexes. Trends Biochem Sci *31*, 553-562.

Ho, Y., Gruhler, A., Heilbut, A., Bader, G.D., Moore, L., Adams, S.L., Millar, A., Taylor, P., Bennett, K., Boutilier, K.*, et al.* (2002). Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry. Nature *415*, 180-183.

Jäger, S., Gulbahce, N., Cimermancic, P., Kane, J., He, N., Chou, S., D'Orso, I., Fernandes, J., Jang, G., Frankel, A.D.*, et al.* (2010). Purification and characterization of HIV-human protein complexes. Methods *53*, 13-19.

König, R., Zhou, Y., Elleder, D., Diamond, T.L., Bonamy, G.M., Irelan, J.T., Chiang, C.Y., Tu, B.P., De Jesus, P.D., Lilley, C.E.*, et al.* (2008). Global analysis of host-pathogen interactions that regulate early-stage HIV-1 replication. Cell *135*, 49-60.

Krogan, N.J., Cagney, G., Yu, H., Zhong, G., Guo, X., Ignatchenko, A., Li, J., Pu, S., Datta, N., Tikuisis, A.P.*, et al.* (2006). Global landscape of protein complexes in the yeast Saccharomyces cerevisiae. Nature *440*, 637-643.

Malim, M.H., and Emerman, M. (2008). HIV-1 accessory proteins--ensuring viral survival in a hostile environment. Cell Host Microbe *3*, 388-398.

Ruepp, A., Waegele, B., Lechner, M., Brauner, B., Dunger-Kaltenbach, I., Fobo, G., Frishman, G., Montrone, C., and Mewes, H.W. (2009). CORUM: the comprehensive resource of mammalian protein complexes--2009. Nucleic Acids Res *38*, D497-501.

Schilling, O., and Overall, C.M. (2008). Proteome-derived, database-searchable peptide libraries for identifying protease cleavage sites. Nat Biotechnol *26*, 685-694.

Shapira, S.D., Gat-Viks, I., Shum, B.O., Dricot, A., de Grace, M.M., Wu, L., Gupta, P.B., Hao, T., Silver, S.J., Root, D.E.*, et al.* (2009). A physical and regulatory map of host-influenza interactions reveals pathways in H1N1 infection. Cell *139*, 1255-1267.

Siridechadilok, B., Fraser, C.S., Hall, R.J., Doudna, J.A., and Nogales, E. (2005). Structural roles for human translation factor eIF3 in initiation of protein synthesis. Science *310*, 1513-1515.

Sowa, M.E., Bennett, E.J., Gygi, S.P., and Harper, J.W. (2009). Defining the human deubiquitinating enzyme interaction landscape. Cell *138*, 389-403.

Stark, C., Breitkreutz, B.J., Reguly, T., Boucher, L., Breitkreutz, A., and Tyers, M. (2006). BioGRID: a general repository for interaction datasets. Nucleic Acids Res *34*, D535-539.

Stelzl, U., Worm, U., Lalowski, M., Haenig, C., Brembeck, F.H., Goehler, H., Stroedicke, M., Zenkner, M., Schoenherr, A., Koeppen, S.*, et al.* (2005). A human protein-protein interaction network: a resource for annotating the proteome. Cell *122*, 957-968.

Strack, P.R., Frey, M.W., Rizzo, C.J., Cordova, B., George, H.J., Meade, R., Ho, S.P., Corman, J., Tritch, R., and Korant, B.D. (1996). Apoptosis mediated by HIV protease is preceded by cleavage of Bcl-2. Proc Natl Acad Sci U S A *93*, 9571-9576.

Tarassov, K., Messier, V., Landry, C.R., Radinovic, S., Serna Molina, M.M., Shames, I., Malitskaya, Y., Vogel, J., Bussey, H., and Michnick, S.W. (2008). An in vivo map of the yeast protein interactome. Science *320*, 1465-1470.

Valente, S.T., Gilmartin, G.M., Mott, C., Falkard, B., and Goff, S.P. (2009). Inhibition of HIV-1 replication by eIF3f. Proc Natl Acad Sci U S A *106*, 4071-4078.

Yaffe, M.B., Rittinger, K., Volinia, S., Caron, P.R., Aitken, A., Leffers, H., Gamblin, S.J., Smerdon, S.J., and Cantley, L.C. (1997). The structural basis for 14-3-3:phosphopeptide binding specificity. Cell *91*, 961-971.

Yeung, M.L., Houzet, L., Yedavalli, V.S., and Jeang, K.T. (2009). A genome-wide short hairpin RNA screening of jurkat T-cells for human proteins contributing to productive HIV-1 replication. J Biol Chem *284*, 19463-19473.

Yu, H., Braun, P., Yildirim, M.A., Lemmens, I., Venkatesan, K., Sahalie, J., Hirozane-Kishikawa, T., Gebreab, F., Li, N., Simonis, N.*, et al.* (2008). High-quality binary protein interaction map of the yeast interactome network. Science *322*, 104-110.

Zhou, H., Xu, M., Huang, Q., Gates, A.T., Zhang, X.D., Castle, J.C., Stec, E., Ferrer, M., Strulovici, B., Hazuda, D.J.*, et al.* (2008a). Genome-scale RNAi screen for host factors required for HIV replication. Cell Host Microbe *4*, 495-504.

Zhou, M., Sandercock, A.M., Fraser, C.S., Ridlova, G., Stephens, E., Schenauer, M.R., Yokoi-Fong, T., Barsky, D., Leary, J.A., Hershey, J.W.*, et al.* (2008b). Mass spectrometry reveals modularity and a complete subunit interaction map of the eukaryotic translation factor eIF3. Proc Natl Acad Sci U S A *105*, 18139-18144.

Chapter 4

Determining architectures of macromolecular assemblies using quantitative genetic interactions

Determining architectures of macromolecular assemblies using quantitative genetic interactions

Peter Cimermancic[1,2], Riccardo Pellarin[1], Hannes Braberg[3], Anthony Shiver[4,5], Richard Alexander[3], Dina Schneidman[1], James S. Fraser[1], Carol Gross[5,6,8], Nevan J. Krogan[3,8*], Andrej Sali[1,8*]

Affiliations:

1 Departments of Bioengineering and Therapeutic Sciences,

2 Graduate Group in Bioinformatics,

3 Cellular and Molecular Pharmacology,

4 Graduate Group in Biophysics,

5 Cell and Tissue Biology

6 Microbiology and Immunology

7 Pharmaceutical Chemistry, University of California, San Francisco, California, USA;

8 California Institute for Quantitative Biosciences, QB3, San Francisco, California, USA;

*Corresponding authors:

1700 4th Street, Byers Hall 308D, University of California, San Francisco, San Francisco, CA 94158; tel 415-476-2980; nevan.krogan@ucsf.edu.

1700 4th Street, Byers Hall 503B, University of California, San Francisco, San Francisco, CA 94158; tel 415-514-4227; web http://salilab.org; sali@salilab.org.

Summary

To understand the workings of a living cell, we need to know the structures of its macromolecular assemblies. Determining these structures has required pure samples of the studied assembly. Here, we present an alternative strategy based on *in vivo* measurements of genetic interactions between the assembly proteins. We show that genetic interactions can be sufficient to define the molecular architecture of an assembly, and are thus comparable in their utility to a sparse set of chemical cross-links.

Introduction

A mechanistic understanding of the cell requires structural characterization of the thousands of its constituent biological assemblies (Alber et al., 2008). So far, conventional approaches have provided a valuable but limited window into the structures of these assemblies. For example, X-ray crystallography, nuclear magnetic resonance (NMR) spectroscopy, and electron microscopy (EM) can resolve the atomic details of individual proteins and small complexes. However, it is often difficult to determine structural data, especially when the assembly is difficult to isolate and purify in sufficient quantities, which is generally the case for weak and transient complexes, such as, membrane assemblies and transient assemblies involved in signaling pathways (Herzog et al., 2012). Moreover, even when the structure is determined, it may contain artifacts. Therefore, direct *in vivo* measurements of structural aspects of a wild-type assembly are needed.

Alternative approaches have emerged recently to determine the structures of macromolecules by considering multiple types of spatial information, including restraints from bioinformatics analyses, such as statistical potentials and evolutionary covariance (Marks et al., 2011; Ovchinnikov et al., 2014), structures of assembly components determined by traditional methods, chemical cross-links, other proteomics data, electron microscopy density maps, and small-angle scattering profiles (Alber et al., 2007a; Alber et al., 2008; Russel et al., 2012; Ward et al., 2013). This information is often limited by low resolution, accuracy, and quantity; however, these shortcomings can be minimized by integration of all available information about a given assembly (Alber et al., 2008). The integrative modeling cycles through four stages: gathering information about the structure of the assembly, choosing how to represent the system and how to translate the information into spatial restraints, calculating an ensemble of structures that satisfy these restraints, and analyzing the ensemble (**Figure 1**). Examples of integrative structures include nuclear pore complex (Alber et al., 2007b), yeast 26S proteasome (Lasker et al., 2012), bacterial type III secretion needle (Loquet et al., 2012), and a three-dimensional model of the yeast genome (Duan et al., 2010).

Results and Discussion

Here, we describe how integrative structure determination can benefit critically from spatial restraints computed from the correlation between the functional impacts of two mutations and their distance. This correlation reflects that site-directed mutagenesis of residues within the same functional region (*eg*, active sites, allosteric sites, and protein-protein binding interfaces) is likely to cause more similar phenotypes than mutagenesis of sites that are distant in space (Halabi et al., 2009; Marks et al., 2011). Therefore, phenotypic similarity of a pair of point mutations can be informative about the 3D structure of a macromolecular assembly.

The point-mutant epistatic miniarray profiling (pE-MAP) approach was recently suggested to measure such correlation between structural proximity and phenotypic similarity (Braberg et al., 2013). In pE-MAP, each point mutation in the target macromolecule is crossed against a gene deletion mutant allele, followed by measuring the growth phenotype of the resulting double mutant allele. The genetic interaction between the point mutant and gene deletion mutant is then quantified by comparing this growth phenotype to growth phenotypes of the two individual single mutant alleles (Collins et al., 2010). An array of growth phenotypes, termed a phenotypic profile, is generated for each point mutation crossed against multiple single gene deletions (**Figure 1**); growth phenotypes can be collected in a high-throughput fashion for hundreds of point mutations against thousands of gene deletion alleles of yeast and mammalian cells. In addition, similar phenotypic profiles can be obtained by pairing point mutations with a number of assays that measure a growth phenotype under different conditions (*eg*, presence of antibiotics and other chemicals, change in temperature or pH, and UV radiation), and by measuring phenotypes other than the growth rate (*eg*, nucleus size, number of mitochondria, and response to chemicals).

pE-MAP data is based on phenotypic (hence indirect) structural observations, and, therefore, subject not only to random noise originating from compositional and structural heterogeneities of a macromolecular assembly, but also to random and systematic noise originating from complex cellular networks; for example, distant positions that are part of an allosteric network, mutations at non-functional sites, functionally irrelevant mutations, "destructive" mutations (*ie*, mutations that lead to misfolding of a macromolecule or those that are highly harmful to the cell), and mutations that perturb the gene, mRNA, and their interactions, but not the structure (mutations that perturb expression, translation, or stability of corresponding nucleic acids). While some of these sources of noise could be minimized during site-directed mutagenesis (*eg*, by only selecting point mutations that cause a measurable phenotype, but do not harm the cell too much), many will have to be considered through the construction of spatial restraints and subsequent modeling.

Here, we quantify the utility of pE-MAP data for structure determination by reconstructing the known molecular architecture of RNA polymerase II (RNAPII) (Hahn, 2004; Wang et al., 2006; Ward et al., 2013) based on a previously determined pE-MAP dataset (Braberg et al., 2013), using our open-source *Integrative Modeling Platform* (IMP) package (Russel et al., 2012). The dataset includes quantitative genetic interactions between 53 single point mutants in RNAPII genes and a library of ~1,200 gene-deletion alleles (Braberg et al., 2013). Due to the limited distribution of point mutations in the pE-MAP dataset, a sufficient number of point mutations was only available for subunits Rpb1 and Rpb2; therefore, we focused on modeling the sub-complex of these two subunits. To realistically mimic application of the proposed approach, we split the Rpb1 subunit into two parts, resulting into three components, and used comparative models of the components rather than their X-ray structures.

We encoded the pE-MAP data into a scoring function that restrains the distance between a pair of mutated residues (**Methods**). The similarity between a pair of pE-MAP profiles (*ie*, pE-

MAP link) was computed using maximum information coefficient (MIC) (Reshef et al., 2011). Although MIC values correlate poorly with distances between the mutated positions (correlation coefficient of -0.32), the maximum distances between a pair of phenotypic profiles and MIC values tend to be inversely proportional (**Figure 2a**). In other words, a high MIC value is more likely obtained for a pair of point mutations that are close in structure. Most of the phenotypic profiles, even those that correspond to positions less than 30 Å apart, are highly dissimilar (68% of pairs with MIC lower than 0.3). The lack of high MIC values for a pair of proximal point mutations does not seem to depend on the nature of chemical change caused by mutations (**Figure S1a**); an average BLOSUM62 score of a pair of point mutations does not depend on that pair's MIC value (correlation coefficient of 0.03 between MIC values and the average BLOSUM62 scores for pairs of point mutations within 30 Å of each other). This noise also cannot be explained by structural heterogeneity in the polymerase, such as conformational changes in the trigger loop (**Figure S1b**). These observations justify imposing an upper bound on a distance between two mutations that depends on the MIC value (**Figure 1** and **2**, **Methods**).

Next, we obtained many configurations of the 3 components of the system that minimize the violations of the pE-MAP restraints and the overlap between the components, by using exhaustive Monte Carlo sampling starting with random initial configurations. Finally, we evaluated the accuracy of the resulting models by calculating C-atom RMSD between the C-terminal half of Rpb1 and Rpb2 of the top-scoring models and the native structure superposed on N-terminal half of Rpb1. We defined a successful prediction when a model has the RMSD value lower than 30 Å (corresponding to 1.5 times the shortest pE-MAP restraint). While the resulting models do not reveal the atomic features of the interfaces between the components, even a coarse characterization can be useful for studying evolution and function as well as for a higher-resolution structure determination (Alber et al., 2008).

We find that the pE-MAP dataset can successfully determine the structure of the RNAPII sub-complex. Our approach predicts the architecture of RNAPII with an RMSD error of 27.8 2.5 Å (**Figure 3**), which is significantly better than models computed by the standard protein-protein docking techniques (RMSD error of 61.2 16.8 Å). Moreover, the accuracy of the model is also significantly better than that of models computed by our approach but based on bootstrapped subsamples of the original dataset (56.6 11.6 Å using 26% of the data points), or based on a dataset with randomly shuffled pE-MAP links and the corresponding MIC values (50.1 10.3 Å). To appreciate the value of the pE-MAP data, we compared it with a previously published cross-linking dataset (Chen et al., 2010). Cross-linking is widely used for structural determination of macromolecular assemblies (Herzog et al., 2012; Lasker et al., 2012). Unexpectedly, we find that the pE-MAP dataset determines the structure of RNAPII as accurately as the cross-linking dataset (RMSD of 26.2 1.1 Å; **Figure 3**). Moreover, the accuracy of the model improves even further if both data types are combined (RMSD of 22.4 1.2 Å), indicating complementarity between the two datasets and demonstrating the premise of integrative structure determination (**Figure S2**). In principle, cross-links should carry more structural information than a pE-MAP link (the distance restraint for a cross-link was set to a constant 12 Å), and should therefore result in more accurate models. However, the number of possible cross-links is limited by the number of proximal lysine pairs (20 in our case), whereas the number of pE-MAP pairs grows quadratically with every new point mutation introduced (in our case, 98 pairs with MIC larger than 0.3 from 44 point mutations). Therefore, the larger number of coarse distance restraints can lead to models as accurate as those obtained by finer but sparse cross-link restraints. This observation agrees with our previous work (Erzberger et al., Submitted).

There are at least three caveats in our approach: First, modeling accuracy was assessed with the aid of the assembly used for fitting the parameters of the spatial restraint function; therefore, it also needs to be tested by modeling other assemblies. Second, the components are large (over 1,200 residues per subunit) and it needs to be determined whether or not the success of the approach depends on the protein size. Third, constructing large gene deletion allele libraries and crossing them against point mutations is laborious. Therefore, alternative approaches to pE-MAP design, such as pairing point mutant alleles with chemicals and physical perturbations, need to be explored. To address these three points, we have modeled two additional assemblies based on unpublished pE-MAP datasets: (i) 253 point mutations in H3 and H4 histones paired with ~1,350 gene deletion alleles and (ii) 49 point mutations in subunits RpoB and RpoC of a bacterial RNA polymerase subject to 139 different conditions (*eg*, treatments with chemicals and temperature shocks).

Importantly, the negative correlation between the MIC value and distance upper bound is also apparent in the histone dataset, indicating that our structural interpretation of the pE-MAP data can be applied to different macromolecular assemblies (**Figure 2b**). However, because histones H3 and H4 are small proteins (the length of the second largest axis almost falls within the pE-MAP dataset resolution; ~30 Å), in principle, multiple solutions around the longest axis should satisfy the restraints, resulting in a model of low precision and accuracy. Indeed, the models were not statistically significantly different from those based on the random shuffling of the dataset (21.3   3.4 Å; *P* = 0.78 based on Student's t-test). Thus, the size of the system restrained by the pE-MAP data should be larger than ~30 Å in all dimensions.

We have also assessed the pE-MAP restraint by predicting the architecture of bacterial RNA polymerase based on a "non-standard" dataset for which phenotypes of point mutant alleles were measured under different perturbations (*eg,* antibiotics, change in temperature, and pH). The MIC-distance graph is noisier (about 10% of data points violate the pE-MAP restraint) than those based on the RNAPII and histone datasets. We attribute this noise mostly to the low number of non-redundant perturbations in the dataset; hierarchical clustering of the perturbations revealed only 10 major clusters (data not shown). The pE-MAP dataset, however, was still informative. Both constant distance restraints of 90 Å and random shuffling of the data points resulted in models of significantly lower accuracy than those computed with pE-MAP data (RMSD errors of 73.4 28.2 Å and 71.0 16.2 Å, respectively, compared to 25.6 13.5 Å; **Figure S4**). This result suggests that a relatively small number of orthogonal phenotypes per point mutation can be used to accurately predict the architecture of a macromolecular assembly; for example, 50 representative cluster members of more than 1,000 available gene-deletion alleles were used to calculate MIC values.

In summary, we show that the architectures of macromolecular assemblies can be determined using quantitative genetic interaction data. Remarkably, the precision and accuracy of such modeling are comparable to those of models based on chemical cross-linking. A key advantage is that pE-MAP data is determined *in vivo*, and can thus inform structures of macromolecular assemblies that are difficult or impossible to isolate and purify. Moreover, the approach can also reveal functionally coupled sites that are distant in structure, even when the structure is unknown. Finally, in addition to protein assemblies, the method can be applied to assemblies of nucleic acids and proteins and nucleic acid, thus significantly expanding the scope of integrative structural biology.

Acknowledgments

We thank members of the Sali lab for stimulating discussion. This research was funded by grants from the National Institutes of Health (P50 GM082250 to NJK). P.C. is supported by a Howard Hughes Predoctoral Fellowship; J.S.F. is supported by the National Institutes of Health (DP5 OD009180, R21 GM110580, P30 DK063720) and National Science Foundation (STC-1231306); A.S. is supported by the National Institutes of Health (R01 GM083960, U54 RR022220, U54 GM094662, P01 AI091575, and U01 GM098256).

Figures

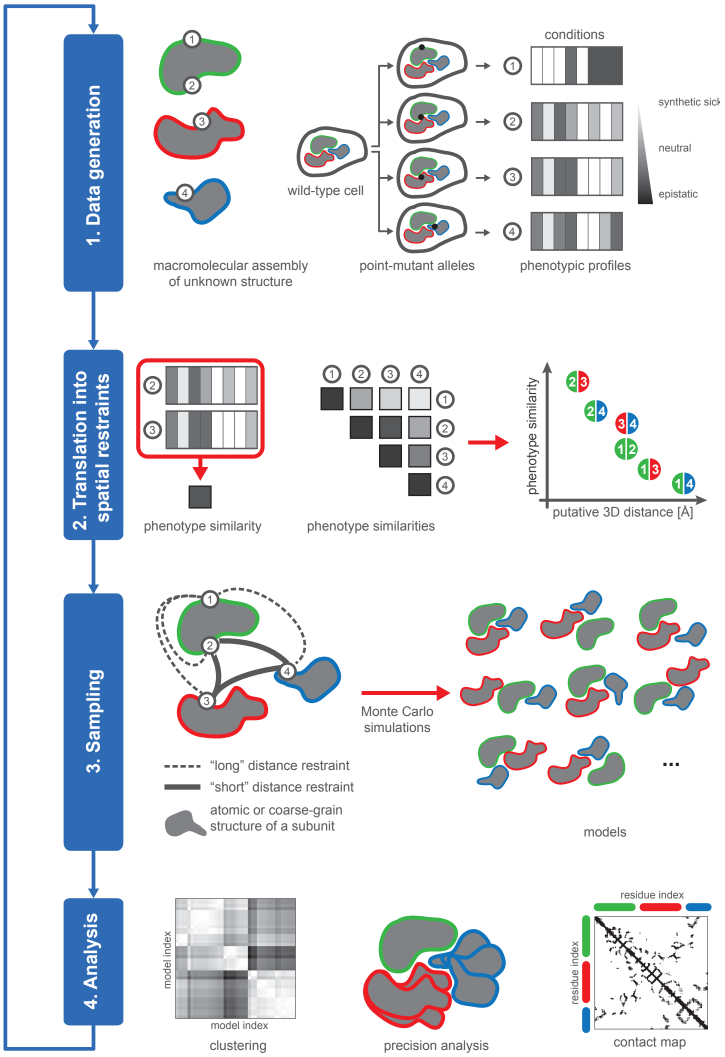


Figure 1: Modeling based on pE-MAP data. First, pE-MAP data is generated by site-directed mutagenesis of genes encoding the subunits in a macromomolecular assembly of interest. These mutants are then crossed against a series of alleles under different conditions (*eg*, chemical perturbation and gene deletions), followed by measurement of the respective phenotypes. Second, the raw phenotypic profiles are translated into spatial restraints by comparing all pairs of phenotypic profiles; a pair of sites with similar phenotypic profiles is

expected to be close in the structure of the assembly. Third, many models are generated by simultaneously minimizing the violations of all pE-MAP and other restraints. Fourth, the ensemble of structural models is clustered, and each cluster analyzed in terms of precision, accuracy, contacts, geometry, and violations of restraints.

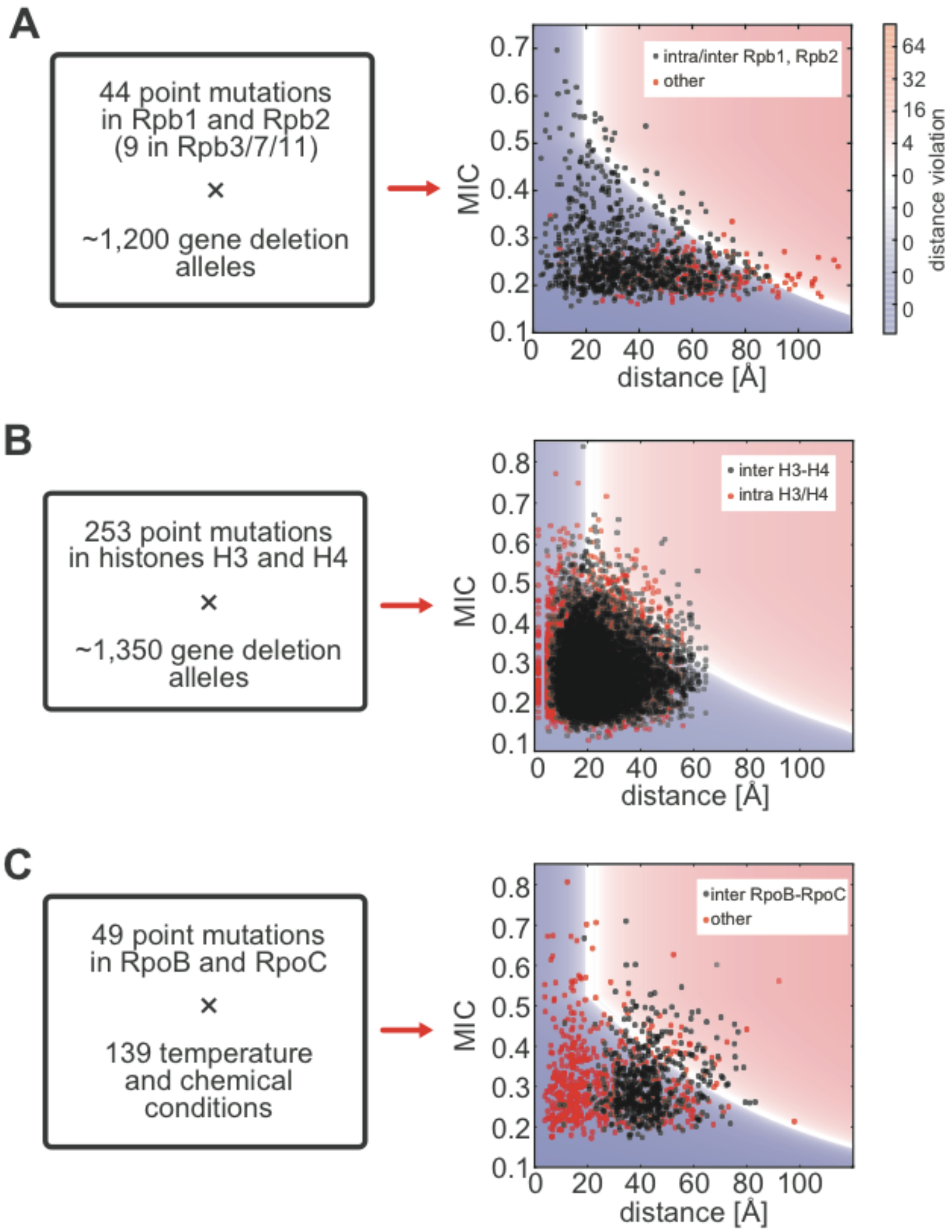


Figure 2: Phenotypic similarity vs. distance plots for 3 assemblies of known structure. pE-MAP datasets were generated for yeast RNAPII complex (**a**) and histones H3 and H4 (**b**), as well as for bacterial RNA polymerase (**c**). Maximal information coefficient (MIC) measure was used to calculate the phenotypic similarities. The pE-MAP distance restraint is represented as a blue-red color map.

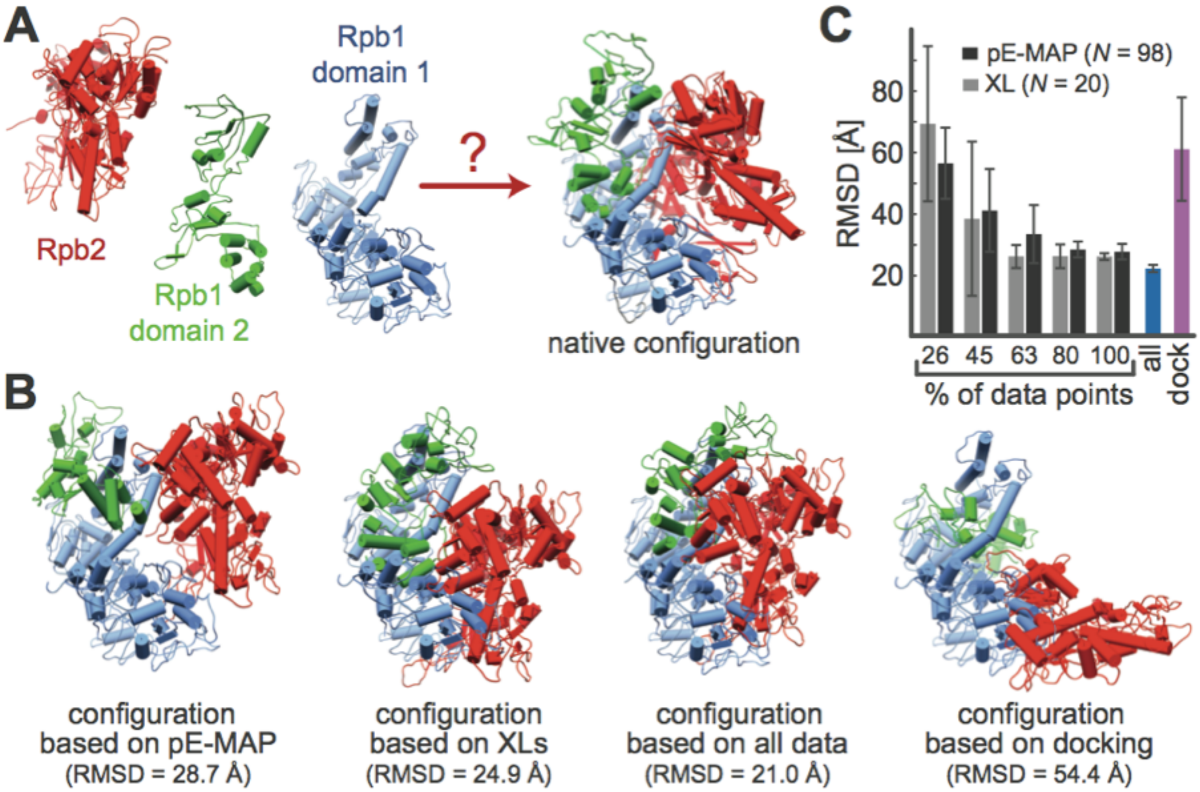


Figure 3: Model of RNAPII sub-complex. (**a**) Three component comparative models were used to represent the sub-complex of Rpb1 and Rpb2 subunits (Methods). (**b**) The panel shows four configurations of the sub-complex based on the pE-MAP data, cross-linking data (XL), the combination of the pE-MAP and cross-linking data, and protein-protein docking. (**c**) The bar chart shows the accuracies of the modeling based on different types and subsets of data points.

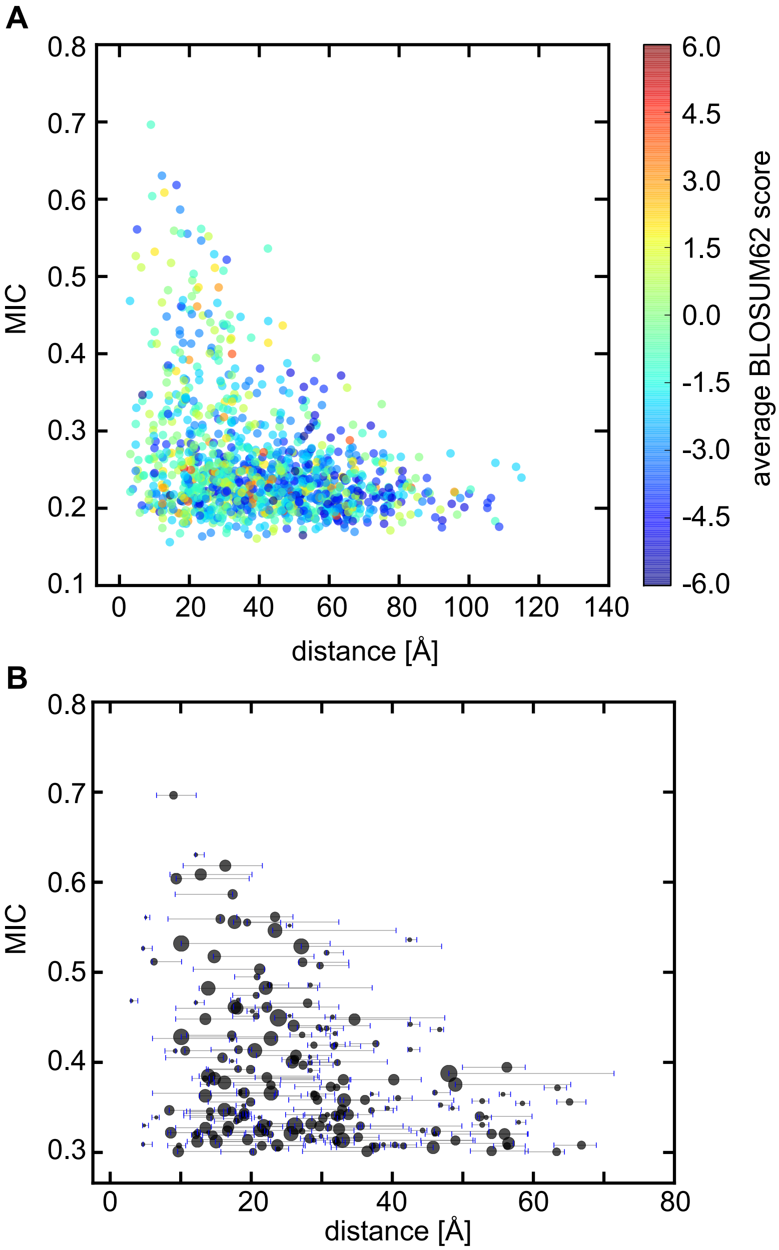


Figure S1: Phenotypic similarity values do not correlate with sequence similarities or structural dynamics. (**a**) The panel shows the phenotypic similarity vs. distance plot based on the yeast RNAPII dataset, with data points colored by the average sequence similarity between a pair of wild-type residues and the respective point mutations. (**b**) This panel also shows the phenotypic similarity vs. distance plot based on the yeast RNAPII dataset (with MIC cutoff of 0.3). The size and ends of an error-bar of each data point denote respectively the average and

the minimum and maximum distances between the pair of point mutations among 97 different structures from the PDB.

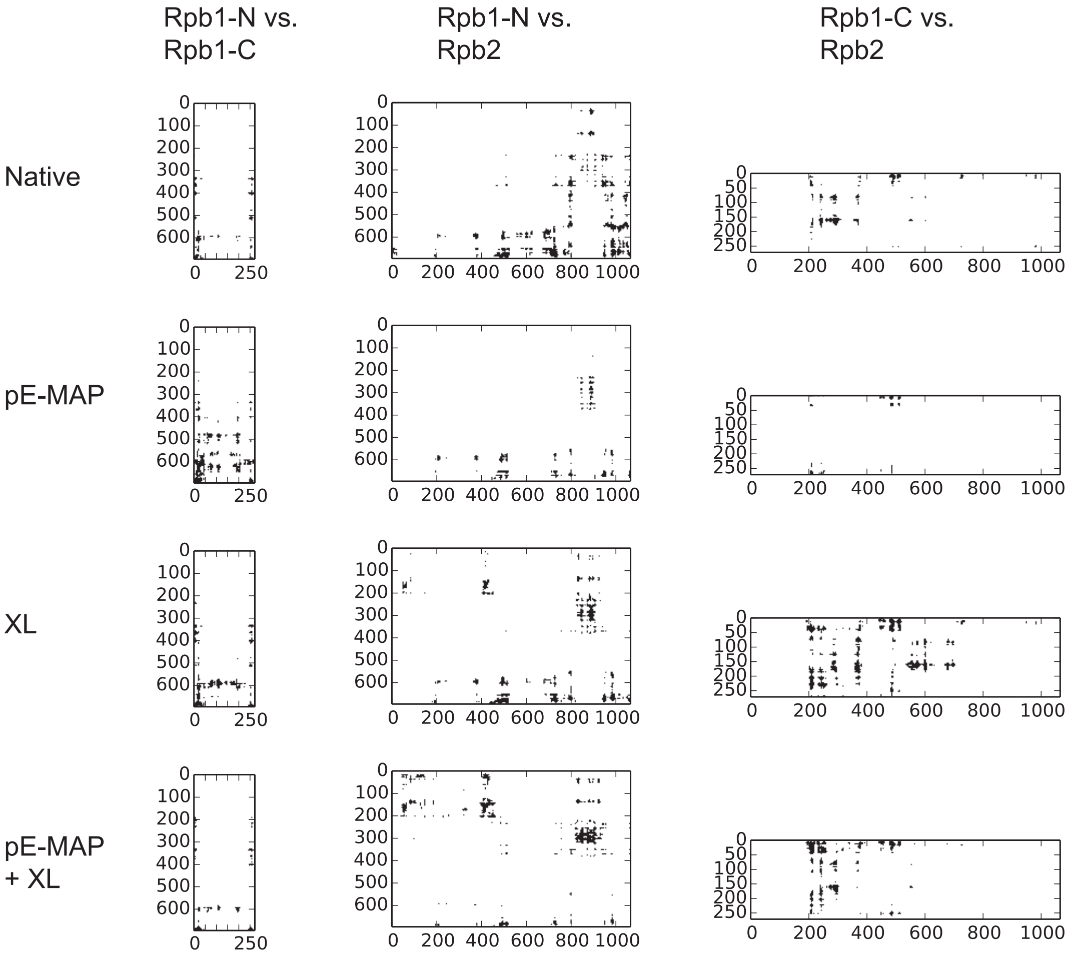


Figure S2: Contact maps. The contacts between C atoms that are less than 25 Å apart are shown for all pairs of subunits.

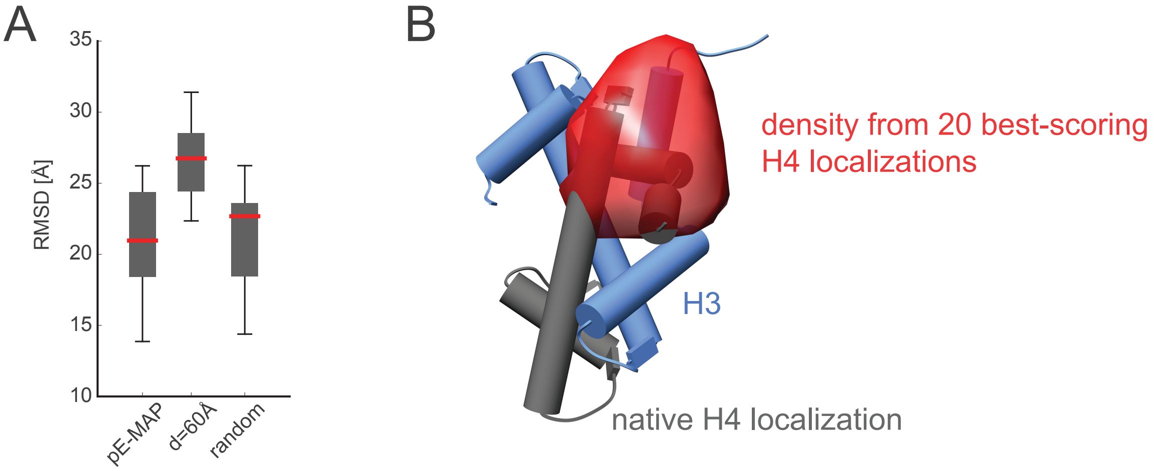


Figure S3: Model of yeast H3-H4 histones interaction. The H3 and H4 histones were represented as two rigid bodies based on their crystal structure (PDB ID: 1id3). (**a**) The chart shows the accuracy as RMSD values of the 20 best-scoring models based on the pE-MAP data, a constant distance restraint, and randomized pE-MAP data. The red lines denote sample medians, whereas the grey boxes and whiskers extend from the lower to upper quartile values of the data, and to the most extreme points within the 1.5th multiple of the lower and upper quartiles, respectively. (**b**) The panel shows the structure of the native complex (blue and grey), overlaid with a density map of H4 (red) from the localizations of the 20 best-scoring models.

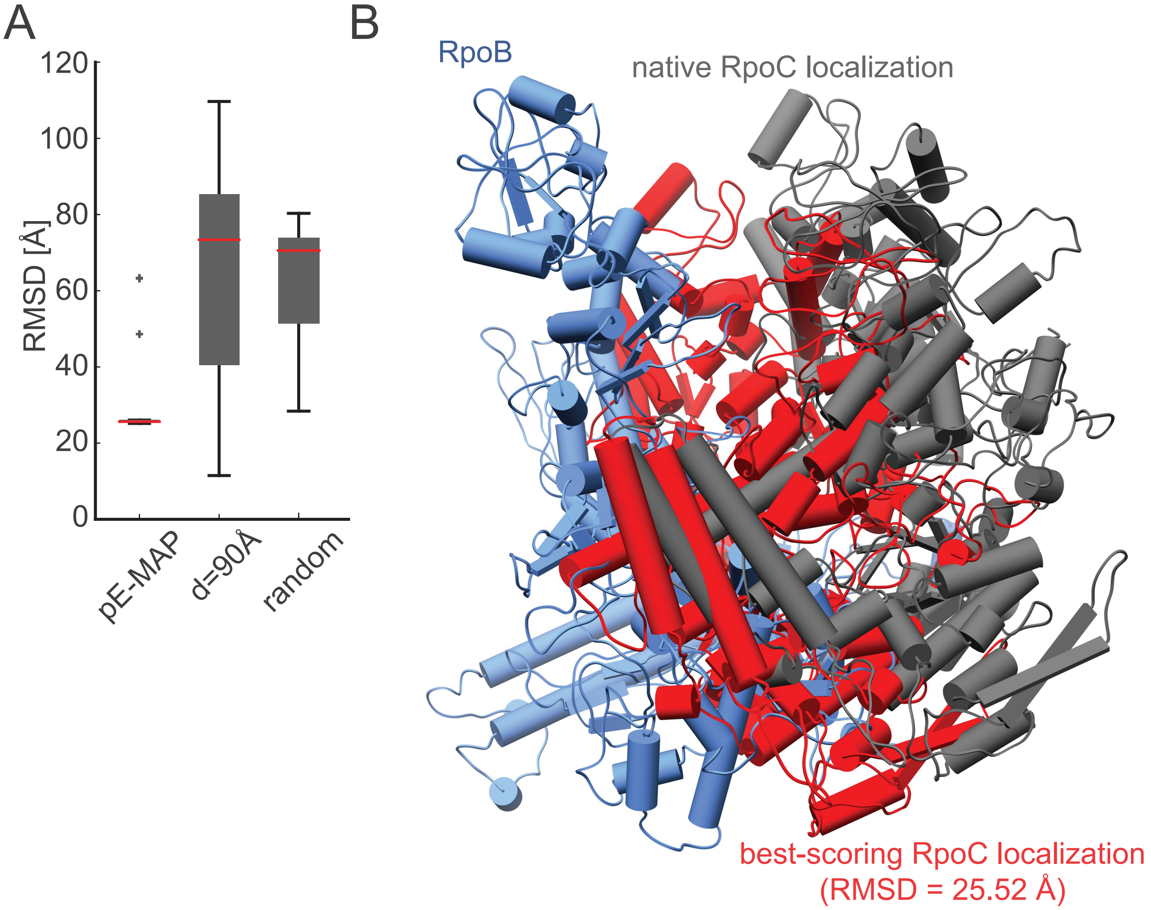


Figure S4: Model of RpoB and RpoC interaction from bacterial RNAP. The subunits were represented as two rigid components based on their crystal structure (PDB ID: 4igc). (**a**) The chart shows the accuracy as RMSD errors of the 20 best-scoring models based on the pE-MAP data, a constant distance restraint, and randomized pE-MAP data. The red lines denote sample medians, whereas the grey boxes and whiskers extend from the lower to upper quartile values of the data, and to the most extreme points within the 1.5th multiple of the lower and upper quartiles, respectively. (**b**) The panel shows the structure of the native complex (blue and grey) and location of RpoC in the best-scoring model (red).

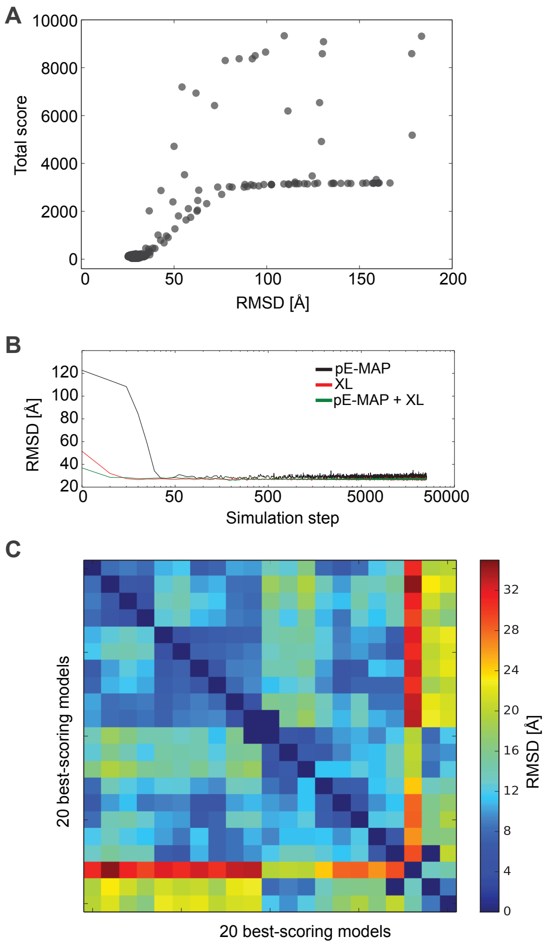


Figure S5: Sampling convergence. The thoroughness of sampling was assessed by the analysis of the RMSD-score landscape (**a**), convergence of RMSD error during sampling (**b**), and clustering of the best-scoring models based on RMSD between all the pairs of models (**c**).

Methods

PE-MAP DATA GENERATION

The pE-MAP datasets were generated as described previously (Braberg et al., 2013). The yeast RNAPII pE-MAP dataset has been previously published (Braberg et al., 2013) and is available in GEO database, under series GSE47429. The other two datasets are not published yet (cite all authors, in preparation; separately for the two datasets).

DESIGN OF PE-MAP SPATIAL RESTRAINT

The distance restraint was designed as follows: First, the gene deletion alleles with E-MAP values larger than 7 against at least one point mutation were removed from the pE-MAP dataset. The remaining data were then clustered by gene deletion alleles using K-mean clustering (Pedregosa et al., 2011) with K = 50 (or K = 10 for the bacterial RNAP dataset), and a random representative gene deletion allele was selected from each cluster; other K values were tested as well, but no significant improvements in correlation between the maximum information coefficient (MIC) (Reshef et al., 2011) values and 3D distances were observed for K larger than 50 (data not shown). The missing values in the pE-MAP dataset were imputed as the mean of the corresponding gene deletion allele E-MAP values. Second, a similarity between a pair of point mutation phenotypic profiles was calculated using MIC; Pearson product-moment correlation coefficient was also tested, but it did not improve the accuracy of the modeling. Third, the upper bound distance threshold was computed from the corresponding MIC values, based on fitting the data points with the largest MIC value from each of 10 distance bins:

$$d = \begin{cases} \frac{log(MIC)-n}{k} & \text{if } MIC < 0.65 \\ 20 & \text{if } MIC \geq 0.65 \end{cases}$$

where *d* is the upper bound distance threshold between the surfaces of a pair of C atoms, and *k* and *n* are -0.0127 and -0.3861, respectively. Fourth, an upper-bound distance restraint for each MIC value is defined as a harmonic function centered at the corresponding distance threshold from the previous step. The harmonic scoring function is truncated to disregard point mutant pairs whose MIC values are much higher than expected:

$$s = \begin{cases} 225w + \frac{w(d+15)^2 - b}{x+o} & \text{if } x < d+15 \\ w(x-d)^2 & \text{if } d < x \leq d+15 \\ 0 & \text{if } x \leq d \end{cases}$$

where *s*, *x*, and *w* are the restraint score, a surface-to-surface distance between a pair of C atoms, and a restraint weight (set to 1 in our case), respectively, and *b* and *o* are constants chosen to make the function smooth and continuous. To improve computational efficiency, we only considered point mutation pairs with MIC values larger than 0.3. Cross-linking data were also transformed into a truncated upper-bound distance restraints using harmonic function centered on a constant upper bound distance threshold of 12 Å.

DATA REPRESENTATION

The Rpb1 and Rpb2 subunits of the yeast RNAPII were represented as three rigid bodies consisting of C atoms, by splitting the structure of Rpb1 into two parts (residues 12-837 and residues 1068-1379). Moreover, instead of using the experimentally determined structures of the three particles, we built the respective comparative models based on templates that are all less than 45% identical in sequence; BLAST (Altschul et al., 1990) was used for alignment and Modeller (Sali and Blundell, 1993) for model building; the template PDB IDs were 2y0s (Rpb1 12-837) and 3iyd (Rpb1 1068-1379) and 2pmz (Rpb2). The C-atom RMSDs between the crystal structures and the comparative models range between 2.8 and 5.5 Å. Crystal structures were used for modeling the other two sub-complexes (Figures S2 and S3).

SAMPLING AND ANALYSIS

The scoring function was defined as a sum of the pE-MAP (or cross-linking) distance restraints and the excluded volume restraints (Lasker et al., 2012). To sample good-scoring models, we used the Monte Carlo algorithm with simulated annealing (Kirkpatrick et al., 1983). The structures of the subunits, represented by their C-atoms, were kept rigid during sampling. For each set of restraints, we ran in parallel 20 optimizations, each starting with random initial positions and orientations. Each optimization consisted of 25,000 Monte Carlo steps, cycling between 50 steps at temperature 3.0 and 10 steps at temperature 10.0. The Monte Carlo steps included random translation and rotation of rigid components (drawn randomly from uniform distributions limited to 2 Å and 0.3 radians, respectively). The accuracy of the resulting 0.5 million models was analyzed by computing C-atom RMSDs between 20 top-scoring models and the native configuration

The landscape of the scoring function including the pE-MAP restraints is funnel-shaped, with a single pronounced minimum at RMSD error of ~27 Å (**Figure S5a**). The sampling is sufficient; an average simulation converges to the 95th percentile of the best score in approximately 50 of 25,000 Monte Carlo steps (**Figure S5b**). The precisions of the best-scoring conformations from 20 different runs based on the pE-MAP, cross-linking, and both datasets are 13.5 Å (**Figure S5c**), 4.9 Å, and 3.9 Å, respectively.

References

Alber, F., Dokudovskaya, S., Veenhoff, L.M., Zhang, W., Kipper, J., Devos, D., Suprapto, A., Karni-Schmidt, O., Williams, R., Chait, B.T.*, et al.* (2007a). Determining the architectures of macromolecular assemblies. Nature *450*, 683-694.

Alber, F., Dokudovskaya, S., Veenhoff, L.M., Zhang, W., Kipper, J., Devos, D., Suprapto, A., Karni-Schmidt, O., Williams, R., Chait, B.T.*, et al.* (2007b). The molecular architecture of the nuclear pore complex. Nature *450*, 695-701.

Alber, F., Forster, F., Korkin, D., Topf, M., and Sali, A. (2008). Integrating diverse data for structure determination of macromolecular assemblies. Annual review of biochemistry *77*, 443-477.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990). Basic local alignment search tool. Journal of molecular biology *215*, 403-410.

Braberg, H., Jin, H., Moehle, E.A., Chan, Y.A., Wang, S., Shales, M., Benschop, J.J., Morris, J.H., Qiu, C., Hu, F.*, et al.* (2013). From structure to systems: high-resolution, quantitative genetic analysis of RNA polymerase II. Cell *154*, 775-788.

Chen, Z.A., Jawhari, A., Fischer, L., Buchen, C., Tahir, S., Kamenski, T., Rasmussen, M., Lariviere, L., Bukowski-Wills, J.C., Nilges, M.*, et al.* (2010). Architecture of the RNA polymerase II-TFIIF complex revealed by cross-linking and mass spectrometry. The EMBO journal *29*, 717-726.

Collins, S.R., Roguev, A., and Krogan, N.J. (2010). Quantitative genetic interaction mapping using the E-MAP approach. Methods in enzymology *470*, 205-231.

Duan, Z., Andronescu, M., Schutz, K., McIlwain, S., Kim, Y.J., Lee, C., Shendure, J., Fields, S., Blau, C.A., and Noble, W.S. (2010). A three-dimensional model of the yeast genome. Nature *465*, 363-367.

Erzberger, J.P., Stengel, F., Pellarin, R., Zhang, S., Schaefer, T., Aylett, C.H.S., Cimermancic, P., Boehringer, D., Sali, A., Aebersold, R.*, et al.* (Submitted). Molecular architecture of the 40S-eIF1-eIF3 translation initiation complex.

Hahn, S. (2004). Structure and mechanism of the RNA polymerase II transcription machinery. Nature structural & molecular biology *11*, 394-403.

Halabi, N., Rivoire, O., Leibler, S., and Ranganathan, R. (2009). Protein sectors: evolutionary units of three-dimensional structure. Cell *138*, 774-786.

Herzog, F., Kahraman, A., Boehringer, D., Mak, R., Bracher, A., Walzthoeni, T., Leitner, A., Beck, M., Hartl, F.U., Ban, N.*, et al.* (2012). Structural probing of a protein phosphatase 2A network by chemical cross-linking and mass spectrometry. Science *337*, 1348-1352.

Kirkpatrick, S., Gelatt, C.D., Jr., and Vecchi, M.P. (1983). Optimization by simulated annealing. Science *220*, 671-680.

Lasker, K., Forster, F., Bohn, S., Walzthoeni, T., Villa, E., Unverdorben, P., Beck, F., Aebersold, R., Sali, A., and Baumeister, W. (2012). Molecular architecture of the 26S proteasome holocomplex determined by an integrative approach. Proceedings of the National Academy of Sciences of the United States of America *109*, 1380-1387.

Loquet, A., Sgourakis, N.G., Gupta, R., Giller, K., Riedel, D., Goosmann, C., Griesinger, C., Kolbe, M., Baker, D., Becker, S.*, et al.* (2012). Atomic model of the type III secretion system needle. Nature *486*, 276-279.

Marks, D.S., Colwell, L.J., Sheridan, R., Hopf, T.A., Pagnani, A., Zecchina, R., and Sander, C. (2011). Protein 3D structure computed from evolutionary sequence variation. PloS one *6*, e28766.

Ovchinnikov, S., Kamisetty, H., and Baker, D. (2014). Robust and accurate prediction of residue-residue interactions across protein interfaces using evolutionary information. eLife *3*, e02030.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., *et al.* (2011). Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research *12*, 2825–2830.

Reshef, D.N., Reshef, Y.A., Finucane, H.K., Grossman, S.R., McVean, G., Turnbaugh, P.J., Lander, E.S., Mitzenmacher, M., and Sabeti, P.C. (2011). Detecting novel associations in large data sets. Science *334*, 1518-1524.

Russel, D., Lasker, K., Webb, B., Velazquez-Muriel, J., Tjioe, E., Schneidman-Duhovny, D., Peterson, B., and Sali, A. (2012). Putting the pieces together: integrative modeling platform software for structure determination of macromolecular assemblies. PLoS biology *10*, e1001244.

Sali, A., and Blundell, T.L. (1993). Comparative protein modelling by satisfaction of spatial restraints. Journal of molecular biology *234*, 779-815.

Wang, D., Bushnell, D.A., Westover, K.D., Kaplan, C.D., and Kornberg, R.D. (2006). Structural basis of transcription: role of the trigger loop in substrate specificity and catalysis. Cell *127*, 941-954.

Ward, A.B., Sali, A., and Wilson, I.A. (2013). Biochemistry. Integrative structural biology. Science *339*, 913-915.

Chapter 5

Expanding the druggable proteome by characterization and prediction of cryptic binding sites

Expanding the druggable proteome by characterization and prediction of cryptic binding sites

Authors: Peter Cimermancic[a,b,*], Patrick Weinkam[a], Justin T. Rettenmaier[c,d], Daniel A. Keedy[a], Rahel A. Woldeyes[a,c], James A. Wells[d,e], James S. Fraser[a], Andrej Sali[a,d,*]

Affiliations:

[a] Departments of Bioengineering and Therapeutic Sciences,

[b] Graduate Group in Biological and Medical Informatics,

[c] Graduate Group in Chemistry and Chemical Biology, and

[d] Pharmaceutical Chemistry,

[e] Cellular and Molecular Pharmacology, and California Institute for Quantitative Biosciences (QB3), University of California, San Francisco, San Francisco, CA 94158.

*Corresponding authors:

1700 4th Street, Byers Hall 501A, University of California, San Francisco, San Francisco, CA 94158; tel 415-696-1455; peter.cimermancic@ucsf.edu.

1700 4th Street, Byers Hall 503B, University of California, San Francisco, San Francisco, CA 94158; tel 415-514-4227; web http://salilab.org; sali@salilab.org.

Abstract

Many proteins have small molecule-binding pockets that are not easily detectable in the ligand-free structures. These cryptic sites require a conformational change to become apparent; a cryptic site can therefore be defined as a site that forms a pocket in a *holo* structure, but not in the *apo* structure. Because many proteins appear to lack druggable pockets, understanding and accurately identifying cryptic sites could expand the set of drug targets. Previously, cryptic sites have been identified experimentally by fragment-based ligand discovery and computationally by long molecular dynamics simulations. Here, we begin by constructing a set of structurally defined *apo*-*holo* pairs with cryptic sites. Next, we comprehensively characterize the cryptic sites in terms of their sequence, structure, and dynamics attributes. We find that cryptic sites tend to be as conserved in evolution as traditional binding pockets, but are less hydrophobic and more flexible. Relying on this characterization, we also use machine learning to predict cryptic sites with relatively high accuracy (for our benchmark, the true positive and false positive rates are 73% and 29%, respectively). We then predict cryptic sites in the entire structurally characterized human proteome (11,201 structures, covering 23% of all residues in the proteome). The method increases the size of the potentially "druggable" human proteome from estimated ~40% to ~78% of disease-associated proteins. Finally, to demonstrate the utility of our approach in practice, we experimentally validate a predicted cryptic site in human protein tyrosine phosphatase 1B using a covalent ligand and NMR spectroscopy.

Introduction

Biological function often involves binding of proteins to other molecules, including small ligands and macromolecules. Usually, these interactions occur at defined binding sites in the protein structure (Nisius et al., 2012). Knowledge of binding site location has a number of applications (Campbell et al., 2003). For example, in drug discovery, binding site localization is often the starting point followed by virtual screening or *de novo* ligand design (Laurie and Jackson, 2005); in cell biology, it facilitates prediction of protein substrates, especially when the target protein cannot be reliably related to homologs of known function (Hermann et al., 2007).

Binding sites, particularly those for small molecules, are often located in exposed concave pockets, which provide an increased surface area that in turn maximizes intra-molecular interactions (Laskowski et al., 1996). A concave pocket can already exist in a ligand-free structure of a protein; such binding sites are called here binding pockets. Sometimes, however, a binding site is flat in the absence of a ligand and only forms in the presence of a ligand; such binding sites are called cryptic sites (**Figure 1a**) (Bowman and Geissler, 2012; Diskin et al., 2008; Durrant and McCammon, 2011; Horn and Shoichet, 2004; Lexa and Carlson, 2011).

Many computational methods have been developed to localize binding pockets on proteins. These methods are based on a variety of principles (Henrich et al., 2010): (*i*) concavity of the protein surface, (*ii*) energy functions including van der Waals terms, (*iii*) geometrical and physico-chemical similarity to known binding pockets, and (*iv*) composite approaches that use a combination of different features (Capra et al., 2009; Le Guilloux et al., 2009; Rossi et al., 2006). Unfortunately, only ~60% of protein structures were judged to have pockets larger than 250 Å^3 (many of which may not be druggable), and could potentially be subjected to ligand discovery based on binding pocket knowledge (Hopkins and Groom, 2002; Sheridan et al., 2010).

In contrast to binding pockets, cryptic sites are not easily detectable in a ligand-free structure of a protein because they by definition require ligand-induced conformational changes to become apparent. For example, large and flat interfaces between interacting proteins were considered undruggable, although several examples of protein interfaces undergoing a conformational change coupled with binding a small molecule were recently described (Arkin and Wells, 2004; Wells and McClendon, 2007). Similarly, allosterically regulated sites are sometimes not apparent in the absence of a small-molecule allosteric regulator (e.g., p38 MAP kinase (Diskin et al., 2008) and TEM1 -lactamase (Horn and Shoichet, 2004)).

Currently, the only two approaches to cryptic site discovery are exhaustive site-directed small-molecule tethering by experiment (Hardy and Wells, 2004; Ostrem et al., 2013; Sadowsky et al., 2011) and long time-scale molecular dynamics simulations by computation (Bowman and Geissler, 2012; Brenke et al., 2009; Durrant and McCammon, 2011; Grove et al., 2013; Lexa and Carlson, 2011), both of which are time-consuming, expensive, and not always successful. Therefore, there is a need for an accurate, automated, and efficient method to predict the location of cryptic pockets in a given ligand-free protein structure. Such a method would offer several advantages. First, a cryptic site may be the only suitable binding site on the target protein; for example, when activation is required and thus the active site cannot be targeted, the active site is not druggable, or active site ligands need to be avoided due to adverse off-target effects. Second, binding sites may be discovered on structures determined or computed at only moderate resolution.

Here, we analyze known cryptic sites and develop a method for predicting cryptic site locations to address a number of questions: What are the sequence, structure, and dynamics attributes of a cryptic site, especially in comparison to binding pockets? Can we accurately, automatically, and efficiently predict cryptic sites? How common are cryptic sites? Are they common enough to significantly expand the druggable proteome? Can we predict cryptic sites in specific proteins of clinical significance?

Results and Discussion

Our analysis proceeded according to **Figure 1b**. In outline, we started by creating a representative dataset of 84 known examples of cryptic binding sites, 92 binding pockets, and 705 concave surface patches from the Protein Data Bank (Bernstein et al., 1977) and the MOAD database (Benson et al., 2008) (**Methods**, **SI Text**, and **Table S1**). We selected cryptic sites and binding pockets whose ligands are biologically relevant (Benson et al., 2008). Next, we designed a set of 29 features that describe sequence, structure, and dynamics of individual residues and their neighbors (**SI Text** and **Table S2**), based on the crystal structures. We then compared these attributes between the three types of a site to better understand the underlying characteristics of each site. Based on these comparisons, we expanded the set of features for proteins containing cryptic sites to 105 (**Table S2**), describing their crystal structures as well as their alternative conformations obtained by molecular dynamics simulations using AllosMod (Weinkam et al., 2013) (**SI Text**). Next, we put to test 11 supervised machine-learning algorithms (Pedregosa et al., 2011; Schaul et al., 2010) to classify residues as belonging to a cryptic site or not. We then predicted cryptic sites in the entire structurally characterized human proteome. Finally, we focused on a detailed characterization of protein tyrosine phosphatase 1B (PTP1B), a protein that is involved in the insulin signaling pathway and is considered a validated therapeutic target for treatment of type 2 diabetes (Combs, 2010).

POCKET FORMATION AT A CRYPTIC SITE IS DRIVEN BY SMALL CHANGES IN THE STRUCTURE, RESULTING IN A CONFORMATIONALLY CONSERVED CRYPTIC SITE REGARDLESS OF THE LIGAND TYPE.

First, we set out to analyze structural changes needed for a binding pocket formation at a cryptic site. The dataset of cryptic sites reveals mostly minor structural changes required for formation of a detectable pocket. The all-atom RMSD of cryptic binding sites between *apo* and *holo* conformations ranges between 0.45 Å and 22.45 Å (**Figure S1a**) with 67% *apo-holo* pairs differing less than 3 Å in RMSD. The only two *apo-holo* pairs whose differences in RMSD exceed 10 Å are calcium ATPase and calmodulin (PDB IDs 1su4–3fgo and 1cll–1ctr, respectively). Loop movement is the most prominent type of conformational changes (detected in 45% of the binding sites), followed by side-chain rotation (18%), domain motion (17%), displacement of secondary structure elements (16%), and N- or C-terminus flexibility (4%).

To determine whether or not a cryptic site assumes the same bound conformation irrespective of the ligand type, we computed similarities between cryptic site conformations in a protein bound to at least 5 different ligands (58 proteins). Interestingly, only 26% of such cases have an average RMSD exceeding 2 Å (**Figure S1b**), even though the average Tanimoto distance (calculated by Open Babel (O'Boyle et al., 2011), **SI Text**) is low (0.8). This finding suggests that the conformation of a given cryptic site generally does not depend strongly on the ligand type (similar analysis of binding pockets yields 9% of cases with an average RMSD exceeding 2 Å, and an average Tanimoto distance of 0.7). Moreover, the magnitude of the conformational difference within a group of *holo* structures is not significantly correlated with ligand similarity (the correlation coefficient between the all-atom binding site RMSD and Tanimoto distance is 0.01; **Figures S1c** and **S1d**). Finally, the average RMSD of 1.7 Å between bound cryptic binding sites is significantly lower than the average RMSD of 3.0 Å between the unbound and bound conformations (P = 1.410^{-3}, based on two-sample Kolmogorov-Smirnov statistics). Thus, the bound form of the cryptic site is surprisingly conformationally conserved with respect to the ligand type (the average RMSD values of bound conformations of cryptic sites and binding pockets are 1.7 and 2.0 Å, respectively). These observations are consistent with a limited number of protein conformational states as well as with the variability in allosterically regulated proteins, where the binding of the effector alters the conformational distribution between two or more conformational states (Gunasekaran et al., 2004). Indeed, 24 of the 58 cryptic sites are found in allosterically regulated proteins, with 17 of the 24 annotated as effector binding sites (Huang et al., 2011). 20 of the remaining 34 cryptic sites are found on proteins with two or more different binding sites that may or may not be allosteric. The remaining 14 cryptic sites occur on enzymes with flexible active sites and receptors for large hydrophobic ligands, where cryptic site residues modulate binding site accessibility (*e.g.*, the “portal” hypothesis for glycolipid transfer protein, lactoglobulin, and adipocyte lipid binding protein) (Jenkins et al., 2002). In other words, a cryptic site does not convert from flat to concave to accommodate a number of different ligands; rather, cryptic sites may have evolved the ability to convert from flat to concave to modulate ligand-binding kinetics, specificity, affinity, and allostery.

CRYPTIC SITES ARE AS FLEXIBLE AS RANDOM CONCAVE SURFACE PATCHES, BUT EVOLUTIONARILY AS CONSERVED AS BINDING POCKETS.

Next, we analyzed the differences between the sequence, structure, and dynamics attributes of cryptic sites, binding pockets, and concave surface patches. While the differences between cryptic sites and binding pockets are generally small, 4 characteristics distinguish a cryptic site from a binding pocket and/or a concave surface patch: First, a cryptic site predominantly localizes at concave protein regions, even though the site itself is not as concave in the unbound form as a binding pocket. For example, while the average number of protruding atoms at a cryptic site and a binding pocket is 170 and 183 (P = 8.010^{-3}) and the average convexity value of 2.4 and 1.9 (*P* = 0.8), the average pocket score is 0.07 and 0.42 (*P* = 1.710^{-31}), respectively (**Table S3**). Second, a cryptic site tends to be less hydrophobic than a binding pocket, due mostly to an increased frequency of charged residues (arginine in particular, *P* = 1.810^{-5}) (**Figure 2a and Table S3**). Third, a cryptic site is more flexible than a binding pocket, as indicated by significantly higher normalized B-factors (**Figure 2b**). Finally, cryptic site residues are evolutionarily as conserved as those of a binding pocket (**Figure 2c**), suggesting a similar degree of evolutionary pressure and selection on the function of many of these two types of binding sites. Evolutionarily conserved residues have been previously associated with low B-factors (Schlessinger and Rost, 2005; Shih et al., 2012; Swapna et al., 2012); low B-factors are an indicator of residue rigidity. Both evolutionarily conserved residues and residues with low B-factors are often found in functionally important regions of a protein, including binding pockets (Bartova et al., 2008; Capra et al., 2009). In contrast to binding pockets, cryptic sites conserve conformational flexibility to convert from flat to concave.

CRYPTIC SITES AND BINDING POCKETS BIND THE SAME TYPES OF LIGANDS.

To find whether or not the differences between cryptic sites and binding pockets are associated with differences between their ligands, we also compared properties of ligands of cryptic sites and binding pockets. We found no statistically significant differences (**Figure S2a and S2b**). However, clustering of *apo* structures of cryptic sites and binding pockets based on the basic ligand properties and the sequence, structure, and dynamics attributes reveals 4 clusters (**Figure S2c**). Two of the clusters are significantly enriched with cryptic sites: one that comprises convex sites with evolutionarily conserved residues and small hydrophilic ligands, and another one that comprises less convex and less conserved sites that bind larger hydrophobic ligands. The third cluster contains an equal number of cryptic sites and binding pockets that are evolutionarily conserved and bind large hydrophilic ligands. The final cluster contains mostly binding pockets that are concave and evolutionarily conserved, and bind small and hydrophobic ligands.

MOLECULAR DYNAMICS SIMULATIONS BASED ON A SIMPLIFIED ENERGY LANDSCAPE, SEQUENCE CONSERVATION, AND DISTANCE TO THE SURFACE ARE SUFFICIENT TO PREDICT CRYPTIC SITES.

To test if cryptic sites could be predicted accurately, automatically, and efficiently, we used the dataset of *apo* structures with cryptic sites to train 10 different machine-predictive models for the prediction of cryptic site residues, based on the extended set of 105 features (**Table S2**). The optimal predictive model and its parameter values were selected by maximizing the sensitivity (true positive rate) and the specificity (true negative rate) of cryptic site residue prediction, using leave-one-out cross validation on the training set of proteins with 84 cryptic binding sites (**Figure S3a**). The optimal predictive model is a support vector machine (SVM) with a quadratic kernel function and full set of features. The area under the receiver-operator curve (AUC), a measure of accuracy, is 0.78. By removing redundant and irrelevant features using greedy-forward selection that maximizes the AUC, we selected a subset of 19 features, resulting in the AUC of 0.81 (**Figure S3b**).

Although an SVM operates as a "black-box", the relative importance of different features is determined by their order of selection, and may be informative about the cryptic site characteristics (Martens et al., 2008). We find the average pocket score from the molecular dynamics simulations is the most informative single feature according to greedy-forward selection (AUC = 0.73) as well as the two-sample Kolmogorov-Smirnov test (*P* = 1.310^{-151}) (**Figure 2*D*** and **Table S2**). This feature alone is almost as informative as the remaining 29 crystal structure features combined (AUC = 0.74) (**Figure S3b**). Therefore, molecular dynamics simulations on a simplified energy landscape, which is significantly more computationally efficient than a traditional all-atom molecular dynamics simulation (Bowman and Geissler, 2012), often provides sufficient information for localizing cryptic sites. The second feature added to the subset of the 19 features by the greedy-forward approach was sequence conservation. Cryptic site residues are significantly more conserved than the rest of a protein (*P* = 2.710^{-68}). The third feature, distance of a residue to the protein surface, also significantly improves the accuracy of the model (AUC = 0.77), even though by itself is one of the least informative features (*P* = 0.34). The remaining 16 selected features further improve the accuracy of the model, but only modestly so (**Figure S3b**). In summary, a cryptic site can be predicted relatively accurately based primarily on pocket formation in molecular dynamics simulations, evolutionary conservation, and proximity to the protein surface.

THE PREDICTIVE MODEL ACCURATELY LOCALIZES OVER 92% OF CRYPTIC BINDING SITES.

To assess the performance of our predictive model, we applied it to the test set of 14 *apo* structures with one or more known cryptic sites that were not used during the training or any of the analyses above. The prediction capability of the SVM model is satisfactory; we measure an overall AUC of 0.79, with respective true positive and false positive rates of 73% and 29% at the residue score threshold of 0.1 (**Figure 3a**). To further dissect the performance of the learning algorithm, we evaluated predictions for individual proteins from our training and test sets (**Figure 3b**). We define a prediction of a cryptic site to be accurate when at least one third of its residues are identified (sensitivity > 33%). Predictions above this threshold can arguably guide small-molecule tethering experiments and more detailed molecular dynamics simulations. Remarkably, all 14 proteins in the test set and 73 out of 79 proteins in the cross-validation/ training set have all of their cryptic sites identified accurately, resulting in 92% recall (**Tables S1** and **S4**); even for 50% sensitivity, the recall is still 85%.

The predictions are particularly accurate when a large and hydrophobic ligand binds to a cryptic site. For example, we identified all cryptic site residues in the acyl-CoA binding site of the fatty acid responsive transcription factor and 94% of cryptic site residues in the lipid-binding site of -lactoglobulin (**Figure S4**). Our predictive model also accurately predicted cryptic sites in 18 out of 20 proteins (including the proteins from the cross-validation set) that undergo domain movements to expose small-molecule binding sites. For example, more than half of the cryptic site residues of GluR2 receptor, exportin-1, and biotin carboxylase were predicted correctly (**Figures 3c and S4**). Generally, the predictive model accurately predicts known cryptic sites that are allosteric or at relatively flat protein-protein interfaces. For example, TEM-1 -lactamase contains one known allosteric cryptic binding site that requires unfolding of a short helix, and was previously studied using extensive molecular dynamics simulations in explicit solvent and Markov state models (Bowman and Geissler, 2012). In comparison, our approach was also accurate (sensitivity of 60%), but significantly faster and completely automated (**Figure S4d**). Additional examples of accurate predictions include a lipid-binding site in MAP kinase insert of p38 kinase (sensitivity of 74%) (Diskin et al., 2008) (**Figures 3b and S4c**) as well as the nucleotide- and substrate-binding sites in this kinase that were not included in our dataset of cryptic sites, but are known binding sites that undergo conformational changes during phosphorylation (Shukla et al., 2014). Our test set also included three examples of cryptic sites at protein-protein interfaces of exportin 1, Bcl-X_L, and interleukin-2 (IL-2) (Sun et al., 2013; Wells and McClendon, 2007); we identified 73%, 58%, and 47% of the cryptic binding site residues, respectively (**Figures 3** and **S4**).

FALSE NEGATIVES RESULT FROM LARGE REARRANGEMENTS.

Next, we analyze false negatives and false positives (defined based on the cryptic sites annotated in MOAD). Our predictive model failed to predict most cryptic sites that undergo extensive conformational changes and whose pockets are difficult to sample with current molecular dynamics approaches, and some sites whose sequence is not conserved in evolution (**Figures 3b and S5**). In particular, we failed at predicting the cryptic site for stabilizing substrates (*eg*, cyclopiazonic acid) in Ca-ATPase (sensitivity of 0%) that resides at the interface between three domains, two of which are ~50 Å apart in the *apo* conformation (**Figure S5a**). Similarly, we also failed at predicting two allosteric sites in the thumb site of HCV RNA polymerase (sensitivities of 0% and 23%, respectively), in glycogen phosphorylase B (sensitivity of 27%), in pyruvate kinase, and in PTP1B (sensitivity of 29%) (**Figure S5**). In the future, inadequate sampling in AllosMod will be addressed by using multiple input structures and/or restraints from experimental data (*e.g.*, small-angle X-ray scattering profiles (Weinkam et al., TBD), chemical cross-links (Molnar et al., 2014), hydrogen/deuterium exchange with mass spectrometry, and electron microscopy density maps (Liao et al., 2013)).

A FALSE POSITIVE PREDICTION CAN BE AN UNKNOWN CRYPTIC SITE.

While it is difficult to be certain that a predicted cryptic site does not bind a ligand, potential false positives include high-scoring isolated residues or terminal regions of truncated proteins, which may not be as flexible in full-length proteins. However, our benchmark probably overestimates the false positive rate, because some predicted cryptic sites are in fact true binding sites, even though they are not annotated as such in the MOAD database (*e.g.*, proteins that bind peptides or other proteins). For example, our predictive model identifies the binding site for the light chain of coagulation factor VII in the heavy chain of coagulation factor VII; the binding site for guanine-nucleotide exchange factor DBS in CDC42 protein; the dimer interfaces in fructose-1,6-bisphosphate aldolase and estrogen-related receptor ; the docking site for its N-terminal motif in Bcl-X$_L$; the phosphate binding site in acid--glucosidase; and the binding sites for Ran, snurportin, and its own loop in exportin-1 **(Figures 3c** and **S4)**. In summary, the analysis of successes and failures demonstrates the potential of our approach to guide the experimental identification of new sites in difficult small-molecule targets.

THE DRUGGABLE PROTEOME IS SIGNIFICANTLY LARGER THAN ESTIMATED PREVIOUSLY.

Given the overall accuracy of our approach (above), a large number of predicted cryptic sites that are not yet annotated as such in our benchmark might also indicate that there are many cryptic sites yet to be discovered. If so, our predictive model could facilitate finding novel binding sites in “undruggable” proteins, and hence expand the druggable proteome space. It has been suggested that the human proteome of approximately 20,000 proteins contains ~3,000 proteins associated with disease and ~3,000 druggable proteins, with the overlap between the two sets of only ~600 – 1,500 (Hopkins and Groom, 2002; Overington et al., 2006; Russ and Lampel, 2005). To predict how much cryptic binding sites expand the druggable proteome space, we first applied a faster version of our predictive model (based only on the features that are not extracted from molecular dynamics simulations, resulting in the speedup factor of 1000) on 4,421 human proteins with at least one domain of known structure (11,201 structures in total). Next, we counted the numbers of cryptic sites and pockets in each structure (**SI Text**). Pockets were predicted in ~1,900 (43%) proteins, and cryptic sites were predicted in ~3,300 (74%) proteins. Among the 1,420 disease-associated proteins of known structure, 40% have pockets in their crystal structures (in agreement with the previous estimate that the fraction of proteins that are both disease-associated and druggable is 20-50% (Schmidtke and Barril, 2010)). In contrast to pockets, cryptic sites were predicted in 72% of the disease-associated proteins, 38% of which have no apparent pockets (**Figure 4**). However, some of the predictions may be false positives (the sites may in fact not bind any ligands). Moreover, for some sites, it may be very difficult to find a ligand (even if it does exist), and even if the ligand is found, it may not be a drug because it does not target the disease-modifying function of a protein or because it does not meet clinical development criteria. Nevertheless, the prediction of cryptic sites on the disease-associated proteins of known structure indicates that small molecules might be used to target significantly more disease-associated proteins than were previously thought druggable.

If cryptic sites are more abundant than previously estimated, why does high-throughput screening not identify them more often than it does? It has been shown that small-molecule libraries are biased towards traditional drug targets, such as G protein-coupled receptors, ion channels, and kinases, while they are not as suitable for antimicrobial targets and those identified from genomic studies (Hert et al., 2009). It is conceivable that the existing libraries are also less suitable for cryptic sites. Moreover, cryptic sites may tend to bind ligands more weakly than binding pockets, due to the need to compensate for the free energy of site formation

(Mobley and Dill, 2009), and may thus be ranked lower on the high-throughput screening lists. Therefore, different approaches based on larger and more diverse chemical libraries, including small fragments (Hardy and Wells, 2004; Makley and Gestwicki, 2013; Wiesmann et al., 2004), peptides, peptidomimetics, and natural products may be needed for more efficient discovery of cryptic site ligands. A case in point is the discovery of a number of ligands for cryptic allosteric sites and cryptic sites at protein-protein interfaces, such as IL-2, caspases, kinase PDK1, and PTP1B, by fragment-based tethering (Hardy and Wells, 2004; Ostrem et al., 2013; Sadowsky et al., 2011; Wiesmann et al., 2004). Our data suggests that cryptic sites are much more prevalent than previously expected. However, while such sites do provide additional opportunities for drug discovery, they may not ultimately lead to drugs.

EXPERIMENTAL CHARACTERIZATION OF A PREDICTED CRYPTIC SITE IN PTP1B BY NMR SPECTROSCOPY.

Finally, to demonstrate the practical utility of our approach, we focused on the clinically significant protein PTP1B. Targeting PTP1B with small molecules has been challenging due to the lack of specificity and bioavailability of substrate mimetics as well as the presence of only a single known allosteric pocket (Combs, 2010; Wiesmann et al., 2004). In addition to identifying 4 of the 14 residues in the known allosteric cryptic site (Wiesmann et al., 2004), out predictive model also suggested another putative cryptic site (**Figure 5**). This site is interesting for several reasons: First, the predicted cryptic site residues form an internal cavity (between residues Ile 67 and Phe 95) in crystal structures of PTP1B that is large enough to accommodate a small molecule (volume of ~150 Å^3). Our molecular dynamics simulations suggest that small conformational changes in the cavity-forming loops could make the cavity accessible to the solvent and expand its size (up to 430 Å^3). Second, the site is in proximity of two cysteine residues (Cys 92 and Cys 121) that could be targeted covalently in small-molecule fragment screening by tethering (Hardy and Wells, 2004). Even though residue Cys 121 seems to be buried in the crystal structure, small conformational changes could expose it to the solvent (Cys 121 is exposed in 0.4% of the molecular dynamics snapshots at 300 K). Third, this cryptic site in PTP1B differs from the corresponding region in the closely related tyrosine-protein phosphatase non-receptor type 2 (TCPTP) at, for example, position 97 (glutamate instead of leucine). This difference between PTP homologs could be exploited to develop selective inhibitors that avoid the serious adverse effects associated with TCPTP inhibition in mice (Wiesmann et al., 2004). Finally, the cryptic site may be allosterically coupled to the catalytic site; examining contacts between pairs of residues (Weinkam et al., 2013) suggests extensive coupling between the cryptic and catalytic sites (**Figure S6a**).

To experimentally test our cryptic site prediction, we studied a previously discovered ligand, ABDF (Hansen et al., 2005), which allosterically inhibits PTP1B through an undefined mechanism. Although PTP1B has three other surface-exposed cysteine residues, ABDF covalently attaches specifically to the side chain of Cys 121, which is adjacent to our predicted cryptic site (**Figure 5** and **Figure S6b**). The Cys 121 side chain points towards the interior of the unlabeled protein, so binding of ABDF likely requires a conformational change in the protein. We were unable to obtain a crystal structure of ABDF-labeled PTP1B, in agreement with other reports that ABDF-labeled PTP1B, unlike *apo* PTP1B, is recalcitrant to crystallization (Hansen et al., 2005). To determine whether or not the covalent label causes specific local conformational changes or globally perturbs the protein, we collected ^{1}H, ^{15}N TROSY HSQC NMR spectra of both *apo* and ABDF-labeled protein (**SI Text** and **Figure S6c-f**). Using previously published backbone resonance assignments (Meier et al., 2002), we observed no perturbation of chemical shifts for a number of residues distal to the predicted cryptic site, indicating that the effects are local and that the protein remains folded. In contrast, a cluster of residues nearby the predicted cryptic site were significantly perturbed (**Figure 5** and **Figure S6c-f**). Many other residues near the predicted cryptic site that would need to move for ligand binding, including the adjacent βsheet and Cys 121 loop, were unassigned due to resonance broadening, which is indicative of conformational exchange. Collectively, these results point to structural flexibility in the vicinity of the predicted cryptic site and the specific perturbation of residues surrounding the predicted binding pocket, validating our prediction.

To conclude, we describe cryptic sites and a method that accurately, automatically, and efficiently predicts their locations in protein structures. Our results support the hypothesis of ubiquitous cryptic sites and suggest many new small-molecule protein targets, including those that are associated with diseases. Moreover, we illustrate how chemical tethering can be used to validate cryptic site predictions by discovering cryptic site ligands. Cryptic sites can also be characterized by experimental techniques that measure protein dynamics, such as NMR spectroscopy and room-temperature X-ray crystallography (Fraser et al., 2011), as well as by discovery of ligands through virtual screening against conformations with pockets computed by AllosMod or molecular dynamics simulations. Our approach provides a convenient first step for such characterizations.

Materials and Methods

We started by finding cryptic sites in the Protein Data Bank (PDB) (Berman et al., 2002; Berman et al., 2000), as follows. First, we gathered structures of protein-ligand complexes as well as structures of proteins in ligand-free (unbound) conformations. We define binding residues as the residues with at least one atom within 5 Å from any atom of a ligand in the bound conformation (a binding site). Second, we removed the redundant protein occurrences in the dataset by applying sequence identity threshold of 40% (**SI Text**). Finally, we evaluated each binding site in the unbound conformation using pocket scores based on two pocket-detection algorithms, Fpocket and ConCavity (Capra et al., 2009; Le Guilloux et al., 2009). Binding sites with bad pocket sores in the unbound conformation and good pocket scores in the bound conformation were defined as cryptic sites, whereas those with good pocket scores in both conformations were defined as binding pockets (**Tables S1** and **S4**). More details and methods are available in **SI Text**. The web server for predicting cryptic binding sites is available at http://salilab.org/cryptosite.

Acknowledgments

The authors thank Hao Fan, Marcus Fischer, Nir London, Avner Schlessinger, Dina Schneidman, and other members of Sali lab for their comments and feedback. P.C. is supported by a Howard Hughes Predoctoral Fellowship; T.J.R. is supported by a predoctoral fellowship from the NIH (F31 CA180378-01) and the Krevans Fellowship; R.A.W. is supported by a National Science Foundation Graduate Research Fellowship; J.S.F. is supported by the National Institutes of Health (DP5 OD009180, R21 GM110580, P30 DK063720) and National Science

Foundation (STC-1231306); A.S. is supported by the National Institutes of Health (R01 GM083960, U54 RR022220, U54 GM094662, P01 AI091575, and U01 GM098256).

References

Arkin, M.R., and Wells, J.A. (2004). Small-molecule inhibitors of protein-protein interactions: progressing towards the dream. Nature reviews Drug discovery *3*, 301-317.

Bartova, I., Koca, J., and Otyepka, M. (2008). Functional flexibility of human cyclin-dependent kinase-2 and its evolutionary conservation. Protein science : a publication of the Protein Society *17*, 22-33.

Benson, M.L., Smith, R.D., Khazanov, N.A., Dimcheff, B., Beaver, J., Dresslar, P., Nerothin, J., and Carlson, H.A. (2008). Binding MOAD, a high-quality protein-ligand database. Nucleic acids research *36*, D674-678.

Berman, H.M., Battistuz, T., Bhat, T.N., Bluhm, W.F., Bourne, P.E., Burkhardt, K., Feng, Z., Gilliland, G.L., Iype, L., Jain, S.*, et al.* (2002). The Protein Data Bank. Acta crystallographica Section D, Biological crystallography *58*, 899-907.

Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., and Bourne, P.E. (2000). The Protein Data Bank. Nucleic acids research *28*, 235-242.

Bernstein, F.C., Koetzle, T.F., Williams, G.J., Meyer, E.F., Jr., Brice, M.D., Rodgers, J.R., Kennard, O., Shimanouchi, T., and Tasumi, M. (1977). The Protein Data Bank: a computer-based archival file for macromolecular structures. Journal of molecular biology *112*, 535-542.

Bowman, G.R., and Geissler, P.L. (2012). Equilibrium fluctuations of a single folded protein reveal a multitude of potential cryptic allosteric sites. Proceedings of the National Academy of Sciences of the United States of America *109*, 11681-11686.

Brenke, R., Kozakov, D., Chuang, G.Y., Beglov, D., Hall, D., Landon, M.R., Mattos, C., and Vajda, S. (2009). Fragment-based identification of druggable 'hot spots' of proteins using Fourier domain correlation techniques. Bioinformatics *25*, 621-627.

Campbell, S.J., Gold, N.D., Jackson, R.M., and Westhead, D.R. (2003). Ligand binding: functional site location, similarity and docking. Current opinion in structural biology *13*, 389-395.

Capra, J.A., Laskowski, R.A., Thornton, J.M., Singh, M., and Funkhouser, T.A. (2009). Predicting protein ligand binding sites by combining evolutionary sequence conservation and 3D structure. PLoS computational biology *5*, e1000585.

Combs, A.P. (2010). Recent advances in the discovery of competitive protein tyrosine phosphatase 1B inhibitors for the treatment of diabetes, obesity, and cancer. Journal of medicinal chemistry *53*, 2333-2344.

Diskin, R., Engelberg, D., and Livnah, O. (2008). A novel lipid binding site formed by the MAP kinase insert in p38 alpha. Journal of molecular biology *375*, 70-79.

Durrant, J.D., and McCammon, J.A. (2011). Molecular dynamics simulations and drug discovery. BMC biology *9*, 71.

Fraser, J.S., van den Bedem, H., Samelson, A.J., Lang, P.T., Holton, J.M., Echols, N., and Alber, T. (2011). Accessing protein conformational ensembles using room-temperature X-ray crystallography. Proceedings of the National Academy of Sciences of the United States of America *108*, 16247-16252.

Grove, L.E., Hall, D.R., Beglov, D., Vajda, S., and Kozakov, D. (2013). FTFlex: accounting for binding site flexibility to improve fragment-based identification of druggable hot spots. Bioinformatics *29*, 1218-1219.

Gunasekaran, K., Ma, B., and Nussinov, R. (2004). Is allostery an intrinsic property of all dynamic proteins? Proteins *57*, 433-443.

Hansen, S.K., Cancilla, M.T., Shiau, T.P., Kung, J., Chen, T., and Erlanson, D.A. (2005). Allosteric inhibition of PTP1B activity by selective modification of a non-active site cysteine residue. Biochemistry *44*, 7704-7712.

Hardy, J.A., and Wells, J.A. (2004). Searching for new allosteric sites in enzymes. Current opinion in structural biology *14*, 706-715.

Henrich, S., Salo-Ahen, O.M., Huang, B., Rippmann, F.F., Cruciani, G., and Wade, R.C. (2010). Computational approaches to identifying and characterizing protein binding sites for ligand design. Journal of molecular recognition : JMR *23*, 209-219.

Hermann, J.C., Marti-Arbona, R., Fedorov, A.A., Fedorov, E., Almo, S.C., Shoichet, B.K., and Raushel, F.M. (2007). Structure-based activity prediction for an enzyme of unknown function. Nature *448*, 775-779.

Hert, J., Irwin, J.J., Laggner, C., Keiser, M.J., and Shoichet, B.K. (2009). Quantifying biogenic bias in screening libraries. Nature chemical biology *5*, 479-483.

Hopkins, A.L., and Groom, C.R. (2002). The druggable genome. Nature reviews Drug discovery *1*, 727-730.

Horn, J.R., and Shoichet, B.K. (2004). Allosteric inhibition through core disruption. Journal of molecular biology *336*, 1283-1291.

Huang, Z., Zhu, L., Cao, Y., Wu, G., Liu, X., Chen, Y., Wang, Q., Shi, T., Zhao, Y., Wang, Y.*, et al.* (2011). ASD: a comprehensive database of allosteric proteins and modulators. Nucleic acids research *39*, D663-669.

Jenkins, A.E., Hockenberry, J.A., Nguyen, T., and Bernlohr, D.A. (2002). Testing of the portal hypothesis: analysis of a V32G, F57G, K58G mutant of the fatty acid binding protein of the murine adipocyte. Biochemistry *41*, 2022-2027.

Laskowski, R.A., Luscombe, N.M., Swindells, M.B., and Thornton, J.M. (1996). Protein clefts in molecular recognition and function. Protein science : a publication of the Protein Society *5*, 2438-2452.

Laurie, A.T., and Jackson, R.M. (2005). Q-SiteFinder: an energy-based method for the prediction of protein-ligand binding sites. Bioinformatics *21*, 1908-1916.

Le Guilloux, V., Schmidtke, P., and Tuffery, P. (2009). Fpocket: an open source platform for ligand pocket detection. BMC bioinformatics *10*, 168.

Lexa, K.W., and Carlson, H.A. (2011). Full protein flexibility is essential for proper hot-spot mapping. Journal of the American Chemical Society *133*, 200-202.

Liao, M., Cao, E., Julius, D., and Cheng, Y. (2013). Structure of the TRPV1 ion channel determined by electron cryo-microscopy. Nature *504*, 107-112.

Makley, L.N., and Gestwicki, J.E. (2013). Expanding the number of 'druggable' targets: non-enzymes and protein-protein interactions. Chemical biology & drug design *81*, 22-32.

Martens, D., Huysmans, J., Setiono, R., Vanthienen, J., and Baesens, B. (2008). Rule Extraction from Support Vector Machines: An Overview of Issues and Application in Credit Scoring. Studies in Computational Intelligence *80*, 33-63.

Meier, S., Li, Y.C., Koehn, J., Vlattas, I., Wareing, J., Jahnke, W., Wennogle, L.P., and Grzesiek, S. (2002). Backbone resonance assignment of the 298 amino acid catalytic domain of protein tyrosine phosphatase 1B (PTP1B). Journal of biomolecular NMR *24*, 165-166.

Mobley, D.L., and Dill, K.A. (2009). Binding of small-molecule ligands to proteins: "what you see" is not always "what you get". Structure *17*, 489-498.

Molnar, K.S., Bonomi, M., Pellarin, R., Clinthorne, G.D., Gonzales, G., Goldberg, S.D., Sali, A., and DeGrado, W.F. (2014). Cys-scanning disulfide crosslinking and Bayesian modeling probe the transmembrane signaling mechanism of the histine kinase, PhoQ. submitted.

Nisius, B., Sha, F., and Gohlke, H. (2012). Structure-based computational analysis of protein binding sites for function and druggability prediction. Journal of biotechnology *159*, 123-134.

O'Boyle, N.M., Banck, M., James, C.A., Morley, C., Vandermeersch, T., and Hutchison, G.R. (2011). Open Babel: An open chemical toolbox. Journal of cheminformatics *3*, 33.

Ostrem, J.M., Peters, U., Sos, M.L., Wells, J.A., and Shokat, K.M. (2013). K-Ras(G12C) inhibitors allosterically control GTP affinity and effector interactions. Nature *503*, 548-551.

Overington, J.P., Al-Lazikani, B., and Hopkins, A.L. (2006). How many drug targets are there? Nature reviews Drug discovery *5*, 993-996.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V.*, et al.* (2011). Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research *12*, 2825–2830.

Pettersen, E.F., Goddard, T.D., Huang, C.C., Couch, G.S., Greenblatt, D.M., Meng, E.C., and Ferrin, T.E. (2004). UCSF Chimera--a visualization system for exploratory research and analysis. Journal of computational chemistry *25*, 1605-1612.

Rossi, A., Marti-Renom, M.A., and Sali, A. (2006). Localization of binding sites in protein structures by optimization of a composite scoring function. Protein science : a publication of the Protein Society *15*, 2366-2380.

Russ, A.P., and Lampel, S. (2005). The druggable genome: an update. Drug discovery today *10*, 1607-1610.

Sadowsky, J.D., Burlingame, M.A., Wolan, D.W., McClendon, C.L., Jacobson, M.P., and Wells, J.A. (2011). Turning a protein kinase on or off from a single allosteric site via disulfide trapping. Proceedings of the National Academy of Sciences of the United States of America *108*, 6056-6061.

Schaul, T., **Bayer, J.**, **Wierstra, D.**, **Sun, Y.**, **Felder, M.**, **Sehnke, F.**, **Ru¨ckstieß, T.**, and **Schmidhuber, J.** (2010). PyBrain. Journal of Machine Learning Research *11*, 743-746.

Schlessinger, A., and Rost, B. (2005). Protein flexibility and rigidity predicted from sequence. Proteins *61*, 115-126.

Schmidtke, P., and Barril, X. (2010). Understanding and predicting druggability. A high-throughput method for detection of drug binding sites. Journal of medicinal chemistry *53*, 5858-5867.

Sheridan, R.P., Maiorov, V.N., Holloway, M.K., Cornell, W.D., and Gao, Y.D. (2010). Drug-like density: a method of quantifying the "bindability" of a protein target based on a very large set of pockets and drug-like ligands from the Protein Data Bank. Journal of chemical information and modeling *50*, 2029-2040.

Shih, C.H., Chang, C.M., Lin, Y.S., Lo, W.C., and Hwang, J.K. (2012). Evolutionary information hidden in a single protein structure. Proteins *80*, 1647-1657.

Shukla, D., Meng, Y., Roux, B., and Pande, V.S. (2014). Activation pathway of Src kinase reveals intermediate states as targets for drug design. Nature communications *5*, 3397.

Sun, Q., Carrasco, Y.P., Hu, Y., Guo, X., Mirzaei, H., Macmillan, J., and Chook, Y.M. (2013). Nuclear export inhibition through covalent conjugation and hydrolysis of Leptomycin B by CRM1. Proceedings of the National Academy of Sciences of the United States of America *110*, 1303-1308.

Swapna, L.S., Bhaskara, R.M., Sharma, J., and Srinivasan, N. (2012). Roles of residues in the interface of transient protein-protein complexes before complexation. Scientific reports *2*, 334.

Weinkam, P., Chen, Y.C., Pons, J., and Sali, A. (2013). Impact of mutations on the allosteric conformational equilibrium. Journal of molecular biology *425*, 647-661.

Weinkam, P., Schneidman-Duhovny, D., Webb, B.M., Tainer, J.A., Hammel, M., and Sali, A. (TBD). Mapping protein allosteric mechanisms with small angle X-ray scattering profiles. submitted.

Wells, J.A., and McClendon, C.L. (2007). Reaching for high-hanging fruit in drug discovery at protein-protein interfaces. Nature *450*, 1001-1009.

Wiesmann, C., Barr, K.J., Kung, J., Zhu, J., Erlanson, D.A., Shen, W., Fahr, B.J., Zhong, M., Taylor, L., Randal, M.*, et al.* (2004). Allosteric inhibition of protein tyrosine phosphatase 1B. Nature structural & molecular biology *11*, 730-737.

Figures

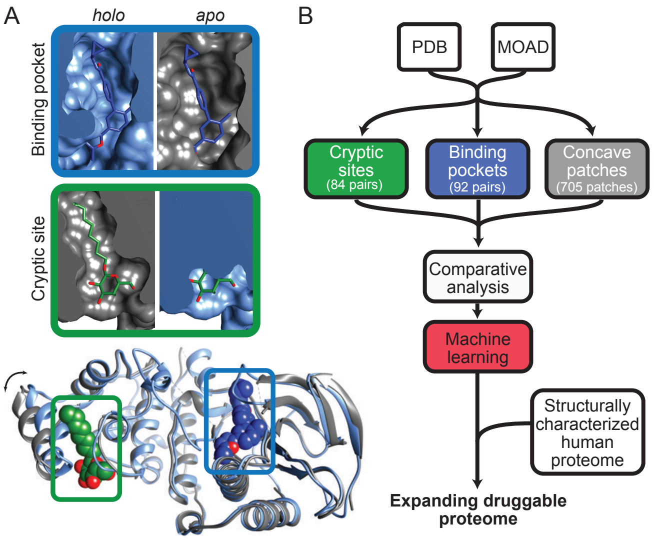

A
holo
apo
Binding pocket
Cryptic site
B
PDB
MOAD
Cryptic sites (84 pairs)
Binding pockets (92 pairs)
Concave patches (705 patches)
Comparative analysis
Machine learning
Structurally characterized human proteome
Expanding druggable proteome

Figure 1: (*A*) Examples of a pocket and cryptic site in p38 MAP kinase. The nucleotide-binding site of the p38 MAP kinase is a pocket visible in both bound (*holo*; blue ribbon; PDB ID: 2zb1) and unbound (*apo*; grey ribbon; PDB ID: 2npq) conformations. The ligand, biphenyl amide inhibitor, is depicted as blue spheres. On the other hand, the site in the C-lobe domain that binds octyglucoside lipid (green spheres) becomes a visible pocket only after the movement of the -helix at the left of the structure (marked with the double-headed arrow). The small molecules are shown as they bind in the *holo* structures. UCSF Chimera software was used for the visualization (Pettersen et al., 2004). (*B*) Flowchart summarizing the analyses in this study (**Materials and Methods, SI Text**).

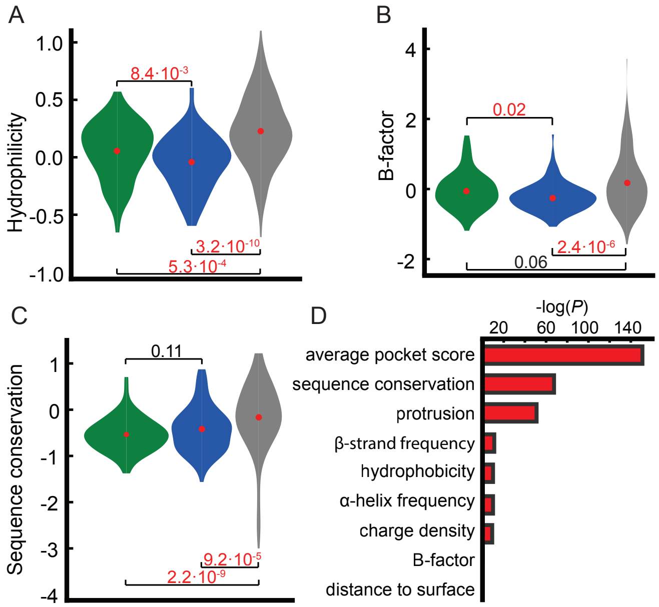


Figure 2: Comparison of cryptic sites, binding pockets, and random concave surface patches. (*A*-*C*) In each panel, the distribution of the feature values of binding site residues are shown as violin plots for cryptic sites (green), binding pockets (blue), and random concave surface patches (grey). The edges between distributions denote P-values based on Kolmogorov-Smirnov two-sample statistics; numbers/letters in red are statistically significant ($P < 0.05$). (*D*) For a few selected residue-based features, the distributions of their values for the cryptic sites and the rest of residues in our dataset are compared. The bars denote statistical significance (P-value) from the two-sample Kologorov-Smirnov non-equality test (**Table S2** for the P-values of other features).

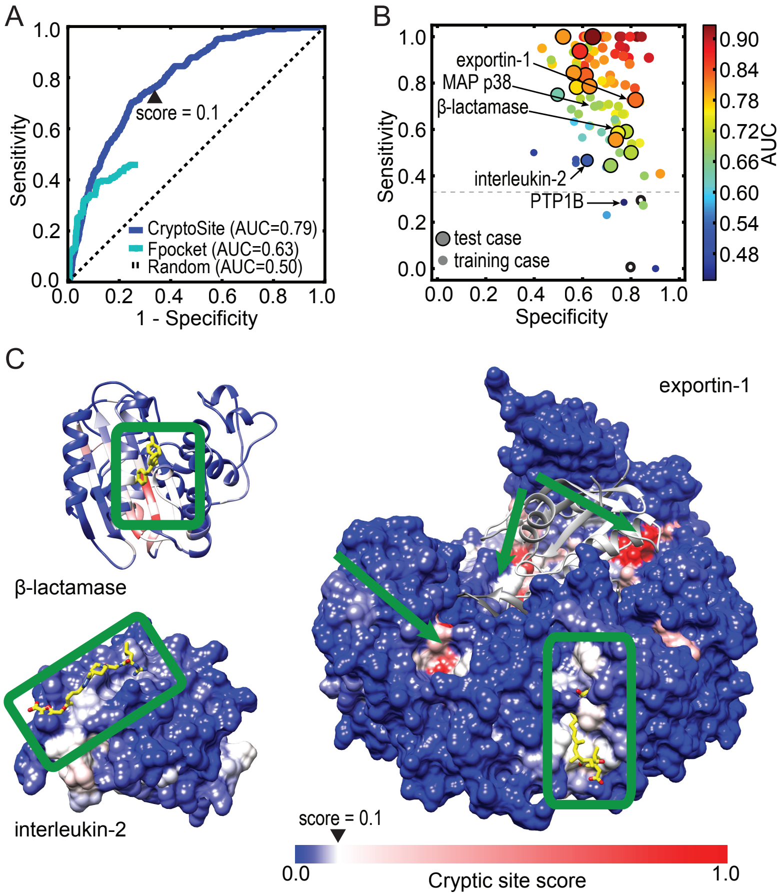


Figure 3: The accuracy of our predictive model or Fpocket is measured as the area under the receiver-operating characteristic (ROC) curve based on predictions on all proteins in the test set (*A*), as well as based on sensitivity (true positive rate) and specificity (true negative rate) values from predictions on individual proteins (*B*). (*A*) Only ~45% of cryptic site residues were detected by Fpocket; the area under the ROC curve was calculated by connecting the end of the ROC curve and the upper-right corner as a straight line. (*B*) Sensitivities and specificities were determined for each protein in our test set (larger data points with black circle) and training set (smaller data points) based on leave-one-out cross-validation. The classification of the residues is based on the score threshold of 0.1. The two empty circles in the lower third of the graph denote two failed predictions of cryptic sites in proteins with more than one cryptic site. (*C*) The cryptic sites from our dataset are marked by green rectangles, and the computed scores that a

residue is in a cryptic site are shown on the blue-to-red color scale. The small molecules that bind into the known cryptic sites are superposed from the alignment to the bound conformations and represented as yellow sticks. The predictive model also identifies additional binding sites in the structure of exportin-1 that are known binding sites for other proteins (for example, Ran represented as white ribbon), or parts of the same protein (grey loop) that bind to a site after a conformational change (all examples are marked with green arrows).

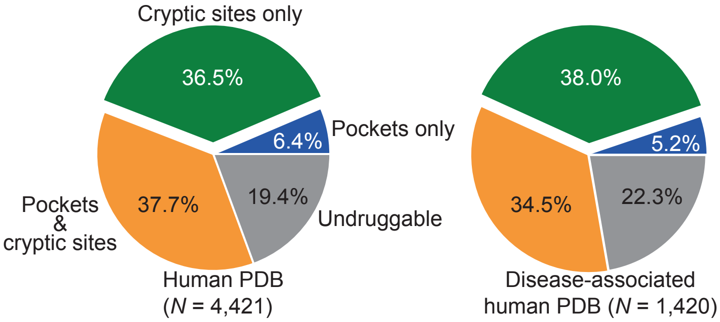


Figure 4: Cryptic binding sites are predicted to expand the size of the druggable proteome. The percentage of proteins for which no binding sites (grey), only cryptic sites (green), only binding pockets (blue), and both cryptic sites and binding pockets (orange) were predicted for all human proteins with known structure (left pie chart) and for a subset of disease-associated proteins (right pie chart). Shown are the results of the fast version of our predictive model that does not take into account features based on molecular dynamics simulations.

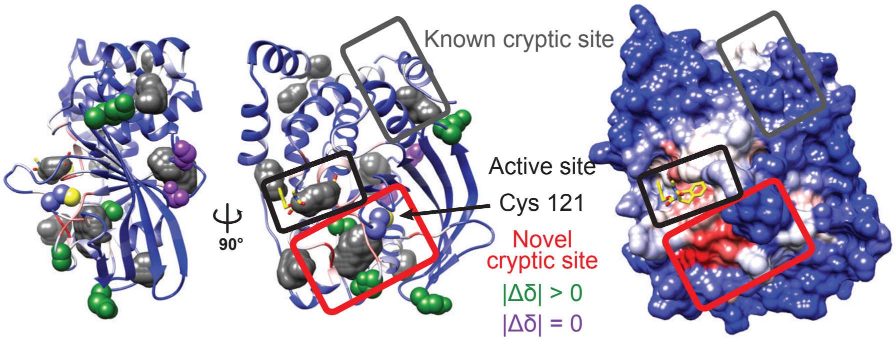


Figure 5: Cryptic binding sites in PTP1B. Ribbon (left and center) and surface (right) representations of the PTP1B structure (PDB ID: 2f6v) are colored based on the cryptic site score as in **Figure 3*C***. Residues with definitive chemical shift changes (II) upon ABDF labeling (green) cluster around the cryptic and ABDF binding sites, whereas residues whose chemical shifts definitively do not change (purple) are more distal. The panel also shows positions and average volumes of the pockets (grey densities) that are at least partially open more than 50% of the time, as observed in the molecular dynamics simulation at 300 K.

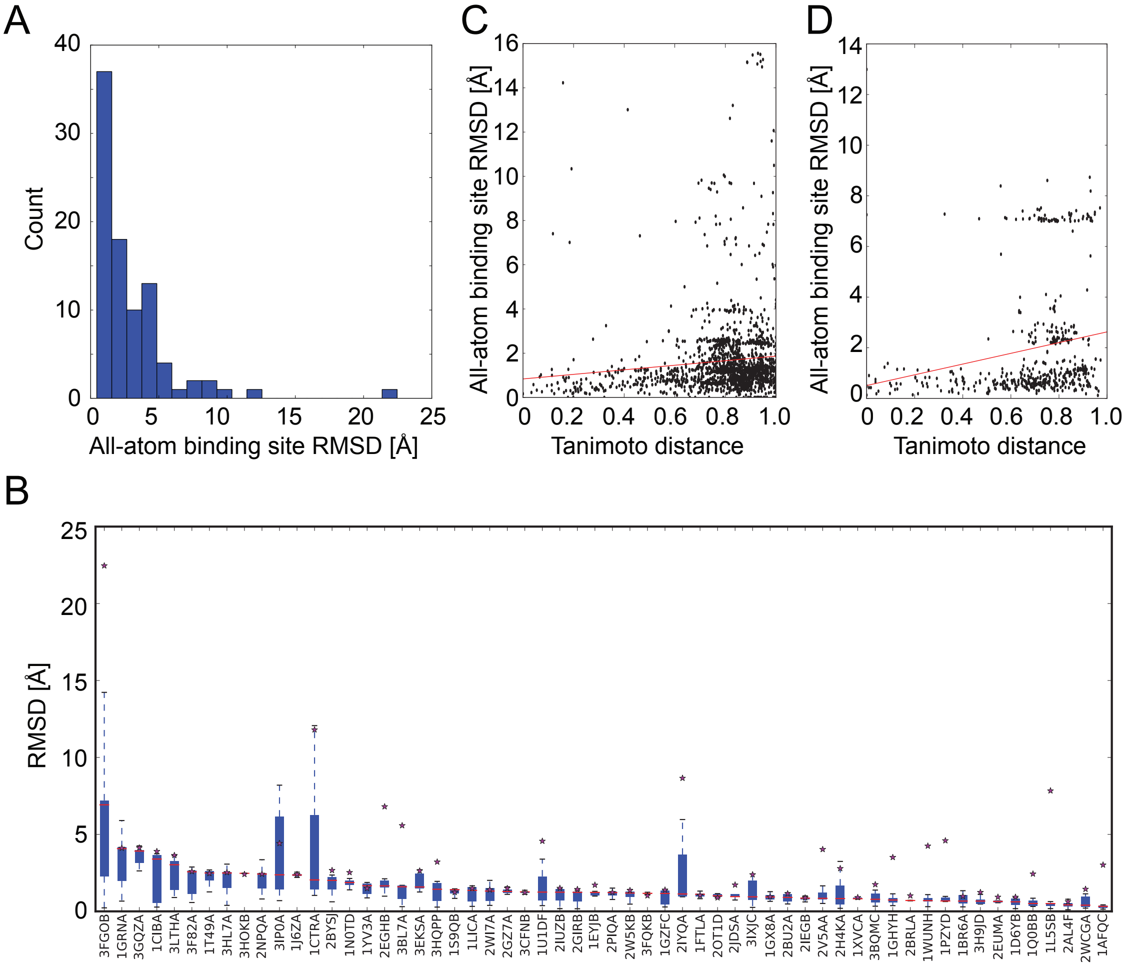


Figure S1: (*A*) Histogram of all-atom binding site RMSDs between *apo* and *holo* conformations. (*B*) Structural similarity (all-atom binding site RMSD) between cryptic site structures bound to at least 5 different ligands. Boxes, whiskers, and red lines denote 10th and 90th percentile, 5th and 95th percentile, and the median of the distribution. The similarities between unbound and bound conformations from our dataset are denoted by star symbols. The degree of structural similarity between bound cryptic sites (*C*) or binding pockets (*D*) is independent of the 2D structural similarity between the bound ligands. Linear path fingerprints (FP2) and Open Babel package were used to calculate the Tanimoto distances. The red line denotes linear fit, with a slope parameter that is not significantly different (R-value < 0.01) from the horizontal regression.

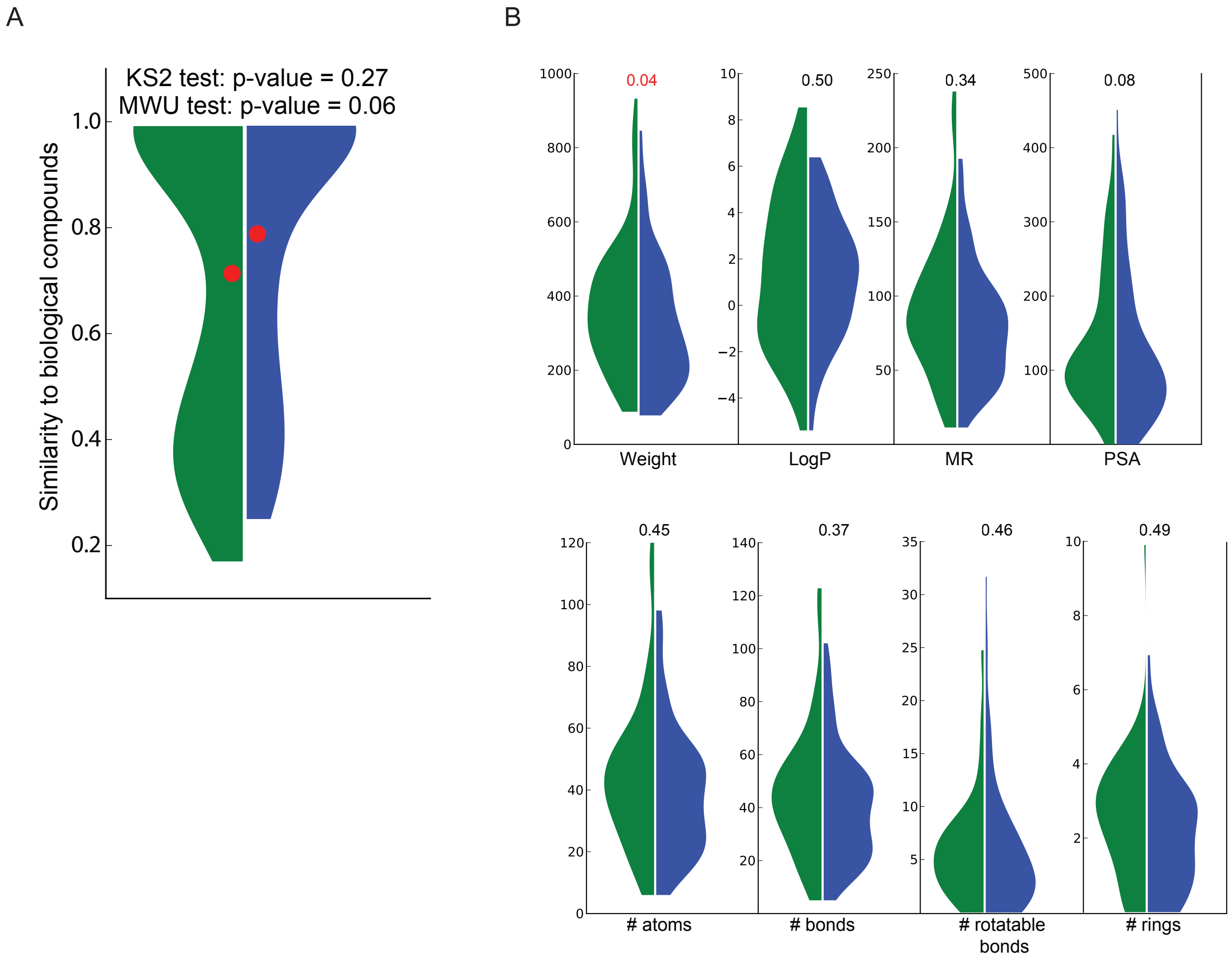


Figure S2: Comparison of small molecule-based features between ligands in cryptic sites (green half-violin plots), and ligands in pockets (blue-half violin plots). (*A*) The distributions of ligand similarities to biological compounds collected from the KEGG database of biological processes. (*B*) Distributions of several ligand descriptors, as determined by Open Babel. (*C*) 2-dimensional clustering of ligand and binding site features as well as binding sites identifies 4 clusters. Two of the clusters are significantly enriched with cryptic sites. One cluster includes convex sites with evolutionarily conserved residues and small hydrophilic ligands (cluster 4), and another one includes less convex and less conserved sites that bind larger hydrophobic ligands (cluster 3). The third cluster contains an equal number of cryptic sites and binding pockets that are evolutionarily conserved and bind large hydrophilic ligands (cluster 2). The final

cluster contains mostly binding pockets that are concave and evolutionarily conserved, and bind small and hydrophobic ligands (cluster 1).

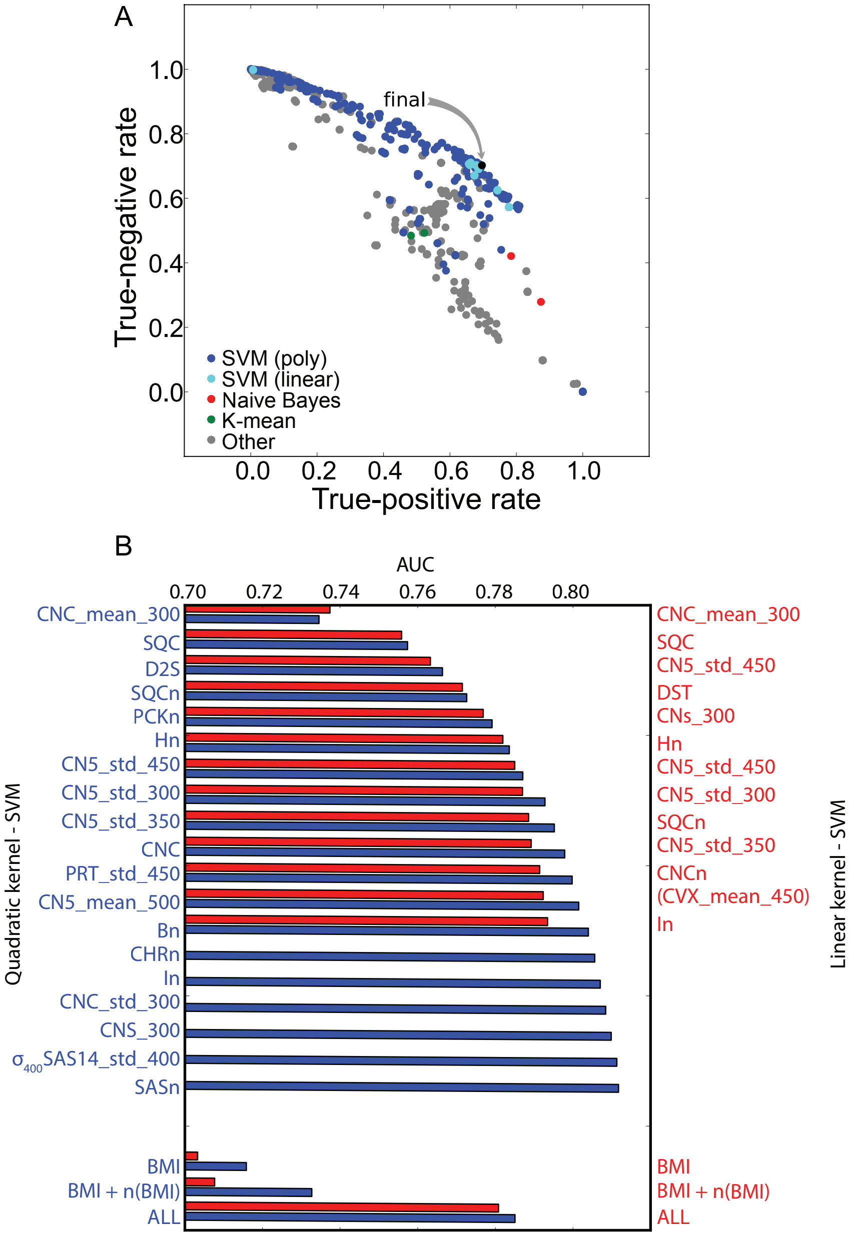


Figure S3: (*A*) Search for the most accurate machine-learning algorithm, data pre-processing method, and the corresponding set of parameters. The most accurate predictive model and its parameter values were selected by maximizing the sensitivity (true-positive rate) and the specificity (true-negative rate) of cryptic site residue classification, using leave-one-out cross validation on the training set of proteins with 84 cryptic binding sites. The arrow points to the most accurate algorithm. (*B*) Feature selection using greedy-forward approach. Feature

selection approach was used for two versions of SVMs, one with a quadratic kernel function (blue) and another one with a linear kernel function (red). Due to the higher accuracy, only the SVM with a quadratic kernel function was used here. See **SI Table 3** for a description of feature labels.

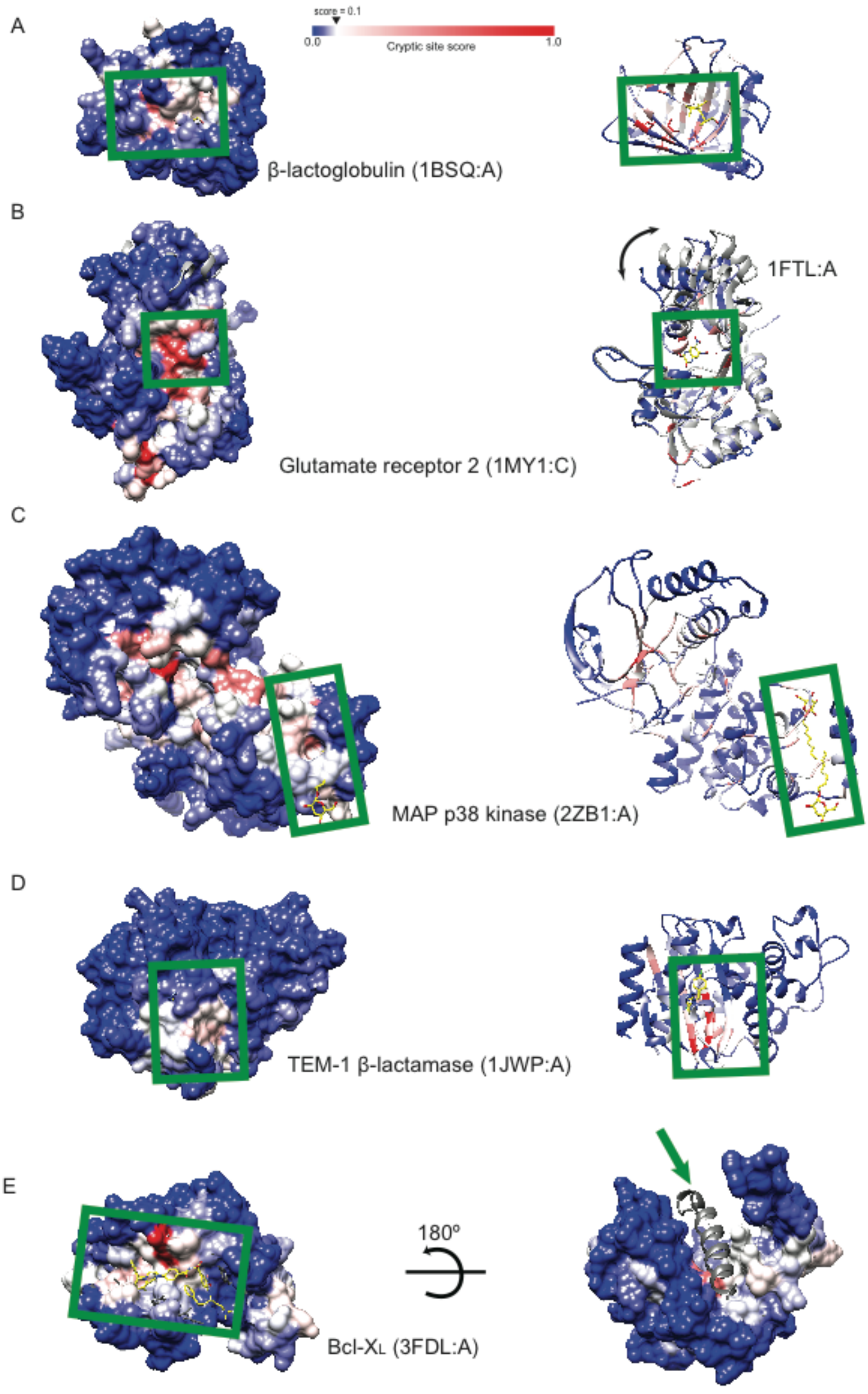


Figure S4: Examples of accurate predictions, shown in surface and ribbon representation of *apo* conformations. Ligands (yellow sticks) are superposed from the alignments with the *holo* conformations. (*A*) 94% of cryptic site residues are predicted accurately in the -lactoglobulin

(PDB ID: 1bsq). To demonstrate the ability of our method to correctly identify the cryptic binding site residues, a few residues on -strands are shown as sticks. These residues are predicted as a cryptic site with high scores and correctly point towards the binding site, whereas the neighboring residues on the -strands that point in the other direction have low scores (the same pattern is observed in other proteins where a cryptic binding site includes -strands). (*B*) Binding to the cryptic site of glutamate receptor 2 requires domain opening (indicated by a black double-headed arrow). The ribbon representation shows both the *apo* (PDB ID: 1my1) and *holo* conformations (in grey; PDB ID: 1ftl). (*C*) Binding into the cryptic site of MAP p38 kinase requires -helix translocation (**Fig. 1**). (*D*) Cryptic site residues that are not solvent accessible in the *apo* conformation of TEM-1 -lactamase are correctly predicted (red patches on -strands). (*E*) Cryptic site in Bcl-X$_L$ is located at the protein-protein interface. The predictive model predicts another cryptic site at the interface of the Bcl-X$_L$ core and its terminal -helix (denoted by green arrow). Proteins are shown in scale.

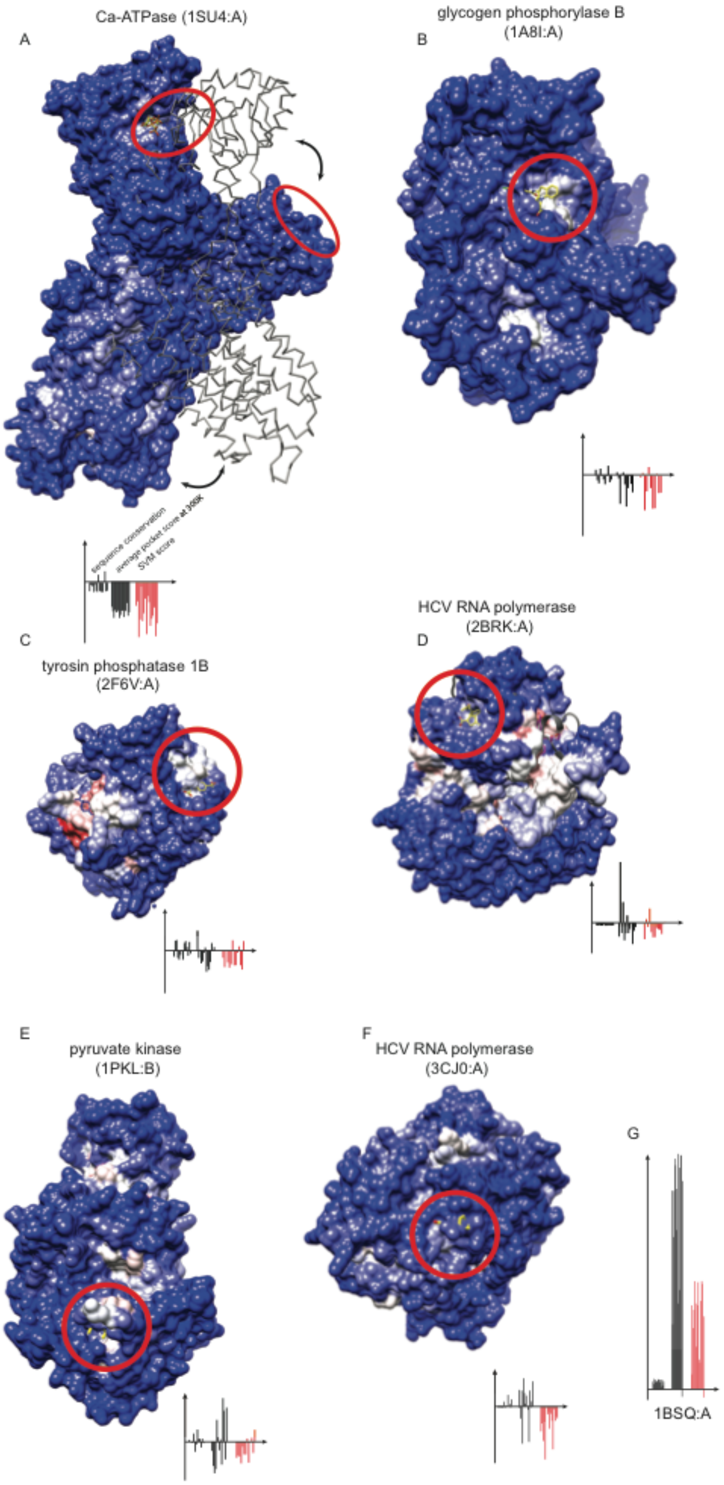


Figure S5: Six inaccurately predicted cryptic sites (marked by red ovals). To characterize the failures, we plot the values of the two most informative features for each cryptic site residue (black bars) as well as SVM scores from the prediction (red bars). Negative values denote

values that disfavor identification of cryptic site residues. (*A*) The cryptic site in Ca-dependent ATPase requires large conformational changes (denoted by black arrows and the *holo* conformation represented by grey trace), not sampled by our molecular dynamics simulations (PDB ID: 1su4). (*B*) Cryptic site scores for the binding site residues in glycogen phosphorylase B are higher then in the rest of the protein, but below our threshold (PDB ID: 1a8i). (*C*) Similarly as in B, the cryptic binding site residues in tyrosin phosphatase 1B were predicted with scores higher than those for most of the protein, but are below our threshold for most of the binding site residues (PDB ID: 2f6v). The predictive model identifies two additional cryptic sites, one that is a known active site and an unannotated site next to the active site. (*D*) The panel shows the structure of HCV RNA polymerase (PDB ID: 2brk). The red patch on the right to the cryptic binding site is also a known cryptic site, and was predicted correctly. Two more failures include a cryptic site in pyruvate kinase (PDB ID: 1pkl) (*E*) and the third cryptic site in HCV RNA polymerase (PDB ID: 3cj0) (*F*). The last panel (*G*) shows an example of an accurate cryptic site prediction.

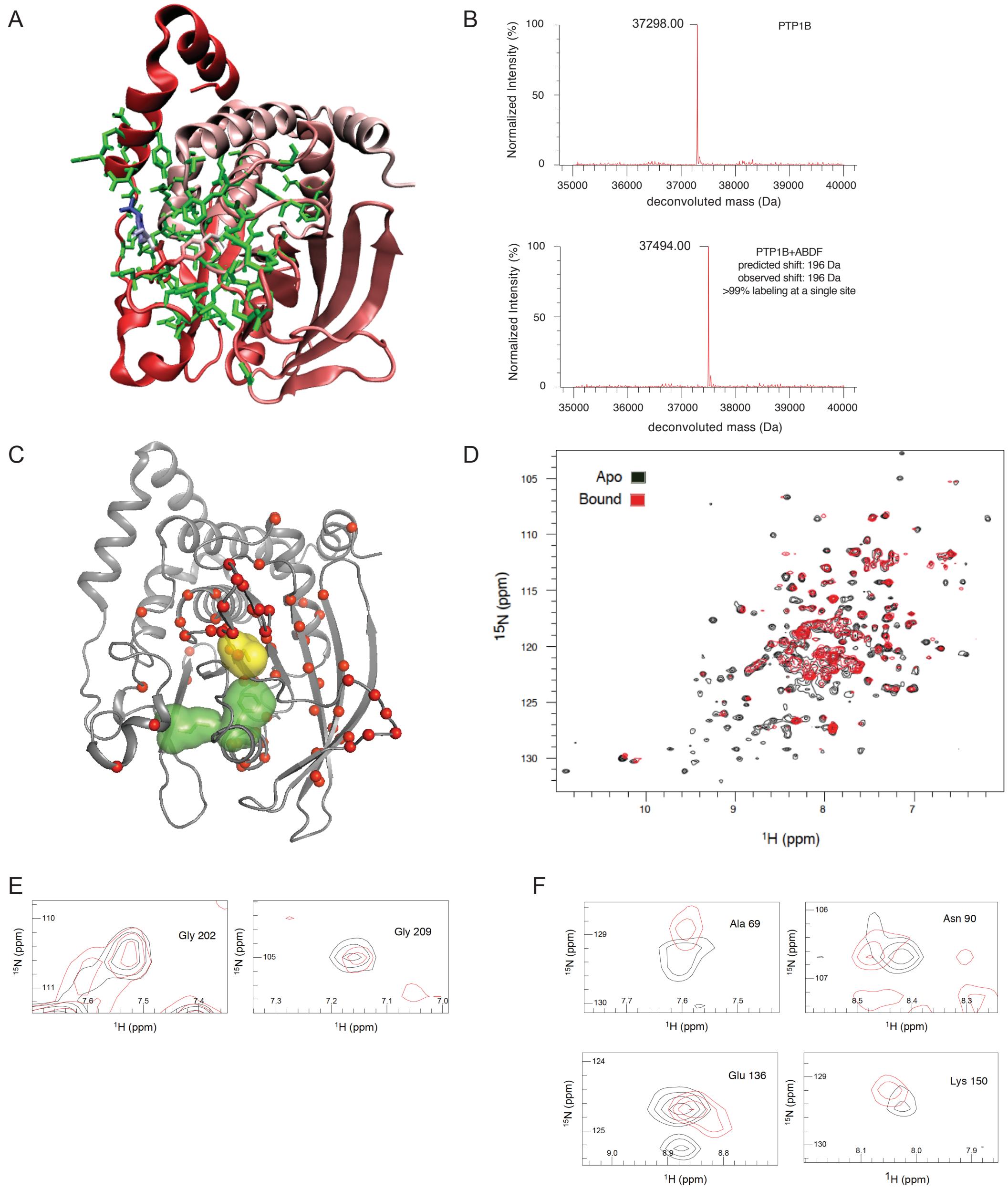


Figure S6: (*A*) Residues coupled with the active site of PTP1B are shown as green sticks (Weinkam et al., 2013). (*B*) Mass spectra of non-modified (top) and ABDF-modified PTP1B (bottom). The difference in mass (196) corresponds to the mass of the ABDF modification (197). (*C*) Many residues in PTP1B surrounding the predicted cryptic site (green surface) and the ABDF labeling site, Cys 121 (yellow surface), are unassigned due to broadened resonances (red spheres) (Meier et al., 2002). (*D*) Overlay of ^{1}H, ^{15}N TROSY HSQC spectra of PTP1B with

(black) and without (red) labeling by ABDF. PTP1B residues with no significant (*E*) or significant (*F*) chemical shift perturbations upon ABDF binding. Resonances are colored using the same color scheme as in *D*.

Supporting information

SI TEXT

The data set generation. We started by collecting all crystal structure PDB IDs of protein-ligand complexes from Binding MOAD (Benson et al., 2008) (downloaded on 2-27-2012); we only considered as ligands organic small molecules of biological relevance, excluding water and other solvent molecules, counterions, buffer components, metal ions, and crystallographic additives. We defined a binding site by selecting residues with at least one atom less than 5 Å away from any of the ligand atoms. Next, we searched for the structures of the same protein without any ligands at a given binding site, following these steps and criteria:

(i) we aligned all protein chain sequences from the Binding MOAD database to all protein chain sequences from PDB that are longer than 50 residues using the *blastp* algorithm (Altschul et al., 1990), and then selected pairs with 100% sequence identity as *apo-holo* pair candidates (504,647 pairs);

(ii) we removed pairs for which either of the two structures was determined at worse than 2.5 Å resolution;

(iii) we removed pairs with ligands in *apo* structures that have at least one atom closer than 10 Å to any atom in the *holo* binding site;

(iv) we grouped *apo-holo* pairs with identical sequences into clusters and for each cluster selected a single pair with the lowest all-atom binding site RMSD as the cluster representative (this resulted in 46,436 pairs);

(v) we further removed *apo* structures that contain other proteins, peptides, or nucleic acids bound within 10 Å from the ligand of interest, superimposed from the *holo* structure;

(vi) we removed *apo-holo* pairs that contained multiple copies of a ligand at the *holo* binding site, that contained amino acid ligands, or pairs whose *holo* binding sites contained less than 5 residues (21,928 pairs remained);

(vii) we removed *apo-holo* pairs with sequence gaps in *apo* structures longer than 3 residues or less than 5 Å away from the binding site;

(viii) we grouped protein sequences into clusters of 40% protein sequence identity, and then further split these clusters into groups of proteins that bind similar ligands (we defined ligand similarity by the Tanimoto distance using linear path fingerprints (FP2) from Open Babel (O'Boyle et al., 2011), followed by selecting the pair with the lowest all-atom RMSD from each group as the cluster representative;

(ix) and finally, we removed all *apo-holo* pairs with C-RMSD > 10 Å. This filtering resulted in a set of 4,766 *apo-holo* structure pairs.

We next utilized two pocket detection algorithms, ConCavity (Capra et al., 2009) and Fpocket (Le Guilloux et al., 2009), to evaluate the “goodness” of pockets in the *apo* and *holo* structures. The output of the Fpocket algorithm is a list of pockets with corresponding druggability scores, with each pocket defined as a set of coordinates depicting centers of fitting (alpha) spheres. We define the Fpocket residue pocket score as the maximum druggability score among the alpha spheres within 5 Å of the residue, or 0 if there are no alpha spheres (and hence pockets) in its neighbourhood. In contrast, ConCavity already provides a score on a per-residue basis, which we define as the ConCavity residue pocket score without additional processing. We use both Fpocket and ConCavity residue pocket scores to define cryptic sites and binding pockets. Cryptic sites are defined as sites with an average residue pocket score of less than 0.1 in the *apo* form and more than 0.4 in the *holo* form. Similarly, we defined binding pockets as binding sites with an average residue pocket score of more than 0.4 for the *apo* and *holo* forms, and Qi (Weinkam et al., 2013) between the *apo* and *holo* forms larger than 0.95. Such filtering resulted in a dataset of 468 *apo-holo* pairs with cryptic sites (190 unique *apo* structures), and 839 *apo-holo* pairs with binding pockets (191 unique *apo* structures).

We had to manually inspect both datasets of binding sites because of the high false-positive rate of pocket detection algorithms (the state-of-the-art algorithms are only ~70% sensitive (Schmidtke and Barril, 2010; Schmidtke et al., 2010) when applied to the unbound conformation of a protein), which resulted in the final datasets of 89 cryptic sites and 92 binding pocket *apo-holo* pairs. 10 randomly chosen cryptic *apo-holo* pairs were put aside for testing purposes. Also for testing purposes, we additionally selected 4 proteins with known cryptic sites from the literature (exportin-1, TEM1 -lactamse, IL-2, and Bcl-X) (**Tables S1** and **S4**).

In summary, the sequence similarity between a pair of two *apo* structures never exceeds 40%, except for 7 proteins that contain 2 different cryptic sites each, and a protein that contained 3 different cryptic sites. Moreover, out of 79 proteins in total, we obtained 59 groups of proteins with putative unique folds based on protein structure alignment (TM-align and TM-score thresholds of more than 0.7) (Xu and Zhang, 2010). Similarly, we retrieved a non-redundant dataset of 92 protein structures with binding pockets; none of the protein sequences is more than 40% identical to any other sequence, and protein structure alignment suggests 69 putative folds.

Pre-processing PDB files. Many PDB files contain more than 1 macromolecule (*ie*, a biologically relevant assembly of multiple macromolecules or an assembly of macromolecules interacting through crystallographic contacts), non-specific solvent molecules, regions of missing density, and modified protein sequences (*eg,* truncated loops or termini). To more accurately assess structural properties (for example, an estimate of surface area would be inaccurate for the residues next to an interacting molecule or a region with a missing density), we deleted from the PDB file all macromolecules except the macromolecule (*ie,* chain) of interest. Furthermore, we filled the gaps in the crystal structures by aligning a PDB structure to the corresponding SEQRES sequence, and then used the loop-modeling routine in Modeller

(Sali and Blundell, 1993) to build a loop conformation while keeping the rest of the protein structure rigid. We built 20 models per chain, and kept the one with the lowest DOPE score (Shen and Sali, 2006) for further analyses.

Molecular dynamics simulations. Standard molecular dynamics simulations are computationally expensive, which makes them impractical for studying the dynamics of the large number of proteins in our dataset. In contrast, AllosMod simulates dynamics more efficiently, by relying on a simplified energy landscape whose minimum is defined by the input native structure (Weinkam et al., 2013). We initialized 50 simulations from the randomized *apo* crystal structure coordinates, each one 6 ns long. The 50 simulations include 10 repeats at 5 different temperatures (300 K, 350 K, 400 K, 450 K, and 500 K), with 3 ps time steps – resulting in a total of 100,000 snapshot conformations.

Feature design. In total, we designed a set of 105 residue-based features that can be grouped into 3 categories: (*i*) features that describe protein sequence conservation, protein shape, and energetics, (*ii*) features that describe sequence conservation, shape, and energetics of neighborhood residues, and (*iii*) features derived from molecular dynamics simulations describing flexibility and dynamics of residues (**Table S2**). Protein shape calculations include p*rotrusion, compactness, convexity, rigidity, hydrophobicity* (using Wimley-White solvent model), and *charge density*, as described previously (Rossi et al., 2006). *Residue surface area* is defined as a sum of surface areas of individual atoms, which was determined by the CHASA algorithm (default probe radius) (Fleming et al., 2005) and Modeller (probe radius of 1.4 Å and 3.0 Å). We define *residue packing* of a given residue as the number of atoms of other residues within 4 Å from any atom in the residue, divided by the number of atoms in the residue. The *number of neighbors* is defined as the number of different residues within the same distance.

Distance to the surface is defined as the smallest distance between any atom of a given residue and the closest atom with surface area > 2 Å^2. *Pocket score* is derived from pocket prediction by Fpocket as explained above (*Data set generation* section).

Sequence conservation of a given sequence position is defined as the Shannon's entropy of reweighted amino-acid frequency counts in a multiple sequence alignment (Morcos et al., 2011). Multiple sequence alignments were obtained by aligning an individual *apo* sequence against the entire Uniprot (UniProt, 2013, 2014) database using the *blastp* algorithm. Clusters of homologous sequences above the 80% sequence identity threshold (used to reweight the amino-acid frequency counts) were calculated using the *usearch* algorithm (Edgar, 2010).

Features derived from molecular dynamics simulations include the mean and standard deviation of the following residue features: pairwise distance similarity metric (*Qi*), surface exposed area (with probe radius of 1.4 Å and 3.0 Å), protrusion, convexity, and pocket score. Additionally, we also calculated the percentage of snapshots with a given residue pocket score higher than 0.4, as well as the mean and standard deviation of the residue pocket scores above the 95th percentile.

Machine learning. To predict whether a given residue belongs to a cryptic site, we utilized Scikit-Learn and PyBrain implementations (Pedregosa et al., 2011; Schaul et al., 2010) of several different supervised machine-learning algorithms. We varied many parameters associated with a given algorithm (*eg*, different kernel functions, a range of different values for penalty parameters, different penalty functions, *etc.*). Furthermore, we mapped the accuracy as a function of scaling the dataset or changing class weights to take into account the unbalanced dataset (only ~5% of residues in our dataset are in cryptic sites). The residue classification accuracy of each combination of scaling, algorithm, and the corresponding set of parameters was evaluated using the confusion matrix with leave-one-out cross-validation (**Fig. S3*A***). The

SVM algorithm with quadratic kernel function, scaling, and penalty parameter *C*, kernel coefficient *gamma*, and independent term in kernel function *coef0* of 0.158, 0.0, 2.154, respectively, was found to perform most accurately. Furthermore, we selected the subset of 19 features that gives the highest accuracy using a greedy-forward approach, evaluating area under the ROC curve and leave-one-out cross-validation (**Fig. S3*B***). The web server for predicting cryptic binding sites is available at http://salilab.org/cryptosite. On average, it takes less than 2 days to predict cryptic sites in a protein of ~300 residues (most of this time is spent on molecular dynamics simulations by AllosMod).

Estimating the size of the druggable proteome. To estimate the size of the druggable proteome, we first retrieved a subset of 11,201 human protein structures from the PDB longer than 50 residues and with X-ray resolution better than 3.5 Å. For each one of these structures, we predicted cryptic sites by using our algorithm without residue-based features that require time-consuming AllosMod simulations. A cryptic site is predicted when at least 5 adjacent residues have the cryptic site score larger than 0.056; two residues are adjacent when any of their atoms are within 3.5 Å of each other. A binding pocket is predicted equivalently, but using the Fpocket-based pocket score with a threshold of 0.5. The two thresholds were chosen to approximately match the sensitivity and specificity of cryptic site and binding pocket prediction (true positive rates of 0.51 and 0.57, and false positive rates of 0.22 and 0.21 for cryptic sites and binding pockets, respectively (Schmidtke and Barril, 2010)). To estimate the number of druggable disease-associated proteins, we first retrieved a dataset of disease-associated genes from OMIM *morbidmap* (3,329 genes) (Hamosh et al., 2005). Druggable disease-associated proteins are defined as proteins of known structure that are encoded by these genes and have at least one predicted cryptic site or binding pocket; for proteins with more than one determined structure, we only include into our analysis the structure with the highest number of predicted cryptic sites or pockets.

Protein expression and purification. The short form of the catalytic domain (residues 1-298) of wild-type human PTP1B was cloned into pET24b. BL21 *E. coli* cells were transformed with this construct. 5 mL overnight cultures of the transformed cells were diluted into 1 L of M9 minimal medium with 1 g/L $^{15}NH_4Cl$ and 35 μg/mL kanamycin, and grown at 37ºC until absorbance at 600 nm reached 0.95 (about 7 hours). PTP1B expression was induced by adding isopropyl-β-D-thiogalactoside (IPTG) to a concentration of 0.5 mM and incubating for 16 hours at 18ºC. Cell pellets were harvested by centrifugation and stored at -80ºC.

For purification, cell pellets were resuspended in lysis buffer (100 mM MES pH 6.5, 1 mM EDTA, 1 mM DTT) (Puius et al., 1997) and lysed by homogenization with an Emulsiflex C3 machine. After centrifugation of the lysate, the supernatant was filtered and loaded onto a Sepharase (SP) cation exchange column equilibrated in lysis buffer. The column was run over a gradient from 0-1 M NaCl; PTP1B eluted around 200 mM NaCl. Those fractions were pooled, concentrated by centrifugation, and loaded onto a Superdex 200 (S200) size-exclusion column equilibrated in 100 mM MES pH 6.5, 1 mM EDTA, 1 mM DTT, 200 mM NaCl. PTP1B-containing fractions were pooled, filtered, and dialysed at 4ºC for 1-2 hours into NMR buffer (20 mM Bis-Tris propane, 25 mM NaCl, 3mM DTT, 0.2 mM EDTA, pH 6.5) (Whittier et al., 2013). The protein sample was then concentrated via centrifugation to 230 μM.

Covalent labeling of PTP1B with ABDF. The protein sample was diluted to 25 μM in NMR buffer without DTT. We then added 500 μM ABDF for 1 hour at room temperature. Next, the unreacted ABDF was removed and the protein was exchanged back into NMR buffer with DTT using a PD10 desalting column. Finally, the protein was concentrated via centrifugation to 110 µM.

TROSY NMR data acquisition. We prepared NMR samples with 7% D_2O and 200 and 110 µM of the apo and ABDF-labeled protein species, respectively. ^1H, ^{15}N TROSY HSQC spectra were collected with a Bruker 800 MHz magnet at 293 K for >5 hours and >7 hours, respectively. Although many resonances were too broadened to confidently match with published assignments (Meier et al., 2002) because we used undeuterated protein in contrast to previous work (Krishnan et al., 2014; Meier et al., 2002; Whittier et al., 2013), we were able to confidently monitor the resonances of several residues between the two spectra (**Fig. 5** and **S6**).

SI REFERENCES

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990). Basic local alignment search tool. Journal of molecular biology *215*, 403-410.

Benson, M.L., Smith, R.D., Khazanov, N.A., Dimcheff, B., Beaver, J., Dresslar, P., Nerothin, J., and Carlson, H.A. (2008). Binding MOAD, a high-quality protein-ligand database. Nucleic acids research *36*, D674-678.

Capra, J.A., Laskowski, R.A., Thornton, J.M., Singh, M., and Funkhouser, T.A. (2009). Predicting protein ligand binding sites by combining evolutionary sequence conservation and 3D structure. PLoS computational biology *5*, e1000585.

Edgar, R.C. (2010). Search and clustering orders of magnitude faster than BLAST. Bioinformatics *26*, 2460-2461.

Fleming, P.J., Fitzkee, N.C., Mezei, M., Srinivasan, R., and Rose, G.D. (2005). A novel method reveals that solvent water favors polyproline II over beta-strand conformation in peptides and unfolded proteins: conditional hydrophobic accessible surface area (CHASA). Protein science : a publication of the Protein Society *14*, 111-118.

Hamosh, A., Scott, A.F., Amberger, J.S., Bocchini, C.A., and McKusick, V.A. (2005). Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. Nucleic acids research *33*, D514-517.

Krishnan, N., Koveal, D., Miller, D.H., Xue, B., Akshinthala, S.D., Kragelj, J., Jensen, M.R., Gauss, C.M., Page, R., Blackledge, M.*, et al.* (2014). Targeting the disordered C terminus of PTP1B with an allosteric inhibitor. Nature chemical biology.

Le Guilloux, V., Schmidtke, P., and Tuffery, P. (2009). Fpocket: an open source platform for ligand pocket detection. BMC bioinformatics *10*, 168.

Meier, S., Li, Y.C., Koehn, J., Vlattas, I., Wareing, J., Jahnke, W., Wennogle, L.P., and Grzesiek, S. (2002). Backbone resonance assignment of the 298 amino acid catalytic domain of protein tyrosine phosphatase 1B (PTP1B). Journal of biomolecular NMR *24*, 165-166.

Morcos, F., Pagnani, A., Lunt, B., Bertolino, A., Marks, D.S., Sander, C., Zecchina, R., Onuchic, J.N., Hwa, T., and Weigt, M. (2011). Direct-coupling analysis of residue coevolution captures native contacts across many protein families. Proceedings of the National Academy of Sciences of the United States of America *108*, E1293-1301.

O'Boyle, N.M., Banck, M., James, C.A., Morley, C., Vandermeersch, T., and Hutchison, G.R. (2011). Open Babel: An open chemical toolbox. Journal of cheminformatics *3*, 33.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V.*, et al.* (2011). Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research *12*, 2825−2830.

Puius, Y.A., Zhao, Y., Sullivan, M., Lawrence, D.S., Almo, S.C., and Zhang, Z.Y. (1997). Identification of a second aryl phosphate-binding site in protein-tyrosine phosphatase 1B: a paradigm for inhibitor design. Proceedings of the National Academy of Sciences of the United States of America *94*, 13420-13425.

Rossi, A., Marti-Renom, M.A., and Sali, A. (2006). Localization of binding sites in protein structures by optimization of a composite scoring function. Protein science : a publication of the Protein Society *15*, 2366-2380.

Sali, A., and Blundell, T.L. (1993). Comparative protein modelling by satisfaction of spatial restraints. Journal of molecular biology *234*, 779-815.

Schaul, T., Bayer, J., Wierstra, D., Sun, Y., Felder, M., Sehnke, F., Rueckstieß, T., and Schmidhuber, J. (2010). PyBrain. Journal of Machine Learning Research *11*, 743-746.

Schmidtke, P., and Barril, X. (2010). Understanding and predicting druggability. A high-throughput method for detection of drug binding sites. Journal of medicinal chemistry *53*, 5858-5867.

Schmidtke, P., Le Guilloux, V., Maupetit, J., and Tuffery, P. (2010). fpocket: online tools for protein ensemble pocket detection and tracking. Nucleic acids research *38*, W582-589.

Shen, M.Y., and Sali, A. (2006). Statistical potential for assessment and prediction of protein structures. Protein science : a publication of the Protein Society *15*, 2507-2524.

UniProt, C. (2013). Update on activities at the Universal Protein Resource (UniProt) in 2013. Nucleic acids research *41*, D43-47.

UniProt, C. (2014). Activities at the Universal Protein Resource (UniProt). Nucleic acids research *42*, D191-198.

Weinkam, P., Chen, Y.C., Pons, J., and Sali, A. (2013). Impact of mutations on the allosteric conformational equilibrium. Journal of molecular biology *425*, 647-661.

Whittier, S.K., Hengge, A.C., and Loria, J.P. (2013). Conformational motions regulate phosphoryl transfer in related protein tyrosine phosphatases. Science *341*, 899-903.

Xu, J., and Zhang, Y. (2010). How significant is a protein structure similarity with TM-score = 0.5? Bioinformatics *26*, 889-895.

Supplementary tables

The manuscript is submitted, and the material will become available online upon its publication.

Empty page.

Publishing agreement

It is the policy of the University to encourage the distribution of all theses, dissertations, and manuscripts. Copies of all UCSF theses, dissertations, and manuscripts will be routed to the library via the Graduate Division. The library will make all theses, dissertations, and manuscripts accessible to the public and will preserve these to the best of their abilities, in perpetuity.

I hereby grant permission to the Graduate Division of the University of California, San Francisco to release copies of my thesis, dissertation, or manuscript to the Campus Library to provide access and preservation, in whole or in part, in perpetuity.

Author Signature

6/27/2014
Date