An open library of human kinase domain constructs for automated bacterial expression

Daniel L. Parton,[1] Sonya M. Hanson,[1] Lucelenie Rodríguez-Laureano,[1] Steven K. Albanese,[2] Scott Gradia,[3] Chris Jeans,[3] Markus Seeliger,[4] Nicholas Levinson,[5] and John D. Chodera[1,*]

[1]Computational Biology Program, Sloan Kettering Institute,
Memorial Sloan Kettering Cancer Center, New York, NY 10065[†]
[2]Gerstner Sloan Kettering Graduate School, Memorial Sloan Kettering Cancer Center, New York, NY 10065[‡]
[3]QB3 MacroLab, University of California, Berkeley, CA 94720[§]
[4]Department of Pharmacological Sciences, Stony Brook University Medical School, Stony Brook, NY 11794[¶]
[5]Department of Pharmacology, University of Minnesota, Minneapolis, MN 55455[**]

(Dated: February 8, 2016)

Kinases play a critical role in cellular signaling pathways. Human kinase dysregulation has been linked to a number of diseases, such as cancer, diabetes, and inflammation, and as a result, much of the effort in developing treatments (and perhaps 30% of *all* current drug development effort) has focused on shutting down aberrant kinases with targeted inhibitors. While insect and mammalian expression systems are frequently utilized for the expression of human kinases, they cannot compete with the simplicity and cost-effectiveness of bacterial expression systems, which historically had found human kinases difficult to express. Following the demonstration that phosphatase coexpression could give high yields of Src and Abl kinase domains in inexpensive bacterial expression systems [**?**], we have performed a large-scale expression screen to generate a library of His-tagged human kinase domain constructs that express well in a simple automated bacterial expression system where phosphatase coexpression (YopH for Tyr kinases, lambda for Ser/Thr kinases) is used. Starting from 96 kinases with crystal structures and any reported bacterial expression, we engineered a library of human kinase domain constructs and screened their coexpression with phosphatase, finding 52 kinases with yields greater than 2 mg/L culture. All sequences and expression data are provided online at `https://github.com/choderalab/kinase-ecoli-expression-panel`, and the plasmids are in the process of being made available through AddGene.

I. INTRODUCTION

Kinases play a critical role in cellular signaling pathways. Perturbations to these pathways due to mutation, translocation, or upregulation events can cause one or more kinases to become highly active and cease responding normally to regulatory signals, often with disastrous consequences. Kinase dysregulation has been linked to a number of diseases, such as cancer, diabetes, and inflammation. Cancer alone is the second leading cause of death in the United States, accounting for nearly 25% of all deaths; in 2015, over 1.7 million new cases were diagnosed, with over 580,000 deaths [**?**]. Much of the effort in developing treatments (and perhaps 30% of *all* current drug development effort) has focused on shutting down aberrant kinases with targeted inhibitors.

The discovery of imatinib, which specifically targets the Abl kinase dysregulated in chronic myelogenous leukemia (CML) patients to abate disease progression, was transformative in revealing the enormous therapeutic potential of selective kinase inhibitors, kindling hope that this remarkable success could be recapitulated for other cancers and diseases [**?**]. While there are now 31 FDA-approved selective kinase inhibitors, these molecules were approved for targeting only 13 out of ∼500 human kinases, with the vast majority targeting just a handful of kinases; the discovery of therapeutically effective inhibitors for other kinases has proven remarkably challenging.

The ability to probe human kinase biochemistry, biophysics, and structural biology in the laboratory is essential to making rapid progress in the understanding of kinase regulation and the design of selective inhibitors. While human kinase expression in baculovirus-infected insect cells can achieve high success rates [**? ?**], it cannot compete in cost or convenience with bacterial expression. While a survey of 62 full-length non-receptor human kinases found that over 50% express well in *E. coli* [**?**], there is often a desire to express and manipulate only the soluble kinase domains, since these are the molecular targets of therapy for targeted kinase inhibitors and could be studied even for receptor-type kinases. While removal of regulatory domains can negatively impact expression, coexpression with phosphatase was shown to greatly enhance bacterial kinase expression in Src and Abl tyrosine kinases, presumably by ensuring that kinases remain in an unphosphorylated inactive form [**?**].

The protein databank (PDB) now contains over 100 human kinases that—according to the PDB data records—were expressed in bacteria. Since bacterial expression is often complicated by the need to tailor expression and purification protocols individually for each protein expressed, we wondered whether a simple, uniform, automatable expression and purification protocol could be used to express a large number of human kinases to produce a convenient bacterial expression library to facilitate kinase research and

* Corresponding author; john.chodera@choderalab.org
† daniel.parton@choderalab.org
‡ steven.albanese@choderalab.org
§ Current address: Caribou Biosciences, Berkeley, CA 94720; sgradia@cariboubio.com
¶ markus.seeliger@stonybrook.edu
** nml@umn.edu

selective inhibitor development. As a first step toward this goal, we developed a structural informatics pipeline to use available kinase structural data and associated metadata to select constructs from available human kinase libraries to clone into a standard set of vectors intended for phosphatase coexpression. Automated expression screening in Rosetta2 cells found that 52 human kinase domains express with yields greater than 2 mg/L culture, which should be usable for biochemical, biophysical, screening, and structural biology studies.

All code and source files used in this project can be found at `https://github.com/choderalab/kinase-ecoli-expression-panel`, and a convenient sortable table of results can be viewed at `http://choderalab.github.io/kinome-data/kinase_constructs-addgene_hip_sgc.html`.

II. METHODS

A. Semi-automated selection of kinase construct sequences for E. coli expression

1. Selection of human protein kinase domain targets

Human protein kinases were selected by querying the UniProt API for any human protein with a domain containing the string "protein kinase", and which was manually annotated and reviewed (i.e. a Swiss-Prot entry). The query string used was:

```
taxonomy:"Homo sapiens (Human) [9606]" AND
domain:"protein kinase" AND reviewed:yes
```

Data was returned by the UniProt API in XML format and contained protein sequences and relevant PDB structures, along with many other types of genomic and functional information. To select active protein kinase domains, the UniProt domain annotations were searched using the regular expression `^Protein kinase(?!; truncated)(?!; inactive)`, which excludes certain domains annotated "Protein kinase; truncated" and "Protein kinase; inactive". Sequences for the selected domains were then stored. The sequences were derived from the canonical isoform as determined by UniProt.

2. Matching target sequences with relevant PDB constructs

Each target kinase gene was matched with the same gene in any other species where present, and UniProt data was downloaded for those genes also. The UniProt data included a list of PDB structures which contain the protein, as well as their sequence spans in the coordinates of the UniProt canonical isoform. This information was used to filter out PDB structures which did not include the protein kinase domain; structures were kept if they included the protein kinase domain sequence less 30 residues at each end. PDB coordinate files were then downloaded for each PDB entry. The coordinate files contain various metadata, including an `EXPRESSION_SYSTEM` annotation, which was used to filter PDB entries to keep only those which include the phrase "ESCHERICHIA COLI". The majority of PDB entries returned had an `EXPRESSION_SYSTEM` tag of "ESCHERICHIA COLI", while a small number had "ESCHERICHIA COLI BL21" or "ESCHERICHIA COLI BL21(DE3).

The PDB coordinate files also contain SEQRES records, which should contain the protein sequence used in the crystallography or NMR experiment. According to the PDB documentation (`http://deposit.rcsb.org/format-faq-v1.html`), "All residues in the crystal or in solution, including residues not present in the model (i.e., disordered, lacking electron density, cloning artifacts, HIS tags) are included in the SEQRES records." However, we found that these records are very often misannotated, instead representing only the crystallographically resolved residues. Since expression levels can be greatly affected by insertions or deletions of only one or a few residues at either terminus [**?**], it is important to know the full experimental sequence, and we thus needed a way to measure the authenticity of a given SEQRES record. We developed a crude measure by hypothesizing that a) most crystal structures would be likely to have at least one or a few unresolved residues at one or both termini and b) the presence of an expression tag (which is typically not crystallographically resolved) would indicate an authentic SEQRES record. To achieve this, unresolved residues were first defined by comparing the SEQRES sequence to the resolved sequence, using the SIFTS service to determine which residues were not present in the canonical isoform sequence. Then regular expression pattern matching was used to detect common expression tags at the N- or C-termini. Sequences with a detected expression tag were given a score of 2, while those with any unresolved sequence at the termini were given a score of 1, and the remainder were given a score of 0. This data was not used to filter out PDB structures at this stage, but was stored to allow for subsequent selection of PDB constructs based on likely authenticity. Also stored for each PDB sequence was the number of residues extraneous to the target kinase domain, and the number of residue conflicts with the UniProt canonical isoform within that domain span.

3. Plasmid libraries

As a source of kinase DNA sequences, we purchased three kinase plasmid libraries: the addgene Human Kinase ORF kit , a kinase library from the Structural Genomics Consortium (SGC), Oxford (`http://www.thesgc.org`), and a kinase library from the PlasmID Repository maintained by the Dana-Farber/Harvard Cancer Center. The aim was to subclone the chosen sequence constructs from these plasmids, though we did not use the same vectors. Annotated data for the kinases in each library was used to match them against the human protein kinases selected for this project. A Python script was written which translated the plasmid ORFs into protein sequences, and aligned them against the target kinase domain sequences from UniProt. Also calculated were the number of extraneous protein residues in the ORF, relative to the target kinase domain sequence, and the number of residue conflicts.

4. *Selection of sequence constructs for expression*

Of the kinase domain targets selected from UniProt, we filtered out those with no matching plasmids from our available plasmid libraries and/or no suitable PDB construct sequences. For this purpose, a suitable PDB construct sequence was defined as any with an authenticity score > 0, i.e. those derived from SEQRES records with no residues outside the span of the resolved structure. Plasmid sequences and PDB constructs were aligned against each target domain sequence, and various approaches were then considered for selecting a) the sequence construct to use for each target, and b) the plasmid to subclone it from. Candidate sequence constructs were drawn from two sources - PDB constructs and the SGC plasmid library. The latter sequences were included because the SGC plasmid library was the only one of the three libraries which had been successfully tested for E. coli expression.

For most of the kinase domain targets, multiple candidate sequence constructs were available. To select the most appropriate sequence construct, we sorted them first by authenticity score, then by the number of conflicts relative to the UniProt domain sequence, then by the number of residues extraneous to the UniProt domain sequence span. The top-ranked construct was then chosen. In cases where multiple plasmids were available, these were sorted first by the number of conflicts relative to the UniProt domain sequence, then by the number of residues extraneous to the UniProt domain sequence span, and the top-ranked plasmid was chosen.

This process resulted in a set of 96 kinase domain constructs, which (by serendipity) matched the 96-well plate format we planned to use for parallel expression testing. We therefore selected these construct sequences for expression testing.

A sortable table of results can be viewed at http://choderalab.github.io/kinome-data/kinase_constructs-addgene_hip_sgc.html.

5. *Other notes*

While much of this process was performed programmatically using Python, many steps required manual supervision and intervention. We hope eventually to develop a fully automated software package for the selection of expression construct sequences for a given protein family, but this was not possible within the scope of this article.

B. Expression testing

For each target, the selected construct sequence was subcloned from the selected DNA plasmid. Expression testing was performed by the QB3 MacroLab (QB3 MacroLab, University of California, Berkeley, CA 94720) [http://qb3.berkeley.edu/qb3/macrolab/], a core facility offering automated gene cloning and recombinant protein expression and purification services.

Each kinase domain was tagged with a N-terminal His10-TEV and coexpressed with either the truncated YopH164 for Tyr kinases or lambda phosphatase for Ser/Thr kinases. All construct sequences were cloned into the 2BT10 plasmid, an AMP resistant ColE1 plasmid with a T7 promoter, using LIC (ligation-independent cloning). The inserts were generated by PCR using the LICv1 forward and reverse tags on the primers (LICv1 FW= TACTTCCAATCCAATGCA; LICv1 RV= TTATCCACTTCCAATGTTATTA). Gel purified PCR products were LIC treated with dCTP. Plasmid was linearized, gel purified and LIC treated with dGTP. LIC-treated plasmid and insert were mixed together and transformed into XL1-Blues for plasmid preps.

Expression was performed in Rosetta2 cells grown with Magic Media (Invitrogen autoinducing medium), 100 μg/mL of carbenicillin and 100 μg/mL of spectinomycin. Single colonies of transformants were cultivated with 900 μL of MagicMedia into a gas permeable sealed 96-well block. The cultures were incubated at 37°C for 4 hours and then at 16 °C for 40 hours while shaking. Next, cells were centrifuged and the pellets were frozen at -80 °C overnight. Cells were lysed on a rotating platform at room temperature for an hour using 700 μL of SoluLyse (Genlantis) supplemented with 400 mM NaCl, 20 mM imidazole, 1μg/mL pepstatin, 1μg/mL leupeptin and 0.5 mM PMSF.

For protein purification, 500 μL of the soluble lysate was added to a 25 μL Ni-NTA resin in a 96-well filter plate. Nickel Buffer A (25 mM HEPES pH 7.5, 5% glycerol, 400 mM NaCl, 20 mM imidazole, 1 mM BME) was added and the plate was shaken for 30 minutes at room temperature. The resin was washed with 2 mL of Nickel Buffer A. Target proteins were eluted by a 2 hour incubation at room temperature with 10 μg of TEV protease in 80 μL of Nickel Buffer A per well and a subsequent wash with 40 μL of Nickel Buffer A to maximize protein release. Nickel Buffer B (25 mM HEPES pH 7.5, 5% glycerol, 400 mM NaCl, 400 mM imidazole, 1 mM BME) was used to elute TEV resistant material remaining on the resin. Untagged protein eluted with TEV protease was run on a LabChip GX II Microfluidic system to analyze the major protein species present. Samples of total cell lysate, soluble cell lysate and Nickel Buffer B elution were run on a SDS-PAGE for analysis.

We are currently making the library of kinase domain constructs, generated in this work, available for distribution through the plasmid repository Addgene. In the meantime, requests for plasmids can be directed to requests@choderalab.org.

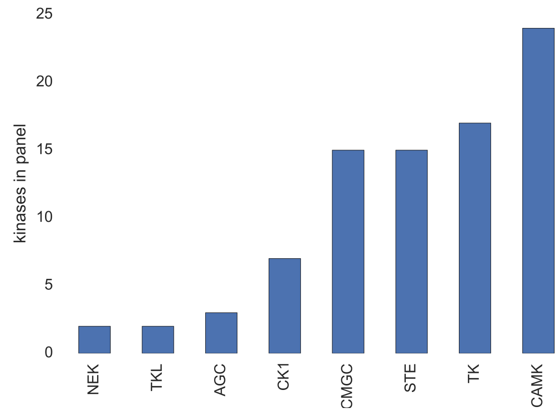

FIG. 1. Distribution of kinases in expression test panel by family. Histogram of the 96 kinases in the expression test panel, separated out by kinase family.

III. RESULTS

A. PDB mining results

Selecting the kinases and their constructs for this expression trial was primarily on the basis of expected success: these specific kinase constructs were bacterially expressed and purified to a degree that a crystal structure could be solved. While the expression protocols used to produce protein for crystallographic studies were likely tailored to maximize expression for individual proteins, we considered these kinases had a high chance of expressing in our semi-automated expression pipeline where the *same* protocol is utilized for all kinases. Statistics of the number of kinases obtained form the PDB mining procedure are shown in Figure 1. Surprisingly, the most highly sampled family was the CAMK family, suggesting that other researchers may have found this family particularly amenable to bacterial expression.

B. Small-scale kinase expression test in *E. coli*

A panel containing the 96 kinase domain constructs selected through our semi-automated method, was tested for expression in *E. coli*. From this initial test, 52 kinase domains showed reasonable expression (yield of more than 2 ng/μL eluate, which corresponds to 2 mg/L culture) (Table I). While the initial panel of 96 kinases was well-distributed across kinase families, the final most highly expressing (yield of more than 12 mg/L kinase) were not as evenly distributed (Figure 2). The 17 most highly expressing kinases showed relatively high purity after elution, though we note that eluting via TEV site cleavage results in a quantity of TEV protease in the eluate (Figure 3).

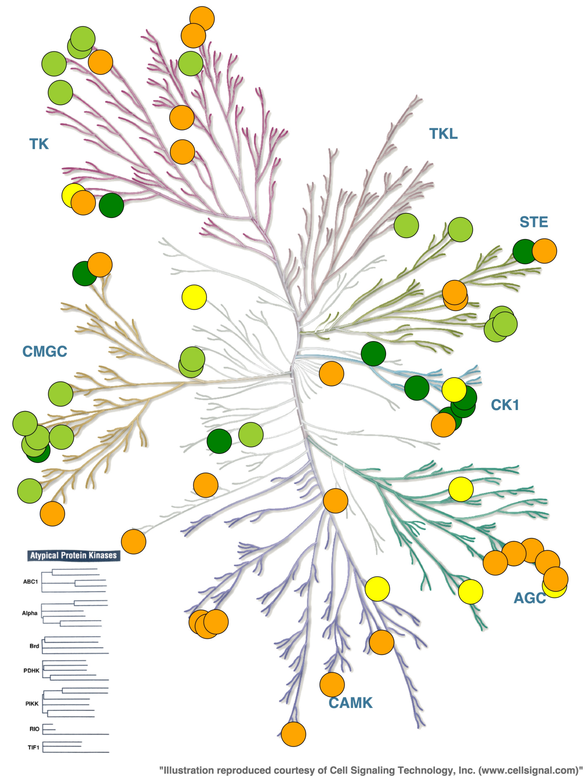

FIG. 2. Representation of kinase domain expression results on phylogenetic tree. Dark green circles represent kinases with expression above 50 mg/L yield. Light green circles represent kinases with expression between 50 and 12 mg/L yield. Yellow circles represent kinases with expression between 12 and 7 mg/L yield. Yellow circles represent kinases with any expression (even below 2 mg/L) up to 7 mg/L yield. Image made with KinMap: http://www.kinhub.org/kinmap.

IV. DISCUSSION

Bacterial coexpression of kinases appears to be a viable approach for studying a wide variety of human kinase domain constructs. We hope that other laboratories find these resources useful in their own work.

V. ACKNOWLEDGMENTS

DLP, SMH, LRL, SKA, and JDC acknowledge support from the Sloan Kettering Institute. JDC and DLP acknowledge partial support from NIH grant P30 CA008748 and the Louis V. Gerstner Young Investigator Award.

kinase expressed	phosphatase co-expressed	expected scale-up culture (mg/L)
MK14_HUMAN_D0	Lambda	70.7
VRK3_HUMAN_D0	Lambda	67.5
GAK_HUMAN_D0	Lambda	64.7
CSK_HUMAN_D0	Truncated YopH164	62.5
VRK1_HUMAN_D0	Lambda	62.3
KC1G3_HUMAN_D0	Lambda	56.3
FES_HUMAN_D0	Truncated YopH164	44.0
PMYT1_HUMAN_D0	Lambda	38.0
MK03_HUMAN_D0	Lambda	36.4
STK3_HUMAN_D0	Lambda	34.3
DYR1A_HUMAN_D0	Lambda	34.1
KC1G1_HUMAN_D0	Lambda	34.1
MK11_HUMAN_D0	Lambda	31.7
MK13_HUMAN_D0	Lambda	31.7
EPHB1_HUMAN_D0	Truncated YopH164	28.9
MK08_HUMAN_D0	Lambda	28.5
CDK16_HUMAN_D0	Lambda	26.9
EPHB2_HUMAN_D0	Truncated YopH164	25.1
PAK4_HUMAN_D0	Lambda	23.9
CDKL1_HUMAN_D0	Lambda	23.2
SRC_HUMAN_D0	Truncated YopH164	22.0
STK16_HUMAN_D0	Lambda	20.7
MAPK3_HUMAN_D0	Lambda	18.8
PAK6_HUMAN_D0	Lambda	18.0
CSK22_HUMAN_D0	Lambda	17.9
MERTK_HUMAN_D0	Truncated YopH164	16.8
PAK7_HUMAN_D0	Lambda	14.7
CSK21_HUMAN_D0	Lambda	14.5
EPHA3_HUMAN_D0	Truncated YopH164	14.1
BMPR2_HUMAN_D0	Lambda	14.1
M3K5_HUMAN_D0	Lambda	14.0
KCC2G_HUMAN_D0	Lambda	13.3
E2AK2_HUMAN_D0	Lambda	11.6
MK01_HUMAN_D0	Lambda	11.2
CSKP_HUMAN_D0	Lambda	10.1
CHK2_HUMAN_D0	Lambda	8.1
KC1G2_HUMAN_D0	Lambda	7.6
DMPK_HUMAN_D0	Lambda	7.6
KCC2B_HUMAN_D0	Lambda	7.1
FGFR1_HUMAN_D0	Truncated YopH164	6.1
KS6A1_HUMAN_D1	Lambda	5.7
DAPK3_HUMAN_D0	Lambda	4.0
STK10_HUMAN_D0	Lambda	3.7
KC1D_HUMAN_D0	Lambda	3.7
KC1E_HUMAN_D0	Lambda	3.5
NEK1_HUMAN_D0	Lambda	3.3
CDK2_HUMAN_D0	Lambda	3.1
ABL1_HUMAN_D0	Truncated YopH164	2.5
DAPK1_HUMAN_D0	Lambda	2.4
DYRK2_HUMAN_D0	Lambda	2.4
HASP_HUMAN_D0	Lambda	2.3
FGFR3_HUMAN_D0	Truncated YopH164	2.3
EPHB3_HUMAN_D0	Truncated YopH164	1.7
SLK_HUMAN_D0	Lambda	1.6
KCC2D_HUMAN_D0	Lambda	1.6
NEK7_HUMAN_D0	Lambda	1.3
PHKG2_HUMAN_D0	Lambda	1.3
VRK2_HUMAN_D0	Lambda	1.2
AAPK2_HUMAN_D0	Lambda	1.1
AURKA_HUMAN_D0	Lambda	1.1
MARK3_HUMAN_D0	Lambda	1.1
KAPCA_HUMAN_D0	Lambda	0.9
STK24_HUMAN_D0	Lambda	0.8
VGFR1_HUMAN_D0	Truncated YopH164	0.5
KCC4_HUMAN_D0	Lambda	0.4
KCC1G_HUMAN_D0	Lambda	0.3
KCC2A_HUMAN_D0	Lambda	0.3
FAK2_HUMAN_D0	Truncated YopH164	0.3

TABLE I. **Expression results by kinase.** Yield (determined by Caliper GX II quantitation of the expected size band) reported in mg/L culture, where total eluate volume was 120 μl from 900 μL bacterial culture.

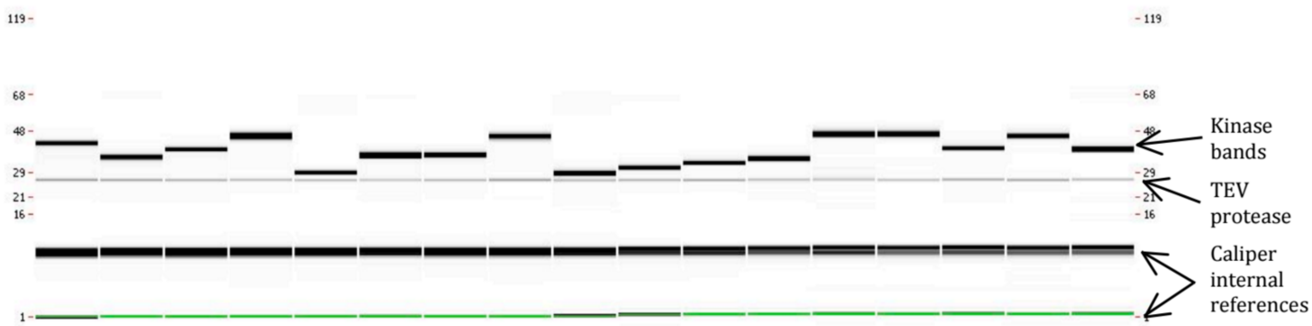

FIG. 3. **Synthetic gel image rendering of highest expressing kinases.** Caliper GX II synthetic gel image rendering of kinases expressing > 25 mg/L culture from microfluidic capillary electrophoresis quantitation.