

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 and John D. Chodera^{1,*}

¹Computational Biology Program, Sloan Kettering Institute,
Memorial Sloan Kettering Cancer Center, New York, NY 10065[†]

²Gerstner Sloan Kettering Graduate School, Memorial Sloan Kettering Cancer Center, New York, NY 10065[‡]

³QB3 MacroLab, University of California, Berkeley, CA 94720[§]

⁴Department of Pharmacological Sciences, Stony Brook University Medical School, Stony Brook, NY 11794[¶]

(Dated: February 2, 2016)

Kinases play a critical role in cellular signaling pathways. Human kinase dysregulation 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 have demonstrated success rates for the expression of human kinases, these expression systems 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 human kinase domain constructs that express well in a simple automated His-tagged bacterial expression system when coexpressed with phosphatase (YopH for Tyr kinases, lambda for Ser/Thr kinases). 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 68 kinases with yields greater than 2 mg/mL 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 target-

ing 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 selective inhibitor development. As a first step toward this goal, we developed a structural informatics pipeline to find

* 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

65 kinases already in the PDB and select constructs from available
 66 human kinase libraries to clone into a standard set of vectors intended for phosphatase coexpression. Automated
 67 expression screening in ROSETTA2 [BL21(DE3)] cells found that 68 human kinase domains express with yields greater
 69 than 2 μ g/mL, which should be usable for biochemical, biophysical, screening, and structural biology studies.

70 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.

78 II. METHODS

79 A. Semi-automated selection of kinase construct sequences 80 for E. coli expression

81 1. Selection of human protein kinase domain targets

82 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:

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

84 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.

100 2. Matching target sequences with relevant PDB constructs

101 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

115 returned had an EXPRESSION_SYSTEM tag of "ESCHERICHIA COLI", while a small number had "ESCHERICHIA COLI BL21" or "ESCHERICHIA COLI BL21(DE3).

116 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.

155 3. Plasmid libraries

156 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.thessgc.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, rela-

170 tive to the target kinase domain sequence, and the number
 171 of residue conflicts.

172 **4. Selection of sequence constructs for expression**

173 Of the kinase domain targets selected from UniProt, we
 174 filtered out those with no matching plasmids from our avail-
 175 able plasmid libraries and/or no suitable PDB construct se-
 176 quences. For this purpose, a suitable PDB construct se-
 177 quence was defined as any with an authenticity score > 0, i.e.
 178 those derived from SEQRES records with no residues out-
 179 side the span of the resolved structure. Plasmid sequences
 180 and PDB constructs were aligned against each target do-
 181 main sequence, and various approaches were then consid-
 182 ered for selecting a) the sequence construct to use for each
 183 target, and b) the plasmid to subclone it from. Candidate se-
 184 quence constructs were drawn from two sources - PDB con-
 185 structs and the SGC plasmid library. The latter sequences
 186 were included because the SGC plasmid library was the only
 187 one of the three libraries which had been successfully tested
 188 for *E. coli* expression.

189 For most of the kinase domain targets, multiple candidate
 190 sequence constructs were available. To select the most ap-
 191 propriate sequence construct, we sorted them first by au-
 192 thenticity score, then by the number of conflicts relative
 193 to the UniProt domain sequence, then by the number of
 194 residues extraneous to the UniProt domain sequence span.
 195 The top-ranked construct was then chosen. In cases where
 196 multiple plasmids were available, these were sorted first by
 197 the number of conflicts relative to the UniProt domain se-
 198 quence, then by the number of residues extraneous to the
 199 UniProt domain sequence span, and the top-ranked plas-
 200 mid was chosen.

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

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

209 **5. Other notes**

210 While much of this process was performed programmat-
 211 ically using Python, many steps required manual supervi-
 212 sion and intervention. We hope eventually to develop a fully
 213 automated software package for the selection of expression
 214 construct sequences for a given protein family, but this was
 215 not possible within the scope of this article.

216 **B. Expression testing**

217 For each target, the selected construct sequence was sub-
 218 cloned from the selected DNA plasmid. Expression testing

219 was performed by the QB3 MacroLab.

220 Each kinase domain was tagged with a N-terminal His10-
 221 TEV and coexpressed with either the truncated YopH164 for
 222 Tyr kinases or lambda phosphatase for Ser/Thr kinases. All
 223 construct sequences were cloned into the 2BT10 plasmid,
 224 an AMP resistant ColE1 plasmid with a T7 promoter, using
 225 LIC (ligation-independent cloning). The inserts were gen-
 226 erated by PCR using the LICv1 forward and reverse tags
 227 on the primers (LICv1 FW= TACTTCCAATCCAATGCA; LICv1
 228 RV=TTATCCACTTCCAATGTTATTAA). Gel purified PCR products
 229 were LIC treated with dCTP. Plasmid was linearized, gel pu-
 230 rified and LIC treated with dGTP. LIC-treated plasmid and in-
 231 sert were mixed together and transformed into XL1-Blues for
 232 plasmid preps.

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

245 For protein purification, 500 µL of the soluble lysate was
 246 added to a 25 µL Ni-NTA resin in a 96-well filter plate. Nickel
 247 Buffer A (25 mM HEPES pH 7.5, 5% glycerol, 400 mM NaCl,
 248 20 mM imidazole, 1 mM BME) was added and the plate was
 249 shaken for 30 minutes at room temperature. The resin was
 250 washed with 2 mL of Nickel Buffer A. Target proteins were
 251 eluted by a 2 hour incubation at room temperature with 10
 252 µg of TEV protease in 80 µL of Nickel Buffer A per well and
 253 a subsequent wash with 40 µL of Nickel Buffer A to maxi-
 254 mize protein release. Nickel Buffer B (25 mM HEPES pH 7.5,
 255 5% glycerol, 400 mM NaCl, 400 mM imidazole, 1 mM BME)
 256 was used to elute TEV resistant material remaining on the
 257 resin. Untagged protein eluted with TEV protease was run
 258 on a LabChip GX II Microfluidic system to analyze the major
 259 protein species present. Samples of total cell lysate, solu-
 260 ble cell lysate and Nickel Buffer B elution were run on a SDS-
 261 PAGE for analysis.

262 We are currently making the library of kinase domain con-
 263 structs, generated in this work, available for distribution
 264 through the plasmid repository Addgene. In the meantime,
 265 you can contact the Chodera Lab for a plasmid request.

266 **III. RESULTS**

267 **A. PDBs mining results**

268 Selecting the kinases and their constructs for this ex-
 269 pression trial was primarily on the basis of expected suc-
 270 cess: these specific kinase constructs previously expressed
 271 and purified easily enough that a crystal structure could
 272 be solved. While the final expression and characterization

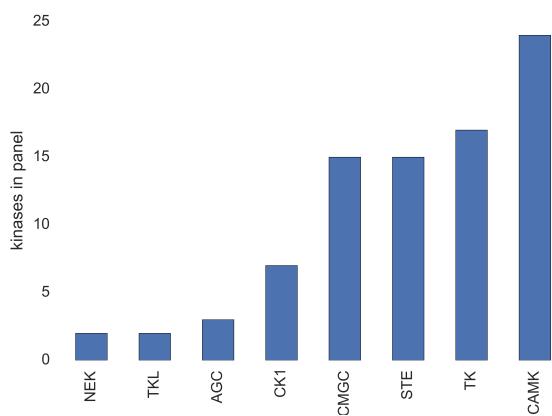


FIG. 1. Distribution of kinases in final expression panel by family. Histogram of the 96 kinases expressed in the expression panel, separated out by kinase family.

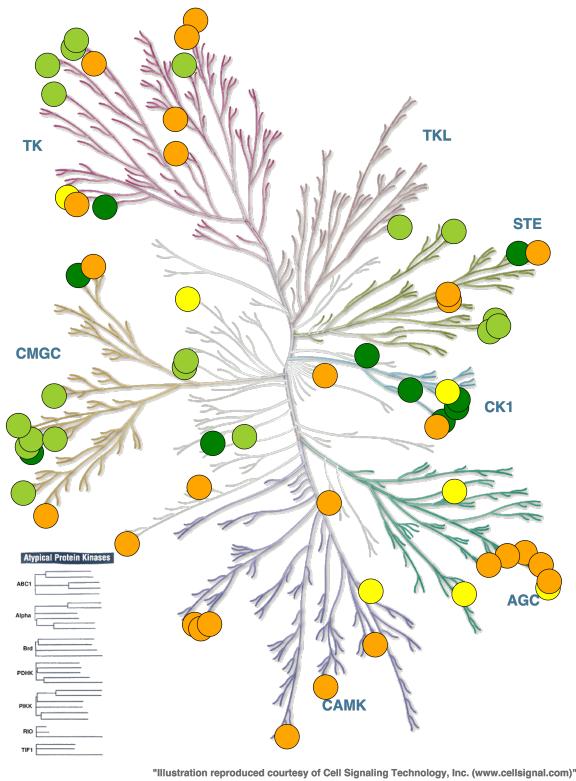
²⁷³ of these kinases was our ultimate goal, the patterns that
²⁷⁴ popped up via the use of our semi-automated pipeline are
²⁷⁵ also worth noting. The most highly sampled family in our
²⁷⁶ final panel, for example, was the CAMK family (Figure ??).

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

²⁷⁷ A panel containing the 96 kinase domain constructs se-
²⁷⁸ lected through our semi-automated method, was tested for
²⁷⁹ expression in *E. coli*. From this initial test, 68 kinase do-
²⁸⁰ mains expressed successfully (yield of more than 2 ng/ μ L
²⁸¹) (Table ??). While the initial panel of 96 kinases was well-
²⁸² distributed across kinase families, the final most highly ex-
²⁸³ pressing (yield of more than 100 ng/ μ L) were not as evenly
²⁸⁴ distributed (Figure ??). The 17 most highly expressing ki-
²⁸⁵ nases all were quite pure with some TEV contaminants still
²⁸⁶ present in Calliper gel images after elution with Imidazole
²⁸⁷ (Figure ??).

kinase expressed	phosphatase co-expressed	concentration (ng/μl)
MK14_HUMAN_D0	Lambda	530
VRK3_HUMAN_D0	Lambda	506
GAK_HUMAN_D0	Lambda	485
CSK_HUMAN_D0	Truncated YopH164	469
VRK1_HUMAN_D0	Lambda	467
KC1G3_HUMAN_D0	Lambda	422
FES_HUMAN_D0	Truncated YopH164	330
PMYT1_HUMAN_D0	Lambda	285
MK03_HUMAN_D0	Lambda	273
STK3_HUMAN_D0	Lambda	257
DYR1A_HUMAN_D0	Lambda	256
KC1G1_HUMAN_D0	Lambda	256
MK11_HUMAN_D0	Lambda	238
MK13_HUMAN_D0	Lambda	238
EPHB1_HUMAN_D0	Truncated YopH164	217
MK08_HUMAN_D0	Lambda	214
CDK16_HUMAN_D0	Lambda	202
EPHB2_HUMAN_D0	Truncated YopH164	188
PAK4_HUMAN_D0	Lambda	179
CDKL1_HUMAN_D0	Lambda	174
SRC_HUMAN_D0	Truncated YopH164	165
STK16_HUMAN_D0	Lambda	155
MAPK3_HUMAN_D0	Lambda	141
PAK6_HUMAN_D0	Lambda	135
CSK22_HUMAN_D0	Lambda	134
MERTK_HUMAN_D0	Truncated YopH164	126
PAK7_HUMAN_D0	Lambda	110
CSK21_HUMAN_D0	Lambda	109
EPHA3_HUMAN_D0	Truncated YopH164	106
BMPR2_HUMAN_D0	Lambda	106
M3K5_HUMAN_D0	Lambda	105
KCC2G_HUMAN_D0	Lambda	100
E2AK2_HUMAN_D0	Lambda	87
MK01_HUMAN_D0	Lambda	84
CSKP_HUMAN_D0	Lambda	76
CHK2_HUMAN_D0	Lambda	61
KC1G2_HUMAN_D0	Lambda	57
DMPK_HUMAN_D0	Lambda	57
KCC2B_HUMAN_D0	Lambda	53
FGFR1_HUMAN_D0	Truncated YopH164	46
KS6A1_HUMAN_D1	Lambda	43
DAPK3_HUMAN_D0	Lambda	30
STK10_HUMAN_D0	Lambda	28
KC1D_HUMAN_D0	Lambda	28
KC1E_HUMAN_D0	Lambda	26
NEK1_HUMAN_D0	Lambda	25
CDK2_HUMAN_D0	Lambda	23
ABL1_HUMAN_D0	Truncated YopH164	19
DAPK1_HUMAN_D0	Lambda	18
DYRK2_HUMAN_D0	Lambda	18
HASP_HUMAN_D0	Lambda	17
FGFR3_HUMAN_D0	Truncated YopH164	17
EPHB3_HUMAN_D0	Truncated YopH164	13
SLK_HUMAN_D0	Lambda	12
KCC2D_HUMAN_D0	Lambda	12
NEK7_HUMAN_D0	Lambda	10
PHKG2_HUMAN_D0	Lambda	10
VRK2_HUMAN_D0	Lambda	9
AAPK2_HUMAN_D0	Lambda	8
AURKA_HUMAN_D0	Lambda	8
MARK3_HUMAN_D0	Lambda	8
KAPCA_HUMAN_D0	Lambda	7
STK24_HUMAN_D0	Lambda	6
VGFR1_HUMAN_D0	Truncated YopH164	4
KCC4_HUMAN_D0	Lambda	3
KCC1G_HUMAN_D0	Lambda	2
KCC2A_HUMAN_D0	Lambda	2
FAK2_HUMAN_D0	Truncated YopH164	2

TABLE I. Expression results by kinase



"Illustration reproduced courtesy of Cell Signaling Technology, Inc. (www.cellsignal.com)"

FIG. 2. Representation of kinase domain expression results on phylogenetic tree. Dark green circles represent kinases with expression above 250 ng/μl. Light green circles represent kinases with expression between 100 and 250 ng/μl. Yellow circles represent kinases with expression between 50 and 100 ng/μl. Orange circles represent kinases with any expression up to 50 ng/μl. Image made with KinMap: <http://www.kinhub.org/kinmap>.

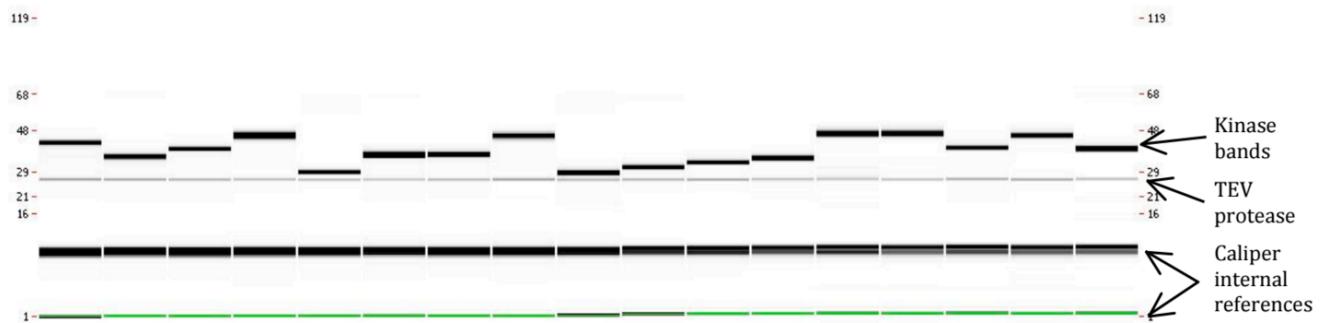


FIG. 3. Gel image of highest expressing kinases. Calliper gel image of kinases expressing > 200 ng/μl.