

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 1, 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 [1], 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 [2]. 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 [3]. 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 [4, 5], 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* [4], 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 [1].

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
 68 than 2 μ g/mL, which should be usable for biochemical, biophysical, screening, and structural biology studies.

72 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
 73 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 meta-data, 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

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

118 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 [DLP: ?CITE, or reference our 96-construct Abl1 expression panel], 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 (CITE) 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.

156 3. Plasmid libraries

157 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 calcu-

lated were the number of extraneous protein residues in the ORF, relative to the target kinase domain sequence, and the number of residue conflicts.

173 4. Selection of sequence constructs for expression

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

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

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

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

210 5. Other notes

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

217 B. Expression testing

218 For each target, the selected construct sequence was sub-
 219 cloned from the selected DNA plasmid. Expression testing
 220 performed by QB3 MacroLab.

221 All genes were cloned into the 2BT10 plasmid, an AMP
 222 resistant ColE1 plasmid with a T7 promoter. Each kinase
 223 domain was tagged with a N-terminal His10-TEV and co-
 224 expressed with either the truncated YopH164 (for Tyr ki-
 225 nases) or lambda phosphatase (for Ser/Thr kinases). Ex-
 226 pression was performed in Rosetta2 cells grown with Magic
 227 Media (Invitrogen autoinducing medium), 100 µg/mL of car-
 228 benicillin and 100 µg/mL of spectinomycin. Single colonies
 229 of transformants were cultivated with 900 µL of MagicMe-
 230 dia into a gas permeable sealed 96-well block. The cultures
 231 were incubated at 37°C for 4 hours and then at 16 °C for 40
 232 hours while shaking. Next, cells were centrifuged and the
 233 pellets were frozen at -80 °C overnight. Cells were lysed on
 234 a rotating platform at room temperature for an hour using
 235 700 µL of SoluLyse (Genlantis) supplemented with 400 mM
 236 NaCl, 20 mM imidazole and protease inhibitors.

237 For protein purification, 500 µL of the soluble lysate was
 238 added to a 25 µL Ni-NTA resin in a 96-well filter plate. Nickel
 239 Buffer A (25 mM HEPES pH 7.5, 5% glycerol, 400 mM NaCl,
 240 20 mM imidazole, 1 mM BME) was added and the plate was
 241 shaken for 30 minutes at room temperature. The resin was
 242 washed with 2 mL of Nickel Buffer A. Target proteins were
 243 eluted by a 2 hour incubation at room temperature with 10
 244 µg of TEV protease in 80 µL of Nickel Buffer A per well and
 245 a subsequent wash with 40 µL of Nickel Buffer A to maxi-
 246 mize protein release. Nickel Buffer B (25 mM HEPES pH 7.5,
 247 5% glycerol, 400 mM NaCl, 400 mM imidazole, 1 mM BME)
 248 was used to elute TEV resistant material remaining on the
 249 resin. Untagged protein eluted with TEV protease was run
 250 on a LabChip GX II Microfluidic system to analyze the major
 251 protein species present. Samples of total cell lysate, soluble
 252 cell lysate and Nickel Buffer B elution were run on a SDS-
 253 PAGE for analysis.

254 We are currently making the library of kinase domain con-
 255 structs, generated in this work, available for distribution
 256 through the plasmid repository [Addgene](#). In the meantime,
 257 you can contact the [Chodera Lab](#) for a plasmid request.

III. RESULTS

A. PDBs mining results

258
 259
 260 Selecting the kinases and their constructs for this ex-
 261 pression trial was primarily on the basis of expected suc-
 262 cess: these specific kinase constructs previously expressed
 263 and purified easily enough that a crystal structure could
 264 be solved. While the final expression and characterization
 265 of these kinases was our ultimate goal, the patterns that
 266 popped up via the use of our semi-automated pipeline are
 267 also worth noting. The most highly sampled family in our
 268 final panel, for example, was the CAMK family (Figure 1).

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

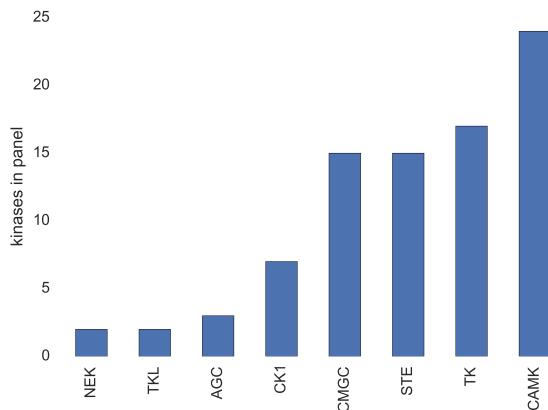


FIG. 1. Distribution of kinases in final expression panel by family. Histogram of the 96 kinases expressed in the expression panel, separated out by kinase family.

270 A panel containing the 96 kinase domain constructs se-
 271 lected through our semi-automated method, was tested for
 272 expression in *E. coli*. From this initial test, 68 kinase do-
 273 mains expressed successfully (yield of more than 2 ng/ μ L
 274) (Table 1). While the initial panel of 96 kinases was well-
 275 distributed across kinase families, the final most highly ex-
 276 pressing (yield of more than 100 ng/ μ L) were not as evenly
 277 distributed (Figure 2). The 17 most highly expressing kinases
 278 all were quite pure with some TEV contaminants still present
 279 in Calliper gel images after elution with Imidazole (Figure 3).

280 [1] M. A. Seeliger, M. Young, M. N. Henderson, P. Pellicena, D. S.
 281 King, A. M. Falick, and J. Kuriyan, *Protein Sci.* **14**, 3135 (2005).
 282 [2] American Cancer Society, *Cancer Facts & Figures* 2015, 2015.

283 [3] F. Stegmeier, M. Warmuth, W. R. Sellers, and M. Dorsch, *Clin.
 284 Pharm. & Therap.* **87**, 543 (2010).
 285 [4] S. P. Chambers, D. A. Austen, J. R. Fulghum, and W. M. Kim, *Pro-
 286 tein Expression and Purification* **36**, 40 (2004).
 287 [5] L. Wang, M. Foster, Y. Zhang, W. R. Tschantz, L. Yang, J. Worrall,
 288 C. Loh, and X. Xu, *Protein Express. Pur.* **61**, 204 (2008).

kinase expressed	phosphatase co-expressed	concentration (ng/ μ l)
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
KS6AI_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

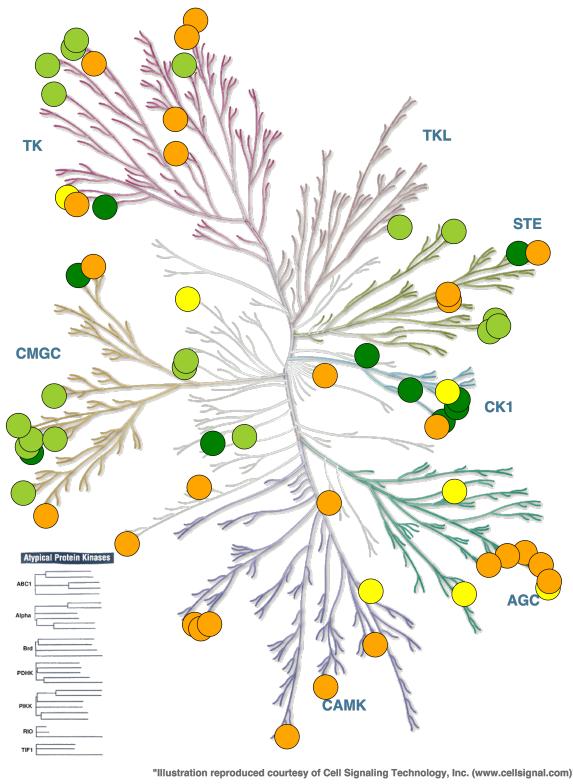


FIG. 2. Representation of kinase domain expression results on phylogenetic tree. Dark green circles represent kinases with expression above 250 ng/μl. Light green circles represent kinases with expression between 100 and 250 ng/μl. Yellow circles represent kinases with expression between 50 and 100 ng/μl.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

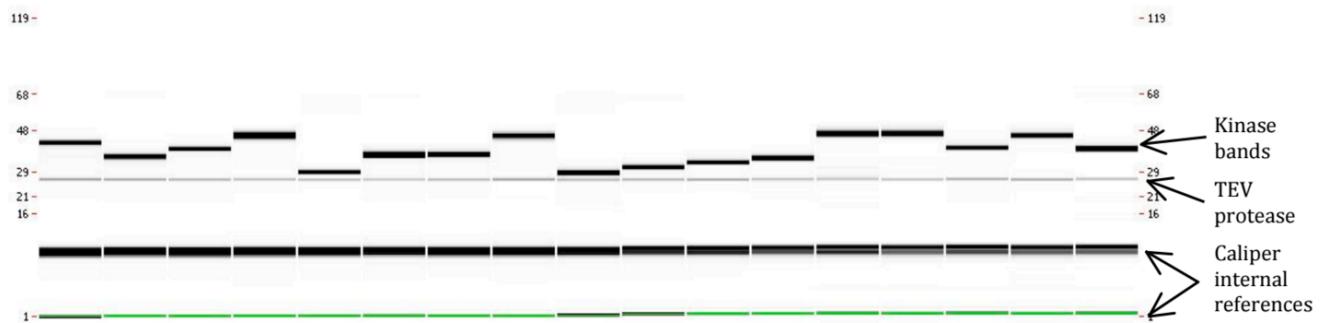


FIG. 3. Gel image of highest expressing kinases. Calliper gel image of kinases expressing > 200 ng/μl.