

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

Daniel L. Parton,<sup>1</sup> Sonya M. Hanson,<sup>1</sup> Lucelenie Rodríguez-Laureano,<sup>1</sup> Steven K. Albanese,<sup>2</sup> Scott Gradia,<sup>3</sup> Chris Jeans,<sup>3</sup> Markus Seeliger,<sup>4</sup> and John D. Chodera<sup>1,\*</sup>

<sup>1</sup>Computational Biology Program, Sloan Kettering Institute,  
Memorial Sloan Kettering Cancer Center, New York, NY 10065<sup>†</sup>

<sup>2</sup>Gerstner Sloan Kettering Graduate School, Memorial Sloan Kettering Cancer Center, New York, NY 10065<sup>‡</sup>

<sup>3</sup>QB3 MacroLab, University of California, Berkeley, CA 94720<sup>§</sup>

<sup>4</sup>Department of Pharmacological Sciences, Stony Brook University Medical School, Stony Brook, NY 11794<sup>¶</sup>

(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 51 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].

Surprisingly, the protein databank (PDB) now contains over 100 human kinases that—according to the PDB data records—were expressed in bacteria. Mindful that bacterial expression is often complicated by the need to tailor expression and purification protocols individually for each protein being 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

\* 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 pipeline to filter these structures and select constructs from  
 66 available human kinase library for cloning into a standard  
 67 set of vectors intended for phosphatase coexpression. Auto-  
 68 mated expression screening in ROSETTA2 [BL21(DE3)] cells  
 69 found that 51 human kinase domains express with yields  
 70 greater than 2 µg/mL, which should be usable for biochemical,  
 71 biophysical, screening, and structural biology studies.  
 72 All code and source files used in this project can  
 73 be found at <https://github.com/choderalab/kinase-ecoli-expression-panel>, and a con-  
 74 venient sortable table of results can be viewed at  
 75 [http://choderalab.github.io/kinome-data/kinase\\_constructs-addgene\\_hip\\_sgc.html](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  
 83 UniProt API for any human protein with a domain contain-  
 84 ing the string "protein kinase", and which was manually  
 85 annotated and reviewed (i.e. a Swiss-Prot entry). The query  
 86 string used was:  
 87 taxonomy: "Homo sapiens (Human) [9606]" AND  
 88 domain: "protein kinase" AND reviewed:yes  
 89 Data was returned by the UniProt API in XML format and  
 90 contained protein sequences and relevant PDB structures,  
 91 along with many other types of genomic and functional  
 92 information. To select active protein kinase domains, the  
 93 UniProt domain annotations were searched using the reg-  
 94 ular expression ^Protein kinase(?!; truncated)(?!;  
 95 inactive), which excludes certain domains annotated  
 96 "Protein kinase; truncated" and "Protein kinase; inactive".  
 97 Sequences for the selected domains were then stored. The  
 98 sequences were derived from the canonical isoform as  
 99 determined by UniProt.

#### 100 2. Matching target sequences with relevant PDB constructs

101 Each target kinase gene was matched with the same gene  
 102 in any other species where present, and UniProt data was  
 103 downloaded for those genes also. The UniProt data in-  
 104 cluded a list of PDB structures which contain the protein,  
 105 as well as their sequence spans in the coordinates of the  
 106 UniProt canonical isoform. This information was used to  
 107 filter out PDB structures which did not include the pro-  
 108 tein kinase domain - structures were kept if they included  
 109 the protein kinase domain sequence less 30 residues at  
 110 each end. PDB coordinate files were then downloaded for  
 111 each PDB entry. The coordinate files contain various meta-  
 112 data, including an EXPRESSION\_SYSTEM annotation, which  
 113 was used to filter PDB entries to keep only those which in-  
 114 clude the phrase "ESCHERICHIA COLI". The majority of PDB

115 entries returned had an EXPRESSION\_SYSTEM tag of "ES-  
 116 CHERICHIA COLI", while a small number had "ESCHERICHIA  
 117 COLI BL21" or "ESCHERICHIA COLI BL21(DE3).

118 The PDB coordinate files also contain SEQRES  
 119 records, which should contain the protein se-  
 120 quence used in the crystallography or NMR ex-  
 121 periment. According to the PDB documentation  
 122 (<http://deposit.rcsb.org/format-faq-v1.html>),  
 123 "All residues in the crystal or in solution, including residues  
 124 not present in the model (i.e., disordered, lacking electron  
 125 density, cloning artifacts, HIS tags) are included in the  
 126 SEQRES records." However, we found that these records  
 127 are very often misannotated, instead representing only the  
 128 crystallographically resolved residues. Since expression  
 129 levels can be greatly affected by insertions or deletions of  
 130 only one or a few residues at either terminus [DLP: ?CITE,  
 131 or reference our 96-construct Abl1 expression panel], it is  
 132 important to know the full experimental sequence, and  
 133 we thus needed a way to measure the authenticity of a  
 134 given SEQRES record. We developed a crude measure by  
 135 hypothesizing that a) most crystal structures would be  
 136 likely to have at least one or a few unresolved residues at  
 137 one or both termini, and b) the presence of an expression  
 138 tag (which is typically not crystallographically resolved)  
 139 would indicate an authentic SEQRES record. To achieve  
 140 this, unresolved residues were first defined by comparing  
 141 the SEQRES sequence to the resolved sequence, using the  
 142 SIFTS service (CITE) to determine which residues were not  
 143 present in the canonical isoform sequence. Then regular  
 144 expression pattern matching was used to detect common  
 145 expression tags at the N- or C-termini. Sequences with a  
 146 detected expression tag were given a score of 2, while those  
 147 with any unresolved sequence at the termini were given  
 148 a score of 1, and the remainder were given a score of 0.  
 149 This data was not used to filter out PDB structures at this  
 150 stage, but was stored to allow for subsequent selection of  
 151 PDB constructs based on likely authenticity. Also stored for  
 152 each PDB sequence was the number of residues extraneous  
 153 to the target kinase domain, and the number of residue  
 154 conflicts with the UniProt canonical isoform within that  
 155 domain span.

#### 156 3. Plasmid libraries

157 As a source of kinase DNA sequences, we purchased three  
 158 kinase plasmid libraries: the **addgene Human Kinase ORF**  
 159 **kit**, a kinase library from the Structural Genomics Consor-  
 160 tium (SGC), Oxford (<http://www.thesgc.org>), and a ki-  
 161 nase library from the **PlasmID Repository** maintained by  
 162 the Dana-Farber/Harvard Cancer Center. The aim was to  
 163 subclone the chosen sequence constructs from these plas-  
 164 mids, though we did not use the same vectors. Annotated  
 165 data for the kinases in each library was used to match them  
 166 against the human protein kinases selected for this project.  
 167 A Python script was written which translated the plasmid  
 168 ORFs into protein sequences, and aligned them against the  
 169 target kinase domain sequences from UniProt. Also calcu-

170 lated were the number of extraneous protein residues in the  
 171 ORF, relative to the target kinase domain sequence, and the  
 172 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 (i.e. those with detected expression tags  
 194 were ranked above those with any other sequence extrane-  
 195 ous to the domain span; while those with no extraneous se-  
 196 quence had already been filtered out), then by the number  
 197 of conflicts relative to the UniProt domain sequence, then  
 198 by the number of residues extraneous to the UniProt do-  
 199 main sequence span. The top-ranked construct was then  
 200 chosen. In cases where multiple plasmids were available,  
 201 these were sorted first by the number of conflicts relative  
 202 to the UniProt domain sequence, then by the number of  
 203 residues extraneous to the UniProt domain sequence span,  
 204 and the top-ranked plasmid was chosen.

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

210 A sortable table of results can be viewed at  
 211 <http://choderalab.github.io/kinome-data/>  
 212 [kinase\\_constructs-addgene\\_hip\\_sgc.html](http://choderalab.github.io/kinome-data/kinase_constructs-addgene_hip_sgc.html).

213 **TODO maybe include a figure to help illustrate the above  
 214 (but may be too complicated):**

215 **5. Other notes**

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

222 **B. Expression testing**

223 [JDC: This protocol is missing crucial information, like ex-  
 224 actly which cell type was used for expression!]

225 For each target, the selected construct sequence was sub-  
 226 cloned from the selected DNA plasmid. Expression testing  
 227 performed by QB3 MacroLab.

228 All genes were cloned into the 2BT10 plasmid, an AMP  
 229 resistant ColE1 plasmid with a T7 promoter. Each kinase  
 230 domain was tagged with a N-terminal His10-TEV and co-  
 231 expressed with either the truncated YopH164 (for Tyr ki-  
 232 nases) or lambda phosphatase (for Ser/Thr kinases). Ex-  
 233 pression was performed in Rosetta2 cells grown with Magic  
 234 Media (Invitrogen autoinducing medium), 100 µg/mL of car-  
 235 benicillin and 100 µg/mL of spectinomycin. Single colonies  
 236 of transformants were cultivated with 900 µL of MagicMe-  
 237 dia into a gas permeable sealed 96-well block. The cultures  
 238 were incubated at 37 °C for 4 hours and then at 16 °C for 40  
 239 hours while shaking. Next, cells were centrifuged and the  
 240 pellets were frozen at -80 °C overnight. Cells were lysed on  
 241 a rotating platform at room temperature for an hour using  
 242 700 µL of SoluLyse (Genlantis) supplemented with 400 mM  
 243 NaCl, 20 mM imidazole and protease inhibitors.

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

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

265 **III. RESULTS**

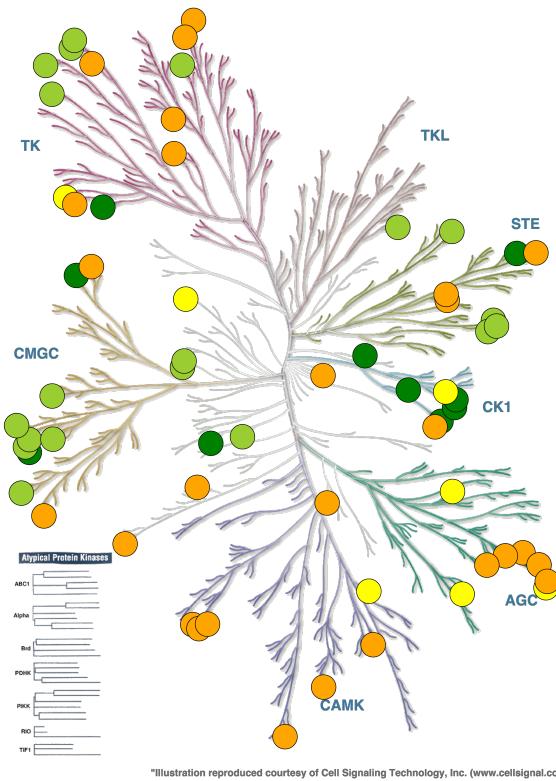
266 **A. PDBs mining results**

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

268 A panel containing the 96 kinase domain constructs se-  
 269 lected through our semi-automated method, was tested for  
 270 expression in *E. coli*. From this initial test, 68 kinase do-  
 271 mains expressed successfully (yield of more than 2 ng/µL).

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
CSPK_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



**FIG. 1. 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>.

- 272 [1] M. A. Seeliger, M. Young, M. N. Henderson, P. Pellicena, D. S. 275 [3] F. Stegmeier, M. Warmuth, W. R. Sellers, and M. Dorsch, Clin.  
273 King, A. M. Falick, and J. Kuriyan, *Protein Sci.* **14**, 3135 (2005). 276 Pharm. & Therap. **87**, 543 (2010).  
274 [2] American Cancer Society, *Cancer Facts & Figures 2015*, 2015. 277 [4] S. P. Chambers, D. A. Austen, J. R. Fulghum, and W. M. Kim, *Pro-  
278 tein Expression and Purification* **36**, 40 (2004).  
279 [5] L. Wang, M. Foster, Y. Zhang, W. R. Tschantz, L. Yang, J. Worrall, 280 C. Loh, and X. Xu, *Protein Express. Pur.* **61**, 204 (2008).