Curation Principles Derived from the Analysis of the SBOL iGEM Data Set

Jeanet Mante,[†] Nicholas Roehner,[‡] Kevin Keating,[¶] James Alastair McLaughlin,[§] Eric Young,[¶] Jacob Beal,[‡] and Chris J. Myers[*,†]

[†]*University of Colorado Boulder, Boulder CO 80309, USA*

[‡]*Raytheon BBN Technologies, Cambridge MA 02138, USA*

[¶]*Worcester Polytechnic Institute, Worcester MA 01609, USA*

[§]*EMBL-EBI, Cambridge CB10 1SD, UK*

E-mail: chris.myers@colorado.edu

Abstract

As an engineering endeavor, synthetic biology requires effective sharing of genetic design information that can be reused in the construction of new designs. While there are a number of large community repositories of design information, curation of this information has been limited. This in turn limits the ways in which design information can be put to use. The aim of this work was to improve this situation by creating a curated library of parts from the *International Genetically Engineered Machines* (iGEM) registry data set. To this end, an analysis of the *Synthetic Biology Open Language* (SBOL) version of the iGEM registry was carried out using four different approaches—simple statistics, SnapGene auto-annotation, SYNBICT auto-annotation, and expert analysis—the results of which are presented herein. Key challenges encountered include the use of free text, insufficient part provenance, part duplication, lack of part removal, and insufficient continuous curation. Based on these analyses, the focus has

shifted from the creation of a curated iGEM part library to instead the extraction of a set of lessons, which are presented here. These lessons can be exploited to facilitate the creation and curation of other part libraries using a simpler and less labor intensive process.

Keywords

SBOL, iGEM, SynBioHub, SYNBICT, Annotation, Curation, Automation, Analysis

One of the key aims of synthetic biology is to standardize genetic engineering and make it more reproducible. This engineering endeavor requires, among other things, effective sharing of genetic design information that can be reused in the construction of new designs. To this end, a number of large community repositories of design information have been created over the years. One of the largest and oldest repository of parts is the *International Genetically Engineered Machines* (iGEM) Registry of Standard Biological Parts (`http://parts.igem.org/`). Since its inception in 2003, the iGEM competition has followed principles aimed at the advancement of synthetic biology via education, competition, and development of an open and collaborative community,[1] including the submission of parts created by iGEM participants to the iGEM registry. There are now tens of thousands of parts, the vast majority of which use the *BioBricks* format and are thus, at least in principle, able to be composed in a modular fashion.[2]

Due to the wealth of genetic parts deposited in the registry and the role of the iGEM registry in training undergraduate synthetic biologists, the registry has been used by some as an indicator to measure the progress of the field of synthetic biology.[3] When parts are found and reused they can significantly reduce the cost of creating new circuits.[4] While there is part reuse, a small core set of parts accounts for most of the reuse, and this core set remains constant from year to year.[5] Likewise, while it is possible to create full circuits based only off of registry parts,[6] finding reliable parts is difficult. The large number of parts, a variety of issues around assembly methods, and issues with quality control mean that there

is uncertainty with part reuse, leading many iGEM teams to choose to create and submit new parts rather than reuse old ones.[7] The use of a manual submission process until 2010 and the time pressure associated with the competition contribute to the variation of quality level in documentation and annotations.[8]

Well characterized biological parts are needed to increase the accuracy of network level simulation,[9] allow programmatic access to registry databases from multiple client applications,[10] and allow the design of scalable circuits without looking at individual reactions.[11] Several attempts have been made to define what a well characterized part is. These include the Provisional BioBrick Lanugage (PoBoL),[12] the use of electrical engineering inspired data sheets for genetic devices,[11,13] behavioral characteristics of a part (for example, polymerase operations per second),[14] and atomic (non-composite/basic) parts.[10] There have also been a number of reviews of the iGEM data set that highlight issues such as: the lack of part reuse,[7,10,15] the lack of annotation of sub-sequences,[10] incomplete or inaccurate part descriptions,[16] and lack of part validation.[7,17] However, of these papers only[10] provides in depth statistics about the iGEM data set and an attempt to create a library of basic iGEM parts, though it does not analyze the descriptions of the parts.

This paper builds on previous work, setting out to address the issues with the iGEM data set by creating a library of thoroughly documented, well annotated, and easily searchable basic parts from the iGEM data set. The aim of a library is to be a set of parts that are well documented enough that researchers can make judgements about the usefulness of components to their work with confidence. Additionally, the library must be encoded in such a way that even at a large scale the appropriate parts, if they exist, can still be easily found. To this end, data records must be machine readable. Machine readability also increases the ease of data set curation.[18]

As a first step towards a well-curated, machine readable iGEM library, the registry was converted to the *Synthetic Biology Open Language* (SBOL)[19] data format, a standard language for describing genetic designs. This paper presents the results of an analysis of the

SBOL iGEM registry. This analysis revealed difficult-to-overcome challenges to distilling the data set into a better annotated and documented form. Thus, this paper proposes using other libraries of parts to further the analysis of the iGEM data set. Additionally, while the registry is a very large and well respected repository of synthetic biology information, it does have several design flaws. The flaws which hinder machine readability, and the curation of an iGEM library are: 1) its over-reliance on free text, 2) insufficient part provenance, 3) part duplication, 4) the lack of part removal, and 5) insufficient continuous curation. This paper suggests that to create a useful library of genetic parts from the iGEM registry, or any other library, these five concerns must be considered and addressed. This conclusion caused a shift in focus; from creating a library from the iGEM registry to applying our lessons learned to develop a process for creating well-curated part libraries. Additionally, we suggest a submission structure to nudge participants towards the submission of well characterised parts.

Results

Analysis of the iGEM Data Set

From an initial exploratory analysis, it was known that the iGEM data set contained spurious parts. Thus, sequence length analysis was carried out as the existence of a sequence over a certain length was considered to be a good heuristic for the trustworthiness of a part. Next, presence of sub-annotations was investigated as the initial aim was to create a library of basic parts, thus any composite devices (explicit or implicit) were not wanted in the basic library. Two kinds of sequence annotation analysis were carried out (SnapGene and SYNBICT) as one uses a common library and the other allowed the use of more libraries specifically selected as potentially being present in the iGEM data set. The final analysis type was looking at the part metadata. Once a set of basic parts were isolated, the presence of good metadata was looked for to indicate that the part was not spurious, and had the associated information

needed to reuse it. The analysis of description field lengths was an initial fast heuristic. The more in depth expert curation of description fields was the next step. These methods are not sufficient to curate a part library, but are initial steps towards that goal. The difficulty of these steps, in particular the last step led to the conclusion that proactive curation rather than continued retrospective analysis should be the next step. If retrospective curation was to continue, the next steps would be looking at the presence and absence of particular kinds of information in the description fields, e.g. source organism mentions, characterization data, and references.

Simple Statistics: As a step towards the development of an iGEM library, we analyzed the SBOL version of the iGEM data set that was created in 2017 (`https://synbiohub.org/public/igem/igem_collection/1`). The initial analysis provides an estimate of the number of unique sequences that are useful in future genetic engineering designs. Some simple data set breakdowns are shown below:

- Total number of BioBrick parts: 38,464
- Number of parts with no sequence: 1,769 (4.6%)
- Parts with a sequence between 1 and 40 bp: 2,391 (6.2%)
- Number of unique sequences: 33,323 (86.6%)
- Number of non-unique sequences: 5,141 (13.4%)
- Parts with the same SO type (e.g., terminator) as their subparts: 1,893 (4.9%)
- Parts with the discontinued flag: 2,568 (6.7%)

A detailed analysis is shown in Table 1. This table is based on the sequential application of filters based on the type of part represented using the *Sequence Ontology* (SO).[20] The first filter applied was the minimum length filter ("Sequences Over Minimum Length" column). The minimum length used is shown in the column "Minimum Length Parameter". Initially, the minimum length for many parts is set to 6 base pairs (bp) or the equivalent of 2 codons for amino acids (aa). However, for CDS 40 bp (about 13 amino acids) was used as the shortest human enzyme found in UniProt (Cytochrome P450 2A7) is 20aa (60bp). As plasmid and

plasmid vectors generally contain a CDS, their minimum length was also increased to 40bp. This simple filter removes almost 2,000 components which had no sequences associated with them or very short sequences. The next filter that was applied looked at unique sequences per SO Type, removing any exact sequence copies. However, as it worked by SO type, the same sequence may still be repeated as both a terminator and CDS, for example. In other words, there are 33,113 unique sequences over a minimum length whereas by role the total is 33,588. Thus, 475 sequences are repeated exactly but given different SO types. The final filter considers looking for basic parts. Components may be 'basic' or they may be 'composite'. Basic parts do not contain any sub-parts whilst composite parts have one or more sub-parts. The set of 'basic unique' parts created by application of the 'basic' filter are used as the basis for the other types of analyses.

Annotation by SnapGene: A SnapGene auto-annotation server was used to annotate the "basic unique" parts of type CDS, Promoter, RBS, and Terminator. SnapGene auto-annotation requires at least 96% sequence match (with features pulled from their most popular plasmid sequences), however the algorithm does allow for codon optimisation.[21] This annotation is summarized in Figure 1. It shows that despite the parts having no sub-annotations from the iGEM data set, many of them can be annotated using the SnapGene library. This indicates that they are not a novel basic part. There were also quite a lot of 'unexpected feature Type(s)' indicating that the type of part entered in the iGEM data set does not always match the features found in the sequence. For example, there are parts of iGEM type RBS which have a terminator annotated in the sequence.

A closer look was taken at how annotations were made by SnapGene in the "basic unique" RBS part set (see Table 2). This table indicates that several of the annotations are used multiple times (e.g., the strong bacterial ribosome binding site (Elowitz and Leibler, 2000) was used 16 separate times). This suggests that there may be different cut outs of the same part. This is further supported by a high percentage cover. If there are several parts all annotated with a high percentage cover, it is likely that there is small variation

Table 1: iGEM statistics. A sequential filtering was used to determine the 'useful unique entities' in the SBOL iGEM data set. Analysis was carried out per SO type as expectations for different types are different (e.g., RBSs would be expected to be shorter than chromosomes). Sequence count are the initial counts by role, sequences over the minimum length is the number of sequences with a length greater than specified by the minimum length parameter, unique sequences over a minimum length takes the previous count but removes any duplicate sequences within the category. Finally, an analysis was carried out to filter out composite parts; Note: this assumption is too strict for Engineered Regions.

SO type	iGEM types	Minimum length parameter	Sequence count	Sequences over minimum length	Unique sequences over minimum length	Unique over minimum length basic parts
CDS	Basic, Coding	40	7,689	7,198	6,788	6,188
Chromosome	Cell	6	73	13	13	10
Engineered Region	All other iGEM types	6	20,171	19,477	17,664	3,700
Mature Transcript Region	RNA	6	595	556	538	485
oriT	Conjugation	6	41	39	39	20
Plasmid	Plasmid	40	609	526	484	398
Plasmid Vector	Plasmid_Backbone	40	404	379	369	353
Polypeptide Domain	Protein_Domain	6	769	718	700	665
Primer	Primer	6	582	574	567	567
Promoter	Regulatory	6	3,106	2,965	2,770	2,495
Restriction Enzyme Assembly Scar	Scar	6	40	26	25	24
Ribosome Entry Site	RBS	6	525	494	454	448
Sequence feature	DNA, Other, Terminator	6	3,149	2,734	2,581	1,923
T7 RNA Polymerase Promoter	T7	6	35	32	28	24
Tag	Tag	6	288	263	233	222
Terminator	Terminator	6	388	381	335	329
Total			**38,464**	**36,375**	**33,588**	**17,851**

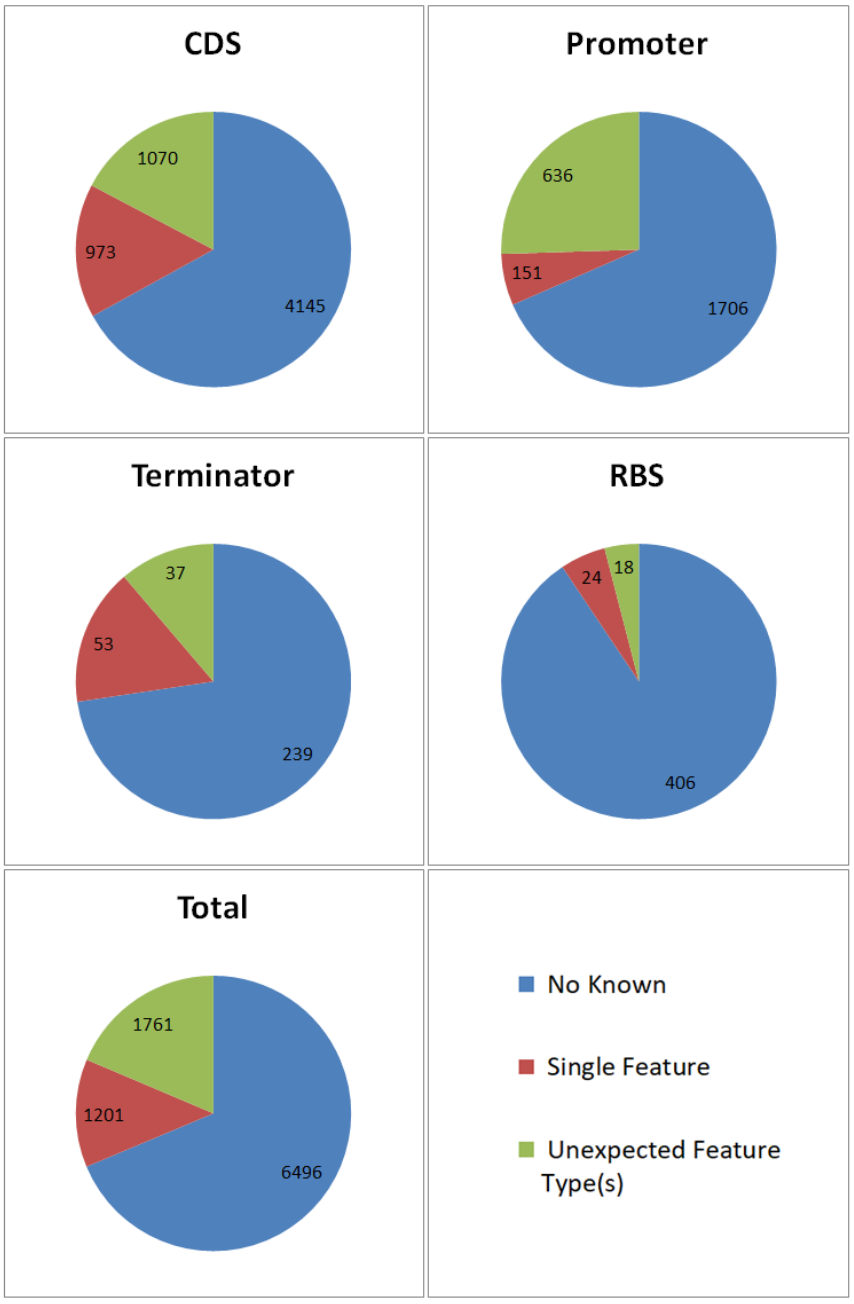

Figure 1: Overview of the SnapGene annotation of "basic unique" parts. The top four pie charts show break downs of the kinds of annotation per SO Type analysed. The final chart shows the summary of annotation types over all four SO Types. The kinds of annotation are: no annotations made using the SnapGene library, a single feature annotation of the type expected, or an annotation of an unexpected type (e.g., a terminator part with a CDS annotation by SnapGene or a CDS which contains a promoter and CDS). The numbers in each segment indicate the number of parts per category. Further analysis can be carried out on the no annotations category as these are potentially new basic parts in addition to those found in the SnapGene library. If the whole iGEM data set were used, the SnapGene library would be expected to annotate a larger percent of the data set.

at the ends of the parts. This hypothesis was substantiated by using a similarity analysis which showed 276 of the "basic unique" RBS parts had at least 80% similarity with at least one other RBS. However, there are also cases where there is a very low percentage cover. This can mean that the exact location is unknown so a large chunk of sequence is taken, or it can mean that it is only one of several 'parts' within the sequence. The other parts may be known to SnapGene or not. Figure 2 shows how coverage can change. Some parts are understandably difficult to classify in the sequence ontology without some thought. For example, the annotation of BBa_M36361 and BBa_M36521 by SnapGene revealed that those parts are combinations of two parts. Upon investigation, it is clear that these parts are bicistronic 5' UTRs, which should not be strictly considered unique RBS parts, but can understandably fall within a loose RBS category as they contain two RBS and are placed between a promoter and terminator. This is an indication that guidance around SO classification should be provided in the data entry field, especially for common ambiguous cases. Another example revealed by SnapGene is BBa_K1231002, which is tagged as an RBS but appears to contain a promoter and RBS (the large gap infront of the RBS was investigated and found to contain a promoter). While there may be good design reasons to maintain these two parts together as a composite part (e.g., spacing sensitivity), this part would likely function poorly if it were incorporated into an automated design framework as an RBS and used as is. Though annotation tools like SnapGene can signal to curators that some parts require inspection, this manual curation process is labor-intensive and relies heavily on interpretation of sometimes ambiguous documentation.

Annotation by SYNBICT: SYNBICT was used to annotate the 'basic unique' RBS (Table 3). SYNBICT uses several libraries of parts mined from 'toolkit' papers. Resources mined for yeast parts included the Yeast Toolkit,[22] Pichia Toolkit,[23] and a combinatorial design paper.[24] Parts for Gram-negative bacteria were drawn from the CIDAR MoClo kit,[25] the CIDAR Extension Kit Volume I,[26] the Voigt Lab terminator collection,[27] and the *Bascillus subtilis* collection.[28] Unlike SnapGene, SYNBICT uses an exact match algorithm. Whilst

Table 2: SnapGene annotations seen in the 'basic unique' ribosome binding sites. Note that most annotations are reused which suggests that some of the ribosome binding sites are not as unique as expected. Additionally, the annotation names suggest that not all of the sequences annotated are RBSs, e.g., finding the lambda t0 terminator in an RBS would be unexpected.

Annotation Name	Number of Uses	Annotation Length	% Cover Min, Average, Max
strong bacterial ribosome binding site (Elowitz and Leibler, 2000)	16	11	0.17, 0.48, 0.85
efficient ribosome binding site from bacteriophage T7 gene 10 (Olins and Rangwala, 1989)	2	22	0.27, 0.53, 0.79
Shine-Dalgarno sequence	8	8	0.02, 0.14, 0.57
efficient ribosome binding site from bacteriophage T7 gene 10 (Olins and Rangwala, 1989)	3	20	0.27, 0.53, 0.79
minP	1	31	0.49, 0.49, 0.49
Factor Xa site	1	11	0.01, 0.01, 0.01
TEV site	1	20	0.01, 0.01, 0.01
TEE	1	11	0.01, 0.01, 0.01
vertebrate consensus sequence for strong initiation of translation (Kozak, 1987)	1	9	0.53, 0.53, 0.53
Csy4 site	1	14	0.13, 0.13, 0.13
BioBrick suffix	6	20	0.01, 0.02, 0.03
mRFP1	1	677	0.78, 0.78, 0.78
BioBrick prefix	7	21	0.01, 0.01, 0.02
his operon terminator	3	57	0.02, 0.03, 0.03
bacterial terminator	3	43	0.02, 0.02, 0.02
CmR	3	659	0.29, 0.3, 0.3
lambda t0 terminator	3	94	0.04, 0.04, 0.04
ori	3	588	0.26, 0.27, 0.27
CMV enhancer	1	379	0.55, 0.55, 0.55
CMV promoter	1	203	0.29, 0.29, 0.29
IRES2	1	586	1, 1, 1
IRES	1	571	0.92, 0.92, 0.92
T7 promoter	1	18	0.22, 0.22, 0.22
rrnB T1 terminator	1	71	0.17, 0.17, 0.17
T7Te terminator	3	27	0.06, 0.06, 0.06
strong bacterial ribosome binding site (Elowitz and Leibler, 2000)	2	11	0.17, 0.48, 0.85
Average	**2.9±3.3**	**161±240**	**0.26±0.29, 0.30±0.29, 0.35±0.33**

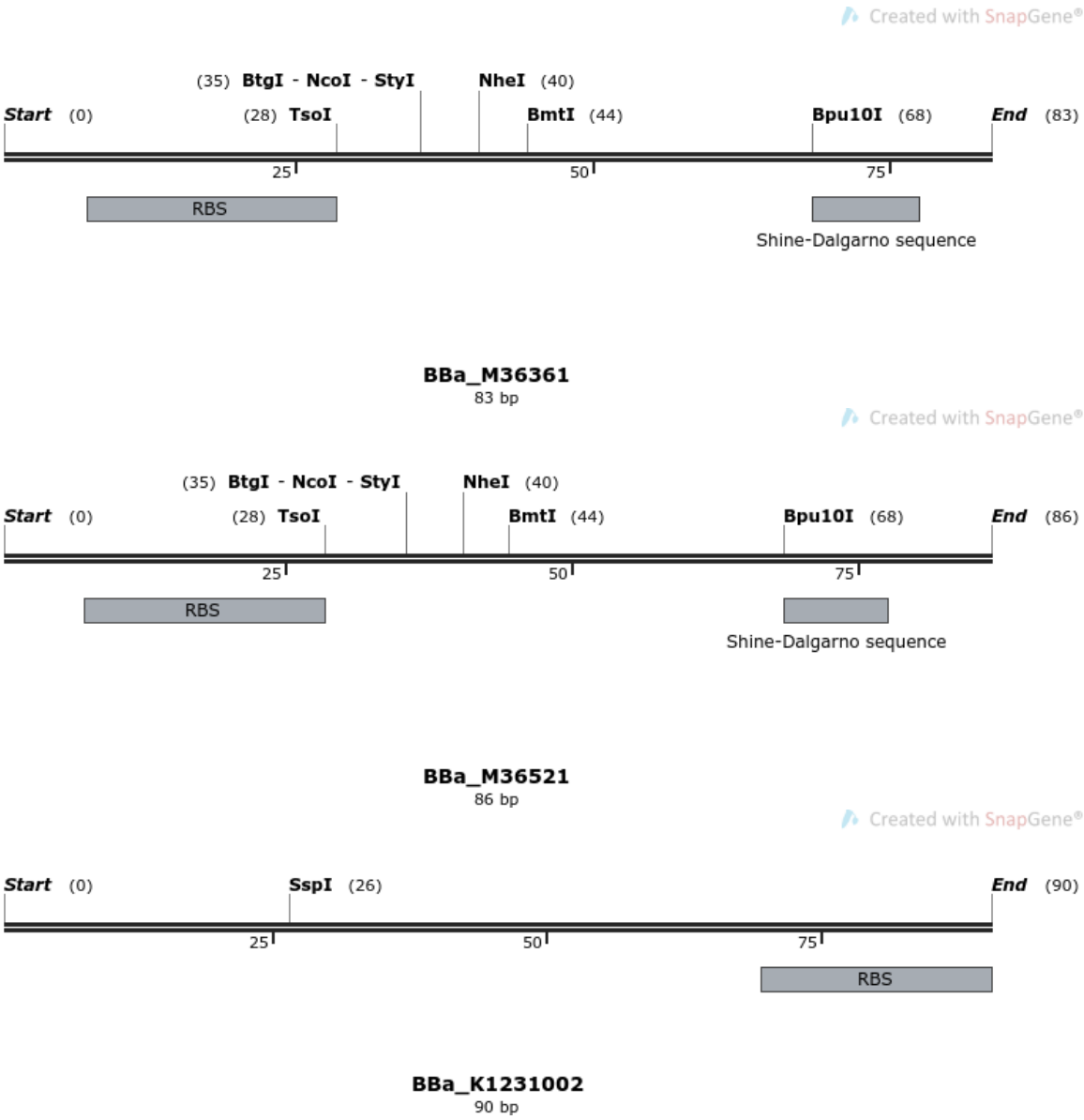

Figure 2: Three examples of the same 20bp RBS annotation. All three sequences were annotated with the same 20bp annotation by SnapGene. However, the parts are not all equivalent. The presence of two matching annotations in the top two (BBa_M36361 and BBa_M36521) as well as their similar length (83 vs 86bp respectively) suggest that these two might be almost equivalent. The bottom part though seems like a wide cut out of the RBS or the RBS with an additional part infront that is unknown to SnapGene, the latter assumption proved true after a promoter was found using BLAST.

most RBS parts were not annotated (92%), 8% were annotated, and seven sequences even had multiple annotations. There are nine annotations for a single sequence labelled as an RBS, which is surprising as RBS are typically quite short. Inspection of this sequence reveals that it is a composite rather than an RBS. The un-annotated RBS sequences may be unique new parts, or may indicate a need to expand the libraries used for SYNBICT annotation. Further manual inspection of these is required to determine whether they are suitable for adding to the annotation library.

Table 3: SYNBICT annotations seen in the 'basic unique' ribosome binding sites. Note that most RBS seem to not be annotated (92%), but 4 RBS sequences have 9 annotations.

Number of Annotations per Sequence	Frequency
0	411
1	30
2	2
3	1
9	4
Total	**448**

Parts in the 'basic unique' sequence set that had at least one feature annotated by SYNBICT were further inspected (see Table 4). Like the SnapGene annotations, several annotations were used multiple times (e.g., BBa_B0034 was used 13 times). It seems that BBa_B0034 was cut out in several different ways as the percentage cover of the annotation was quite high. On the other hand J23100 was used 12 times but had a very low percentage cover so is likely used in combination with other parts/features, some of which may be unknown. It is important to note that there are also non-RBS type annotations. In particular J23100 (promoter), the CamR Terminator, and the CamR Promoter stand out as not being RBS. This suggests that there was mislabeling of the original iGEM part type.

SYNBICT was also used to analyze whether larger, potentially composite parts contained unannotated features. In particular, SYNBICT was used to annotate all 16,715 non-CDS parts in the iGEM data set greater than or equal to 900 bp in length. The feature library used for this analysis was derived from RegulonDB[29] and contained 8,824 promoter features

Table 4: SYNBICT annotations seen in the 'basic unique' ribosome binding sites. Note that most annotations are reused which suggests that some of the ribosome binding sites are not as unique as expected. Additionally, the annotation names suggest that not all of the sequences annotated are RBSs, e.g., finding the CamR terminator in an RBS is unexpected.

Annotation Name	Number of Uses	Annotation Length	% Cover Min, Average, Max
BCD12	2	83	0.91, 0.93, 0.95
B0034m	3	20	0.69, 0.73, 0.8
BCD2	3	83	0.91, 0.95, 0.99
BCD8	2	83	0.91, 0.93, 0.95
BBa_B0034	13	11	0.17, 0.49, 0.85
BCD14	1	83	0.99, 0.99, 0.99
BCD13	1	83	0.91, 0.91, 0.91
B0033m	2	19	0.7, 0.75, 0.79
BBa_B0034	2	11	0.23, 0.51, 0.79
B0015	3	128	0.3, 0.3, 0.3
BBa_B0011	4	45	0.1, 0.1, 0.1
BBa_B0012	2	40	0.09, 0.09, 0.09
T7 consensus	1	22	0.27, 0.27, 0.27
UTR1	1	33	0.4, 0.4, 0.4
J23100	12	34	0.01, 0.02, 0.04
B0032m	2	21	0.72, 0.77, 0.81
ChlorR	3	659	0.29, 0.3, 0.3
BBa_B0057	3	41	0.02, 0.02, 0.02
CamR Promoter	6	104	0.05, 0.05, 0.05
BBa_B0062-R	3	40	0.02, 0.02, 0.02
CamR Terminator	3	108	0.05, 0.05, 0.05
BCD16	1	83	0.91, 0.91, 0.91
Average	**3.3±3.2**	**83±133**	**0.44±0.37, 0.48±0.37, 0.52±0.39**

and 218 CDS features for transcription factors from *E. coli*. SYNBICT made a total of 5,524 annotations (5,233 promoter annotations and 291 CDS annotations) over 3,883 parts, or an average of 1.4 annotations per part. Figure 3 shows the distribution of annotations per part. The two features most commonly annotated on any part were the lacZp4 promoter (2,009 annotations) and the core of the araCp promoter (834 annotations), together making up over half of all promoter annotations made by SYNBICT in this case.

In order to determine whether features were previously unannotated, SYNBICT was used to analyze how many of its annotations had similar locations to those of iGEM annotations. If the offset of a SYNBICT annotation to an iGEM annotation was less than 80 bp (the minimum length for a RegulonDB promoter), then the SYNBICT annotation was marked as potentially novel. Figure 3 shows the distribution of all such novel annotations per part. This method indicated that up to 71% of the annotations made by SYNBICT are for previously unannotated regulatory features, which highlights the need for automated sequence annotation and design curation.

Expert Annotation: Expert annotation was used to take a closer look at the 'basic unique' iGEM data sets. While the SnapGene and SYNBICT auto-annotation were focused on the sequences and the annotations, the expert annotation was used to look at the information associated with the sequences. Despite some sequences being long enough to be plausible for a part type, it might not actually be a plausible sequence. Deciding what kind of sequence is plausible is difficult, but looking at the sequence description fields often give a better idea. For example, at the time of the iGEM to SBOL conversion BBa_K1740001 has a sequence of length 672, however the description fields say 'This is a test', 'none', and '.' which suggests that the sequence likely is not 'real'. For this reason, a mostly qualitative analysis was carried out to identify patterns in the sequence descriptions which could form the basis for further curation efforts. The main observations are as follows:

- Sequence descriptions and the relevant information is split across four fields: Long Description, Notes, Source, and description (Table 6). The split is by no means equal

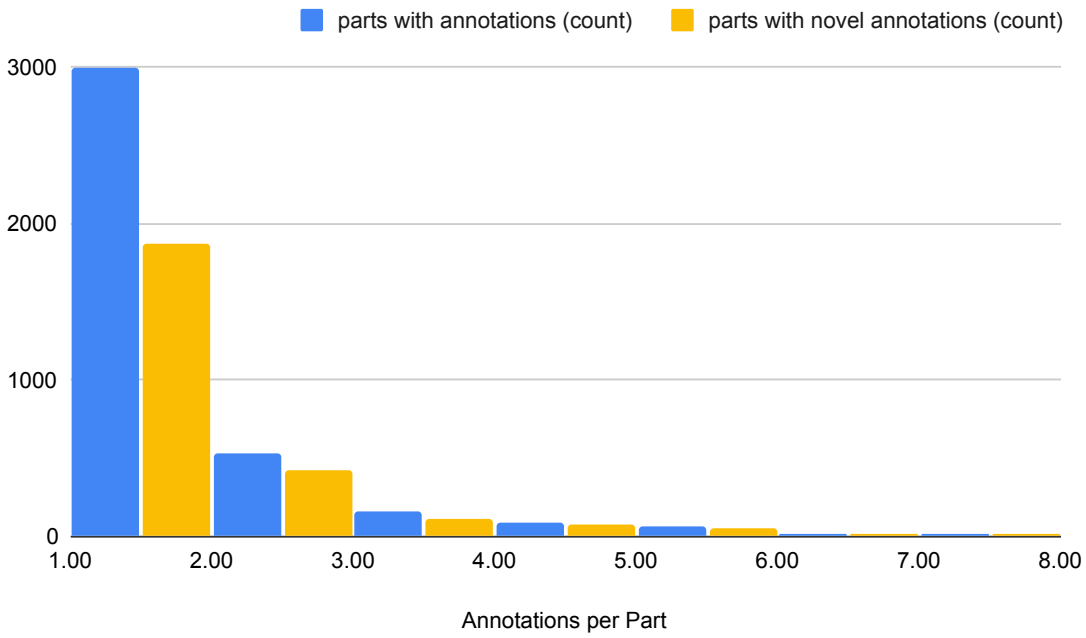

Figure 3: Side-by-side histograms of counts for parts with 1 to 7 annotations per part (blue) and 1 to 7 novel annotations per part (orange). Each bin is for a number of annotations per part greater than or equal to the lower bound and less than the upper bound. Not shown are several outliers for 9, 11, 15, and 16 annotations per part (one part each) and 10 annotations per part (four parts).

though with 7530 (39%) having no text in at least one of the fields. The length of the text varies widely between parts and does not necessarily correlate with valuable information (for example, a single Addgene reference conveys much more than 'test test test test'). Additionally, the descriptions do not appear to improve over the years, or over the months per year as the iGEM jamboree gets closer (see the supplemental for the related figures).

- Many of the parts are temporary/test parts. They are often indicated via descriptions including 'test', 'temporary', 'none', 'kill', 'bla', 'blah', or a keyboard smash.
- Subheadings have been implemented within single fields to break out more clearly information such as: notes, references, source, see also, design notes, mutagenicity, assembly, and functional parameters.
- In many cases a lot of information is present in descriptions (e.g., source organism, target organism, paper citations, assembly methods) however how it is presented varies widely (e.g., species information may include genus, species, or be a common name, as shown in Figure 4).
- There are many descriptions stating they will be edited or finished later when more information is known/gathered, or the part has been tested.

To compare such spurious parts to 'real' parts, 400 of the spurious parts were randomly selected and 400 of the 'real' parts were randomly selected (full list is given in the supplemental). A significant difference was seen in the length of descriptive fields between the two sample groups. The 'real' parts had more information in each of the fields and generally had significantly longer sequences too (Table 5). These results are not surprising as part of the method of determining whether a part is 'real' is looking at the description provided. Additionally, spurious parts are expected to generally have shorter made up sequences. SYNBICT was used to annotate the sequences for the two groups. In the 'real' group, the average number of annotations was 0.340 ± 1.133, whilst in the spurious group it was 0.375 ± 1.246. The student's t-test indicates no significant difference between the two groups with a p-value

of 0.678. The lack of difference in annotation may be explained by the annotation library used in SYNBICT (a larger set of libraries might have led to more annotations). It should be noted that all parts selected had no annotations by the authors so the lack of significant difference between 'real' and spurious annotation number does not indicate a lack of correlation between 'good' descriptions and 'good' sequence annotations in general.

Table 5: Comparison of a random sample of 400 'real' vs 400 spurious parts. The length of each of the descriptive fields is given in characters (mean ± standard deviation). The p-value for the student's t-test between the two groups is given. Each of the fields show a significant difference between the 'real' and spurious group at the 0.05 alpha level.

	Long Description	**Notes**	**Sequence Length**	**Description**
'Real' Sample	264 ± 320	117 ± 208	1028 ± 2298	32 ± 20
Spurious Sample	30 ± 69	11 ± 43	629 ± 1216	23 ± 21
p-value	<0.00001	<0.00001	0.00234	<0.00001

Table 6: Statistics about the length of sequence description fields for the iGEM 'basic unique' data set. There are 4 sequence description fields: Long Description, Notes, Source, and Description. The 'Fields Std.Dev.' column indicates the standard deviation between the length of the four description column fields, and 'Fields Mean' calculates the mean length of the four fields.

	Long Description	Notes	Source	Description	Fields Std.Dev.	Fields Mean
Mean	263	102	67	31	135	116
Std.Dev.	736	218	149	20	371	210

Lessons from the iGEM Data Set

The analysis carried out to create a library of reusable parts from the iGEM data set highlighted several broad issues which made the creation of a library difficult. These issues are discussed below.

Over Reliance on Free Text: The original format of the parts was the form found in the iGEM registry, which relies heavily on several free text sections including the 'main page' wiki. While the information present on a wiki page is often extensive and informative, it is difficult to search over due to the variation in terms used to indicate the same thing

(e.g., *E. coli*, *E coli*, *Escherichia coli* or *H. sapiens*, human, *Homo sapiens*). The word cloud in Figure 4 illustrates the variety and the context-dependent nature of searches. Additional ambiguity lies in the precise meaning of the species specified (e.g., is the sequence derived from the *E. coli* genome, is it being used to transform *E. coli*, or was *B. subtilis* chosen due to problems with *E. coli*). Thus, while simple string searches are possible, more detailed faceted searches (e.g., organism designed for = *E. coli* or author type = undergraduate) are not possible.

The use of free text is additionally problematic as the information provided by students varies widely. Without separate input fields for each piece of information required, students may forget to provide some types of information or not know that the information would be useful for other people. For example, expert annotation of descriptions in the basic iGEM parts (i.e., parts with no sub-parts) indicates that of these 19,285 parts, 10,978 (55%) gave no species information in the description of the component.

Additionally, due to the many different free text fields, it can be difficult to know where to look for information. There is a long description, short description, and notes section and it is the case that some of the parts that had no species in the long description did have a species listed in the notes section.

Finally, though the basic part subset of the iGEM data set is supposed to include parts with no BioBrick sub-parts, there are several descriptions of sub-parts in the free text fields, sometimes even with detailed descriptions of the start and end positions of the sub-parts. This information is not accessible for search and requires manual curation to find fields that contain sub-part descriptions and then add the sub-part annotations to the file in the 'standard' way.

Over Reliance on Free Text (solution): To increase the machine readability of fields and make them more curatable, we suggest the use of *ontologies* (controlled vocabularies) to pull out information. For example the NCBI taxonomy[30,31] can be used to reference species in a uniform manner. Additionally, creating more fields that are specifically defined would

help encourage students to input information whilst also leading to more uniform free text where free text is needed. For example, rather than notes, description, and long description, the fields could be: design notes (notes about codon optimisation and restriction site removal of the part) and references (any labs you wish to acknowledge e.g., Genetic Logic Lab at CU Boulder, USA). The use of more fields and example text for fields will help create more uniform input without much additional effort to users. This is somewhat substantiated by the fact that some users already use headings in the different fields to separate out the different kinds of information they are entering into it. For description, the common headings are: notes, references, source, see also, design notes, mutagenicity, assembly, and functional parameters.

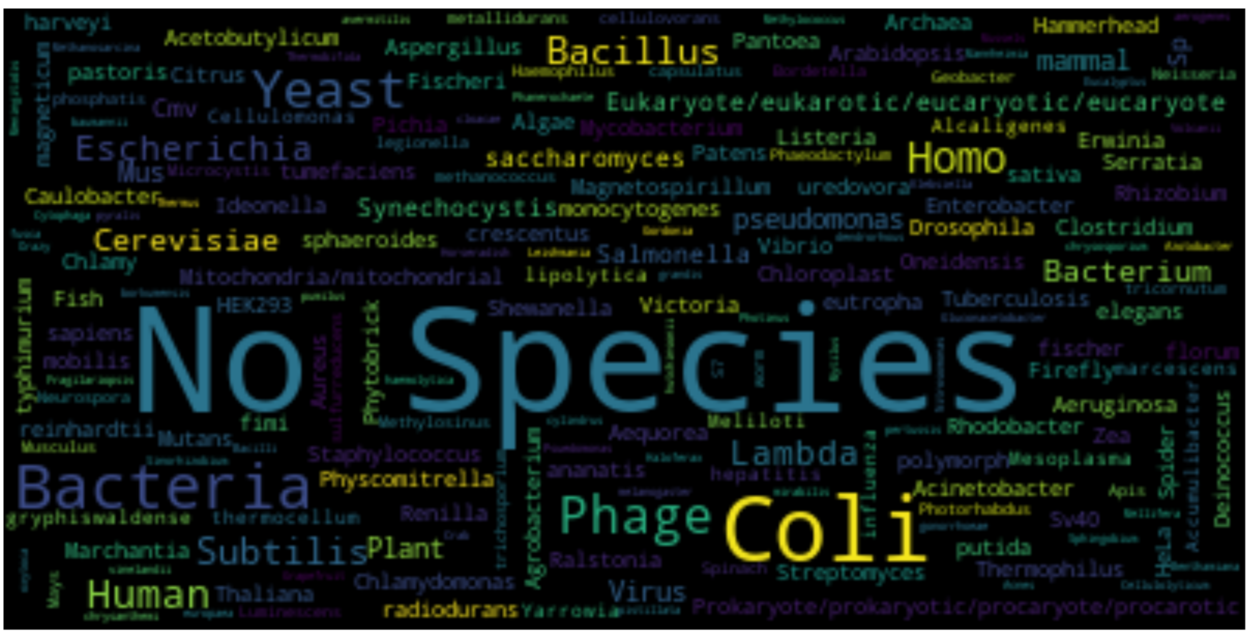

Figure 4: This is a word cloud of ‘species’ found in the long description of basic iGEM parts. The frequency of a genus name versus a species name were counted separately to account for *E. coli* or *Clostridium sp.* There are 271 species (not including ‘No species’) across the data set. However, of the 19825 parts 10978 have no species mentioned (55%). On average every species is mentioned by 41 ± 159 parts.

Insufficient Part Provenance: The second challenge faced by the iGEM data set is that there is insufficient part provenance. This is both the metadata provenance (who created it) and the physical provenance of the part (which species does it come from).

Whilst there is a creator field, this field is again free text and does not seem to be standardized in any way. This means that descriptions of the part creators vary from: 'iGEM TU_Darmstadt 2019' to 'Valentina Herbring, Sebastian Palluk, Andreas Schmidt' and the field being left completely blank. This poses an issue as it means that filtering based on 'trusted' authors is not possible, and that if additional information or clarification is required, it is nearly impossible to reach out to the original creators.

The creator field only indicates who created the record and not who created the part. Most parts were likely found in literature or other data repositories, and this provenance information should be recorded as well. Many of the descriptions include references to external sources that provide this type of provenance information for the sequences used to create parts. Sometimes these is sufficient provenance, but it is just difficult to find in a free text field, e.g., 'Novel gentamicin resistance gene. Derived from GenBank DQ208936, pTEX5500ts, kind gift of Barbara E. Murray. This, in turn was derived from GenBank AF016483. See pmid=16391062, pmid=9593155' for BBa_P1014. Other times the provenance given is not sufficient to trace it back, e.g., 'comes from original sequence form lu lab' for 'BBa_K1681001' (note this no longer exists on the current wiki only in the 2017 snapshot of the wiki which highlights the problem of the dynamic nature of the wiki and lack of version control discussed below).

The lack of sufficient provenance is also seen in the derivation of parts from other parts. For example, of the 19,825 basic iGEM components, 2,193 mention 'BBa_' in their description suggesting that they were somehow derived from a different existing component. Making clear that a part is a different instance of another part, and what changes were made to create the new component (see Figure 5 for possible changes to parts) is important to increase confidence in the component's utility.

Provenance (solution): Provenance would be greatly improved by more rigid guidance on how creators are added (e.g., via ORCiD or email address associated with the iGEM account). Additionally, having a method to separately add references (particularly Addgene,

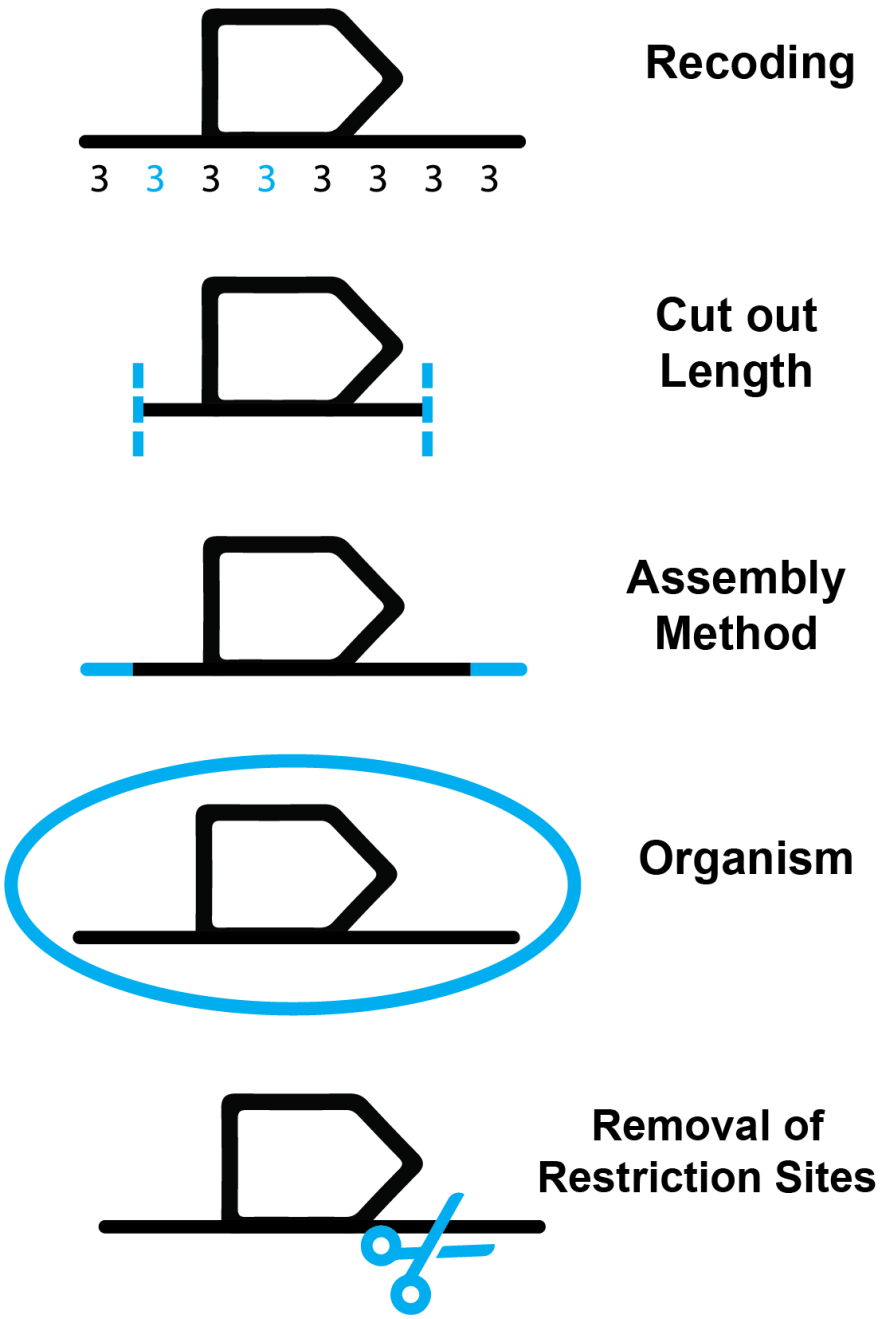

Figure 5: There are numerous ways that an original DNA sequence can be changed from its source during part derivation. For CDS, common examples include recoding of parts to use different codons to better work in the target organism, cutting out different lengths of the surrounding code, adding flanking regions for different assembly methods, the organism in which the part is used, and removal of restriction sites.

GenBank, iGEM registry parts, DOI references, and PubMed IDs) would be beneficial. This could be paired with the *Citation Typing Ontology* (CiTo)[32] to indicate what kind of reference is being given (e.g., data source, related, recommended reading) and the SO genetic variant description terms[33] (e.g., frameshift, stop gained). Finally, having an easy way to reference iGEM registry parts would make citing the use of parts easier and make provenance from the iGEM registry, as well as within the iGEM registry, easier.

Part Duplication: As mentioned in the provenance section, there is part duplication contained in the iGEM data set. This includes both a complete sequence match (a twin) and a partial sequence match (similar). Figure 6 shows a single 'cluster' that can be found just within the basic unique terminators. Overall there were 32 such clusters in the set of iGEM terminators and more among different part types. This suggests that apart from direct sequence re-uploading with different metadata, there is also a significant amount of similar part uploading. This is borne out by the SnapGene auto-annotation (described above) of basic parts which indicates that many of the parts are slightly longer than parts in the SnapGene library and show the same SnapGene parts making up 90% of several 'unique' part sequences. Some potential reasons for uploading similar parts are suggested in Figure 7. Some of these reasons are intentional differences, like codon optimization, but others, like typos, are unintended differences.

Part Duplication (solution): Part duplication could be combated in a number of ways: improving search, adding more editing capabilities, creating new versions, and organizing parts by function. To reduce the re-uploading of already existing parts, the reason for the re-uploading must be considered. If the re-uploading is occurring because the original cannot be found, then the creation of better search functionalities offers a solution. If re-uploading is occurring because the editing of existing part metadata is difficult, then the way to address this is by increasing editing capabilities. The ability to edit a part after it has been uploaded may cause problems if other parts reference the part to be edited, thus versioning might provide a better solution. Versioning is the easy creation of a linked or derived-from part

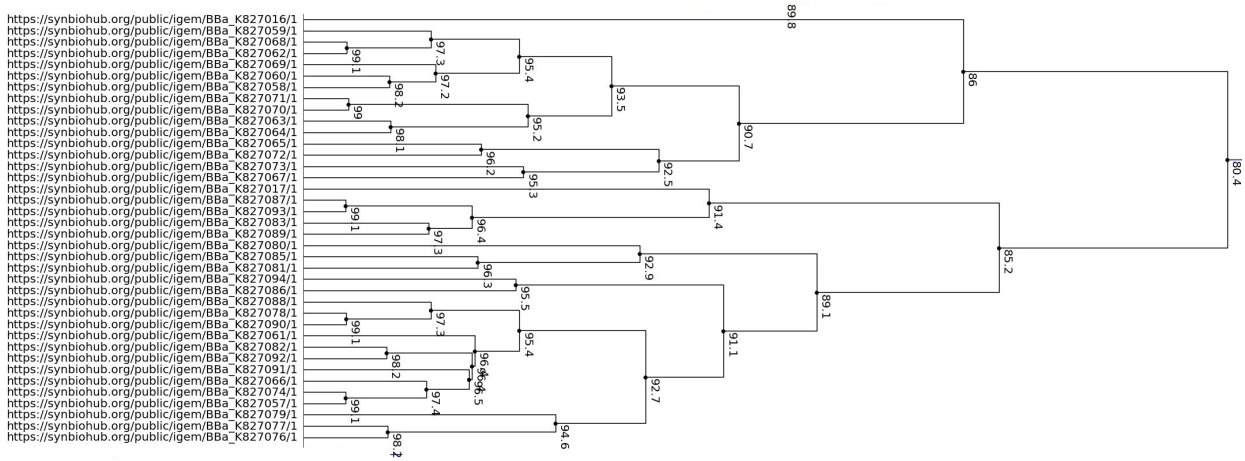

Figure 6: This is a dendrogram of some of the 'unique' basic terminators found in the iGEM data set. Whilst all exact twins have been removed and there are no subparts annotated (so no shared subparts) this cluster of 38 terminators still share at least 80.4% sequence similarity (the numbers noted on the diagram are percent sequence similarity). Overall there were 32 such clusters (though some comprised of only 2 similar terminators). To compare, we also ran a cluster analysis on the Voigt Terminator Collection,[27] whilst there was some clustering in the Voigt Terminators too, there were fewer clusters for the data set size and the average cluster size was smaller. Additionally, there were many more clusters in other part types (e.g., CDSs and promoters).

which has metadata which can then be edited. Versioning can be linked to curation in that when a new part is uploaded, same and similar sequences are flagged and the option for a new version of an old part is suggested. For similar parts, part duplication may be reduced by creating functional parts. Functional parts describe the way a part acts in a particular situation and all uploads of sequences which act in such a way would be linked to the functional part record. For example, a functional part 'promoter with a relative promoter unit[34] of 0.8' might be linked to sequences which have those properties 'in E. coli at 25C' or the 'Golden Gate Assembly compatible S. cerevisiae promoter'. Whilst this might not actually reduce the number of similar sequences, it would help indicate when sequence changes have an impact on functionality. This is important as a small sequence difference can have more impact than a large sequence difference if the changes are made in the start codon of a CDS or create a frame shift, compared to a codon optimization change in a CDS, which could change a lot of the sequence without having any impact on the final protein.

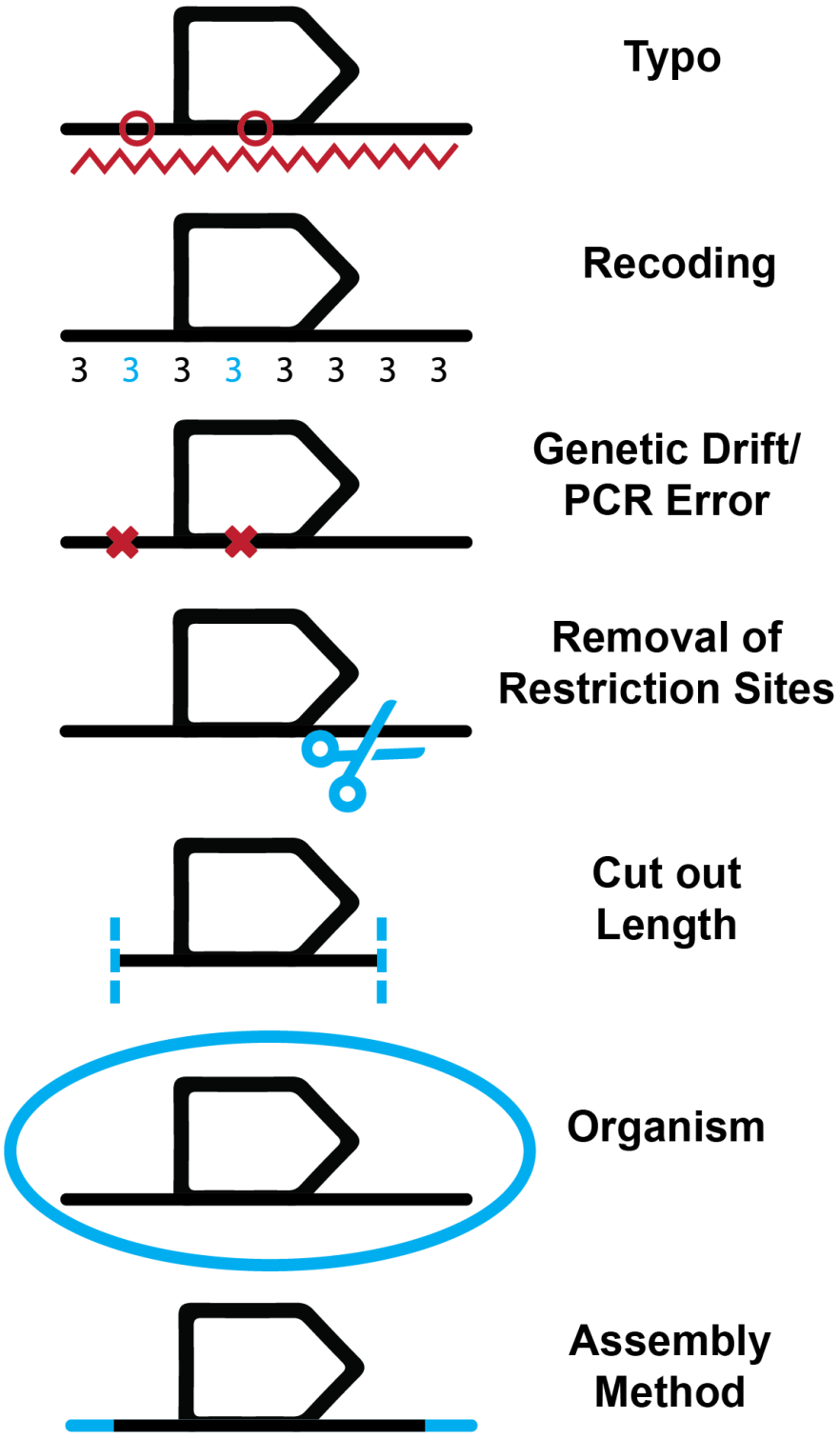

Figure 7: Reasons for similar components that are not twins. There are seven broad changes that could have been made to a part: a typo in the sequence, recoding of parts to use different codons (i.e., codon optimization) to better work in the target organism, genetic drift or a PCR error, intentional mutations to remove restriction sites, cutting out different lengths of the surrounding code, the organism from which the part was taken, and the addition of flanking sequences for varying assembly methods.

Lack of Part Removal: Currently there appears to be no way to remove parts from the iGEM registry. This is problematic when combined with a lack of versioning/editing as the same parts can, and are, re-uploaded with more metadata added. Additionally, since the iGEM registry is used by many students and researchers new to the field of synthetic biology, there are many 'test' parts uploaded. These are designed to see how fields display on the page, test search capabilities, or try the upload process. Whilst many of these tests are obvious from the multiple fields filled out as 'test' there is a surprising variety and ingenuity of test text. For example: 'This is a long description of the part, wow!!!!!.', 'bla bla', 'none', 'kill this part', 'just testing', and various keyboard smashes are used. These are often immediately obvious to a person, but the variety makes any automated filtering difficult. Additionally, there are aspiring authors among the iGEM students which results in fields like: "This is a protease that cleaves iGEM members' hands from their pipettes after 24 hours of consecutive pipetting.", 'Wayne Rooney's head dependent promoter is activated by proximity of Wayne Rooney's head. Is inhibited by broken metatarsal protein right before the World Cup.', and 'This part will be used to turn young bicycles into mature motorcycles ready to take on the world. VROOM!'. While these satirical parts do inject humor into the perusal of the data set, they can be more difficult for non-experts to catch, particularly because biological names can sometimes be somewhat nonsensical. For example, the Wayne Rooney's head dependent promoter does not exist, however SpyTag, SpyCatcher and Sonic hedgehog do exist.

Part Removal (solution): Along with the methods for reducing part duplication, further methods to tackle part duplication are suggested. For example, allowing archiving of an outdated version of a part could help make the iGEM registry more useful. Archiving is particularly helpful when compared to a simple deletion as it allows any references to the archived part to be maintained.

Additionally, having a flag for a test part which would allow removal either by a student or automatically after a period of time would greatly tidy the data set. Even if the 'test

part' flag was not related to part removal, it would still make automatic filtering of test parts easier. An even better solution, might be to have separate private and public repositories, so that testing and editing could be done in the private repository, and only well curated parts could be released into the public repository.

Insufficient Continuous Curation: The lack of curation is an overarching theme relating to free text, provenance, part duplication, and part removal. Currently, the iGEM data set is curated (to create part distributions and lists of best documented parts), but this is insufficient. The metric used to determine the documentation level of parts is the length of the page's HTML (`http://parts.igem.org/Well_Documented_Parts`). As there are several parts that have long sequences embedded in the description or repetitive 'this is a test' descriptions the use of length of HTML is insufficient to determine the level of documentation. Additionally, as the plans for iGEM distributions are currently in flux, the use of 'was in distribution in 20xx' may not remain reliable as a metric that can be applied for curation or data filtering.

Furthermore, there is a particular problem of parts that have varying levels of complete descriptions. If a part has a description field reading 'An RNA thermometer that theoretically switches on translation at 27 degrees Celsius and switched of translation below that temperature. Still to be tested.' but a notes field 'a lot, will be added soon.' Is this a part that should be trusted? Was the part properly tested and useful but the documentation was not finished or was the documentation not finished because the part did not do what was intended? Alternatively, if there is a long and complete description but all the other fields contain 'N/A' is that a student having filled out all the information in one field and not wanting to split it across fields or does it mean the part is not well documented?

Continuous Curation (solution) There are several steps necessary to better curate the iGEM registry: ability to check that information is provided, checking information is useful, integrating curation into the part publication process, and automating curation. To check if information is provided, the use of more fields that each have a specific purpose

rather than fewer broader fields can help. To check that the information is useful, several types of field checks can be used: ontologies (controlled vocabularies), regex field checkers, multiple choice inputs, and length checkers. The integration of curation can happen by searching for similar or twin sequences when a part is being submitted and asking the user to curate their part and the part metadata before submission. Finally, the automation of curation can occur if the iGEM registry entries are more machine readable. This would mean that fewer secondary heuristics have to be used and curation can happen more continuously and at a larger scale. Automated curation may also include cross-referencing with other databases (e.g., GenBank,[35] KEGG,[36] and Uniprot[37]) and the creation of part statistics for user filtering (e.g., how many times has this part been used, or how often has this part been cited). Finally, as mentioned earlier, having separate private and public repositories would allow parts to be developed over time and enable a process in which a curator gatekeeper could ensure that only well-documented parts are published, though it may be difficult to make such a process compatible with the goals of iGEM.

Discussion

The results section highlights several points that need to be addressed to allow the iGEM data set to be transformed into a well-curated and machine-accessible library. These include: 1) the over reliance on free text, 2) insufficient part provenance, 3) part duplication, 4) the lack of part removal, and 5) insufficient continuous curation. Several solutions are also suggested to address these issues: 1) the use of ontologies and controlled vocabularies for some fields including species, citations, and references, 2) more specific fields rather than fewer vague/expansive fields including author fields, 3) more cross-referencing with other databases, 4) a standard way of citing iGEM parts, 5) increase editing capabilities or allow versioning, 6) functional components, 7) a test part flag, 8) private repositories for parts under development, and 9) integration of automatic curation with the submission process.

A curated iGEM library is the end goal due to the difficulty of defining a 'trustworthy' part. Rather than attempting to construct one metric for part usefulness that fits all workflows, a curated library allows searching and filtering based on a users preferences. Thus, a curated library allows for the many different types of synthetic biology that are being carried out. However, to make filtering and searching useful, there has to be a sufficient number of well filled out fields that match what users are looking for. One of the proposed solutions to the difficulty of creating a curated iGEM data set is more specific data fields combined with ontologies. However, if the data set is going to be re-used, there must not only be more specific data fields, but also an alignment between data fields for part submission and the data fields users want to search over. To this end, it would be helpful to carry out a survey of which fields people from across the synthetic biology community would like to search over. Creating a list of fields which 80% of the community or workflows would search over would help ensure that the work towards a curated data set is aligned with what would be most useful to the community.

Along the same lines, the input of data should be aligned with what is needed for part reuse. In particular, if ontologies are used for fields, this should be made easy for those entering data. The use of human readable names for ontology components is helpful. This might be accomplished through the use of spreadsheet data entry tools that are linked to ontologies such as ObjTables,[38] Rightfield,[39] and Excel2SBOL.[40]

A suggested solution for part duplication issues is the use of functional parts. Functional parts describe the way a part acts in a particular situation and all uploads of sequences which act in such a way would be linked to the functional part record. For example, functional part 'promoter with a relative promoter unit (RPU)[34] of 0.8' might be linked to sequences which have those properties 'in E. coli at 25C' or the 'Golden Gate Assembly compatible S. cerevisiae promoter'. However, to implement this, there needs to be agreement about what the functionality of a component is and how it is measured. Does functionality include the temperature or not, e.g., RPU at 25C or just RPU? Additionally, how does one measure

the functionality of different components? RPU can be used for promoters and binding efficiency for ribosomes, but choosing easily measurable heuristics for all the different part types is difficult.

Another difficulty is the information about part authors. Being able to contact part authors for more information is similar to the standard practice of a corresponding author in scientific publications. However, some of the students participating in iGEM are high school students. Thus, additional care and consideration should be taken about the privacy and safety of their contact information. One possible solution might be to supply the contact information of a team mentor instead. This also has the advantage of being more likely to be a stable email address as advisors are less likely to frequently change academic institutions and associated academic email addresses.

Cross-referencing of the iGEM data set with other databases is one of the suggestions for incorporation into continuous curation. This would help provide additional provenance information, and confirmation of the data entered. There have already been attempts to better cross-reference the iGEM data set. These include the creation of the BioMaster BioBrick data base[41] and the creation of Uniprot (`https://github.com/watturkara/Plugin-Visual-UniprotLink`) and GenBank (`https://github.com/helloSeen/Plugin-Visual-Genban`) cross-referencing plugins for SynBioHub. Continuous curation can similarly be integrated into the iGEM data set via the use of curation plugins. For example, on SynBioHub there are SnapGene auto-annotation plugins (`https://github.com/SynBioHub/Plugin-Submit-SnapGene` and `https://github.com/SynBioHub/Plugin-Download-SnapGene`), and SYNBICT auto-annotation could be similarly integrated via a plugin interface. Such plugins could either be run only on demand, or be run whenever a new part is submitted and highlight areas where more information is required. Whilst the iGEM registry does not currently appear to have such plugins, they could be implemented into their submission procedure.

The lessons from this paper can also be used to create a completely new 'living' database of useful parts. A living database is one that would allow the continual submission and

editing of parts without compromising the usefulness of the parts it contains. To make such a database compatible with the current iGEM competition, strict versioning would need to be in place so that judging may be carried out on the version/state of a part at the submission deadline without hampering the continued updating of the database. The creation of a living data base must consider the workflows of its users and integrate the lessons presented in this paper to make their experience better. These lessons could be integrated in two different aspects of the database: part retrieval and part submission.

Part retrieval is the finding of parts and referencing that they were used. The creation of a database that allows cardinal faceted search (i.e. a series of numbered filters along a dimension/facet of the data set, e.g. Species: *E. coli* 2982, *C. elegans* 809, *D. melanogaster* 304, *H. sapiens* 72) will reduce the duplication of parts as it will be easier to reuse parts rather than 'reinventing' them. This search would be based on machine readable fields such as species, and curation flags like 'test' and 'metadata completion score', rather than free text. This kind of search may also leverage ontologies to make query expansion easier. Additionally, a strong reference system (for example via part DOIs) would allow the incorporation of parts into methods and reference sections of papers making finding parts via the literature easier as well. This would be complemented by relational searches that allow searching for all parts from the same or similar papers, organisms, etc. Part submission is the other half of the database. The integration of curation with submit is the most important way of implementing the lessons learned.

The three forms of curation to implement are duplicate highlighting, metadata completion nudging, and free text named entity recognition (NER). Duplicate highlighting would be done via sequence matching, it presents the user with parts that use the same or similar sequence and presents the option to use metadata from those parts or to create an additional version of the part rather than a new submission. Metadata completion nudging is the attempt to ensure a 'complete' set of metadata is captured as needed by the search half of the database. This may be done via the use of drop down menus to make choices about certain fields (e.g.

species or role), the use of multiple text boxes to remind the user about all the different kinds of information that they should provide (such as references about part provenance), and the automated suggestions of tags or additional fields based on information the user has already provided or similar parts (automated fields might include cross-references with other databases). Finally, the use of NER for free text is the immediate grounding of terms used in free text. This ensures that users can still use free text but already starts to ground terms and make the field more machine readable. For example, if a user writes "testing was done with E. coli in mind" the highlighting of the term E. coli and connecting it to the ontology term for E. coli allows linking to the free text via the 'object' E. coli despite spelling variations and free text which allows a less restrictive context than standard drop down fields. An example of what a submission interface might look like is shown in Figure 8.

Most of the suggestions are focused on changes to the iGEM registry to make future curation easier. However, these steps will not make the current wealth of iGEM data any easier to curate. Thus, to still be able to use the current iGEM data set, further curation is required. This could include an investigation of the 'realness' of parts over time for each team. As hand curation or simple statistical curation is insufficient, machine learning could be applied to make further progress. Natural Language Processing techniques could be used to identify key information in long text fields, which would then be presented to experts for verification. In this way, free text could be used to populate several data fields, potentially even using ontologies if the terms are properly grounded.

There is still a lot of work to be done to curate the iGEM data set, and to decrease the effort required to curate future data sets. The sooner the work is done, the less effort that will need to be expended retroactively in curation. Additionally, this would make the wealth of information contained in the iGEM registry more accessible for researchers and likely increase part reuse. In this way, continuously curating the iGEM registry will help advance the field of synthetic biology.

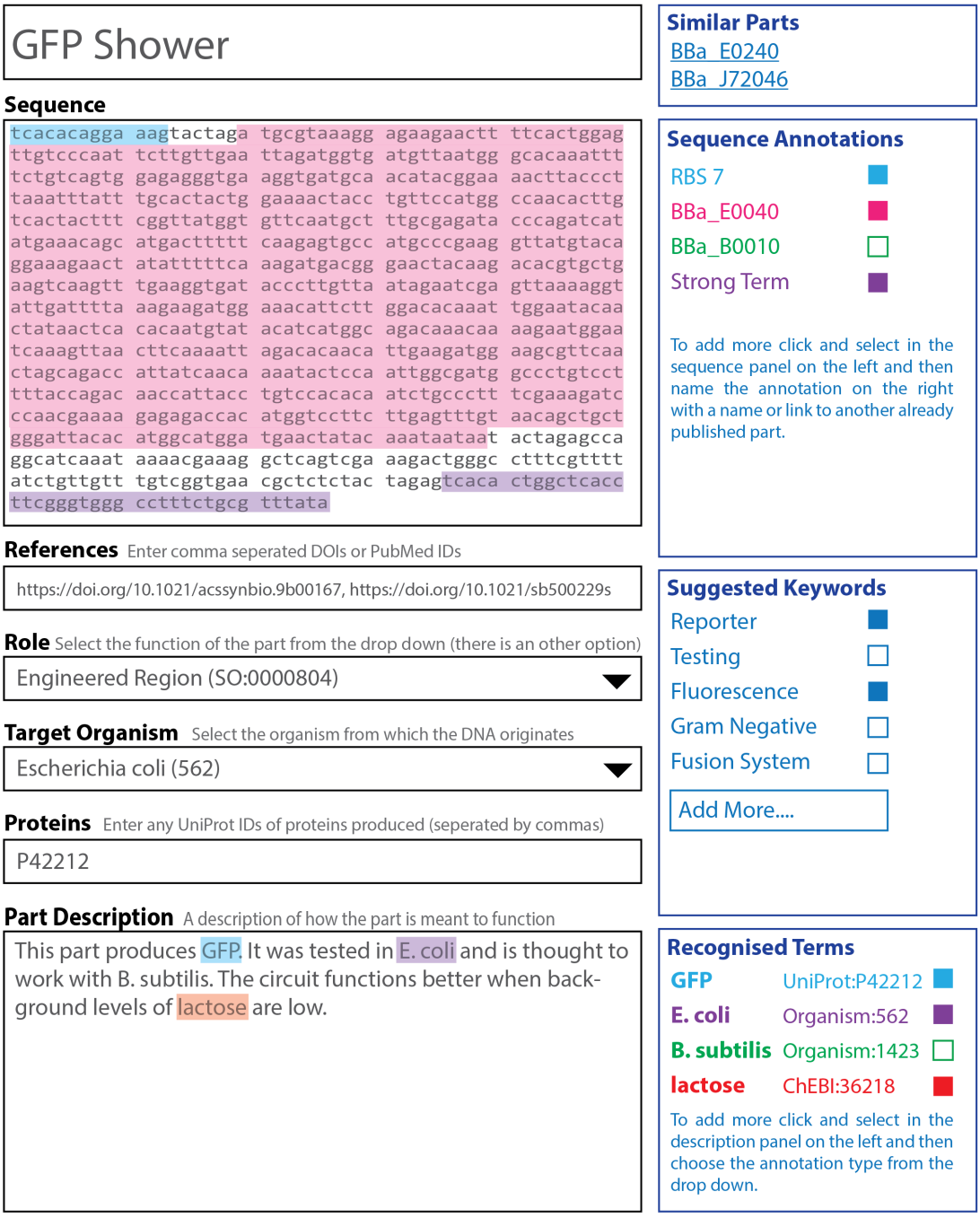

Figure 8: This shows a potential submit interface for a new iGEM database. The left hand side (black) tackles data input, whilst the right hand side (blue) integrates machine aided curation with submission. This submission interface decreases the reliance on free text, ensures part provenance is captured, reduces part duplication (by the similar parts box up at the top right), and integrates continuous curation. The one aspect that this does not address is the lack of part removal, however this would not be expected to be addressed in part submission. Key features include: the incorporation of linkage to other parts (sequence annotations), the inclusion of references, the use of ontologies (for role, target organism), the cross-linking with other databases (the protein link with UniProt), the grounding of terms in Free text with NER, and the prompting of the user to provide all the relevant bits of information with separate boxes and explanations for different kinds of information.

Methods

Several python scripts were used for the analysis of the iGEM data set. These can be found in two github repositories: `https://github.com/JMante1/iGem-Data-Cleaning` and `https://github.com/synbioks/sequence_supplementals`.

iGEM Registry to SBOL Conversion: To convert the iGEM Registry to SBOL a simple automatic conversion method was used:

1. Each part is converted into a ComponentDefinition.
2. The Sequence Ontology (SO) role for that part is mapped from the iGEM part type as defined in columns 1 and 2 of Table 1
3. A composite part composed of other BioBricks is constructed using Component instantiations.
4. A composite part not composed of BioBricks, but rather simply described with annotations has these annotations converted into SequenceAnnotations with roles.
5. The categories are been converted into Collections. Each sub-category is mapped into a member of its parent category. All 'top-level' categories are mapped into a category_collection, and this collection and all iGEM parts are members of an iGEM collection. Each collection that a part is a member of has been annotated as an iGEM annotation within the SBOL record for the part.
6. Most fields available in the iGEM SQL database that have not mapped as above have been mapped into iGEM custom annotations. There are a few exceptions, but these are mostly fields that map into some other table that is not currently accessible. One example is there the GroupId field that somehow maps to a Group who provided the part, but this mapping has not been shared by iGEM.

The result of this procedure was: 372 Collections, 38,365 ComponentDefinitions, and 36,595 Sequences. These data were uploaded to `https://synbiohub.org`.

Simple Statistics: An initial analysis was carried out to collect simple statistics about the iGEM data set. For this analysis, the SPARQL interface of SynBioHub was used to calculate overview statistics (`https://www.w3.org/TR/2013/REC-sparql11-overview-20130321/`). Then, the data was pulled using SPARQL queries called from a Python script. The Python script was used to calculate more in depth statistics, including the number of unique sequences, sequence length statistics, and the analysis of sequence similarity and the creation of dendograms. Figure 9 shows the filtering steps to find 'basic unique' parts.

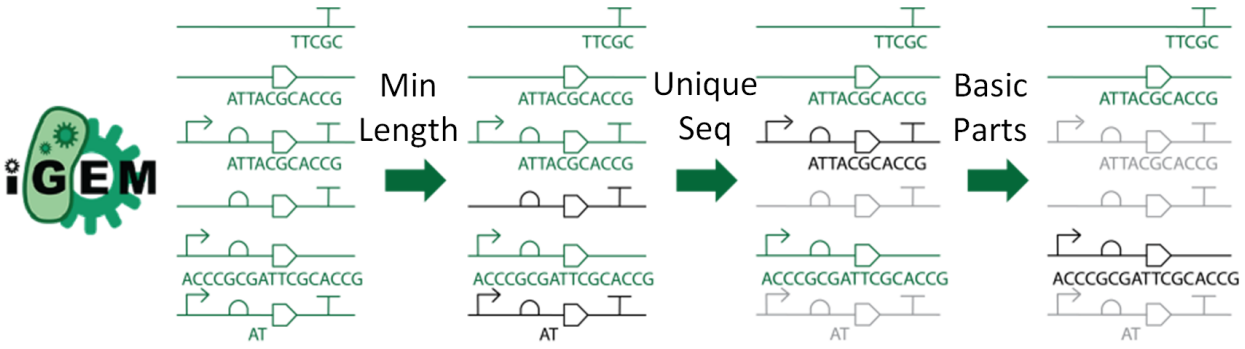

Figure 9: As part of the simple statistics a series of filters were applied to the iGEM Data Set. First any parts with a sequence less than the minimum length (see Table 1) were filtered out. Next any duplicate/twin sequences were removed from the count (the most annotated sequence was kept). Finally, any sequences with annotations were filtered out to leave 'basic unique' parts which were a starting point for creating a library of building blocks. The iGEM is the property of the iGEM Foundation and used under the Creative Commons Attribution.

SnapGene: For the SnapGene auto-annotation, a reduced data set was used based on the 'basic unique' sequences (removed sequence duplicates and any which had sequence annotations) found in the simple statistics analysis. The sequences whee then annotated using a SnapGene server hosted at `snapgene.sbolstandard.org`. Files were automatically uploaded to the SnapGene server and the annotated GenBank format and plasmid map was pulled using a python script. The python script then converted the GenBank to SBOL and used it to find the annotations added by the SnapGene server. Python was then used to carry out further statistics on the returned annotations.

SYNBICT: For the annotation with SYNBICT, papers were mined for parts. Resources mined for yeast parts included the Yeast Toolkit,[22] Pichia Toolkit,[23] and a combinatorial de-

sign paper.[24] Parts for Gram-negative bacteria were drawn from the CIDAR MoClo kit,[25] the CIDAR Extension Kit Volume I,[26] the Voigt Lab terminator collection,[27] and the *Bascillus subtilis* collection.[28]

The data mined from these papers was input into Excel spreadsheets and converted to SBOL using Excel2SBOL (`https://github.com/SynBioDex/Excel-to-SBOL`). The collections were uploaded to a SynBioHub instance and subsequently used as the feature libraries in SYNBICT. The SYNBICT sequences_to_features module was used to annotate the 'basic unique' ribosome binding sites with a minimum feature length of 10 bp. An overview of the annotations was created using a Python script.

For annotation with SYNBICT's feature library derived from RegulonDB,[29] the sequences_to_features module was also used, but with a minimum feature length of 80 bp instead of 10 bp to avoid annotating with sub-features of promoters such as their operator binding sites. The minimum target length was set to 899 bp. For identifying overlapping annotations, a cover offset of 80 bp was used and sequences_to_features was run in non-interactive mode. To perform the statistical analysis of how many annotations SYNBICT made and how many of these annotations were for potentially novel features, a Python script was used that takes the log file produced by sequences_to_features as input.

Expert Curation: For expert curation, an abbreviated version of the SBOL version of the iGEM data set was used. It contained only parts classified as 'basic unique' by earlier analysis. The data was converted to a CSV format with the fields/properties being converted to columns. The columns included: long description, short description, notes, and source as the free text fields. The CSV was viewed and analysed using OpenRefine `https://github.com/OpenRefine`. This allowed the reading of all of the descriptions to compile a list of all mentioned species. This list was then used together with the text filter functionality to count the number of rows/parts that contained mentions of a particular species. The ability to flag rows meant any particularly silly descriptions were flagged for further analysis and discussion.

For the analysis of the difference between real and spurious parts, a sample of 400 random ones of each was taken. The random selection was done by using Excel to generate a random number column and sorting based on that. The first 400 'real' parts and first 400 spurious parts were then selected for the analysis.

Acknowledgement

CM and JM are supported by the National Science Foundation under Grant No. 1939892. JM is additionally supported by a Dean's Graduate Assistantship at the University of Colorado Boulder. NR and JB are supported by DARPA award HR0011-15-C-0084. EY and KK are supported by the National Science Foundation under Grant No. 1939860. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the funding agencies.

Author Contributions

JAM and CM converted the iGEM data set to SBOL. JM and NR carried out the SYNBICT analysis. KK created the libraries used with SYNBICT. The rest of the analysis was carried out by JM. EY, JB, and CM supervised the project. All authors contributed to the writing of this manuscript.

Conflicts of Interest

The authors declare no conflicts of interest.

Supporting Information Available

Supplemental Files:

- teams_over_time.xlsx: the data behind the preliminary analysis of whether part metadata submitted by a team improves over time
- Part Descriptions Per Year.xlsx: data behind the analysis that part descriptions do not seem to improve over time or by month in a year
- Spurious_vs_actual.xlsx: List of sequences used in the spurious vs real analysis

Supplemental Figures:

- Variation in different part description field lengths over different iGEM years.
- Variation in part submission and description by month and year. a) The number of submissions being made per month. The spikes indicate the approximate time of the jamboree each year. b) Part Description Variation by Month.
- Variation in the average part description field lengths over time for three different iGEM submission groups. No clear pattern is visible here or in any of the simple scatter plots.

Link to the SynBioHub repository containing the SBOL iGEM Data Set:

- `https://synbiohub.org/public/igem/igem_collection/1`

Github Code Repositories for the analysis:

- `https://github.com/JMante1/iGem-Data-Cleaning`
- `https://github.com/synbioks/sequence_supplementals`

SnapGene Server Instance for SnapGene autoannotation:

- `snapgene.sbolstandard.org`

Links to the SBOL files of the annotation libraries used with SYNBICT:

- `https://synbiohub.org/public/bsu/bsu_collection/1`

- https://synbiohub.programmingbiology.org/public/Cello_Parts/Cello_Parts_collection/1
- https://synbioks.org/public/Pichia_MoClo_Toolkit_Lu_Lab/Pichia_MoClo_Toolkit_Lu_Lab_collection/1
- https://synbioks.org/public/CIDAR_MoClo_Toolkit_Densmore_Lab/CIDAR_MoClo_Toolkit_Densmore_Lab_collection/1
- https://synbioks.org/public/EcoFlex_MoClo_Toolkit_Freemont_Lab/EcoFlex_MoClo_Toolkit_Freemont_Lab_collection/1
- https://synbioks.org/public/MoClo_Yeast_Toolkit_Dueber_Lab/MoClo_Yeast_Toolkit_Dueber_Lab_collection/1
- https://synbioks.org/public/Anderson_Promoters_Anderson_Lab/Anderson_Promoters_Anderson_Lab_collection/1
- https://synbioks.org/public/Itaconic_Acid_Pathway_Voigt_Lab/Itaconic_Acid_Pathway_Voigt_Lab_collection/1
- https://synbioks.org/public/CIDAR_MoClo_Extension_Kit_Volume_I_Murray_Lab/CIDAR_MoClo_Extension_Kit_Volume_I_Murray_Lab_collection/1
- https://synbioks.org/public/Natural_and_Synthetic_Terminators_Voigt_Lab/Natural_and_Synthetic_Terminators_Voigt_Lab_collection/1

References

(1) Smolke, C. D. Building outside of the box: iGEM and the BioBricks Foundation. *Nature Biotechnology* **2009**, *27*, 1099 – 1102.

(2) Shetty, R. P.; Endy, D.; Knight, T. F. Engineering BioBrick vectors from BioBrick parts. *Journal of Biological Engineering* **2008**, *2*, 5.

(3) Kahl, L. J.; Endy, D. A survey of enabling technologies in synthetic biology. *Journal of Biological Engineering* **2013**, *7*, 13.

(4) Timmons, J. J.; Densmore, D. Repository-based plasmid design. *PLOS ONE* **2020**, *15*, e0223935.

(5) Siyari, P.; Dilkina, B.; Dovrolis, C. Evolution of Hierarchical Structure and Reuse in iGEM Synthetic DNA Sequences. Computational Science - ICCS 2019. 2019; pp 468 – 482.

(6) Barone, F.; Dorr, F.; Marasco, L. E.; Mildiner, S.; Patop, I. L.; Sosa, S.; Vattino, L. G.; Vignale, F. A.; Altszyler, E.; Basanta, B. et al. Design and evaluation of an incoherent feed-forward loop for an arsenic biosensor based on standard iGEM parts. *Synthetic Biology* **2017**, *2*.

(7) Vilanova, C.; Porcar, M. iGEM 2.0 - refoundations for engineering biology. *Nature Biotechnology* **2014**, *32*, 420 – 424.

(8) MÃ¼ller, K. M.; Arndt, K. M. In *Synthetic Gene Networks: Methods and Protocols*; Weber, W., Fussenegger, M., Eds.; Methods in Molecular Biology; Humana Press, 2012; pp 23 – 43.

(9) Alterovitz, G.; Muso, T.; Ramoni, M. F. The challenges of informatics in synthetic biology: from biomolecular networks to artificial organisms. *Briefings in Bioinformatics* **2010**, *11*, 80–95.

(10) Peccoud, J.; Blauvelt, M. F.; Cai, Y.; Cooper, K. L.; Crasta, O.; DeLalla, E. C.; Evans, C.; Folkerts, O.; Lyons, B. M.; Mane, S. P. et al. Targeted development of registries of biological parts. *PloS One* **2008**, *3*, e2671.

(11) Matsuoka, Y.; Ghosh, S.; Kitano, H. Consistent design schematics for biological systems: standardization of representation in biological engineering. *Journal of the Royal Society Interface* **2009**, *6*, S393–S404.

(12) Galdzicki, M.; Chandran, D.; Nielsen, A.; Morrison, J.; Grünberg, R.; Sleight, S.; Sauro, H. BBF RFC 31: Provisional BioBrick Language (PoBoL). 12.

(13) Canton, B.; Labno, A.; Endy, D. Refinement and standardization of synthetic biological parts and devices. *Nature Biotechnology* **2008**, *26*, 787–793.

(14) Purnick, P. E. M.; Weiss, R. The second wave of synthetic biology: from modules to systems. *Nature Reviews Molecular Cell Biology* **2009**, *10*, 410–422.

(15) Porcar, M.; Peretó, J. Are we doing synthetic biology? *Systems and Synthetic Biology* **2012**, *6*, 79–83.

(16) Frischmann, B. M.; Madison, M. J.; Strandburg, K. J. *Governing Medical Knowledge Commons*; Cambridge University Press, 2017; p 441, Google-Books-ID: Ai02DwAAQBAJ.

(17) Kwok, R. Five hard truths for synthetic biology. *Nature* **2010**, *463*, 288–290.

(18) *New Horizons for a Data-Driven Economy*; Springer International Publishing, 2016; p 103.

(19) Galdzicki, M.; Clancy, K. P.; Oberortner, E.; Pocock, M.; Quinn, J. Y.; Rodriguez, C. A.; Roehner, N.; Wilson, M. L.; Adam, L.; Anderson, J. C. et al. The Synthetic Biology Open Language (SBOL) provides a community standard for communicating designs in synthetic biology. *Nature Biotechnology* **2014**, *32*, 545–550.

(20) Eilbeck, K.; Lewis, S. E.; Mungall, C. J.; Yandell, M.; Stein, L.; Durbin, R.; Ashburner, M. The Sequence Ontology: a tool for the unification of genome annotations. *Genome biology* **2005**, *6*, R44.

(21) Blogger, G. Identifying Sequence Elements with SnapGene's Feature Database. https://blog.addgene.org/identifying-sequence-elements-with-snapgenes-feature-database.

(22) Lee, M. E.; DeLoache, W. C.; Cervantes, B.; Dueber, J. E. A Highly Characterized Yeast Toolkit for Modular, Multipart Assembly. *ACS Synthetic Biology* **2015**, *4*, 975–986, Publisher: American Chemical Society.

(23) Obst, U.; Lu, T. K.; Sieber, V. A Modular Toolkit for Generating *Pichia pastoris* Secretion Libraries. *ACS Synthetic Biology* **2017**, *6*, 1016–1025.

(24) Young, E. M.; Zhao, Z.; Gielesen, B. E. M.; Wu, L.; Benjamin Gordon, D.; Roubos, J. A.; Voigt, C. A. Iterative algorithm-guided design of massive strain libraries, applied to itaconic acid production in yeast. *Metabolic Engineering* **2018**, *48*, 33–43.

(25) Iverson, S. V.; Haddock, T. L.; Beal, J.; Densmore, D. M. CIDAR MoClo: Improved MoClo Assembly Standard and New E. coli Part Library Enable Rapid Combinatorial Design for Synthetic and Traditional Biology. *ACS Synthetic Biology* **2016**, *5*, 99–103, Publisher: American Chemical Society.

(26) Addgene: CIDAR MoClo Extension, Volume I. https://www.addgene.org/kits/murray-cidar-moclo-v1/#protocols-and-resources.

(27) Chen, Y.-J.; Liu, P.; Nielsen, A. A. K.; Brophy, J. A. N.; Clancy, K.; Peterson, T.; Voigt, C. A. Characterization of 582 natural and synthetic terminators and quantification of their design constraints. *Nature Methods* **2013**, *10*, 659–664, Number: 7 Publisher: Nature Publishing Group.

(28) Misirli, G.; Wipat, A.; Mullen, J.; James, K.; Pocock, M.; Smith, W.; Allenby, N.; Hallinan, J. S. BacillOndex: An Integrated Data Resource for Systems and Synthetic Biology. *Journal of Integrative Bioinformatics* **2013**, *10*, 103 – 116.

(29) Santos-Zavaleta, A.; Salgado, H.; Gama-Castro, S.; Sánchez-Pérez, M.; Gómez-Romero, L.; Ledezma-Tejeida, D.; García-Sotelo, J. S.; Alquicira-Hernández, K.; Muñiz-Rascado, L. J.; Peña-Loredo, P. et al. RegulonDB v 10.5: Tackling Challenges to Unify Classic and High Throughput Knowledge of Gene Regulation in E. coli K-12. *Nucleic Acids Res.* **2018**, *47*, D212–D220.

(30) Sayers, E. W.; Cavanaugh, M.; Clark, K.; Ostell, J.; Pruitt, K. D.; Karsch-Mizrachi, I. GenBank. *Nucleic Acids Research* **2019**, *47*, D94 – D99.

(31) Schoch, C. L.; Ciufo, S.; Domrachev, M.; Hotton, C. L.; Kannan, S.; Khovanskaya, R.; Leipe, D.; Mcveigh, R.; O'Neill, K.; Robbertse, B. et al. NCBI Taxonomy: a comprehensive update on curation, resources and tools. *Database: The Journal of Biological Databases and Curation* **2020**, *2020*, baaa062.

(32) Shotton, D. CiTO, the Citation Typing Ontology. *Journal of Biomedical Semantics* **2010**, *1*, S6.

(33) Cunningham, F.; Moore, B.; Ruiz-Schultz, N.; Ritchie, G. R.; Eilbeck, K. Improving the Sequence Ontology terminology for genomic variant annotation. *Journal of Biomedical Semantics* **2015**, *6*, 32.

(34) Kelly, J. R.; Rubin, A. J.; Davis, J. H.; Ajo-Franklin, C. M.; Cumbers, J.; Czar, M. J.; de Mora, K.; Glieberman, A. L.; Monie, D. D.; Endy, D. Measuring the activity of BioBrick promoters using an in vivo reference standard. *Journal of Biological Engineering* **2009**, *3*, 4.

(35) Benson, D. A.; Cavanaugh, M.; Clark, K.; Karsch-Mizrachi, I.; Lipman, D. J.; Ostell, J.; Sayers, E. W. GenBank. *Nucleic Acids Research* **2013**, *41*, D36–42.

(36) Kanehisa, M.; Goto, S. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research* **2000**, *28*, 27 – 30.

(37) Consortium, T. U. UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Research* **2019**, *47*, D506 – D515.

(38) Karr, J. R.; Liebermeister, W.; Goldberg, A. P.; Sekar, J. A. P.; Shaikh, B. ObjTables: structured spreadsheets that promote data quality, reuse, and integration. *arXiv:2005.05227 [cs, q-bio]* **2020**, arXiv: 2005.05227.

(39) Wolstencroft, K.; Owen, S.; Horridge, M.; Krebs, O.; Mueller, W.; Snoep, J. L.; du Preez, F.; Goble, C. RightField: embedding ontology annotation in spreadsheets. *Bioinformatics* **2011**, *27*, 2021 – 2022.

(40) Pötzsch, I. M.; Mante, J.; Beal, J.; Myers, C. J. Creating SBOL Designs with Excel. Computational Modeling in Biology Network Forum (COMBINE 2020). 2020.

(41) Wang, B.; Yang, H.; Sun, J.; Dou, C.; Huang, J.; Guo, F.-B. BioMaster: An Integrated Database and Analytic Platform to Provide Comprehensive Information About BioBrick Parts. *Frontiers in Microbiology* **2021**, *12*.

Graphical TOC Entry