16.11.2014 submit your revision within forty-five days of the date of this decision.   
  
0. EH, HHR: From what you told me, you used some new methodes to achieve calculations for longer peptides than, initially possible. I guess if we use them in Pelica, we have to describe what was done in the paper.

1. HHR, TS, EH: Rebuttal letter that responds to each point brought up by the academic editor and reviewer(s). This letter should be uploaded as a 'Response to Reviewers' file

2. HHR, TS, EH: A clean revised manuscript as your 'Manuscript' file.

~~3. EH: A marked-up copy of the changes made from the previous article file as a 'Revised Manuscript with Track Changes' file. This can be done using 'track changes' in programs such as MS Word and/or highlighting any changes in the new document.~~

4. EH: Please include separate legends for each figure, at the end of the main text in your manuscript.  
  
5. EH: Remove all Supplementary Tables from your manuscript file and upload them as separate files, with the file type of Supporting Information.   
  
6. EH: Include separate titles and legends for all uploaded Supporting Information files at the end of your manuscript beneath the references, and update your in-text citations and item descriptions in Editorial Manager to match these exactly. Please see our Supporting Information guidelines for more information: [http://www.plosone.org/static/supportingInformation](https://3c.web.de/mail/client/dereferrer?redirectUrl=http%3A%2F%2Fwww.plosone.org%2Fstatic%2FsupportingInformation" \t "_blank).  
  
~~7. EH, HHR: 1. The (approximate) expression for Var[Z] given in theorem 1 (p. 6) is wrong. The expression for E[Z] is correct, but its derivation is not. As a result, the expressions for the coverage variance (p. 9) and for the relative efficiency variance (p. 10) are also wrong.  
  
In more detail: In the proof of theorem 1, when deriving the expression for E[Z\_{i+1}] (bottom of p. 21), the authors write E[Z\_i]^2 instead of E[Z^2\_i]. In this part of the proof the terms cancel out so the final result is correct. However, this is not the case with Var[Z] - here (middle of p. 22), the correct second moment is   
  
E[Z^2\_{i+1}] = (b - 2)/b \* E[Z^2\_i] + (2b - 1)/b \* E[Z\_i] + 1  
  
and not  
  
E[Z^2\_{i+1}] = (b - 2)/b \* E[Z\_i]^2 + (2b - 1)/b \* E[Z\_i] + 1  
  
Therefore, one cannot substitute here the expression for E[Z\_i] derived earlier. The resulting variance is therefore wrong (a small simulation check can easily confirms this).  
  
I am not sure whether the current, recursive approach can be fixed to yield a valid expression for Var[Z]. However, Kong (J. Theo Bio, 2009, vol. 259, pp. 641-645) already derived correctly the variance (in the context of DNA randomization, which is mathematically identical) using a different approach, namely, by representing Z as a sum of indicators. The current approach is also somewhat problematic since it does not take into account the boundary condition Z\_i = b, though this technical difficulty can be easily fixed.~~We completely agree with reviewer, and we are sorry for the earlier oversight. We have restated the variance results. The recursion does, indeed, allow for a derivation of the variance (based on a conditional random variable $Z\_{i+1}$ given $Z\_i$. The results now match up with the results presented by Kong (2009), and we have referenced this earlier finding accordingly.

~~8. EH: p. 3, line 1: “andclonal” should be “and clonal”~~  
 ~~9. EH: p. 5, last paragraph: “describing the fraction” should be “describing the expected fraction”~~  
  
10. TS, EH: The term “peptide diversity” is defined in p. 5, yet it used beforehand, throughout the introduction. It is better to define a term before using it. In addition, if I understand correctly, the random variable Z, defined in p. 6, is exactly “peptide diversity”; if this is so, it is better to explicitly state that.  
  
Reply: The meaning of the term peptide diversity was in fact defined earlier in the manuscript, but it was again defined on page 5. We changed the wording to clarify this point and define the term at its first use on page two.

~~11. EH, HHR: p. 6, Theorem 1: it should be stated in the theorem (and not only in the preceding text) that the b possible peptide types are equally likely.~~  
  
~~12. EH, HHR: p. 6, Theorem 1: “relative standard deviation” is not the square root of the variance, but rather, the ratio between this square root and the mean (as stated correctly in the bottom of p. 22).~~  
~~13. TS, EH: p. 8, 3rd line after table 2: I believe that “number of codons” should be “number of amino acids”~~  
This refers to the definition of n – this is the number of elements from each amino acid class. We refered to them as codons, but it could also be aas, so it doesn’t matter. I’ve changed it to aa.  
  
14: HHR, EH: ~~p. 9, middle: the quantity Np\_i is only the expected number of peptides from class i in the library. The actual number is random, and may be lower or higher. Therefore, the expression b\_i(1 – exp{-Np\_i/b\_i}) is only an approximation of the true quantity of interest, and it is unclear how precise are the subsequent approximations for D(N, k) etc. Therefore, first, it should be stated clearly that this is an approximation (on top of the approximations from equations 1 and 2);~~ second, it will be illuminating to conduct a simulation study to investigate this point, i.e., to compare the reported approximate mean and variance of the peptide diversity with the ones estimated in simulation.  
  
15. HHR, EH: p. 9, ~~bottom: expected coverage can never equal 1, since coverage, as a random variable, is bounded by 1 and has a positive probability of being < 1. Therefore, the remark “for this, the size of ...” (p. 10) is irrelevant and should be omitted.~~  
  
~~16. EH: p. 10, line 4: “profound effect on coverage” should be “profound effect on expected coverage”~~  
  
We agree, and generally replaced “coverage” by “expected coverage”.  
  
17. TS: ~~p. 25, Table S4: The table shows that for some sequences (SLRSLRS), the inclusion probabilities are the highest under NNN randomization and lowest under trimer randomization, whereas for others (MMMMMMM) the order is reversed. This point needs to be explained and discussed: briefly, S, L, and R are the most probable amino acids under NNN, with probability of 6/64 for each, and M is the least probable (1/64). The probabilities under trimer are always 1/19, between these two values; hence the phenomenon.~~  
  
Text altered on page 12: The high variability introduced by schemes with varying codons per amino acid ratios causes libraries to be biased towards peptides with a high number of possible encodings at the cost of rare ones. This makes the chance of success in selections strongly dependent on the question, if the a priori unknown “best" peptide has many possible encodings or not. Therefore, the inclusion probability for some peptides is maximal in biased schemes like NNN and exceeds that achievable with 20/20 encoding (See "Examples for inclusion probability" and table S4 in the supplementary). However, for about 75% of all possible peptides the highest inclusion probability is reached when an unbiased coding scheme like trimer is used (See Figure 3).

Text added on page 24:

The exemplary peptides in Table S4 cover the maximal range of inclusion probabilities. While the amino acids S, L and R are represented by 6 codons each in NNN (probability of 6/64), M is only encoded by ATG (probability of 1/64). The inclusion probability of “SLRSLRS” is therefore much higher than that of “MMMMMMM” if the NNN encoding scheme is used. However, with encoding schemes like Trimer in which each amino acid has only one representation, all individual peptides have the same inclusion probability (probability for each amino acid is 1/19, if cysteines and stops are excluded).

18. ~~EH, HHR: p. 25, Table S7: at least some of the values reported in the table are incorrect. For example, for NNS and library size of ten million, GLUE-IT reports completeness of 0.1081, and not 0.9016 (as written in the table). Another server, Toplib ([http://stat.haifa.ac.il/~yuval/toplib](https://3c.web.de/mail/client/dereferrer?redirectUrl=http%3A%2F%2Fstat.haifa.ac.il%2F%7Eyuval%2Ftoplib" \t "_blank)), agrees with GLUE-IT. If the results of PeLiCa are indeed as reported in Table S8, there is probably an error in the server implementation, or perhaps even with the entire analysis. Table S7 and S8 values are all mixed up by confusing the assignment of the randomisation scheme and the library size.~~

There was a problem with tables S7 and S8 – we believe that the tables were transposed. This problem is now fixed. We double-checked the results given by PeLiCa, and they are identical to the ones given in the manuscript. We have by now consolidated tables S7 and S8 into one, to make a comparison between GLUE-IT and PeLiCa easier.

~~19. EH, HHR: See #18…Similar problems may be present in Tables S5 and S6.~~  
  
~~20. EH: Figure 1: the meaning of the “maximum” curve is not defined.~~  
  
21. EH: Figure 2 and Figure 3 got mixed up  
  
22. TS, EH, HHR: The language of the paper is too specific. The phrasing and scientific language used to write the article have clear origin in the field of mathematics, making reading and understanding of the text difficult for researchers from other backgrounds.

See #37.

23. ~~TS, EH, HHR: Further have the authors coined a couple of terms, where other phrases already existed and are easier to understand (e.g multiplicity ≙ redundancy of the genetic code; see minor remarks).~~   
Replaced “multiplicity” in the text with “redundancy” whenever applicable.

24. EH; HHR: The authors state on page 15 “(*Data not shown)*”. In times with unlimited space in the SI, please add the data. Especially since the conclusions of the authors is incorrect. NNK/S (32 codons: 19 aa, 1 C: 1 stop) is always significant advantages over NNB (48:19:1:1). NNK/S provides higher probabilities that rare sequences are included due to the lower redundancy as NNB, therefore having per se a higher diversity irrespective of library size. The numbers outputted by PeLiCa for these two cases do not reflect this.   
We agree with the reviewer – we should have gone into more detail in the discussion of this example in the first place. However, we have to respectfully disagree that our conclusions are wrong.   
Expected coverage depends strongly on the amount of initial loss due to invalid sequences. When cysteines are included in the schemes, 31 out of 32 codons in the NNK/S scheme lead to valid sequences (= 96.9%), whereas 47 out of 48 codons lead to valid sequences in the NNB scheme (= 97.6%). For larger libraries, when an expected coverage of about 50% is reached, the expected coverage under the NNK/S scheme does win out over the NNB scheme, due to the higher redundancy of NNK/S, we completely agree on this point.

25. TS, EH: There are no figure legends and the order of the figures is mixed up (especially 2 and 3).

We are sorry for the omission – the captions for the figures got lost during the submission process.

26. Insert citations EH: Missing Citations:

*Algorithm, statistical or mathematical approaches:*

2. Makowski, L. and Soares, A. (2003) Estimating the diversity of peptide populations from limited sequence data. Bioinformatics **19**, 483–489   
cite as makowski03

(\*)3. M. J. Volles and P. T. Lansbury Jr, A computer program for the estimation of protein and nucleic acid sequence diversity in random point mutagenesis libraries. Nucleic Acid Research 2005, 33 (11): 3667-3677. doi: 10.1093/nar/gki669

cite as volles:2005

(\*\*)5. M. Denault and J. N. Pelletier "Protein library design and screening: Working out the probabilities". Methods in Molecular Biology: Protein Engineering Protocols, Humana Press (2007)

cite as denault:2007

(\*)5. Y. Kong, Calculating complexity of large randomized libraries. J Theor Biol. 2009, 259(3):641-5. doi: 10.1016/j.jtbi.2009.04.008.   
cite as kong:2009

6. Y. Nov, Probabilistic Methods in Directed Evolution: Library Size, Mutation Rate, and Diversity, 2014, DOI 10.1007/978-1-4939-1053-3\_18,

cite as nov:2014

*Removal of redundancy libraries/technologies and conclusion thereof:*

(\*)7. Hughes, M. D., Nagel, D. A., Santos, A. F., Sutherland, A. J., and Hine, A. V. (2003) Removing the redundancy from randomised gene libraries. J. Mol. Biol. 331, 973−979.  
cite as hughes:2003

(\*)8. M. Ashraf, L Frigotto, M E. Smith, S Patel, M D. Hughes, A J. Poole, H R.M. Hebaishi, C G. Ullman, and A V. Hine, ProxiMAX randomization: a new technology for non-degenerate saturation mutagenesis of contiguous codons 1,2 Biochem Soc Trans. Oct 1, 2013; 41(Pt 5): 1189–1194. doi: 10.1042/BST20130123   
ashraf:2013

(\*)9. S Kille, C G. Acevedo-Rocha, L P. Parra, Z-G Zhang, D J. Opperman, M T. Reetz, and J P Acevedo, Reducing Codon Redundancy and Screening Effort of Combinatorial Protein Libraries Created by Saturation Mutagenesis, ACS Synthetic Biology 2012 2 (2), 83-92   
kille:2013

9a. L Tang, H Gao, X Zhu, X Wang, M Zhou, Construction of “small-intelligent” focused mutagenesis libraries using well-designed combinatorial degenerate primers, BioTechniques, Vol. 52, No. 3, March 2012, pp. 149–158.   
tang:2012

(\*)10. S Hoebenreich, F E. Zilly, C G. Acevedo-Rocha, M Zilly, and M T. Reetz, Speeding up Directed Evolution: Combining the Advantages of Solid-Phase Combinatorial Gene Synthesis with Statistically Guided Reduction of Screening Effort, ACS Synthetic Biology, DOI: 10.1021/sb5002399 Publication Date (Web): June 12, 2014

hoebenreich:2014

(\*)11. Reetz MT, Kahakeaw D, Lohmer R., Addressing the numbers problem in directed evolution. Chembiochem. 2008 Jul 21;9(11):1797-804. doi: 10.1002/cbic.200800298

reetz:2008

12. Neuner, P., Cortese, R., and Monaci, P. (1998) Codon-based mutagenesis using dimer-phosphoramidites. Nucleic Acids Res. 26, 1223−1227.   
neuner:1998

13. Paul Gaytán and Abigail Roldán-Salgado, Elimination of Redundant and Stop Codons during the Chemical Synthesis of Degenerate Oligonucleotides. Combinatorial Testing on the Chromophore Region of the Red Fluorescent Protein mKate. ACS Synth. Biol., 2013, 2 (8), pp 453–462 DOI: 10.1021/sb3001326  
gaytan:2013

27. TS: The authors correctly state that “determination of diversity of a library is difficult.” (page 2) and can only be accessed by sequencing. In the field of saturation mutagenesis, the need of cost-intensive single sequencing to asses library quality was replace by the “Quick-Quality-Control (QQC)”, a method using a single sequencing run to detect significant errors in the DNA sequence of a library, e.g. bad primer/cassette synthesis, primer-dimer insertions, percentage of randomisation over wild-type background etc. It was introduced and shown to be a valid guidance by the group of M. T Reetz (10.1002/adsc.200900644, 10.1021/sb5002399). It should be as useful in libraries for phage display, since the underlying laboratory techniques are the same. It should be mentioned in the introduction to improve article quality.

We inserted the following text:

A cheap, simple and powerful way to investigate if the production of a library was successful is the “Quick-Quality-Control (QQC)” (Reference: C1; C2). In short, library material is pooled and used in a single sanger sequencing run to uncover undesired imbalances in the ratios of inserted bases as well as production errors like primer-dimer insertions etc. which might lead to a reduced library diversity.

28. TS, EH, HHR: Clonal Diversity should be called what it is, the number of clones obtained on plate after transformation. It has nothing to do with diversity and gives no value of how divers the dna/peptide library is. If the cassettes/primers used to create the clones were results of bad solid-phase synthesis (and that happens) the clones may just contain non-sense sequences or wild-type sequences. The diversity can only be accessed by sequencing (either single clones or pools of clones as in the QQC). The QQC provides a tool to asses if the statistical performed analysis of library characteristics by PeLiCa applies to the physically created library.

Colony numbers (not clonal diversity) is indeed one of the caps for overall physically created diversity, but this assumes that the diversity creation processes before were perfect or near perfect and without any greater flaws. Often the yield of the PCR and how much DNA was used to infect/transform is an earlier restriction as discussed by Denault and Pelletier 2007. Please reflect this in the introduction on page 2.

In the field of viral peptide libraries the number of random bacterial colonies observed in the library production process is often used as a measure for the diversity of the library (e.g. Michelfelder S, Kohlschutter J, Skorupa A, Pfennings S, Muller O, et al. (2009) Successful expansion but not complete restriction of tropism of adeno-associated virus by in vivo biopanning of random virus display peptide libraries. PLoS ONE 4: e5122.; Maersch S, Huber A, Büning H, Hallek M, Perabo L (2009) Optimization of stealth adeno-associated virus vectors by randomization of immunogenic epitopes. Virology: 167-175.).

We agree with reviewer 2 that the concept of a clonal diversity is only of limited value and that the number of clones is only a measure for the maximum possible diversity in the library which can only be achieved if all steps in the production performed optimally (See page 2 “This number [clonal diversity] is easily assessable, and represents the maximally achievable diversity for the phage/virus library…”).

To clarify this point we now omitted the use of the term “clonal diversity” in the manuscript.

We now altered our previous statement that “Particular precautions must be taken to avoid - or at least, to minimise - losses to diversity in subsequent steps of the library production [after counting of the bacterial clones] to make the clonal diversity a valid qualifier for the peptide library.” to “Particular precautions must be taken to avoid - or at least, to minimise - losses to diversity in all steps of the library production to make the number of bacterial colonies a valid qualifier for the peptide library (see also Denault and Pelletier 2007).”

We now explicitly recommend the use of the quick quality control (QQC) as suggested by reviewer 2, as a method to confirm that no greater flaws occurred in the library production (Page 2: “To confirm that the library does not suffer from greater flaws, a QQC (quick quality control) can be performed (Reference: C1; C2).”

29. TS: On page 3 the authors discuss the different used randomisation schemes. In 2012, two new schemes replacing NNN and NNK/S in commercially available, primer based methods, were introduced by Tang et al and Kille et al. and should be mentioned. Also various other techniques beside the cited Trimer paper exist for solid-phase synthesis as stated above, most importantly the so called MAX method, and its update ProxiMAX, by Hughes and Hines. Please cite accordingly.

The methods introduced by Tang et al and Kille et al as well as the MAX method are not suitable for the introduction of longer randomized peptides and therefore for peptide libraries. ProxiMAX by Hughes and Hines offers an alternative approach to trimer that is useable also in peptide libraries. Currently the trimer concept is more widely known in the field of peptide library production. Therefore we only stated this method in the manuscript, but now also discussed the requested additional schemes.

We added the following sentences on page 3:

“An alternative to the trimer approach to generate peptide libraries with a one codon per amino acid ratio is the “ProxiMAX” system (Reference: Ashraf et al. 2013).”

and

“Apart from the mentioned, a number of further encoding schemes exist. Special attention in saturation mutagenesis received the “MAX randomization” (Reference: Hughes et al. 2003), the “22c trick” (Reference: Kille et al. 2012) and the “small-intelligent libraries” (Reference: Tang et al. 2012). However, as these techniques are not suited to produce long stretches of randomized sequences (Reference: Ashraf et al. 2013) (e.g. 5 or more amino acid positions) they are not used for the production of peptide libraries.”

30. TS, EH: The authors wrote on page 4 “ …concepts of coverage and efficiency”, as well as „We *define* and discuss three measures of library quality: peptide diversity…expected coverage… . “ on page 5. They failed to give credit in this two chapters to Patrick and Firth 2003, who originally coined and introduced the terms *expected coverage and diversity*. *Relative efficiency* was computed by Volles et al. for NNK/S. Even though the authors’ web interface is now allowing a more extended and way easier use of this concept compared to the work of Volles et al, they forgot to give credit.

Patrick and Firth introduced in their publication of 2008 (Firth AE, Patrick WM (2008) GLUE-IT and PEDELAA: new programs for analyzing protein diversity in randomized libraries. Nucleic Acids Research 36: W281{285.) concepts for amino acid diversity and completeness of protein libraries. In their publication from 2003 (together with JM Blackburn) they used the same terms (diversity and completeness) to investigate libraries on DNA level. Furthermore, Makowski. and Soares (2003) use the term “technical diversity” to describe completeness/coverage on DNA level. To prevent confusion, we chose to use the terms peptide diversity and expected coverage in our manuscript. We did however give credit to Patrick and Firth on page 5 (“The “GLUE-IT" software (available at http://guinevere.otago.ac.nz/stats.html; Patrick and Firth 2008) generates values for diversity and coverage for protein libraries with up to six modified codons per protein.”) and page 17 (“The validity of our approach was successfully confirmed by direct comparison with \GLUE-IT" [27]. GLUE-IT determines protein diversity and coverage for small libraries”). For benefit of understandability we used here the term “coverage” over “completeness”.

To further emphasize their work, we now added this additional comment on Page 5: “The terms of diversity and completeness used by Patrick and Firth (Patrick and Firth 2008) for saturation mutagenesis experiments are equivalent to the concepts peptide diversity and expected coverage we use here for peptide libraries.”

Volles and Lansbury 2005, use a “screening efficiency” they defined as:” The efficiency of sequence space coverage as a function of number of sequences screened. The efficiency is the expected number of unique elements… which will be covered by screening one additional clone.” The relative efficiency we use in our manuscript is defined as “Relative efficiency is defined as the ratio of expected peptide diversity of a library relative to its overall number of oligonucleotides”. It is therefore distinct from the screening efficiency defined by Volles and Lansbury.

31. TS EH: Please tone down the sentence “our framework allows for the first time…” on page 4. Part of what has been claimed as novel was shown and discussed by several researchers, whose work was not cited, see above (\*).

Done

32. TS, EH: The authors wrote on page 5: „In this publication, we regard them as composed of peptides assembled from 19 distinct codons (one per amino acid; excluding cysteine). “ But actually later in their calculation and on their webpage treat the trimer randomisation scheme as containing 20 codons and have the Trimer-C for the 19 codon set. Please clarify.

We have adjusted the notation of all library schemes to correctly specify the basis we are using, and both the paper and the website are using the same notation.

“NNK/S” now means both in the paper and on the PeLiCa website an NNK/S scheme based on 20 aa, while NNK/S-C means 19aa.

“Trimer” is now renamed to 20/20 to reflect its property of using 20 codons for 20 amino acids rather than referring to its manufacturing process. ProxiMAX (see #29) is therefore also covered by the 20/20 library scheme.

33. EH, HHR: On page 6 towards the end, the authors missed to discuss their approach with the work of Pelletier (2007), who presented a way to calculate certain library statistics incooperating the different probabilities of highly- or low- redundant codons and Hughes et al 2003, who discussed the probabilities of rare and over abundant sequences in randomized libraries.   
  
XXX

34.TS: On page 10, the observation that the used encoding scheme (randomisation scheme) influences the coverage has also been observed and discussed extensively by Denault and Pelletier 2007, Reetz et all 2008, Nov 2012 and Kille et al 2013. Please include in your discussions.

done

35. TS, EH, HHR: Page 10: How can the percentage of the initial loss (arising from 2 codons) be as high as the percentage loss due to redundancy (10 superfluous/redundant codons e.g. in NNK)?

We agree that the comparison between redundancy in 20/20 libraries and the initial loss in NNK/S libraries is confusing and have removed this sentence from the manuscript.

36. TS, EH: Page 10: What is the difference between ideal (maximum) libraries or randomisation schemes and trimer libraries? XXX Please clarify (page 10). Why is MAX not selectable in the web interface?

  
MAX encoding is not well suited to build longer randomized peptides used in peptide libraries. ProxiMAX would be a possibility. For our calculations every method leading to a ratio of one nucleotide trimer/codon per amino acid will give the same result. Therefore, the scheme “20/20” can be used to analyze peptide libraries as well as saturation mutagenesis libraries generated with any current or future method that encodes one codon for each amino acid.

To emphasize this and to make PeLiCa more accessible to use for researchers working with saturation mutagenesis, we now explicitly point out this information at PeLiCa.

37. EH, HHR: ~~Page 10: Figure 1 is not reproducible with PeLiCa. Please update with the plots created by PeLiCa, which are much easier to understand.~~ Expected coverage was calculated by PeLiCa for a k=7, NNK, N= 1 000 000 000 and resulted in 0.33. Using the same characteristics and the equation for expected coverage (fractional completeness) provided by Patrick and Firth, coverage is calculated as 0.028. How can such significant differences be explained for the same concept?   
We double-checked the calculations and believe that for NNK, N= 1 000 000 000 and k=7 the result of 0.33 is, in fact, the correct number. However, for NNK, N= 1 000 000 000 and sequences of length k=8, coverage is calculated as 0.028. We hope that this clears up the difference in results.

Guided by relative efficiency it appeals (especially at the web interface) that increasing of screening effort (library size) is bad, since the value for relative efficiency drops. This easy to make misinterpretation, especially by unexperienced users, should be avoided by clearer explaining the interpretation of the values of coverage and relative efficiency.   
It is important to optimize conditions of library creation and having a high screening effort/library size striving for high expected coverage numbers (>45%) to ensure high probabilities of finding an improved variant. Höbenreich et al 2014.

We have tried to address this by including easier to interpret definitions and some examples of coverage and relative efficiency (see also #22).

To further stress the point we now state on page 11 of the manuscript: “Increasing library size always improves coverage until 100% coverage is reached. However the ratio between necessary effort and benefit always decreases as indicated by the relative efficiency.”

On PeLiCa we now state: “The relative efficiency is a quality marker for your library that reflects the relation between the number of distinct valid peptides in the library and the total number of sequences at the plasmid/transformant level (N). For example, a library of 100 sequences that encodes only 50 distinct valid sequences has a relative efficiency of 0.5. As the chance that a random peptide added to the library is novel decreases with increasing library size, the relative efficiency of the library also decreases. **Increasing library size always improves coverage until 100% coverage is reached. However the ratio between necessary effort and benefit always decreases as indicated by the relative efficiency.**”

38. TS: Page 12: the conclusion reached in the last chapter of “inclusion probabilities” was also reached by Huges and Vines. Please discuss this in the text.

done

39. TS, HHR: ~~What does the authors mean by “Further, the leveled distribution of encodings actually causes the inclusion probability for about 75% of all possible peptides to increase if trimer.”? Please clarify, what is meant with leveled distribution of encodings? Equal ratio of codons to aa without redundancy? Compared to which other randomisation schemes?~~   
‘leveled distribution’ should actually read ‘uniform distribution’, as in the uniform probabilities of peptides in a trimer library. We have completely rewritten the passage to clarify our conclusions.

40. TS, EH, HHR: The authors analyse first and second degree conservative neighbourhoods (peptide sequences deviated from the assumed one by one or two exchanges, HENNING -> HDNNING -> HDNNLNG. This might be sufficient for short peptides (2-4/6) but for longer once, also 3rd or 4 th degrees might play a significant role. It would be good if the authors expand this calculation and allow also 3rd and fourth (maybe even more) degree of neighbourhood calculations (also on the web platform) in those cases where it makes sense.

We agree, that it would make sense to compute neighborhoods of degrees higher than 2, but neighborhood calculations at the moment are computationally intense (of polynomial order, where the exponent is determined by the degree of the neighborhood), which makes a computation of even degree 3 neighbors prohibitive. We are working on the issue and hope to find a simplification that will allow us to compute higher degree neighborhoods. We have included a statement to that effect in the paper.

41. TS: In discussions on page 14+15, the authors need to tone down their claims and add more citations, see the mentioned missed citations. E.g. With ProxiMAX, the Tang trick and the 22codon trick available, removal of redundancy isn’t as expensive anymore as it once was.  
  
The discussion has been modified as requested and additional references have been incorporated. In our original manuscript we focused on literature which directly referred to the field of peptide libraries. As reviewer 2 pointed out, our manuscript and PeLiCa might also be of high interest for researchers in the area of protein evolution. We therefore included some information regarding this field as well.

**42. EH: Remarks web interface**

It is clear that the authors have developed their PeLiCa web interface to analyse the statistical aspects of peptide/protein libraries from the viewpoint of phage display. On the other hand, in their conclusion they hint that PeLiCa could be of use for other related areas, which also deal with peptide/protein libraries, such as directed evolution by saturation mutagenesis. This is indeed true, but their programmed web interface settings need modifications to reflect this, making PeLiCa even a more general tool with great potential and usefulness.

~~• The minimum physical library size (clones on plate) is set to 1 \* 10^6, whereas when using reduced codon sets and a small number of randomized codons, this number can drop significantly on paper (statistics) and physically (colonies on plate). Therefore, it would be good if a smaller minimum is implemented. Maybe with a warning that small numbers are risky in statistics.~~

~~• The web interface allows peptide length of 6-10, or, in other words, randomisation of 6 to 10 codons/aa simultaneously. The authors itself write on page 5 of the manuscript, that randomisation shemes in phage display are seen from “five up to twenty or more“. Therefore, and from the viewpoint to create a more universal web interface, I suggest implementing the option to broaden the „peptide length settings“ from 1 up to 12 (at least, as in Yuval Nov’s Toblib web interface) and maybe even 15-20.~~ The current state of the art is the removal of redundancy and the usage of reduced alphabets while increasing the number of randomized positions. An ongoing trend where the authors could provide, from the beginning, a suitable web interface that will not suffer from the limitation to be outdated rapidly as it was with GLUE-IT. It also would be good if the term „peptide length“ is extended by e.g. „aa positions randomized“, since also in phage display randomisation sites may not be mandatory consecutive. This will have no effect of the sequence analysis, since for the math it is irrelevant if the sequence has gabs or not.

~~• Library shemes: Usage of NNN is outdated since several years and also NNK/S randomisations have been show 2012 as overcome, the standard setting of the drop down menu is suggested to be “trimer” or at least NNK/S. In addition, providing a few more standard drop downs for smaller, non-~~redundant amino acid sets would be improving the versatility of PeLiCa significantly, e.g combinations of 12:12, 5:5 and 4:4. Or, at least, allow the user to save the prepared costume schemes temporarily in the drop down list or make them exportable/importable for reuse.

~~• Currently standard setting of the plotted graphs in the web interface is to display all schemes, even though only one was selected in the basic menu. This is irritating. If the user wishes to see the other schemes, he/her should actively add them by going to the advanced settings and not vice versa. Same for the peptide length settings.~~

A suggestion for improvement: Could the authors additionally provide in the neighbourhood panel and/or Summary panel of the web interface an answer to the question: “How many sequences are actually expected in the 1st or 2nd degree neighbourhood of e.g. HENNING, when a library of size N with scheme X is explored?

• The authors have created great potential with the web interface PeLiCa but missed to highlight it (also in the manuscript). The ability to create custom sets of randomization schemes coupled with the ability to calculate the fraction of a library with non-productive peptides (as in stop or C) holds great analysis power. For protein engineering it would be interesting if the introduced concept of trimer-C would be more generalized (trimer-X) and expanded (trimer-2X) and make selectable as a predefined custom set in the drop down menu. To use the custom settings to define such classes manually was complicated.

• Asking to display inclusion probabilities crashes the web application or the java, independent on the chosen schemes or peptide lengths. Also switching/clicking fast through the randomisation schemes in the advanced settings crashes the application while being on the „inclusion“ tab. Trials to reproduce the crash did not succeed on the following day.

We were trying to reproduce the crashes of PeLiCa, but were not successful. We suspect that the internet connection might have been interrupted at some point between the server and the client machine, which leads to a ‘grey-out’ of the site. Unfortunately, this is something outside of our control.

~~43. TS, EH: Please move “(PubMed query April 2014 on “peptide library“) to a endnote.~~

Done

44. ~~HHR, EH: Redesign figure 3+4, it is impossible to distinguish the lines. Maybe it would be clearer when having one box per randomisation scheme instead of having one box per peptide length.~~

45. ~~TS: The term~~ **~~multiplicity~~** ~~was introduced but the correct phrasing for the described phenomenon is redundancy (of the genetic code).~~

See #23

46. TS: The term **Trimer** was introduced as randomisation scheme. This term causes distraction, because it has not been used in the literature for the non-redundant encoding scheme of all 20 amino acids in 20 codons. Instead the literature has coined the termed MAX (from Hughes et al 2003). Also this is irritating and maybe a better term can be found in general, also for the web application that reflect the 20:20 setting.

“Trimer” is now renamed to 20/20 to reflect its property of using 20 codons for 20 amino acids rather than referring to its manufacturing process. ProxiMAX (see #29) is therefore also covered by the 20/20 library scheme, see also #30.

47. TS: ~~When defining the neighbourhood’s best and worst scenarios, please direct the reader to Table S4 for an example.~~

Done

48. ~~TS: The used abbreviation AAV is not explained~~.

Done

49. ~~EH: Please replace Ref 15, with this earlier one: Ono, A., Matsuda, A., Zhao, J., and Santi, D. V. (1995) The synthesis of blocked triplet-phosphoramidites and their use in mutagenesis. Nucleic Acids Res. 23, 4677−4682.~~

Done, thank you.

50. TS, EH, HHR: The concept of relative efficiency needs a more clear description. Easier to understand for non-mathematics. Volles provide a very good example: “in a NNS 10mer library, diversity predictions are superfluous; we know that essentially every additional clone examined will be unique.”

51. TS, EH: ~~The sentence on page 9 “Within this class, all peptides have the same number of oligonucleotides” would be clearer if written as “ … each amino acid is encoded by one codon only.”~~   
A class of amino acids is a set of amino acids that are all encoded by the same number of codons. This number is in general not equal to 1 – e.g. in NNN the amino acids S, L, and R make up on class. They are all encoded by 6 codons. Therefore any peptides of length k containing only S, L, and R are encoded by 6k codons.

52. TS, EH: Page 12: “Non-trimer” is the wrong scientific term, the correct is “non-redundant”.

We wanted to describe with the term “non-trimer”, encoding schemes that possess a bias for encoding certain amino acids due to unbalances in their encodings. We now clarified the sentence: “The high variability introduced by schemes with varying codons per amino acid ratios causes libraries to be biased towards peptides with a high number of possible encodings at the cost of rare ones.”

53. ~~TS, EH: What is “dfn 2”?~~

Dfn 2 was a reference to ‘Definition 2’ – we have changed it accordingly in the manuscript.

54. ~~TS, EH: Add in the headings for tables S5, S6, S7 and S8 that hexapeptides were analysed.~~

Done. There were a set of other changes to the tables, so that tables S5 and S6 are now combined in table S6 and the former tables S7 and S8 are now combined into table S7.