Electronic Supplementary Material (ESI) for Integrative Biology.


# Predicting essential genes in prokaryotic genomes using a linear method: ZUPLS

## Supplementary data

**Kai Song[§], Tuopeng Tong, Fang Wu**

School of Chemical Engineering and Technology, Tianjin University, Tianjin, 300072, China

[§]Corresponding author
E-mail address:
KS: ksong@tju.edu.cn

**Table S1 Details of the samples of the 12 prokaryotic genomes***

| | Prokaryotic genomes | Abbreviations | Ess (our study) | Ess (Ref.) |
|---|---|---|---|---|
| 1. | *Acinetobacter baylyi* | AB | 498 | 499 [1] |
| 2. | *Bacillus subtilis 168* | BS | 267 | 271 [2] |
| 3. | *Caulobacter crescentus* | CC | 480 | 480 [3] |
| 4. | *Escherichia coli*** | EC | 296 | 303 [4] |
| 5. | *Francisella novicida* | FN | 390 | 396 [5] |
| 6. | *Mycoplasma genitalium* | MG | 378 | 382 [6] |
| 7. | *Mycoplasma pulmonis* | MP | 309 | 310 [7] |
| 8. | *Pseudomonas aeruginosa UCBPP-PA14* | PA | 335 | 335 [8] |
| 9. | *Staphylococcus aureus N315* | SA315 | 302 | 150/146/122 [9] |
| 10. | *Staphylococcus aureus NCTC 8325* | SA8325 | 351 | 351 [10] |
| 11. | *Salmonella enterica serovar Typhi* | SE | 352 | 356 [11] |
| 12. | *Streptococcus sanguinis* | SS | 218 | 218 [12] |

* Ess (our study): the number of essential genes used in our study; Ess (Ref.): the number of essential genes given by the references. **There are three datasets of *E. coli* in the DEG 6.5 database, Gerdes set contains 609 essential genes [13], Baba set contains 296 essential genes [4] and the union set of the Gerdes set and the Baba set. Gerdes set was identified by transposon mutagenesis. Baba set was identified by the single gene knockout experiments. Gerdes set contains more false positives than Baba set does. Consequently, we only used the Baba set.

## The Z-curve methods

Z-curve is a powerful tool in visualizing and analysing DNA sequences [14]. For convenience, we briefly introduce Z-curve parameters.

***The Z-curve parameters for frequencies of phase-independent mononucleotides***. The frequencies of the bases A, C, G, and T occurring in a fragment of DNA sequence are denoted by $a$, $c$, $g$, and $t$, respectively. Based on the $Z$-curve method, $a$, $c$, $g$, and $t$ are mapped onto a point $\boldsymbol{P}$ in a three-dimensional space $\boldsymbol{V}$, which are denoted by $x$, $y$ and $z$ [14b].

$$\begin{cases} x = (a+g)-(c+t) \\ y = (a+c)-(g+t) \\ z = (a+t)-(c+g) \end{cases} \tag{1}$$

***The Z-curve parameters for frequencies of phase-specific mononucleotides***. The frequencies of bases A, C, G and T occurring in a fragment of DNA sequence at the 1st, 2nd and 3rd codon positions are denoted by $a_i$, $c_i$, $g_i$, $t_i$, $i$=1, 2, 3, respectively. Based on the Z-curve method, $a_i$, $c_i$, $g_i$ and $t_i$ are mapped onto a point $\boldsymbol{P}_i$ in a three-dimensional space $\boldsymbol{V}_i$, $i$=1, 2, 3, which are denoted by $x_i$, $y_i$, $z_i$ [14a].

$$\begin{cases} x_i = (a_i+g_i)-(c_i+t_i) \\ y_i = (a_i+c_i)-(g_i+t_i) \\ z_i = (a_i+t_i)-(c_i+g_i) \\ x_i, y_i, z_i \in [-1,+1], \quad i = 1,2,3 \end{cases} \tag{2}$$

***The Z-curve parameters for frequencies of phase-independent di-nucleotides***. Let the frequency of di-nucleotides $XY$ be denoted by $p(XY)$, where $X$, $Y$=A, C, G and T. Using the Z-curve of DNA sequences, we have

$$\begin{cases} x_X = [p(X\mathrm{A})+p(X\mathrm{G})]-[p(X\mathrm{C})+p(X\mathrm{T})] \\ y_X = [p(X\mathrm{A})+p(X\mathrm{C})]-[p(X\mathrm{G})+p(X\mathrm{T})] \\ z_X = [p(X\mathrm{A})+p(X\mathrm{T})]-[p(X\mathrm{C})+p(X\mathrm{G})] \\ X = \mathrm{A,C,G,T} \end{cases} \tag{3}$$

where $x_X$, $y_X$ and $z_X$ are the coordinates of a point in a three-dimensional space.

***The Z-curve parameters for frequencies of phase-specific di-nucleotides***. Using similar notations, we have

$$\begin{cases} x_X^k = [p^k(X\mathrm{A}) + p^k(X\mathrm{G})] - [p^k(X\mathrm{C}) + p^k(X\mathrm{T})] \\ y_X^k = [p^k(X\mathrm{A}) + p^k(X\mathrm{C})] - [p^k(X\mathrm{G}) + p^k(X\mathrm{T})] \\ z_X^k = [p^k(X\mathrm{A}) + p^k(X\mathrm{T})] - [p^k(X\mathrm{C}) + p^k(X\mathrm{G})] \\ X = \mathrm{A,C,G,T}, \quad k = 1,2,3 \end{cases} \tag{4}$$

where $k$=1,2,3 mean that the nucleotides are situated at the 1st, 2nd and third codon positions, respectively.

***The Z-curve parameters for frequencies of phase-independent tri-***nucleotides. Using similar notations, we have

$$\begin{cases} x_{XY} = [p(XY\mathrm{A}) + p(XY\mathrm{G})] - [p(XY\mathrm{C}) + p(XY\mathrm{T})] \\ y_{XY} = [p(XY\mathrm{A}) + p(XY\mathrm{C})] - [p(XY\mathrm{G}) + p(XY\mathrm{T})] \\ z_{XY} = [p(XY\mathrm{A}) + p(XY\mathrm{T})] - [p(XY\mathrm{C}) + p(XY\mathrm{G})] \\ X = \mathrm{A,C,G,T}, \quad Y = \mathrm{A,C,G,T}, \end{cases} \tag{5}$$

***The Z-curve parameters for frequencies of phase-specific tri-nucleotides***. Similarly, we have

$$\begin{cases} x_{XY}^k = [p^k(XY\mathrm{A}) + p^k(XY\mathrm{G})] - [p^k(XY\mathrm{C}) + p^k(XY\mathrm{T})] \\ y_{XY}^k = [p^k(XY\mathrm{A}) + p^k(XY\mathrm{C})] - [p^k(XY\mathrm{G}) + p^k(XY\mathrm{T})] \\ z_{XY}^k = [p^k(XY\mathrm{A}) + p^k(XY\mathrm{T})] - [p^k(XY\mathrm{C}) + p^k(XY\mathrm{G})] \\ X = \mathrm{A,C,G,T}, Y = \mathrm{A,C,G,T}, \quad k = 1,2,3 \end{cases} \tag{6}$$

A fragment of DNA sequence can be represented by a point in an $n$-dimensional space $\boldsymbol{V}$ by a selective combination of $n$ variables derived from the Z-curve method, where $n \in [1..252]$. Please refer to Gao and Zhang (2004) [14b] for more details.

**Table S2. Descriptions of the selected 93' *Z*-curve variables**

| Variables | Descriptions |
|---|---|
| $x_1, y_1, z_1, x_2, y_2, z_2, x_3, y_3, z_3,$ | Phase-specific parameters of mononucleotide |
| $x_{\mathrm{A}}^1, z_{\mathrm{A}}^1, x_{\mathrm{T}}^1, y_{\mathrm{T}}^1, z_{\mathrm{T}}^1, y_{\mathrm{C}}^1, z_{\mathrm{C}}^1, x_{\mathrm{A}}^2, z_{\mathrm{A}}^2, x_{\mathrm{T}}^2, z_{\mathrm{T}}^2, z_{\mathrm{C}}^2, x_{\mathrm{T}}^3, z_{\mathrm{T}}^3$ | Phase-specific parameters of di-nucleotides |
| $x_{\mathrm{AT}}^1, z_{\mathrm{AT}}^1, z_{\mathrm{AC}}^1, y_{\mathrm{AG}}^1, x_{\mathrm{TA}}^1, x_{\mathrm{TT}}^1, y_{\mathrm{TT}}^1, x_{\mathrm{TG}}^1, y_{\mathrm{TG}}^1, z_{\mathrm{TG}}^1, x_{\mathrm{CA}}^1, z_{\mathrm{CC}}^1, z_{\mathrm{GG}}^1,$ $z_{\mathrm{AA}}^2, z_{\mathrm{AT}}^2, z_{\mathrm{AG}}^2, x_{\mathrm{TA}}^2, x_{\mathrm{TT}}^2, z_{\mathrm{TT}}^2, x_{\mathrm{TG}}^2, y_{\mathrm{CA}}^2, x_{\mathrm{CT}}^2, x_{\mathrm{CC}}^2, z_{\mathrm{CC}}^2, y_{\mathrm{GG}}^2, z_{\mathrm{GG}}^2,$ $x_{\mathrm{TT}}^3, x_{\mathrm{TC}}^3, y_{\mathrm{TC}}^3, y_{\mathrm{TG}}^3, z_{\mathrm{TG}}^3, x_{\mathrm{CT}}^3, y_{\mathrm{CT}}^3, x_{\mathrm{CC}}^3, y_{\mathrm{CC}}^3, x_{\mathrm{GT}}^3$ | Phase-specific parameters of tri-nucleotides |
| *x, y, z* | Phase-independent parameters of mononucleotide |
| $x_{\mathrm{A}}, y_{\mathrm{A}}, z_{\mathrm{A}}, x_{\mathrm{T}}, y_{\mathrm{T}}, z_{\mathrm{T}}, x_{\mathrm{C}}, y_{\mathrm{C}}, y_{\mathrm{G}}$ | Phase-independent parameters of di-nucleotide |
| $x_{\mathrm{AT}}, z_{\mathrm{AT}}, y_{\mathrm{AG}}, x_{\mathrm{TA}}, z_{\mathrm{TA}}, x_{\mathrm{TT}}, y_{\mathrm{TT}}, z_{\mathrm{TT}}, x_{\mathrm{CA}}, x_{\mathrm{CT}}, z_{\mathrm{CT}}, x_{\mathrm{CC}}, z_{\mathrm{CC}}, x_{\mathrm{CG}}, y_{\mathrm{CG}},$ $z_{\mathrm{GA}}, x_{\mathrm{GT}}, x_{\mathrm{GC}}, y_{\mathrm{GC}}, z_{\mathrm{GC}}, y_{\mathrm{GG}}, z_{\mathrm{GG}}$ | Phase-independent parameters of tri-nucleotide |

## *Orthologs*

Previous studies have proven that essential genes tend to be evolutionarily more conserved than non-essential genes in bacterial species [15]. This is because essential genes are more likely involved in basic cellular processes, and the negative selection acting on essential genes are more stringent than for non-essentials. Thus the number of orthologs has been found to be a valuable predictive feature for gene essentiality [16].

We introduced a reciprocal best hit (RBH) [17] method to identify the orthologs between training and target genome. For example, in the case of calculating the number of orthologs between EC and PA: we first queried an $ORF_i$ in PA against all known ORFs in EC by Blastp to yield the set of hits *W*. The corresponding E-value threshold was $10^{-5}$. Then, we queried the hit with the lowest E-value in *W* ($ORF_j$) against all ORFs in PA to yield the set of hits *Y*. A pair of proteins ($ORF_i$, $ORF_j$) was considered putative orthologs if $ORF_i$ was the hit in *Y* with the lowest E-value, and if they also met two strict criteria:

(i) $Length_{ShorterProtein}=Length_{LongerProtein} \geqq 80\%$;

(ii) $Length_{AlignedRegion}=Length_{ShorterProtein} \geqq 80\%$;

to ensure sufficient coverage of aligned regions.

We used 17 genomes given by DEG6.5 database and other 180 genomes [15c] to calculate the orthologs. 185 genomes are finally selected because the overlapped genome between DEG 6.5 dataset and the 180 dataset.

For example, in the case of EC→PA, there are 183 genomes left except for EC and PA. So there are 183 orthologs features for EC, noted as ECorth. Similarly, there are 183 orthologs features for PA, noted as PAorth. We also used the standard deviations and mean values of them as the features.

## *Other sequence-based features*

✧ *Gene size*

There is a trend for proteins to become larger throughout evolution. Gene size is therefore expected to be indicative of essentiality [15c]. We used the number of nucleotides of the gene sequences as the "Gene size" feature.

✧ *Strand bias*

Essential genes are more likely to be encoded on the leading strand of the circular chromosome [18]. If the gene is on the leading strand, the corresponding *strand bias* feature is '1', or else is '0'.

✧ *Codon Adaptation Index*

Codon Adaptation Index (CAI) is a measurement of the relative adaptiveness of the codon usage of a gene towards the codon usage of highly expressed genes [19]. We used the CodonW software (http://codonw.sourceforge.net) to calculate the CAI values.

✧ *Frequency of optimal codons*

Frequency of optimal codons (Fop) is the ratio of optimal codons to synonymous codons (genetic code dependent) [15c]. We also used the CodonW software (http://codonw.sourceforge.net) to calculate the Fop values.

✧ *Frequency of all encoded amino acids*

Lin et al. found that rather than all essential genes, only those with the COG functional category of information storage and process (J, K and L), and subcategories D (cell cycle control), M (cell wall biogenesis), O (posttranslational modification), C (energy production and conversion), G (carbohydrate transport and metabolism), E (amino acid transport and metabolism) and F (nucleotide transport and metabolism) were preferentially situated at the leading strand [20]. Therefore, we used the frequency of encoded amino acids as features.

✧ *Close_stop_ratio*

We also calculated Close_stop_ratio, the number of codons that are one third-base mutation removed from a stop codon. There are five such codons: TAC and TAT encoding Tyr are one third-base mutation away from TAA and TAG; TGC and TGT encoding Cys, and TGG encoding Trp are one third-base mutation away from TGA [21]. Such codons were counted and the ratio can

be calculated as follows: if an ORF contained 200 codons, and the translated sequence contains 3 Tyr, 2 Cys and 1 Trp. Then, the total ratio of the codons that close to the stop codon would be (3+2+1=6)/200=0.03.

- *Paralogs*

Genes that have duplicates are more likely to be unessential [15b]. They are often referred to as paralogs which typically have a similar function. In our study, paralogs were defined as the number of those genes which were present in the same genome. An all-against-all FASTA search was conducted for the proteins coded in an organism to identify the putative paralogs[15c, 17]. The E-value threshold is $10^{-5}$.

- *DES (Domain enrichment score)*

Domain enrichment reflects the conservation of local sequences rather than the entire gene [17]. We introduced the DES feature proposed by Deng et al. 2011[17] as one of the features to predict essential genes.

## Partial least squares classifier

Partial least squares (PLS) algorithm is a key technique for modelling linear relationships between a set of known class-labels and a set of input variables. In PLS model, it is assumed that the investigated pattern is influenced by just a few underlying variables, which are called Latent Variables (LVs). Thus the original data space is projected to a much lower LV space to eliminate the interference of noise and the missing data. The multi-colinearity among the original data is also excluded by the orthogonality among the LVs. Interestingly, the LV assumptions closely correspond to the use of microscopic concepts such as molecules and reactions in chemistry and molecular biology, thus making PLS philosophically suitable for the pattern recognition (PR) of chemical and biological data [22].

**The methodology and procedure of PLS**

In the first step, PLS creates uncorrelated latent variables (LVs) which are linear combinations of the original input variables. The basic point of the procedure is that the weights used to determine the linear combinations of the original variables are proportional to the maximum covariance among input and output variables [23]. Then, the original data space is projected to the much lower LV-space to eliminate the interference of noise, missing data and multi-collinearity. This leads to a biased but lower variance estimate of the regression coefficients compared to the least squares method [23].

Direction in plane defining best correction with $\boldsymbol{L}$

The PLS recognition model can be written in matrix form as:

Given two data matrices $\boldsymbol{X}\in\boldsymbol{R}^{m\times n}$ and $\boldsymbol{L}\in\boldsymbol{R}^{m\times l}$ and assuming they are linearly related by:

$$\boldsymbol{L}=\boldsymbol{XB}+\boldsymbol{F} \tag{7}$$

where $\boldsymbol{B}\in\boldsymbol{R}^{n\times l}$ and $\boldsymbol{F}\in\boldsymbol{R}^{m\times l}$ are coefficient and noise matrices, respectively.

For example, the two-class PLS recognition model can be written as

$$\boldsymbol{l}=\mathrm{sgn}(\boldsymbol{XB}+\boldsymbol{F}) \tag{8}$$

sgn(*) is the Signum function. For each element of *, sgn(*) returns 1 if the element is greater than zero, 0 if it equals zero and -1 if it is less than zero.

The PLS algorithm builds a linear model by decomposing matrices $\boldsymbol{X}$ and $\boldsymbol{L}$ into bilinear terms:

$$\boldsymbol{X}=\hat{\boldsymbol{X}}+\tilde{\boldsymbol{X}}=\sum_{i=1}^{v}\boldsymbol{t}_i\boldsymbol{p}_i^T+\tilde{\boldsymbol{X}}=\boldsymbol{TP}^T+\tilde{\boldsymbol{X}} \tag{9}$$

$$\boldsymbol{L}=\hat{\boldsymbol{L}}+\tilde{\boldsymbol{L}}=\sum_{i=1}^{v}\boldsymbol{u}_i\boldsymbol{q}_i^T+\tilde{\boldsymbol{L}}=\boldsymbol{UQ}^T+\tilde{\boldsymbol{L}} \tag{10}$$

where $\hat{\boldsymbol{X}}$ and $\hat{\boldsymbol{L}}$ are the modeling matrices of $\boldsymbol{X}$ and $\boldsymbol{L}$, respectively; $\tilde{\boldsymbol{X}}$, $\tilde{\boldsymbol{L}}$ are corresponding model error matrices; $\boldsymbol{t}_i$ and $\boldsymbol{u}_i$ are the $i$th LV score vectors; and $\boldsymbol{p}_i$ and $\boldsymbol{q}_i$ are their loading vectors. The deflation of $\boldsymbol{L}$ is not always necessary.

The above two equations formulate a PLS outer model. A least squares regression is then performed on the subset of extracted orthogonal latent variables (LVs). Latent vectors are then related by a linear inner model:

$$\boldsymbol{u}_i=c_i\boldsymbol{t}_i+\boldsymbol{r}_i \qquad (i=1,\ldots,v) \tag{11}$$

where $c_i$ is an inner coefficient determined by minimizing the residual $\boldsymbol{r}_i$.

After performing the first LV calculation, the second LV is calculated by decomposing the residuals $\tilde{\boldsymbol{X}}$ and $\tilde{\boldsymbol{L}}$ using the same procedure as for the first LV. The total number of LVs, $v$, required in the model is usually determined by cross-validation, although elsewhere an F-test is suggested [24].

The regression coefficients matrix/vector $\boldsymbol{B}$ has the form

$$\boldsymbol{B}=\boldsymbol{W}(\boldsymbol{P}^{\mathrm{T}}\boldsymbol{W})^{-1}\boldsymbol{Q}^{\mathrm{T}} \tag{12}$$

where $\boldsymbol{P}$ ($n\times v$) and $\boldsymbol{Q}$ ($1\times v$) are the matrices/vectors consisting of loading vectors $\boldsymbol{p}_i$ and $\boldsymbol{q}_i$ ($i=1,\ldots,v$), and $\boldsymbol{W}$ is the ($n\times v$) weighting matrix consisting of weighting vectors $\boldsymbol{w}_i$ ($i=1,\ldots,v$).

There exist several different modifications of the basic algorithm for PLS regression originally developed by Wold [25]. In its basic form, a special case of the nonlinear iterative partial least squares (NIPALS) algorithm is used [26]. NIPALS is a robust procedure for solving singular value decomposition problems and is closely related to the power method.

The classical NIPALS-PLS algorithm training procedure is:

1. Scale the data $\{\boldsymbol{X}, \boldsymbol{L}\}$ to zero mean and unit variance, or as otherwise specified. $\boldsymbol{E}_0=$

$\boldsymbol{X}$, $\boldsymbol{F}_0 = \boldsymbol{L}$;

2. Randomly initialize $\boldsymbol{u}_i$ (in two-class problems, always set $\boldsymbol{u}=\boldsymbol{l}$, $\boldsymbol{l}$ is the label vector)
3. Build the outer model of PLS
   Repeat the following steps until $\boldsymbol{t}_i$ convergence
   $\boldsymbol{w}_i = \boldsymbol{E}_{i-1}^{\mathrm{T}}\boldsymbol{u}_i / \boldsymbol{u}_i^{\mathrm{T}}\boldsymbol{u}_i$
   $\boldsymbol{t}_i = \boldsymbol{E}_{i-1}\boldsymbol{w}_i / \left\|\boldsymbol{E}_{i-1}\boldsymbol{w}_i\right\|$
   $\boldsymbol{q}_i = \boldsymbol{F}_{i-1}^{\mathrm{T}}\boldsymbol{t}_i / \left\|\boldsymbol{F}_{i-1}^{'}\boldsymbol{t}_i\right\|$
   $\boldsymbol{u}_i = \boldsymbol{F}_{i-1}\boldsymbol{q}_i$
4. Calculate the loading vectors of $\boldsymbol{X}$: $\boldsymbol{p}_i = \boldsymbol{E}_{i-1}^{\mathrm{T}}\boldsymbol{t}_i / \boldsymbol{t}_i^{\mathrm{T}}\boldsymbol{t}_i = \boldsymbol{E}_{i-1}^{\mathrm{T}}\boldsymbol{t}_i$
5. Build the inner model of PLS: $c_i = \boldsymbol{u}_i^{\mathrm{T}}\boldsymbol{t}_i / \boldsymbol{t}_i^{\mathrm{T}}\boldsymbol{t}_i = \boldsymbol{u}_i^{\mathrm{T}}\boldsymbol{t}_i$
6. Deflate $\boldsymbol{E}$, $\boldsymbol{F}$ matrices: $\boldsymbol{E}_i = \boldsymbol{E}_{i-1} - \boldsymbol{t}_i\boldsymbol{p}_i^{\mathrm{T}}$ (the deflation of $\boldsymbol{F}$ is not always necessary)
   $\boldsymbol{F}_i = \boldsymbol{F}_{i-1} - \boldsymbol{b}_i\boldsymbol{t}_i\boldsymbol{q}_i^{\mathrm{T}}$
7. Repeat steps 3-6 untill given number of LVs have been extracted or cross-validation indicates that there is no more significant information in $\boldsymbol{X}$ about $\boldsymbol{L}$.

where $(*)^{\mathrm{T}}$ is the transpose of *; ||*|| is the norm of *; $\boldsymbol{t}_i$ and $\boldsymbol{u}_i$ are the $i$th LV score vectors; and $\boldsymbol{p}_i$ and $\boldsymbol{q}_i$ ($i$=1,…,$v$) are the corresponding loading vectors.

## References


1. V. de Berardinis, D. Vallenet, V. Castelli, M. Besnard, A. Pinet, C. Cruaud, S. Samair, C. Lechaplais, G. Gyapay, C. Richez, M. Durot, A. Kreimeyer, F. Le Fevre, V. Schachter, V. Pezo, V. Doring, C. Scarpelli, C. Medigue, G. N. Cohen, P. Marliere, M. Salanoubat, J. Weissenbach, A complete collection of single-gene deletion mutants of Acinetobacter baylyi ADP1. *Molecular systems biology* 2008, *4*. 174, DOI: 10.1038/msb.2008.10.
2. K. Kobayashi, S. D. Ehrlich, A. Albertini, G. Amati, K. K. Andersen, M. Arnaud, K. Asai, S. Ashikaga, S. Aymerich, P. Bessieres, F. Boland, S. C. Brignell, S. Bron, K. Bunai, J. Chapuis, L. C. Christiansen, A. Danchin, M. Débarbouillé, E. Dervyn, E. Deuerling, K. Devine, S. K. Devine, O. Dreesen, J. Errington, S. Fillinger, S. J. Foster, Y. Fujita, A. Galizzi, R. Gardan, C. Eschevins, T. Fukushima, K. Haga, C. R. Harwood, M. Hecker, D. Hosoya, M. F. Hullo, H. Kakeshita, D. Karamata, Y. Kasahara, F. Kawamura, K. Koga, P. Koski, R. Kuwana, D. Imamura, M. Ishimaru, S. Ishikawa, I. Ishio, D. Le Coq, A. Masson, C. Mauël, R. Meima, R. P. Mellado, A. Moir, S. Moriya, E. Nagakawa, H. Nanamiya, S. Nakai, P. Nygaard, M. Ogura, T. Ohanan, M. O'Reilly, M. O'Rourke, Z. Pragai, H. M. Pooley, G. Rapoport, J. P. Rawlins, L. A. Rivas, C. Rivolta, A. Sadaie, Y. Sadaie, M. Sarvas, T. Sato, H. H. Saxild, E. Scanlan, W. Schumann, J. F. M. L. Seegers, J. Sekiguchi, A. Sekowska, S. J. Séror, M. Simon, P. Stragier, R. Studer, H. Takamatsu, T. Tanaka, M. Takeuchi, H. B. Thomaides, V. Vagner, J. M. van Dijl, K. Watabe, A. Wipat, H. Yamamoto, M. Yamamoto, Y. Yamamoto, K. Yamane, K. Yata, K. Yoshida, H. Yoshikawa, U. Zuber, N. Ogasawara, Essential Bacillus subtilis genes. *Proceedings of the National Academy of Sciences* 2003, *100*. 4678-4683, DOI: 10.1073/pnas.0730515100.
3. B. Christen, E. Abeliuk, J. M. Collier, V. S. Kalogeraki, B. Passarelli, J. A. Coller, M. J. Fero, H. H. McAdams, L. Shapiro, The essential genome of a bacterium. *Molecular systems biology* 2011, *7*. 528, DOI: 10.1038/msb.2011.58.
4. T. Baba, T. Ara, M. Hasegawa, Y. Takai, Y. Okumura, M. Baba, K. A. Datsenko, M. Tomita, B. L. Wanner, H. Mori, Construction of Escherichia coli K-12 in-frame, single-gene knockout mutants: the Keio collection. *Molecular systems biology* 2006, *2*. 2006 0008, DOI: 10.1038/msb4100050.
5. L. A. Gallagher, E. Ramage, M. A. Jacobs, R. Kaul, M. Brittnacher, C. Manoil, A comprehensive transposon mutant library of Francisella novicida, a bioweapon surrogate. *Proceedings of the National Academy of Sciences of the United States of America* 2007, *104*. 1009-14, DOI: 10.1073/pnas.0606713104.
6. J. I. Glass, N. Assad-Garcia, N. Alperovich, S. Yooseph, M. R. Lewis, M. Maruf, C. A. Hutchison, 3rd, H. O. Smith, J. C. Venter, Essential genes of a minimal bacterium. *Proc Natl Acad Sci U S A* 2006, *103*. 425-30, DOI: 10.1073/pnas.0510013103.
7. C. T. French, P. Lao, A. E. Loraine, B. T. Matthews, H. Yu, K. Dybvig, Large-scale transposon mutagenesis of Mycoplasma pulmonis. *Molecular microbiology* 2008, *69*. 67-76, DOI: 10.1111/j.1365-2958.2008.06262.x.
8. N. T. Liberati, J. M. Urbach, S. Miyata, D. G. Lee, E. Drenkard, G. Wu, J. Villanueva, T. Wei, F. M. Ausubel, An ordered, nonredundant library of Pseudomonas aeruginosa strain PA14 transposon insertion mutants. *Proceedings of the National Academy of Sciences of the United States of America* 2006, *103*. 2833-8, DOI: 10.1073/pnas.0511100103.
9. (a) Y. Ji, B. Zhang, S. F. Van, Horn, P. Warren, G. Woodnutt, M. K. Burnham, M. Rosenberg, Identification of critical staphylococcal genes using conditional phenotypes generated by antisense RNA. *Science* 2001, *293*. 2266-9, DOI: 10.1126/science.1063566; (b) R. A. Forsyth, R. J. Haselbeck, K. L. Ohlsen, R. T. Yamamoto, H. Xu, J. D. Trawick, D. Wall, L. Wang, V. Brown-Driver, J. M. Froelich, K. G. C, P. King, M. McCarthy, C. Malone, B. Misiner, D. Robbins, Z. Tan, Z. Y. Zhu Zy, G. Carr, D. A. Mosca, C. Zamudio, J. G. Foulkes, J. W. Zyskind, A genome-wide strategy for the identification of essential genes in Staphylococcus aureus. *Molecular microbiology* 2002, *43*. 1387-400; (c) S. O. O. Ko Kwan, L. E. E. Ji Young, S. Jae-Hoon, B. Jin Yang, O. H. Won Sup, C. Jongsik, Y. Ha Sik, Screening of essential genes in Staphylococcus aureus N315 using comparative genomics and allelic replacement mutagenesis. *Journal of microbiology and biotechnology* 2006, *16*. 623-632.
10. R. R. Chaudhuri, A. G. Allen, P. J. Owen, G. Shalom, K. Stone, M. Harrison, T. A. Burgis, M. Lockyer, J. Garcia-Lara, S. J. Foster, S. J. Pleasance, S. E. Peters, D. J. Maskell, I. G. Charles, Comprehensive identification of essential Staphylococcus aureus genes using Transposon-Mediated Differential Hybridisation (TMDH). *BMC Genomics* 2009, *10*. 291, DOI: 10.1186/1471-2164-10-291.
11. G. C. Langridge, M. D. Phan, D. J. Turner, T. T. Perkins, L. Parts, J. Haase, I. Charles, D. J. Maskell, S. E. Peters, G. Dougan, J. Wain, J. Parkhill, A. K. Turner, Simultaneous assay of every Salmonella Typhi gene using one million transposon mutants. *Genome Res* 2009, *19*. 2308-16, DOI: 10.1101/gr.097097.109.
12. P. Xu, X. Ge, L. Chen, X. Wang, Y. Dou, J. Z. Xu, J. R. Patel, V. Stone, M. Trinh, K. Evans, T. Kitten, D. Bonchev, G. A. Buck, Genome-wide essential gene identification in Streptococcus sanguinis.

*Sci Rep* 2011, *1*. 125, DOI: 10.1038/srep00125.
13. S. Y. Gerdes, M. D. Scholle, J. W. Campbell, G. Balazsi, E. Ravasz, M. D. Daugherty, A. L. Somera, N. C. Kyrpides, I. Anderson, M. S. Gelfand, A. Bhattacharya, V. Kapatral, M. D'Souza, M. V. Baev, Y. Grechkin, F. Mseeh, M. Y. Fonstein, R. Overbeek, A. L. Barabasi, Z. N. Oltvai, A. L. Osterman, Experimental determination and system level analysis of essential genes in Escherichia coli MG1655. *Journal of bacteriology* 2003, *185*. 5673-84.
14. (a) C. T. Zhang, R. Zhang, Analysis of distribution of bases in the coding sequences by a diagrammatic technique. *Nucleic acids research* 1991, *19*. 6313-7; (b) F. Gao, C. T. Zhang, Comparison of various algorithms for recognizing short coding sequences of human genes. *Bioinformatics (Oxford, England)* 2004, *20*. 673-81, DOI: 10.1093/bioinformatics/btg467; (c) C. T. Zhang, A symmetrical theory of DNA sequences and its applications. *Journal of theoretical biology* 1997, *187*. 297-306, DOI: 10.1006/jtbi.1997.0401.
15. (a) I. K. Jordan, I. B. Rogozin, Y. I. Wolf, E. V. Koonin, Essential Genes Are More Evolutionarily Conserved Than Are Nonessential Genes in Bacteria. *Genome research* 2002, *12*. 962-968, DOI: 10.1101/gr.87702; (b) G. Giaever, A. M. Chu, L. Ni, C. Connelly, L. Riles, S. Veronneau, S. Dow, A. Lucau-Danila, K. Anderson, B. Andre, A. P. Arkin, A. Astromoff, M. El-Bakkoury, R. Bangham, R. Benito, S. Brachat, S. Campanaro, M. Curtiss, K. Davis, A. Deutschbauer, K. D. Entian, P. Flaherty, F. Foury, D. J. Garfinkel, M. Gerstein, D. Gotte, U. Guldener, J. H. Hegemann, S. Hempel, Z. Herman, D. F. Jaramillo, D. E. Kelly, S. L. Kelly, P. Kotter, D. LaBonte, D. C. Lamb, N. Lan, H. Liang, H. Liao, L. Liu, C. Luo, M. Lussier, R. Mao, P. Menard, S. L. Ooi, J. L. Revuelta, C. J. Roberts, M. Rose, P. Ross-Macdonald, B. Scherens, G. Schimmack, B. Shafer, D. D. Shoemaker, S. Sookhai-Mahadeo, R. K. Storms, J. N. Strathern, G. Valle, M. Voet, G. Volckaert, C. Y. Wang, T. R. Ward, J. Wilhelmy, E. A. Winzeler, Y. Yang, G. Yen, E. Youngman, K. Yu, H. Bussey, J. D. Boeke, M. Snyder, P. Philippsen, R. W. Davis, M. Johnston, Functional profiling of the Saccharomyces cerevisiae genome. *Nature* 2002, *418*. 387-91, DOI: 10.1038/nature00935; (c) A. Gustafson, E. Snitkin, S. Parker, C. DeLisi, S. Kasif, Towards the identification of essential genes using targeted genome sequencing and comparative analysis. *BMC Genomics* 2006, *7*. 265.
16. Y. Chen, D. Xu, Understanding protein dispensability through machine-learning analysis of high-throughput data. *Bioinformatics* 2005, *21*. 575-581, DOI: 10.1093/bioinformatics/bti058.
17. J. Deng, L. Deng, S. Su, M. Zhang, X. Lin, L. Wei, A. A. Minai, D. J. Hassett, L. J. Lu, Investigating the predictability of essential genes across distantly related organisms using an integrative approach. *Nucleic Acids Res* 2011, *39*. 795-807, DOI: 10.1093/nar/gkq784.
18. (a) E. P. C. Rocha, A. Danchin, Essentiality, not expressiveness, drives gene-strand bias in bacteria. *Nat Genet* 2003, *34*. 377-378, DOI: http://www.nature.com/ng/journal/v34/n4/suppinfo/ng1209_S1.html; (b) Y. Lin, R. R. Zhang, Putative essential and core-essential genes in Mycoplasma genomes. *Scientific reports* 2011, *1*. 53, DOI: 10.1038/srep00053.
19. P. M. Sharp, W. H. Li, The codon Adaptation Index--a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Res* 1987, *15*. 1281-95.
20. Y. Lin, F. Gao, C. T. Zhang, Functionality of essential genes drives gene strand-bias in bacterial genomes. *Biochemical and biophysical research communications* 2010, *396*. 472-6, DOI: 10.1016/j.bbrc.2010.04.119.
21. (a) Y. C. Hwang, C. C. Lin, J. Y. Chang, H. Mori, H. F. Juan, H. C. Huang, Predicting essential genes based on network and sequence analysis. *Mol Biosyst* 2009, *5*. 1672-8, DOI: 10.1039/B900611G; (b) M. Seringhaus, A. Paccanaro, A. Borneman, M. Snyder, M. Gerstein, Predicting essential genes in fungal genomes. *Genome Res* 2006, *16*. 1126-35, DOI: 10.1101/gr.5144106.
22. (a) A. J. Burnham, J. F. MacGregor, R. Viveros, Latent variable multivariate regression modeling. *Chemometrics and Intelligent Laboratory Systems* 1999, *48*. 167-180, DOI: 10.1016/s0169-7439(99)00018-0; (b) O. M. Kvalheim, The latent variable. *Chemometrics and Intelligent Laboratory Systems* 1992, *14*. 1-3, DOI: 10.1016/0169-7439(92)80088-l.
23. R. Rosipal, L. J. Trejo, Kernel partial least squares regression in Reproducing Kernel Hilbert Space. *J Mach Learn Res* 2002, *2*. 97-123.
24. (a) P. Geladi, B. R. Kowalski, Partial least-squares regression: a tutorial. *Analytica Chimica Acta* 1986, *185*. 1-17, DOI: 10.1016/0003-2670(86)80028-9; (b) A. Höskuldsson, PLS regression methods. *Journal of Chemometrics* 1988, *2*. 211-228, DOI: 10.1002/cem.1180020306.
25. H. Wold, Causal flows with latent variables : Partings of the ways in the light of NIPALS modelling. *European Economic Review* 1974, *5*. 67-86, DOI: 10.1016/0014-2921(74)90008-7.
26. (a) F. Lindgren, P. Geladi, S. Wold, The kernel algorithm for PLS. *J Chemometr* 1993, *7*. 45-59, DOI: 10.1002/cem.1180070104; (b) S. Rännar, F. Lindgren, P. Geladi, S. Wold, A PLS kernel algorithm for data sets with many variables and fewer objects. Part 1: Theory and algorithm. *J*

*Chemometr* 1994, *8*. 111-125, DOI: 10.1002/cem.1180080204.