Bioinformatics - Final Assignment

Paul Rognon - Class 2019-2020

10/01/2020

Contents

1 Introduction and objectives

In this document, we replicate the work done by DeJesus et Iorger in [1]. They propose a Hidden Markov Model (HMM) for identifying essential regions in bacterial genome from sequencing data obtained by transposon insertion. They evaluate the performance of the model on a sequence dataset of the H37Rv strain of *M.Tuberculosis* (Mtb) transposon mutants constructed by Griffin et al. [2].

2 Background

Transposon mutagenesis is an experimental method in which genes are transferred to a host organism's chromosome, interrupting or modifying the function of an existent gene and causing mutation. The method relies on the ability of transposons, semi-parasitic DNA sequences, to replicate and spread through the host's genome. For example, in our case, Griffin et al. generated 10^5 independent insertion events in the H37Rv genome composed of one chromosome, using a modified Himar1 based transposon. The Himar1 transposon inserts randomly into TA dinucleotide sites. The resulting modified genome is called the library of transposon mutants. Griffin et al. then grew replicates of this library for 12 generations. The authors then used Illumina deep sequencing to obtain the sequence of transposon insertions mutants. Deep sequencing consists in sequencing a genomic region multiple times and it enabled the authors of [2] to map in the genome of H37Rv the TA sites that withstood transposon insertions and those that did not.

Transposon mutagenesis is useful to find essential regions in genome because the mutations on genes essential for the organism growth can prove to be lethal. Therefore, mutant genes that are required for growth or survival should be absent or significantly under-represented in the grown population. On the contrary, mutant genes that survived after insertion are very probably non-essential. The absence or presence of an insertion in a site and the number of reads in the sequencing are then good indicators of the essentiality of region.

On one hand, some methods are based on the presence or absence of insertion. They identify essentiality from the probability that a gene lacking insertions is essential that can be modelled with a Binomial, negative-Binomial or Extreme Value distribution. However, those methods are highly sensitive to spurious reads, such as isolated reads that translate to spurious non-essential region. On the other hand, methods are based on read counts as it reflects the abundance of certain clones in the library and hence the degree to which a region of the genome is essential. Those methods are susceptible to spikes in the data, where there is a massive over-representation of reads at an isolated site, that greatly influence statistics. Both types of approach then lack flexibility and robustness.

The DeJesus et Iorger design a HMM that incorporates information from read counts at individual TA sites. The proposed HMM performs a smoothing that can accommodate such abrupt changes that disrupt afore mentioned methods. Indeed, it is a sequence-dependent model as it predicts the essentiality from the conditional probability of a state conditioned on the previous neighbouring site. By coupling neighbouring sites together, it is able to disambiguate the interpretation of each site. TA sites with no insertion in non-essential regions (e.g. because the insertion process missed the site during the construction of the library) are tolerated because neighbouring sites have insertions.

3 Data

As explained before the data used in this work is a sequence dataset of **M.Tuberculosis** (Mtb) transposon mutants constructed by Griffin et al. [2]. We obtained the data from the project home page referenced in the original article [1]. The dataset is a WIG file composed of two columns: the column "variableStep" indicating the location (TA site) in the genome and the variable "count" that gives the number of read counts at that TA site. DeJesus et Iorger mention 74,605 TA sites in [1] but the provided dataset has only 73,385 sites. Table 1 contains a sample of the data.

Table 1: Sample of data

variableStep	count
3729467	171
668418	186
4336054	0
982712	0
2600989	111
4129126	386
1656462	262
2148332	324
3571573	0
1455885	0

4 HMM model

In addition to the two obvious states, essential (ES) and non-essential (NE). DeJesus et Iorger define two extra states:

- growth-defect (GD), for regions that are not essential but whose disruption leads to impaired growth of the organism, they have low number of read counts.
- growth-advantage (GA), regions that are not essential but also could have a metabolic cost such that their disruption is advantageous for growth in vitro. They have a very high number of read counts.

According to the authors, the addition of these two states to the HMM allows it to distinguish regions in the sequence data with depressed or unusually high read counts in a statistically rigorous way.

4.1 Emission probabilities

DeJesus et Iorger first define the emission probabilities. They chose a geometric distribution to model the read counts with each state having a different parameter θ. The geometric distribution is a discrete distribution such that:

$$P(C_i = c_i|l) = (1-\theta_l)^{c_i}\theta_l$$

where c_i is the count and l is the state in TA site i. θ takes values between 0 and 1, the larger θ the higher the probability of a 0 count. θ can be understood as the Bernoulli probability of an insertion. The maximum likelihood estimator for θ is $\frac{1}{\bar{c}}$ where $\bar{c}$ is the average read count at non-empty read counts. Therefore, for the non-essential state θ is set at $\frac{1}{\bar{c}}$. For the essential state, the authors set $\theta = 0.99$ making 0 count highly probable but allowing for 1-2 read counts. For the growth-defect state, θ is set at $\theta = 1/(0.01\bar{c}+2)$ reflecting the fact that the growth-defect state must represent approximately 100 times lower read counts than the non-essential state. For the growth-advantage state, θ is set to the inverse of five times the mean read count $\frac{1}{5\bar{c}}$ to capture sites with significantly more insertions locally than what is observed on average in the genome. Table 2 shows the $\hat{\theta}$ for each state. In the estimation of $\bar{c}$, the authors exclude the largest 5% for robustness. Just like DeJesus et Iorger, we obtain $\bar{c} = 195$.

4.2 Transition matrix

DeJesus et Iorger then define a transmission matrix T that is symmetric for simplicity. They note the transmission matrix determines the degree of "smoothing" in the boundaries of essential and non-essential regions. According to the authors the probability of self-transition should be nearly 1 for all states, the probability of transitioning from one state to another nearly 0. This ensures that a significant change in read-counts is required to justify a transition and as well as smoothing over spurious reads. They make the

Table 2: Emission distribution parameters

State	θ
ES	9.900e-01
GD	2.531e-01
NE	5.126e-03
GA	1.025e-03

transition matrix depend on the expected minimum length of essential regions. The length of such regions is modelled by a geometric distribution again. Indeed the geometric distribution models the number of failures before the first success. If we define an insertion as the success event, a variable following a geometric distribution with parameter the probability of an insertion will count the number of sites without insertion until the first insertion occurs.

If the entire dataset is used to estimate the probability of insertion, p_{ins}, the sample will include essential regions with insertion probabilities which are not representative of non-essential regions. To alleviate this bias, the authors suggest discarding regions with 10 or more TA sites in a row lacking insertions. The probability is then calculated as the insertion density in the remaining areas. We obtain a probability of insertion of $p_{ins} = 0.664$.

Once the insertion probability is estimated, the minimum length of essential regions, r^*, is taken to be the smallest integer such that the probability of r^* failures, that r^* no insertions, is less than 0.01:

$$(1 - p_{ins})^{r^*} < 0.01$$

We obtain $r^* = 5$ TA sites.

The probability of self-transitioning, remaining in the same state, is finally as taken as:

$$T_{ll} = 1 - P(C = 0|NE)^{r^*}$$

DeJesus et Iorger justify this calibration by a rationale that we could rephrase as follows. We part from the most common state: non-essential. Indeed previous studies [3] showed that only 15% of the genes in the genome of prokaryotic organisms are essential. Then if in a non-essential state, we observe 0 read-counts r^* times, the expected minimum length of essential regions, the model should switch to the essential state. Therefore the probability of remaining in non-essential state is taken as 1 minus the probability of such event. The same probability of self-transitioning is used for all states. As we have the diagonal defined, the rest of the matrix is set such that the matrix is symmetric and rows sum 1.

We obtain the following matrix:

$$\log T = \begin{bmatrix} -1.104e-12 & -2.863e+01 & -2.863e+01 & -2.863e+01 \\ -2.863e+01 & -1.104e-12 & -2.863e+01 & -2.863e+01 \\ -2.863e+01 & -2.863e+01 & -1.104e-12 & -2.863e+01 \\ -2.863e+01 & -2.863e+01 & -2.863e+01 & -1.104e-12 \end{bmatrix}$$

4.3 Algorithm

We use, as done by original authors, the Viterbi algorithm to estimate the most probable state. The algorithm requires the multiplication of small probabilities numerous times so, to overcome underflow issues, the computation are carried out in a logarithmic scale. DeJesus et Iorger set initial probabilities as follows in Table 3 without discussing this choice. The initial probabilities have no impact on the most probable path in HMM. In Table 5 we compare the results obtained with the initial probabilities proposed by the authors and the results obtained with equal initial probabilities, that is 0.25 for all the states.

Table 3: Initial probabilities

State	π_0
ES	0.7
GD	0.1
NE	0.1
GA	0.1

5 Results

5.1 State frequency

Table 4 shows the log-likelihood at the last TA site. The non-essential state is the most likely state in this position, therefore the trace-back in Viterbi algorithm starts from there to get the most probable path.

Table 4: Final site log-likelihood

State	Log-likelihood
ES	-397937.4
GD	-397941.5
NE	-397924.5
GA	-397943.9

We obtained frequencies of states in the most probable path that are consistent with the ones reported by the DeJesus et Iorger in [1]. They are presented in Table 5. The most probable state is non-essential, followed by essential. We obtain 16.5% of essential sites in the genome which matches previously reported estimate of 15% of essentiality in prokaryote genome [3]. The frequencies of growth-advantage and growth-defect states are small.

Table 5: State frequency in TA sites with proposed (left) and equal (right) initial probabilities

State	Total % of genome	State	Total % of genome
ES	16.50	ES	16.50
GD	3.99	GD	3.99
NE	78.29	NE	78.29
GA	1.22	GA	1.22

5.2 Essentiality of regions

DeJesus et Iorger then analyse the mean read counts, mean insertion density and mean number of TA sites by state. The means are computed as averages across all regions belonging to a given state. The authors do not specify in [1] how they define a region but we assumed a region was a sequence of equal values of state. We then split the most probable path in stretches of equal state values and proceeded to compute similar statistics that are reported in Table 6.

The results we obtained are very similar to the one obtained by the authors. There is nevertheless difference in the mean read counts for the growth advantage state, we obtained about 616 while DeJesus et Iorger obtained 701. However, this difference does not impair the consistency of the results and we can give them the same interpretation as the authors. The mean insertion density and the mean read count observed decreases with the level of essentiality as expected. The essential regions have a mean read counts close to

0 and very low insertion rate, while non-essential have a large mean read counts and high insertion rate. The growth-defect and growth-advantage state correctly rank as states with some insertions but in a small number for the former, and very high numbers of insertion and mean read counts for the latter.

For DeJesus et Iorger, those results reflect the fact that the HMM is successfully separating regions with average read counts and insertions from those with counts significantly lower or significantly higher than average. Figure 1 shows how regions with the same degree of essentiality are clearly separated in function of mean read counts and insertion frequency.

Table 6: Statistics for state classification on regions

State	Mean # TA sites	Mean insertion density	Mean read counts
ES	27.2	0.006	0.20
GD	30.5	0.137	25.97
NE	114.7	0.700	225.31
GA	33.0	0.900	615.88

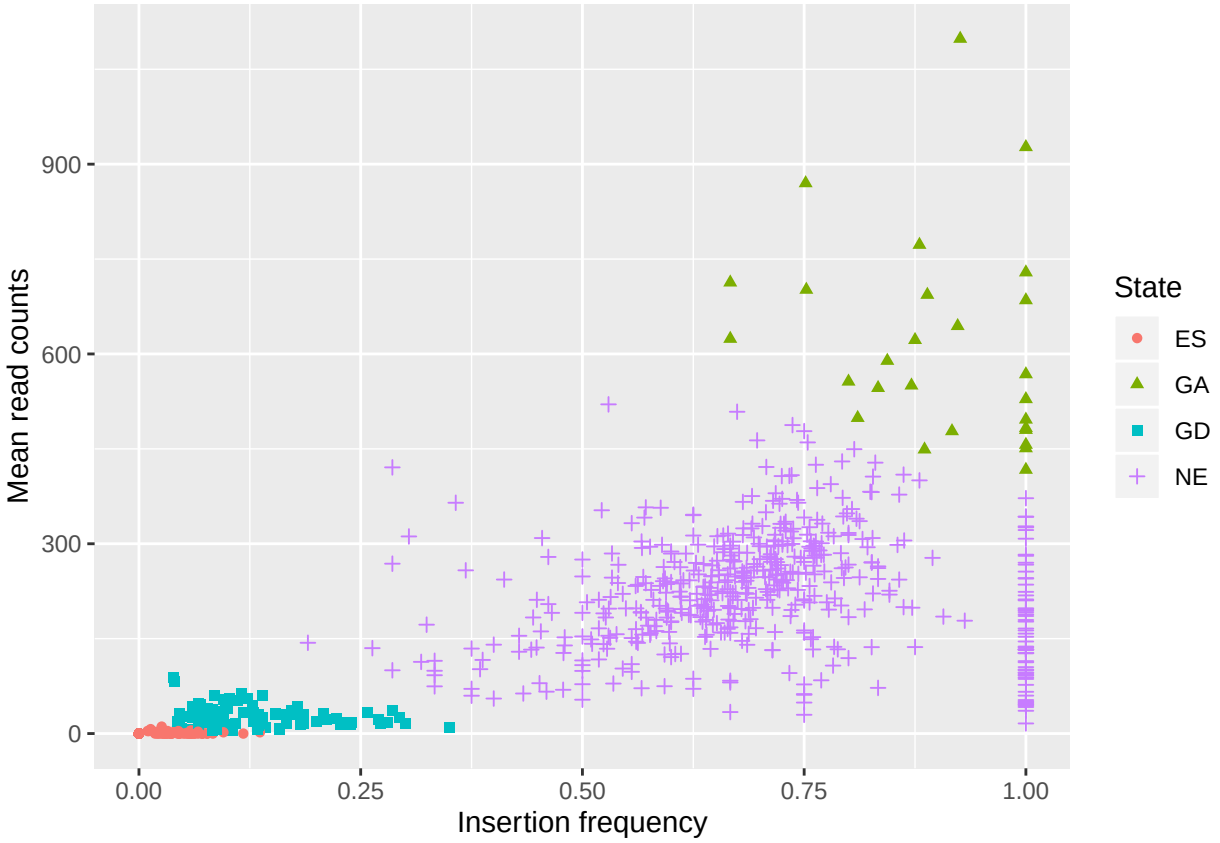

Figure 1: Mean insertion density and read counts for regions

6 Essentiality of individual genes

6.1 Essentiality as most frequent state

Following the methodology developed by the authors we map the sequencing data to identified genes in Mtb genome. We read gene positions from a GFF3 file obtained from EMBL database and match them with the TA sites from the sequencing dataset. We identify 3979 in the original dataset. 39 identified in H37Rv genome were not found in the sequencing data. Some TA sites are not recognised as belonging to any protein-coding gene. We identify such stretches of sites and label them as non-protein coding. We find 2174 of them.

As a first approximation, we can assign genes the most frequent among its the TA sites. If we compute the same statistics as on regions, we obtain the Tables 7 and 8. Those statistics were not reported by DeJesus et Iorger. We find a smaller proportion of essential genes than essential regions, however it is not too far from the 15% reference frequency. The non-essential state is still by far the most frequent degree of essentiality. We also see that the insertion density decreases with essentiality as expected. The mean number of read counts generally decreases with essentiality as well, except between growth-defect and essential states. This expected disordering might not be significant as only a small number of genes were assigned to growth-defect.

Table 7: State frequency in genes

State	Total % of genome
ES	11.99
GD	0.76
NE	3.15
GA	84.09

Table 8: Statistics for state classification on genes

State	Mean # TA sites	Mean insertion density	Mean read counts
ES	18.8	0.046	91.81
GD	19.0	0.154	51.46
NE	15.0	0.669	236.49
GA	31.0	0.805	687.81

6.2 Essentiality with extreme value distribution

Individual genes can mix essential and non-essential regions. As a consequence, to refine the assignment of essentiality of genes, the authors suggest also assigning essentiality to genes that contain sub-sequences of essential TA sites which are statistically longer than expected. The authors rely on asymptotic results for the maximum of n geometric variables to define a threshold for significance. As previously, the length of an essential region can be modelled with a geometric distribution with parameter θ, the probability of insertion. From [4], the expectation and variance of such maximum is:

$$E(max_L) = \mu(n,\theta) = log_{1/\theta}(n(1-\theta)) + \frac{\gamma}{ln(1/\theta)} - 1/2 + r_1(n) + \epsilon_1(n)$$

$$Var(max_L) = \sigma^2(n,\theta) = \frac{\pi^2}{6 * ln(1/\theta)^2} + 1/12 + r_2(n) + \epsilon_2(n)$$

where $\gamma \approx 0.577$ is Euler-Mascheroni constant, $r_1(n)$ and $r_2(n)$ are very small and ϵ_1 and ϵ_2 tend to 0. DeJesus et Iorger set r_1 to 0.000016 and r_2 to 0.00006, without giving any specific justification. In [5], Schilling shows that the distribution of the maximum converges to the Gumbel distribution called Extreme Value Distribution in [1]. The authors of [1] approximate high quantiles of the Gumbel distribution by:

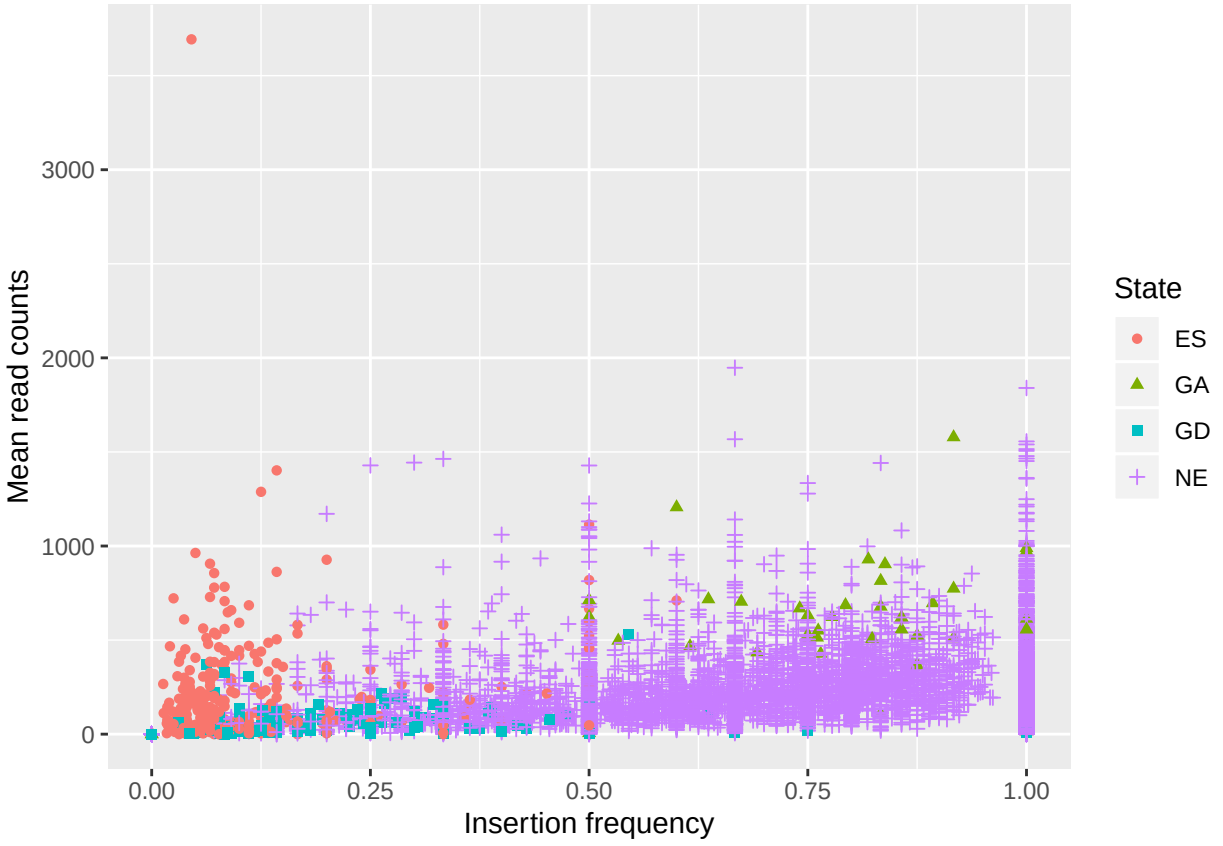

Figure 2: Mean insertion density and read counts for regions

$$q(n, \theta) = \mu(n, \theta) + 3\sigma(n, \theta)$$

The probability θ is estimated as the proportion of sites with non-zero count which gives a value of 0.542. We note that such estimation of the probability insertion is not consistent with the previous estimation performed to define the transition matrix, p_{ins}. The latter removed sequences of 0 count longer that 10 sites from the estimation sample. Therefore θ is smaller than p_{ins}. Finally if a gene of length n contains more than $q(n, \theta)$ essential sites it is considered essential.

Such refined assignment causes the number of essential gene to rise from 738 to 770, with 4 growth-advantage and 28 non-essential genes becoming essential. The detail of cross-comparison of previous and refined assignments is presented in Table 9.

Table 9: Contingency table of essentiality assignment

	Refined assignment			
	ES	GD	NE	GA
Original				
ES	738	0	0	0
GD	4	190	0	0
NE	28	0	5146	0
GA	0	0	0	47

DeJesus et Iorger compare their results to those obtained by Sassetti et al. in [5] who used a completely different method, Transposon Site Hybridization (TraSH). We also cross-compare our results with the results from Sassetti et al in Table 10. Sasetti et al. define a growth-defect class of regions and genes but DeJesus et Iorger deem that it is not comparable to the growth-defect definition they use in [1] and therefore do not compare their results for growth-defect genes. We, however, include the class in our comparison for reference. We see the assignments are matching for a large majority of genes. They are not matching for the growth-defect class of genes. Nevertheless our results are in line with DeJesus et Ioger's as they mention in [1] that the majority of growth-defect genes identified by Sasetti et al. were classified as non-essential by their HMM.

Table 10: Contingency table of essentiality assignment with TraSH

	HMM			
	ES	GD	NE	GA
TraSH				
ES	423	54	134	0
GD	8	3	30	1
NE	83	36	2408	20

6.3 Notable growth-defect and growth-advantage genes

Finally, DeJesus et Iorger identify a list of notable growth-defect and growth-advantage genes. For those genes the labels "growth-defect" or "growth-advantage" have a biological explanation. Those genes are listed in Tables 11 and 12. We find that most of them have a corresponding label in our results. The exceptions are genes which state we assigned on an individual basis while they are grouped in the authors results.

Table 11: Notable Growth-Defect genes

Orf Ids	State	Included genes	Insertion density	Length	Average reads
Rv0015c	GD	pknA	0.062	16	372.0
Rv0016c	GD	pbpA	0.333	36	39.6
Rv0126	GD	treS	0.226	31	38.2
Rv0467	GD	icl1	0.263	19	62.2
Rv1097c	NE	NA	0.353	17	124.0
Rv1098c	GD	fum	0.000	13	0.0
Rv1099c	GD	glpX	0.143	14	47.0
Rv2379c	ES	mbtF	0.281	64	43.8
Rv2380c	GD	mbtE	0.418	79	47.0
Rv2381c	GD	mbtD	0.182	44	20.3
Rv2382c	GD	mbtC	0.375	16	28.8
Rv3841	GD	bfrB	0.300	10	34.7

Table 12: Notable Growth-Advantage genes

Orf Ids	State	Included genes	Insertion density	Length	Average reads
Rv0479c	ES	NA	0.077	13	91.0
Rv0480c	GA	NA	0.875	16	524.6
Rv0481c	GA	NA	0.833	12	815.0
Rv0483	GA	lprQ	0.893	28	696.4
Rv0554	GA	bpoC	0.857	14	620.5
Rv1843c	NE	guaB1	0.870	23	885.6
Rv1844c	NE	gnd1	0.944	18	487.2
Rv2411c	GA	NA	0.917	24	773.9
Rv2930	GA	fadD26	0.600	40	1206.3
Rv2931	GA	ppsA	0.741	81	668.8
Rv2932	GA	ppsB	0.761	71	513.6
Rv2933	GA	ppsC	0.762	84	550.9
Rv2934	GA	ppsD	0.750	68	633.3
Rv2935	GA	ppsE	0.838	68	903.8
Rv2939	GA	papA5	0.793	29	686.0
Rv2940c	GA	mas	0.819	83	929.0
Rv2941	GA	fadD28	0.674	46	704.1
Rv3295	GA	NA	0.917	12	1578.5
Rv3296	NE	lhr	0.768	56	369.6

7 References

[1] DeJesus and Ioerger: A Hidden Markov Model for identifying essential and growth-defect regions in bacterial genomes from transposon insertion sequencing data. BMC Bioinformatics 2013 14:303.

[2] Griffin JE, Gawronski JD, DeJesus MA, Ioerger TR, Akerley BJ, et al. (2011) High-Resolution Phenotypic Profiling Defines Genes Essential for Mycobacterial Growth and Cholesterol Catabolism. PLoS Pathog 7(9): e1002251. doi:10.1371/journal.ppat.1002251

[3] Gerdes SY, Scholle MD, Campbell JW, Balazsi G, Ravasz E, Daugherty MD, Somera AL, Kyrpides NC, Anderson I, Gelfand MS, Bhattacharya A, Kapatral V, D'Souza M, Baev MV, Grechkin Y, Mseeh F, Fonstein MY, Overbeek R, Barabasi AL, Oltvai ZN, Osterman AL: Experimental determination and system level analysis of essential genes in Escherichia coli MG1655. J Bacteriol 2003, 185(19):5673–5684.

[4] M. F. Schilling, The longest run of heads, The College Mathematics Journal 21(3) (1990) 196–207 http://dx.doi.org/10.2307/2686886.

[5] Sassetti CM, Boyd DH, Rubin EJ: Genes required for mycobacterial growth defined by high density mutagenesis. MolMicrobiol 2003, 48:77–84. [http://dx.doi.org/10.1046/j.1365-2958.2003.03425.x]