molecular systems biology

The essential genome of a bacterium

Beat Christen, Eduardo Abeliuk, Michael Collier, Virginia Kalogeraki, Ben Passarelli, John Coller, Michael J. Fero, Harley McAdams, Lucy Shapiro

Corresponding author: Lucy Shapiro, Stanford University

Review timeline:

Submission date:	01 June 2011
Editorial Decision:	21 June 2011
Revision received:	13 July 2011
Accepted:	13 July 2011

Transaction Report:

(Note: With the exception of the correction of typographical or spelling errors that could be a source of ambiguity, letters and reports are not edited. The original formatting of letters and referee reports may not be reflected in this compilation.)

1st Editorial Decision 21 June 2011

Thank you again for submitting your work to Molecular Systems Biology. We have now heard back from the three referees who accepted to evaluate the study. As you will see, the referees find the topic of your study of potential interest and are supportive. They make however several suggestions for modifications and clarifications, which we would ask you to carefully address in a revision of the present work. We would also encourage you to include Supplementary Figure 1 in the main text (either as Figure 1 or 'Box 1', with illustration + respective text).

*** PLEASE NOTE *** As part of the EMBO Publications transparent editorial process initiative (see our Editorial at http://www.nature.com/msb/journal/v6/n1/full/msb201072.html), Molecular Systems Biology will publish online a Review Process File to accompany accepted manuscripts. When preparing your letter of response, please be aware that in the event of acceptance, your cover letter/point-by-point document will be included as part of this File, which will be available to the scientific community. More information about this initiative is available in our Instructions to Authors. If you have any questions about this initiative, please contact the editorial office msb@embo.org . Thank you for submitting this paper to Molecular Systems Biology. I look forward to reading your revised study.

Yours sincerely,

Editor
Molecular Systems Biology

http://www.nature.com/msb

Referee reports

Reviewer #1 (Remarks to the Author):

The manuscript by McAdams et.al. present highly interesting data the essential genome of Caulobacter crescents. The data was generated through the application of transposon mutagenesis coupled to multiplex illumina sequencing of viable clones capable of growing on rich media. The authors document outstanding precision in the determination of essentiality of genes and provide some verification of their findings using orthogonal methods.

I find this paper extremely interesting to a broad audience both in terms of the conclusions drawn and also the general technology. I therefore recommend the paper to be published without further changes.

A minor suggestion would be to include figure S1 in main text, since this figure very well illustrates the methodology used.

Reviewer #2 (Remarks to the Author):

This is an important and well done study to map the essential sequences of a bacterial genome. I am particularly impressed with the mapping of promoter elements. I have several relatively minor comments.

1. By not allowing for a mismatch (base change, deletion or insertion) on one or both paired-end reads, insertions near reference genome error sites will be excluded. This could contribute to some small sites that are incorrectly deemed essential sequences. We have yet to find a bacterial reference genome that lacks such many such errors, particularly genomes that were sequenced by shot-gun cloning into E. coli before sequencing of inserts.

2. Does Tn5 has a preference for less transcribed regions of the genome? This would complicate the interpretation of the data, since many essential genes may be highly transcribed. This should be discussed in the manuscript.

3. The strategy in Sup. Table 3 for validating essential genes is somewhat flawed. Even with equal length regions of homology upstream and downstream of the deletion, there is frequently a recombination bias. This results in the plasmid integrating AND excising more frequently via one or the other region of homology. If the bias is extreme, i.e., >95% bias, then this would easily explain the results for some of the genes tested. The lack of Tn5 insertions in the gene is stronger evidence for essentiality, though complicated by cold spots for Tn5. Another validation method is preferred, such as showing that insertions in essential genes occurred soon after mating and transposition into C. crescentus.

Reviewer #3 (Remarks to the Author):

Manuscript ID: MSB-11-2992-T
Title: The essential genome of a bacterium

Authors: Harley McAdams, Beat Christen, Eduardo Abeliuk, Michael Collier, Virginia Kalogeraki, Ben Passarelli, John Coller, Michael Fero and Lucy Shapiro

This manuscript described the full discovery of essential genome including non-coding, regulatory and ORF coding regions at 8 bp resolution of Caulobacter crescentus by saturated transposon insertion mutagenesis. The insertion points have been determined by deep sequencing method. This miscrobe is an important model organism of a bacterial cell cycle and this manuscript will give readers important information about the essentiality of this bacteria. They discovered 1012 genome features as essential components, which includes 480 ORFs, 402 regulatory sequences and 130 non-coding elements.
I think this paper will give benefits to the readers of this journal with some modification, if needed, pointed below.

Construction of mutant library in Supplementary Methods (I)
The schematic drawing of the structure of the Tn delivery plasmid pXMCS2::Tn5Pxyl may be helpful for readers.
In the supplementary section of (I), "1000 - 1500 mutants were arrayed into 96-well plates" is not clear for me. Does this mean the conjugation with 1000 to 1500 single colony donor strains carrying pXMC2::Tn5Pxyl transposon delivery plasmid? Each conjugation generates more than 500 different insertion mutations, is my understanding correct? If not, more clear description may be required.
Three steps for the preparation of DNA fragments adjacent to the Tn5 insertion point has been performed. In total, three times PCR steps are essential or effective?
Main documents
in the section of "Essential protein coding sequences", 60 ORFs had transposon insertions within their 3' region allow the identification of non-essential protein segment even of essential ORF genes, the authors mentioned. Also they mentioned about the insertion mutation in 5' region of target ORF genes, that those mutant suggested the mis-annotation of translational start site. Their transposon insertion fragment was designed to generate fusion gene with artificial translation initiation codon with RBS at the region in I-end. My understanding is that the insertion mutation in the +1 frame at the 5' end of the target ORF gene could produce fusion protein. In these cases, insertion points should be classified by their position of frame. I could not find any of the description about this, or just I missed?
In figure 3, the essential genes of Caulobacter in the red circle contains 469 ORFs, but the authors reported 480 essential ORFs in the main text. In addition to that, the number of essential genes of E. coli in this figure shows 512. 620 and 303 with 25 updated have been reported by Gerdes in 2003 and Baba in 2006 with updated report by Yamamoto in 2009, respectively. Which data was used as essential genes of E. coli?
Ussery and his colleagues have reported the comparison of 61 E. coli genomes and showed the core genes. It is quite reasonable to compare essential genes of reported here with these core genes of E. coli.

1st Revision - authors' response 13 July 2011

Reviewer remarks and author responses
Manuscript Number:MSB-11-2992-T
The essential genome of a bacterium

Reviewer comments	Author responses
Reviewer #1	
The manuscript by Christen et.al. present highly interesting data the essential genome of Caulobacter crescents. The data was generated through the application of transposon mutagenesis coupled to multiplex illumina sequencing of viable clones capable of growing on rich media. The authors document outstanding precision in the determination of essentiality of genes and provide some verification of their findings using orthogonal methods.	
I find this paper extremely interesting to a broad audience both in terms of the conclusions drawn and also the general technology. I therefore recommend the paper to be published without further changes.	
A minor suggestion would be to include figure S1 in main text, since this figure very well illustrates the methodology used.	As suggested, former Supplementary Figure 1A has been moved into the main manuscript as the new Figure 1.
Reviewer #2	
This is an important and well done study to map the essential sequences of a bacterial genome. I am particularly impressed with the mapping of promoter elements.	
I have several relatively minor comments.	

1. By not allowing for a mismatch (base change, deletion or insertion) on one or both paired-end reads, insertions near reference genome error sites will be excluded. This could contribute to some small sites that are incorrectly deemed essential sequences. We have yet to find a bacterial reference genome that lacks such many such errors, particularly genomes that were sequenced by shot-gun cloning into E. coli before sequencing of inserts.	We performed a SNP (single nucleotide polymorphism) and Indel (insertion and deletion) analysis and compared the genome sequence of the *Caulobacter* strain used for this essential genome study against the published reference genome of *Caulobacter crescentus* NA1000. We found no evidence for sequencing error with the reference genome. This might be because the *Caulobacter crescentus* NA1000 genome was recently resequenced with 24x coverage using next-generation sequencing rather than shot-gun cloning and Sanger sequencing. To make this clear, we added the following sentences to the materials and method section (iii): ”A SNPs and Indels analysis was performed to identify any sequencing error sites within the reference *Caulobacter crescentus* NA1000 genome. Sequencing error could lead to some small sites being incorrectly deemed essential. The SNP and Indel analysis results showed that the genome sequence of the *Caulobacter* Δ*recA* strain used in this study was identical to the published reference genome sequence of *Caulobacter crescentus* NA1000 with the exception of the engineered 700bp deletion covering part of the recA gene.”
2. Does Tn5 has a preference for less transcribed regions of the genome? This would complicate the interpretation of the data, since many essential genes may be highly transcribed. This should be discussed in the manuscript.	Even though many highly expressed genes are essential, (for example most of the ribosomal genes are indeed highly expressed and essential), there are many highly expressed genes that are non-essential (i.e., they had many transposon insertions). An example is the *rsaA* gene encoding the S-layer protein and also flagellin genes. To examine this further, we cross-correlated the level of expression of non-essential genes with the number of transposon insertions that landed within the gene. We found no correlation (c = 0.038) between the expressiveness of a gene and the density of transposon insertions observed.

3. The strategy in Sup. Table 3 for validating essential genes is somewhat flawed. Even with equal length regions of homology upstream and downstream of the deletion, there is frequently a recombination bias. This results in the plasmid integrating AND excising more frequently via one or the other region of homology. If the bias is extreme, i.e.,>95% bias, then this would easily explain the results for some of the genes tested. The lack of Tn5 insertions in the gene is stronger evidence for essentiality, though complicated by cold spots for Tn5. Another validation method is preferred, such as showing that insertions in essential genes occurred soon after mating and transposition into C. crescentus.	Cold spots for the Tn5 element would not allow for any insertion of the transposon element. However, within essential ORFs we frequently do observe transposon insertions in the +1 frame within the 5' region of the gene. Because of the outward pointing Pxyl promoter element, a RBS and a start codon within the transposon end, transposon insertions in the +1 frame are not disruptive and will allow for translation of downstream sequences of the essential ORF. Furthermore, we observe a strong bias in transposon insertion orientation within the upstream regions of essential ORFs. Only transposon insertions that point the Pxyl promoter toward the essential gene are found close to the beginning of the ORF. This evidence was in the manuscript on pages 6 and supplementary materials and method section (iv).
Reviewer #3	
This manuscript described the full discovery of essential genome including non-coding, regulatory and ORF coding regions at 8 bp resolution of Caulobacter crescentus by saturated transposon insertion mutagenesis. The insertion points have been determined by deep sequencing method. This microbe is an important model organism of a bacterial cell cycle and this manuscript will give readers important information about the essentiality of this bacteria. They discovered 1012 genome features as essential components, which includes 480 ORFs, 402 regulatory sequences and 130 non-coding elements.	
I think this paper will give benefits to the readers of this journal with some modification, if needed, pointed below.	
Construction of mutant library in Supplementary Methods (I)	
The schematic drawing of the structure of the Tn delivery plasmid pXMCS2::Tn5Pxyl may be helpful for readers.	As the reviewer suggested, we included a feature map of the plasmid pXMCS2::Tn5Pxyl into Supplementary Figure 1.

In the supplementary section of (I), "1000 - 1500 mutants were arrayed into 96-well plates" is not clear for me. Does this mean the conjugation with 1000 to 1500 single colony donor strains carrying pXMC2::Tn5Pxyl transposon delivery plasmid? Each conjugation generates more than 500 different insertion mutations, is my understanding correct? If not, more clear description may be required.	To clarify the methodology used to deliver the transposon plasmid into *Caulobacter* and better describe the pooling schemata used, we added the following sentences in the supplementary material and method section (i): ”PYE plates were incubated at 30 ºC for 7 days till kanamycin resistant colonies formed. Pools of approximately 1000-1500 kanamycin resistant colonies were deposited into individual wells of 96-well plates. The resulting Tn5Pxyl insertion library contained an estimated 8 x 10^5 transposon mutants.”
Three steps for the preparation of DNA fragments adjacent to the Tn5 insertion point has been performed. In total, three times PCR steps are essential or effective?	The reviewer apparently misinterpreted the PCR methodology described in Materials and Methods section (ii) and the Supplementary Figure 1B that illustrate the PCR strategy used to parallel amplify DNA reading outward of transposon elements into adjacent genomic sequences. Specifically, transposon junctions were amplified using just two PCR steps. Both PCR steps are essential. A first round of PCR includes a few touch-down cycles to permit annealing of the semi-arbitrary PCR primer in proximity to transposon elements. A second nested PCR step was then used to specifically amplify the transposon junctions further and add terminal sequences compatible for Illumina sequencing. The resulting DNA library was then loaded onto an Illumina sequencing flow-cell.

Main documents	
in the section of "Essential protein coding sequences", 60 ORFs had transposon insertions within their 3' region allow the identification of non-essential protein segment even of essential ORF genes, the authors mentioned. Also they mentioned about the insertion mutation in 5' region of target ORF genes, that those mutant suggested the mis-annotation of translational start site. Their transposon insertion fragment was designed to generate fusion gene with artificial translation initiation codon with RBS at the region in I-end. My understanding is that the insertion mutation in the +1 frame at the 5' end of the target ORF gene could produce fusion protein. In these cases, insertion points should be classified by their position of frame. I could not find any of the description about this, or just I missed?	Indeed, transposon insertion sites were classified by their position of frame as described in the supplementary materials and method section (iii). We have now clarified in supplementary material (iv) the fact that insertions in the +1 at the 5' end of a target ORF will produce fusion proteins and modified following sentences: "Transposon insertions in the stop codon of an ORF or located in the +1 reading frame were excluded. Insertions targeting the stop codon of an ORF will not abrogate protein production while insertions located in the +1 frame will permit translation of downstream sequences due to the engineered RBS within the transposon element. Transposon insertions in the +1 frame of a target ORF do not necessarily abrogate protein function and could yield functional fusion proteins or split a given protein into two polypeptides fragments that retain the capability to assemble into a function protein."

In figure 3, the essential genes of Caulobacter in the red circle contains 469 ORFs, but the authors reported 480 essential ORFs in the main text.	When analyzing the conservation of genes across *Caulobacter crescentus* and other proteobacteria (including *E. coli*), a slightly different annotation of the *Caulobacter crescentus* genome was used (*Caulobacter crescentus* CB15N), since we based the analysis on the STRING database which is based on that annotation. This caused the small disparity between the numbers of essential genes reported in the main manuscript and the one in the Venn diagram. To clarify this, we added the following to the legend of Figure 4: "Only essential *Caulobacter* ORFs present in the STING database were considered, leading to a small disparity in the total number of essential *Caulobacter* ORFs."
In addition to that, the number of essential genes of E. coli in this figure shows 512. 620 and 303 with 25 updated have been reported by Gerdes in 2003 and Baba in 2006 with updated report by Yamamoto in 2009, respectively. Which data was used as essential genes of E. coli?	As stated in supplementary materials and methods (v), to make the list of essential *E. coli* genes more robust we combined the lists reported by Yamazaki (2008), Gerdes (2003), and, Baba (2006) into a single list.
Ussery and his colleagues have reported the comparison of 61 E. coli genomes and showed the core genes. It is quite reasonable to compare essential genes of reported here with these core genes of E. coli.	Even though the work reported by Ussery and his colleagues is of interest to the *E. coli* community, these genes constitute a set of conserved genes across related species. Because we were interested in comparing essential ORF complements between two different species, we used a list of essential *E. coli* genes rather than a list of conserved *E.coli* core genes. We find that conservation does not necessarily correlate with essentiality. Indeed, we found many essential genes in one species that were not necessarily conserved in the other species, as illustrated in Figure 4C.