The essential genome of a bacterium

Beat Christen1,5, Eduardo Abeliuk1,2,5, John M Collier2, Virginia S Kalogeraki1, Ben Passarelli3,4, John A Coller3, Michael J Fero1, Harley H McAdams1 and Lucy Shapiro1,*

1 Department of Developmental Biology, Stanford University, Stanford, CA, USA, 2 Department of Electrical Engineering, Stanford University, Stanford, CA, USA, 3 Functional Genomics Facility, Stanford University, Stanford, CA, USA and 4 Stem Cell Institute Genome Center, Stanford University, Stanford, CA, USA

* Corresponding author. Department of Developmental Biology, Stanford University, B300 Beckman Center, Stanford, CA 94305, USA. Tel.: +1 650 725 7678; Fax: +1 650 725 7739; E-mail: shapiro@stanford.edu

Introduction

In addition to protein-coding sequences, the essential genome of any organism contains essential structural elements, non-coding RNAs and regulatory sequences. We have identified the Caulobacter crescentus essential genome to 8 bp resolution by performing ultra-high-resolution transposon mutagenesis followed by high-throughput DNA sequencing to determine the transposon insertion sites. A notable feature of C. crescentus is that the regulatory events that control polar differentiation and cell-cycle progression are highly integrated, and they occur in a temporally restricted order (McAdams and Shapiro, 2011). Many components of the core regulatory circuit have been identified and simulation of the circuitry has been reported (Shen et al., 2008). The identification of all essential DNA elements is essential for a complete understanding of the regulatory networks that run a bacterial cell.

Accurate identification of the entire essential genome of any bacterial species at a resolution of a few base pairs.

Results and discussion

We engineered a Tn5 derivative transposon (Tn5Pxyl) that carries at one end an inducible outward pointing Pxyl promoter (Christen et al., 2010; Supplementary Figure 1A; Materials and methods). Thus, the Tn5Pxyl element can activate or disrupt transcription at any site of integration, depending on the insertion orientation. About $8 \times 10^5$ viable Tn5Pxyl transposon insertion mutants capable of colony formation on rich media (PYE) plates were pooled. Next, DNA from hundreds of thousands of transposon insertion sites reading outwards into flanking genomic regions was parallel PCR amplified and sequenced by Illumina paired-end sequencing (Figure 1; Supplementary Figure 1B; Materials and methods). A single sequencing run yielded 118 million raw sequencing reads. Of these, >90 million (80%) read out from the transposon element into adjacent genomic DNA regions (Supplementary Figure 1C) and were subsequently mapped to the 4-Mbp genome, allowing us to determine the location and orientation of 428735 independent transposon insertions with base-pair accuracy (Figure 2A; Materials and methods).

Eighty percent of the genome sequence showed an ultrahigh density of transposon hits; an average of one insertion event...
by standard paired-end sequencing. Subsequent parallel amplification of individual transposon junctions by a nested arbitrary PCR yields DNA fragments reading out of transposon elements into adjacent genomic DNA sequences. DNA fragments carry terminal adapters (orange, blue) compatible to the Illumina flow-cell and are sequenced in parallel.

Within non-coding sequences of the Caulobacter genome, we detected 130 small non-disruptable DNA segments between 90 and 393 bp long (Materials and methods; Supplementary Data-DT1). (Tables in the Excel file of Supplementary Data are designated DT1, DT2 and so on.) Owing to the uniform distribution of transposition across the genome (Materials and methods), such non-disruptable DNA regions are highly unlikely (Supplementary Figure 2). Among 27 previously identified and validated sRNAs (Landt et al, 2008), three (annotated as R0014, R0018 and R0074 in Landt et al, 2008) were contained within non-disruptable DNA segments while another three (R0005, R0019 and R0025) were partially disruptive. Figure 2B shows one of the three (Supplementary Data-DT1) non-disruptable sRNA elements, R0014, that is upregulated upon entry into stationary phase (Landt et al, 2008). Two additional small RNAs found to be essential are the transfer-messenger RNA (tmRNA) and the ribozyme RNaseP (Landt et al, 2008). In addition to the 8 non-disruptable sRNAs, 29 out of the 130 essential non-coding sequences contained non-redundant tRNA genes (Figure 2C); duplicated tRNA genes were found to be non-essential. We identified two non-disruptable DNA segments within the chromosomal origin of replication (Figure 2D). A 173-bp long essential region contains three binding sites for the replication repressor CtrA, as well as additional sequences that are essential for chromosome replication and initiation control (Marczynski et al, 1995). A second 125 bp long essential DNA segment contains a binding motif for the replication initiator protein DnaA. Surprisingly, between these non-disruptable origin segments there were multiple transposon hits suggesting that the Caulobacter origin is modular with possible DNA looping compensating for large insertion sequences. Thus, we resolved essential non-coding RNAs, tRNAs and essential replication elements within the origin region of the chromosome. Although 90 additional non-disruptable small genome elements were identified (Supplementary Data-DT1), they cannot be explained within the context of the current genome annotation. Eighteen of these are conserved in at least one closely related species. Only two could encode a protein of over 50 amino acids.

***Essential protein-coding sequences***

For each of the 3876 annotated open reading frames (ORFs), we analyzed the distribution, orientation and genetic context of transposon insertions. We identified the boundaries of the essential protein-coding sequences and calculated a statistically robust metric for ORF essentiality (Materials and methods; Supplementary Data-DT2). There are 480 essential ORFs and 3240 non-essential ORFs. In addition, there were 156 ORFs that severely impacted fitness when mutated, as evidenced by a low number of disruptive transposon insertions (Supplementary methods). Figure 2E shows the distribution of transposon hits for a subregion of the genome encoding essential and non-essential ORFs. Genome-wide transposon insertion frequencies for the annotated Caulobacter ORFs are shown in Figure 2F. In all, 145/480 essential ORFs lacked transposon insertions across the entire coding region, suggesting that the full length of the encoded protein up to the last amino acid is essential. The 8-bp resolution allowed a dissection of the essential and non-essential regions of the coding sequences. Sixty ORFs had transposon insertions within a significant portion of their 3’ region but lacked insertions in the essential 5’ coding region, allowing the identification of non-essential protein segments. For example, transposon insertions in the essential cell-cycle regulatory gene divL, a tyrosine kinase, showed that the last 204 C-terminal amino acids did not impact viability (Figure 2G), confirming previous reports that the C-terminal ATPase domain of DivL is dispensable for viability (Reisinger et al, 2007; Iniesta et al, 2010). Our results show that the entire C-terminal ATPase domain, as well as the majority of the adjacent kinase domain, is non-essential while the N-terminal region including the first 25 amino acids of the kinase domain contain essential DivL functions.

Conversely, we found 30 essential ORFs that tolerated disruptive transposon insertions within the 5’ region while no insertion events were tolerated further downstream (Supplementary Table 1). One such example, the essential histidine phosphotransferase gene chpT (Biondi et al, 2006), had 12 transposon insertions near the beginning of the annotated ORF (Figure 2H). These transposon insertions would prevent the production of a functional protein and should not be detectable within chpT or any essential ORF unless the translational start site is mis-annotated. Using LacZ-reporter assays (Supplementary methods), we found that the promoter element as well as the translational start site of chpT was located downstream of the annotated start codon (Figure 2H). Cumulatively, >6% of all essential ORFs (30 out of 480) appear to be shorter than the annotated ORF (Supplementary Table 1), suggesting that these are probably mis-annotated, as well. Thus, 145 ORFs showed all regions were essential, 60 ORFs showed non-essential
The essential genome of a bacterium

B Christen et al

The essential cell-cycle gene dispensible region encodes parts of the histidine kinase domain as well as an ATPase domain that is non-essential. (downstream of the mis-annotated start codon as confirmed by observed are below. (the sRNA (Landt contains a stationary phase expressed non-coding sRNA (Landt 3876 annotated ORFs shown in the inner (minus strand) and middle (plus strand) tracks by black lines. (2011 EMBO and Macmillan Publishers Limited Molecular Systems Biology 2011

Among the 480 essential ORFs, there were 10 essential transcriptional regulatory proteins (Supplementary Table 4), including the cell-cycle regulators ctrA, gcrA, ccrM, sciP and C-termini and the start of 30 ORFs were mis-annotated. The remaining 245 ORFs tolerated occasional insertions within a few amino acids of the ORF boundaries (Supplementary Figure 3; Materials and methods).

The majority of the essential ORFs have annotated functions. They participate in diverse core cellular processes such as ribosome biogenesis, energy conversion, metabolism, cell division and cell-cycle control. Forty-nine of the essential proteins are of unknown function (Table I; Supplementary Table 2). We attempted to delete 11 of the genes encoding essential hypothetical proteins and recovered no in-frame deletions, confirming that these proteins are indeed essential (Supplementary Table 3).
The essential genome of a bacterium

B Christen et al

Table 1  The essential Caulobacter genome

| Category                        | Quantity | Size (bp) | Fraction of genome (%) |
|--------------------------------|----------|-----------|-------------------------|
| Essential non-coding elements  | 130      | 14991     | 0.37                    |
| Non-coding elements, unknown   | 91       | 10893     | 0.27                    |
| tRNAs                          | 29       | 2312      | 0.06                    |
| Small non-coding RNAs          | 8        | 1488      | 0.04                    |
| Genome replication elements    | 2        | 298       | 0.01                    |
|                               |          |           |                         |
| Essential ORFs                 | 480      | 444417    | 11.00                   |
| Metabolism                     | 176      | 160011    | 3.96                    |
| Ribosome function              | 95       | 76420     | 1.89                    |
| Cell wall & membrane biogenesis| 54       | 63393     | 1.57                    |
| Proteins of unknown function   | 49       | 30469     | 0.75                    |
| Cell cycle, division and DNA   | 43       | 52322     | 1.29                    |
| replication                    |          |           |                         |
| Other cellular processes       | 39       | 36017     | 0.89                    |
| Transcription                  | 15       | 16417     | 0.41                    |
| Signal transduction            | 9        | 9368      | 0.23                    |
| Essential promoter regions     | 402      | 33533     | 0.83                    |
| Contained within intergenic    | 210      | 131500    | 0.33                    |
| sequences                      |          |           |                         |
| Extending into upstream ORFs   | 101      | 11428     | 0.28                    |
| Driving operons                | 91       | 8955      | 0.22                    |
| Essential Caulobacter genome   | 492941   |           | 12.19                   |

The identified essential non-coding, protein-coding and regulatory genome features are shown.

*aFraction of the entire Caulobacter genome encoding essential elements.

dnaA (McAdams and Shapiro, 2003; Holtzendorff et al, 2004; Collier and Shapiro, 2007; Gora et al, 2010; Tan et al, 2010), plus 5 uncharacterized putative transcription factors. We surmise that these five uncharacterized transcription factors either comprise transcriptional activators of essential genes or repressed genes that would move the cell out of its replicative state. In addition, two RNA polymerase sigma factors RpoH and RpoD, as well as the anti sigma factor ChrR, which mitigates rpoE-dependent stress response under physiological growth conditions (Lourenco and Gomes, 2009), were also found to be essential. Thus, a set of 10 transcription factors, 2 RNA polymerase sigma factors and 1 anti sigma factor comprise the essential core transcriptional regulators for growth on rich media.

Essential promoter elements

To characterize the core components of the Caulobacter cell-cycle control network, we identified essential regulatory sequences and operon transcripts (Supplementary Data-DT3 and DT4). Figure 3A illustrates the transposon scanning strategy used to locate essential promoter sequences. The promoter regions of 210 essential genes were fully contained within the upstream intergenic sequences, and promoter regions of 101 essential genes extended upstream into flanking ORFs (Table I). We also identified 206 essential genes that are co-transcribed with the corresponding flanking gene(s) and experimentally mapped 91 essential operon transcripts (Table I; Supplementary Data-DT4). One example of an essential operon is the transcript encoding ATPase synthase components (Figure 3B). Altogether, the 480 essential protein-coding and 37 essential RNA-coding Caulobacter genes are organized into operons such that 402 individual promoter regions are sufficient to regulate their expression (Table I). Of these 402 essential promoters, the transcription start sites (TSSs) of 105 were previously identified (McGrath et al, 2007).

We found that 79/105 essential promoter regions extended on average 53 bp upstream beyond previously identified TSS (Figure 3C; McGrath et al, 2007). These essential control elements accommodate binding sites for transcription factors and RNA polymerase sigma factors (Supplementary Table 5). Of the 402 essential promoter regions, 26 mapped downstream of the predicted TSS. To determine if these contained an additional TSS, we fused the newly identified promoter regions with lacZ and found that 24 contained an additional TSS (Supplementary Table 6). Therefore, 24 genes contain at least 2 TSS and only the downstream site was found to be essential during growth on rich media. The upstream TSS may be required under alternative growth conditions.

Cell cycle-regulated essential genes

Of the essential ORFs, 84 have a cell cycle-dependent transcription pattern (McGrath et al, 2007; Supplementary Data-DT5). The cell cycle-regulated essential genes had statistically significant longer promoter regions compared with non-cell cycle-regulated genes (median length 87 versus 41 bp, Mann–Whitney test, P-value 0.0018). The genes with longer promoter regions generally have more complex transcriptional control. Among these are key genes that are critical for the commitment to energy requirements and regulatory controls for cell-cycle progression. For example, the cell-cycle master regulators ctrA, dnaA and gcrA (Collier et al, 2006) ranked among the genes with the longest essential promoter regions (Figure 3D and E; Supplementary Data-DT5). Other essential cell cycle-regulated genes with exceptionally long essential promoters included ribosomal genes, gyrB encoding DNA gyrase and the ftsZ cell-division gene (Figure 3E). The essential promoter region of ctrA extended 171 bp upstream of the start codon (Figure 3F) and included two previously characterized promoters that control its transcription by both positive and negative feedback regulation (Domian et al, 1999; Tan et al, 2010). Only one of the two upstream SciP binding sites in the ctrA promoter (Tan et al, 2010) was contained within the essential promoter region (Figure 3F), suggesting that the regulatory function of the second SciP binding site upstream is non-essential for growth on rich media.

Altogether, the essential Caulobacter genome contains at least 492 941 bp. Essential protein-coding sequences comprise 90% of the essential genome. The remaining 10% consists of essential non-coding RNA sequences, gene regulatory elements and essential genome replication features (Table I). Essential genome features are non-uniformly distributed along the Caulobacter genome and enriched near the origin and the terminus regions, indicating that there are constraints on the chromosomal positioning of essential elements (Figure 4A). The chromosomal positions of the published E. coli essential
The question of what genes constitute the minimum set required for prokaryotic life has been generally estimated by comparative essentiality analysis (Carbone, 2006) and for a few species experimentally via large-scale gene perturbation studies (Akerley et al., 1998; Hutchinson et al., 1999; Kobayashi et al., 2003; Salama et al., 2004). Of the 480 essential Caulobacter ORFs, 38% are absent in most species outside the α-proteobacteria and 10% are unique to Caulobacter (Figure 4B). Interestingly, among 320 essential Caulobacter proteins that are conserved in E. coli, more than one third are non-essential (Figure 4C). The variations in essential gene complements relate to differences in bacterial physiology and life style. For example, ATP synthase components are essential for Caulobacter, but not for E. coli, since Caulobacter cannot produce ATP through fermentation. Thus, the essentiality of a gene is also defined by non-local properties that not only depend on its own function but also on the functions of all other essential elements in the genome. The strategy described here provides a direct experimental approach that, because of its simplicity and general applicability, can be used to quickly determine the essential genome for a large class of bacterial species.

Materials and methods
Supplementary information includes descriptions of (i) transposon construction and mutagenesis, (ii) DNA library preparation and sequencing, (iii) sequence processing, (iv) essentiality analysis and (v) statistical data analysis.
Acknowledgements

This research was supported by DOE Office of Science grant DE-FG02-05ER64136 to HM; NIH grants K25 GM070972-01A2 to MF, R01, GM51426k, R01 GM32506 and GM073011-04 to LS; Swiss National Foundation grant PA00P3-126243 to BC and the L&Th. La Roche Foundation Fellowship to BC.

Author contributions

BC designed the research. BC, EA, VSK and MJF performed the experiments and analysis. BC, JMC, BP and JAC performed the sequencing and related analysis. BC, EA, HHM and LS wrote the manuscript.

Conflict of interest

The authors declare that they have no conflict of interest.

References

Akerley BJ, Rubin EJ, Camilli A, Lampe DJ, Robertson HM, Mekalanos JJ (1998) Systematic identification of essential genes by in vitro mariner mutagenesis. Proc Natl Acad Sci USA 95: 8927–8932
Baba T, Ara T, Hasegawa M, Takai Y, Okumura Y, Baba M, Datsenko KA, Tomita M, Wanner BL, Mori H (2006) Construction of Escherichia coli K-12 in-frame, single-gene knockout mutants: the Keio collection. Mol Syst Biol 2: 2006 0008
Biondi EG, Reisinger SJ, Skerker JM, Arif M, Perchuk BS, Ryan KR, Laub MT (2006) Regulation of the bacterial cell cycle by an integrated genetic circuit. Nature 444: 899–904
Carbone A (2006) Computational prediction of genomic functional cores specific to different microbes. J Mol Evol 63: 733–746
Christen B, Fero MJ, Hillson NJ, Bowman G, Hong SH, Shapiro L, McdAdams HH (2010) High-throughput identification of protein localization dependency networks. Proc Natl Acad Sci USA 107: 4681–4686
Collier J, Murray SR, Shapiro L (2006) DnaA couples DNA replication and the expression of two cell cycle master regulators. EMBO J 25: 346–356
Collier J, Shapiro L (2007) Spatial complexity and control of a bacterial cell cycle. Curr Opin Biotechnol 18: 333–340
Domian UJ, Reisenauer A, Shapiro L (1999) Feedback control of a master bacterial cell-cycle regulator. Proc Natl Acad Sci USA 96: 6648–6653
Glass JI, Assad-Garcia N, Alperovich N, Yooseph S, Lewis MR, Marul M, Hutchison III CA, Smith HO, Venter JC (2006) Essential genes of a minimal bacterium. *Proc Natl Acad Sc USA* **103**: 425–430

Gora KG, Tsokos CG, Chen YE, Srinavasan BS, Perchuk BS, Laub MT (2010) A cell-type-specific protein-protein interaction modulates transcriptional activity of a master regulator in *Caulobacter crescentus*. Mol Cell **39**: 455–467

Holzendorff J, Hung D, Brende P, Reisenauer A, Viollier PH, McAdams HH, Shapiro L (2004) Oscillating global regulators control the genetic circuit driving a bacterial cell cycle. *Science* **304**: 983–987

Hutchison CA, Peterson SN, Gill SR, Cline RT, White O, Fraser CM, Smith HO, Venter JC (1999) Global transposon mutagenesis and a minimal *Mycoplasma* genome. *Science* **286**: 2165–2169

Iniesta AA, Hillson NJ, Shapiro L (2010) Cell pole-specific activation of a critical bacterial cell cycle kinase. *Proc Natl Acad Sci USA* **107**: 7012–7017

Jacobs MA, Alwood A, Thaipisuttikul I, Spencer D, Haugen E, Ernst S, Will O, Kaul R, Raymond C, Levy R, Chun-Rong L, Guenthner D, Bovee D, Olson MV, Manoil C (2003) Comprehensive transposon mutant library of *Pseudomonas aeruginosa*. *Proc Natl Acad Sci USA* **100**: 14339–14344

Kobayashi K, Ehrlich SD, Albertini A, Amati G, Andersen KK, Arnaud M, Asai K, Ashikaga S, Aymerich S, Bessieres P, Boland F, Brignell SC, Bron S, Bunai K, Chapuis J, Christiansen LC, Danchin A, Debarbouille M, Dervyn E, Deuerling E et al. (2003) Essential *Bacillus subtilis* genes. *Proc Natl Acad Sci USA* **100**: 4678–4683

Landt SG, Abeliu E, McGrath PT, Lesley JA, McAdams HH, Shapiro L (2008) Small non-coding RNAs in *Caulobacter crescentus*. *Mol Microbiol* **68**: 600–614

Langridge GC, Phan MD, Turner DJ, Perkins TT, Parts L, Haase J, Charles I, Maskell DJ, Peters SE, Dougan G, Wain J, Parkhill J, Turner AK (2009) Simultaneous assay of every *Salmonella typhi* gene using one million transposon mutants. *Genome Res* **19**: 2308–2316

Loureiro RF, Gomes SL (2009) The transcriptional response to cadmium, organic hydroperoxide, singlet oxygen and UV-A mediated by the sigmaE-ChrR system in *Caulobacter crescentus*. *Mol Microbiol* **72**: 1159–1170

Marczynski GT, Lentine K, Shapiro L (1995) A developmentally regulated chromosomal origin of replication uses essential transcription elements. *Genes Dev* **9**: 1543–1557

McAdams HH, Shapiro L (2003) A bacterial cell-cycle regulatory network operating in time and space. *Science* **301**: 1874–1877

McAdams HH, Shapiro L (2011) The architecture and conservation pattern of whole-cell control circuitry. *J Mol Biol* **409**: 28–35

McGrath PT, Lee H, Zhang L, Iniesta AA, Hottes AK, Tan MH, Hillson NJ, Hu P, Shapiro L, McAdams HH (2007) High-throughput identification of transcription start sites, conserved promoter motifs and predicted regulons. *Nat Biotechnol* **25**: 584–592

Reisinger SJ, Huntwork S, Viollier PH, Ryan KR (2007) DivL performs critical cell cycle functions in *Caulobacter crescentus* independent of kinase activity. *J Bacteriol* **189**: 8308–8320

Rocha EP (2004) The replication-related organization of bacterial genomes. *Microbiology* **150**: 1609–1627

Salama NR, Shepherd B, Falkow S (2004) Global transposon mutagenesis and essential gene analysis of *Helicobacter pylori*. *J Bacteriol* **186**: 7926–7935

Shen X, Collier J, Dill D, Shapiro L, Horowitz M, McAdams HH (2008) Architecture and inherent robustness of a bacterial cell-cycle control system. *Proc Natl Acad Sci USA* **105**: 11340–11345

Tan MH, Kazdon JB, Shen X, Shapiro L, McAdams HH (2010) An essential transcription factor, SciP, enhances robustness of *Caulobacter* cell cycle regulation. *Proc Natl Acad Sci USA* **107**: 18985–18990

van Opinjen T, Bodi KL, Camilli A (2009) Tn-seq: high-throughput parallel sequencing for fitness and genetic interaction studies in microorganisms. *Nat Methods* **6**: 767–772

Molecular Systems Biology is an open-access journal published by European Molecular Biology Organization and Nature Publishing Group. This work is licensed under a Creative Commons Attribution-Noncommercial-Share Alike 3.0 Unported License.