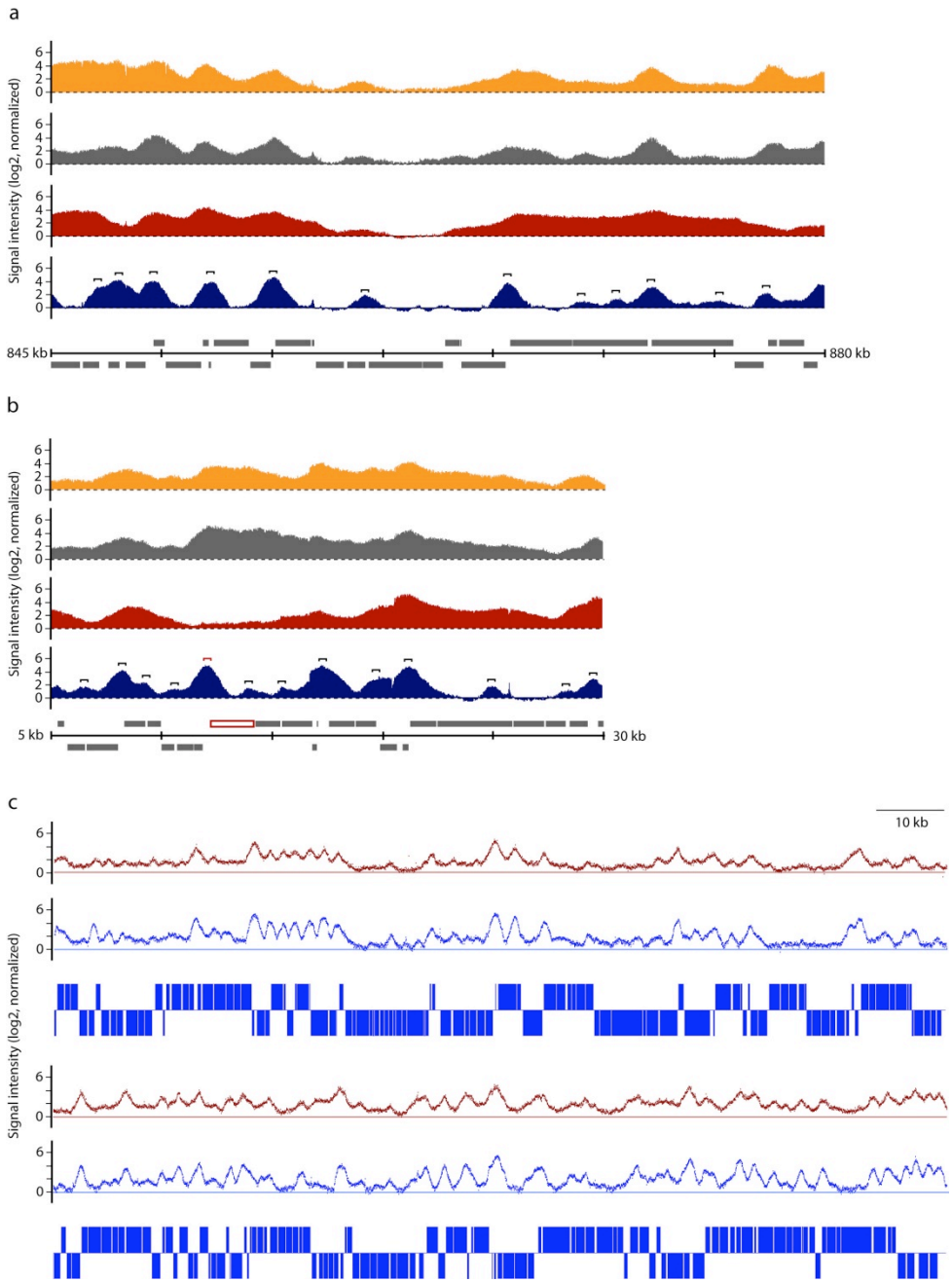
Supplementary Information

Elucidation of the Transcription Unit Architecture of the *Escherichia coli* K-12 MG1655 Genome

Byung-Kwan Cho[1], Karsten Zengler[1], Yu Qiu[1], Young Seoub Park[1], Eric M Knight[1], Christian Barrett[1], Yuan Gao[2], and Bernhard Ø. Palsson[1]

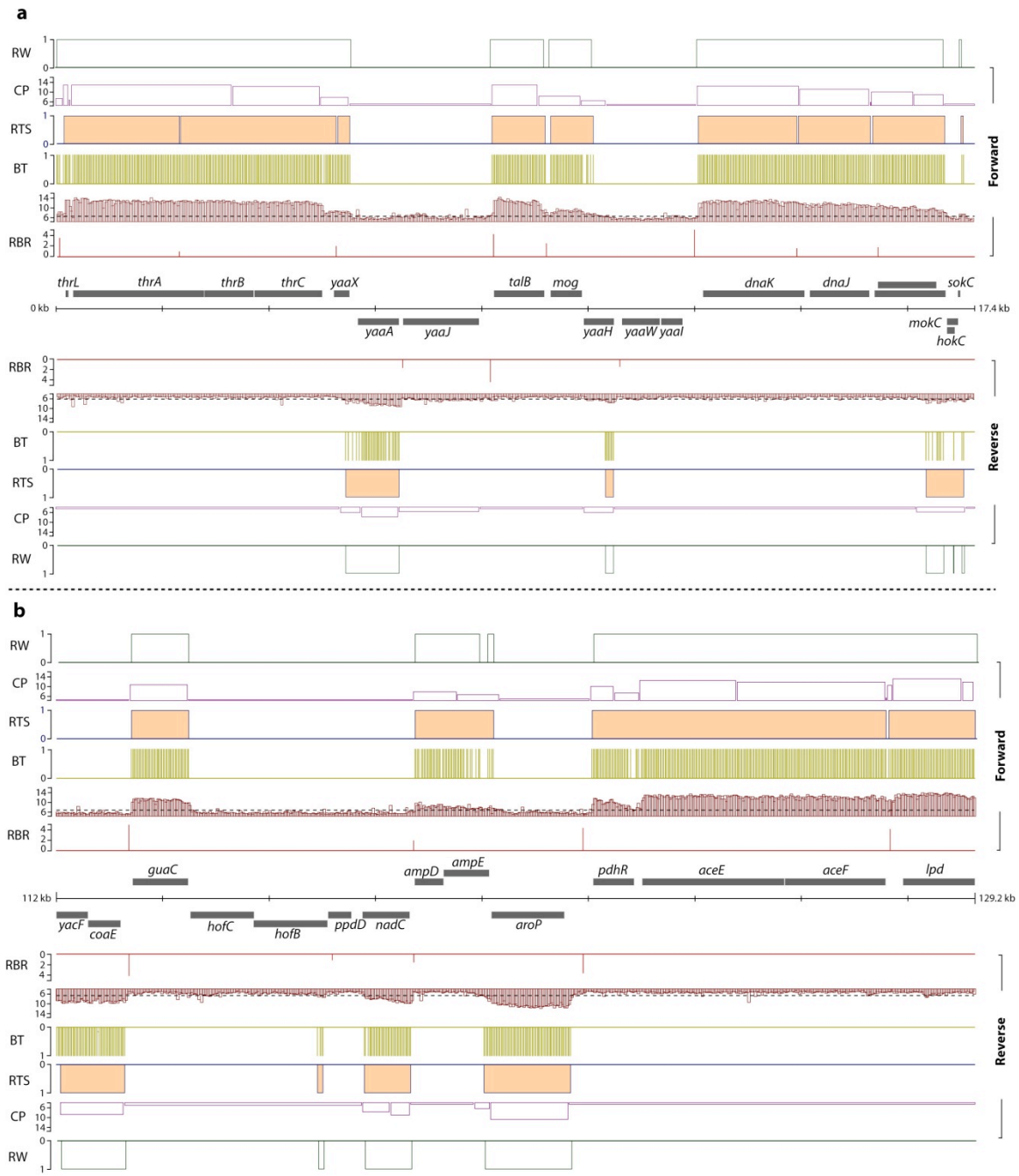
[1]Department of Bioengineering, University of California, San Diego, La Jolla, CA 92093, USA. [2]Center for the Study of Biological Complexity and Department of Computer Science, Virginia Commonwealth University, Richmond, VA 23284, USA.

This supplementary information contains a total of nine supplementary figures supporting the main manuscript. The supplementary figure titles are as follows:

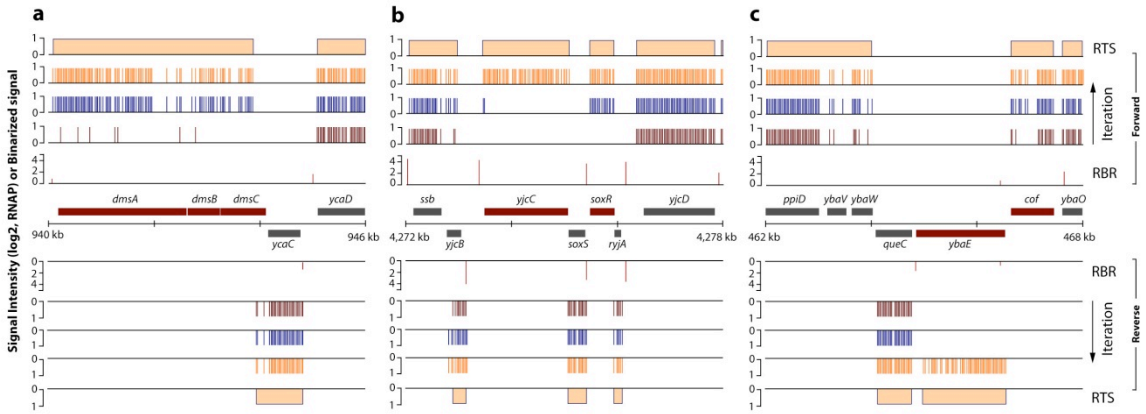


Supplementary Figure 1 | Static and dynamic map of RNA polymerase binding. Determination of the binding locations of RNA polymerase was nearly condition dependent. Although we observed the differential binding levels of RNA polymerase under different conditions, the binding locations (i.e., promoter regions) were nearly identical. **(a, b)** Examples of RNA polymerase (RNAP) binding under different growth conditions (log phase, red; heat-shocked, grey; stationary phase, orange). Binding of RNAP was determined by the static map although regions of log phase cells or log phase and heat-shocked cells did not show RNAP binding under the dynamic map. Regions of differential binding are highlighted in red. **(c)** Static RNAP-binding maps of log phase

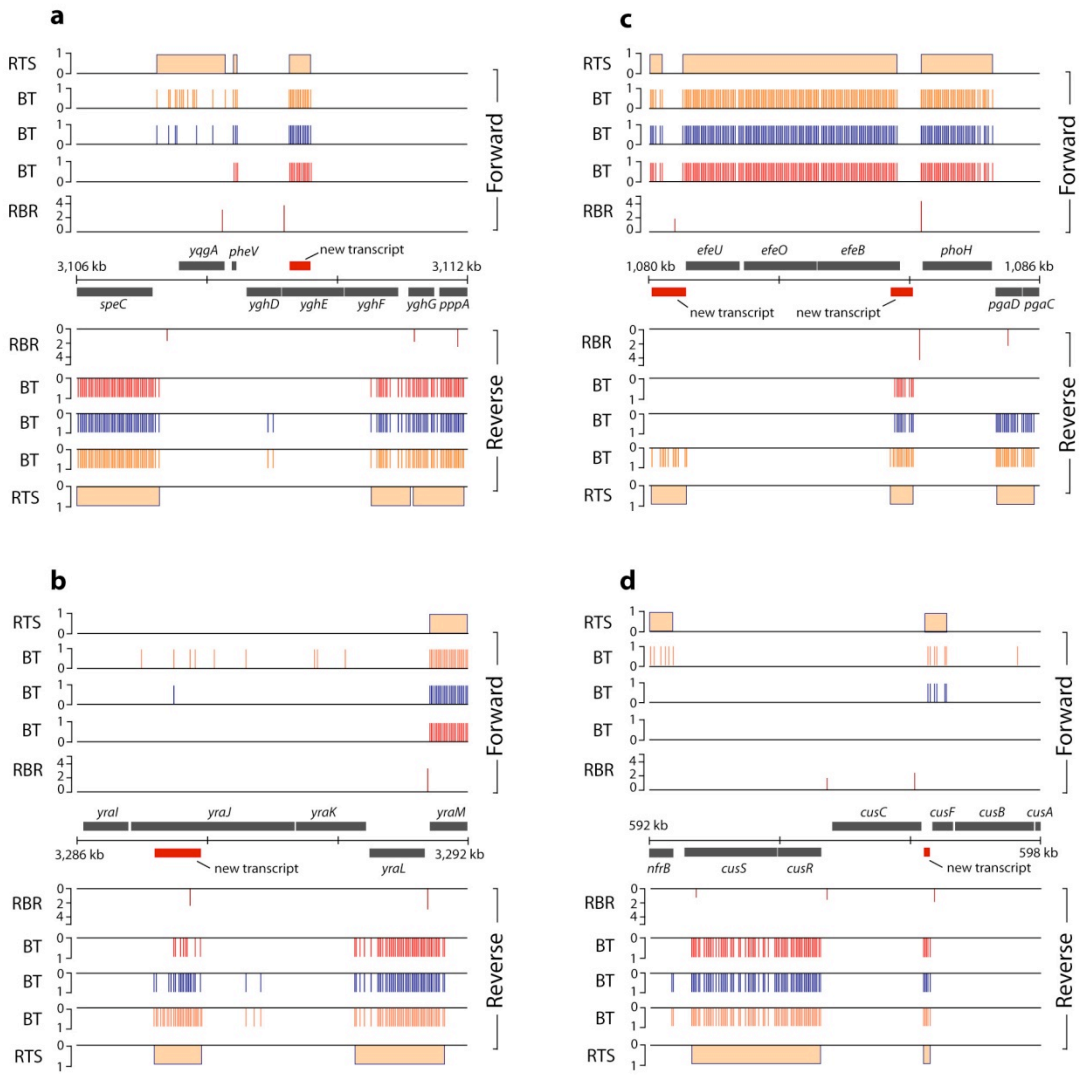
(red) and leucine condition (blue). We observed differential RNAP-binding levels, however, the binding locations of RNAP was nearly identical.



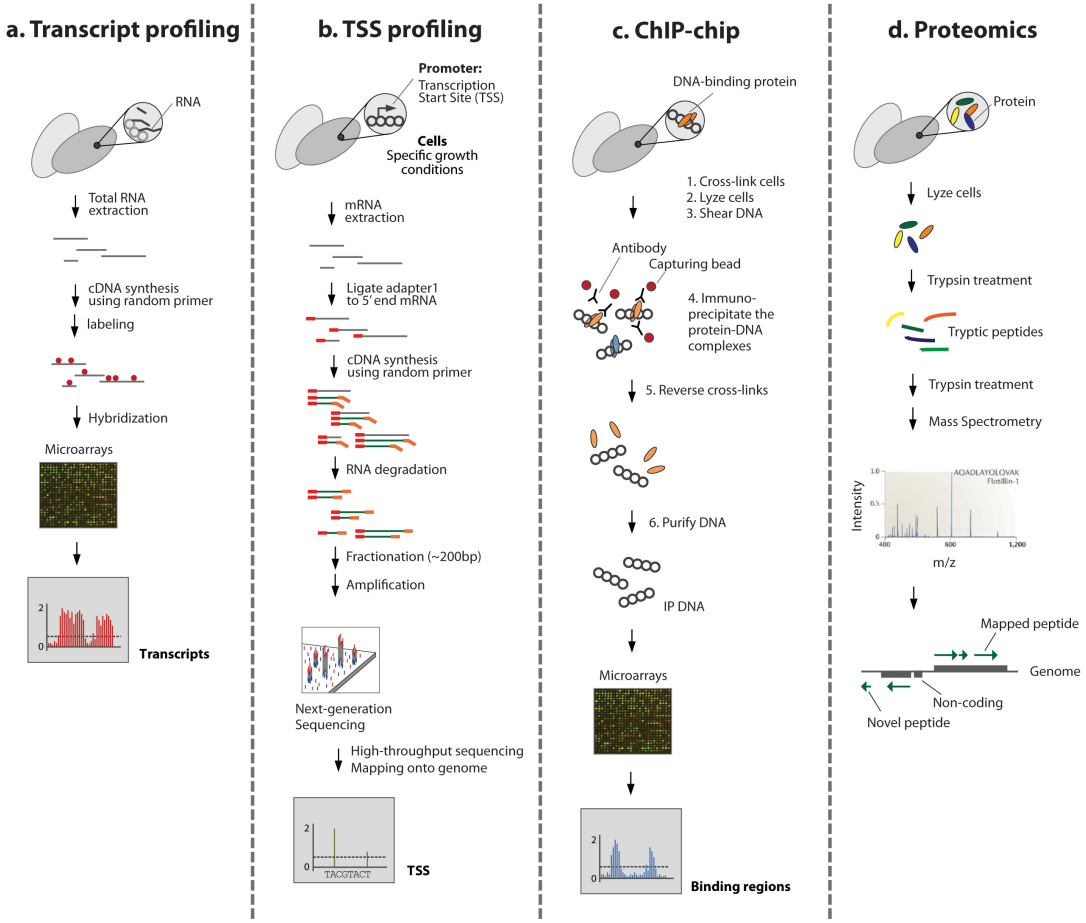
Supplementary Figure 2 | Comparison of RNAP-guided transcript segment (RTS) to change point algorithm and running-window approach. Integration of RNA polymerase binding regions (RBRs) with binary transcript calls (BT) lead to RTSs. RTS, based on integration of two experimental derived genome-wide data sets, yielded the best results when compared to change point algorithm (CP) and running window approach (RW). Two examples **(a, b)**, representative for all data, demonstrate that determination of transcription fragments using CP resulted in too many fragments (too sensitive), whereas the RW yielded too few fragments (less sensitive).



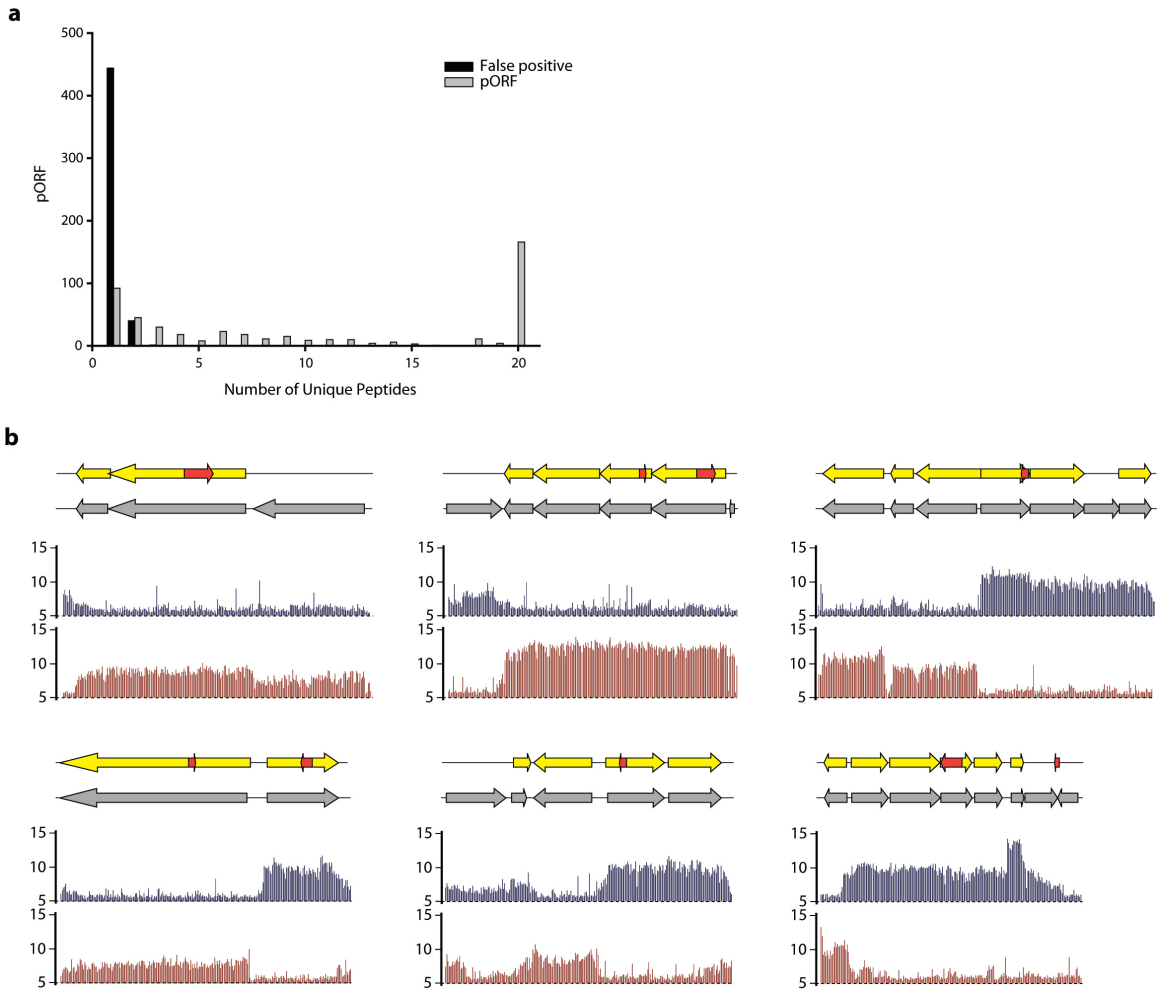
Supplementary Figure 3 | Increase of genomic coverage and accuracy by iterative integration. Iterative integration of transcripts, derived from various growth conditions, with RNA polymerase binding regions (RBRs) resulted in increased genomic coverage and accuracy **(a, b, c)**, genes of interest are highlighted in red. Iteration of data from various growth conditions (log phase, red; heat-shocked, blue; stationary phase, orange) also allowed for determination of condition-specific transcripts, such as *yjcC* **(b)** and *ybaE* **(c)** from stationary growth phase, and *soxR* **(b)** from heat-shocked cells.



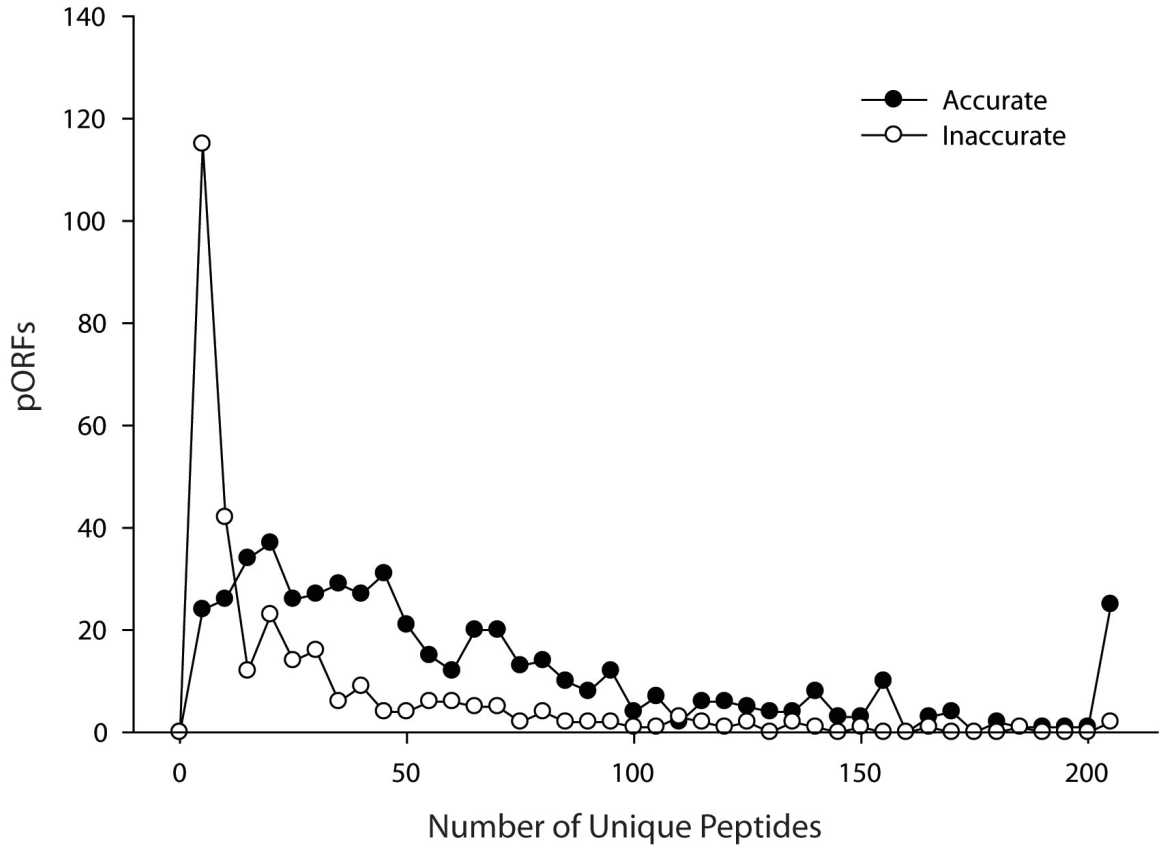
Supplementary Figure 4 | Discovery of new transcripts. New transcripts were determined by systematic and iterative integration of RNA polymerase binding regions (RBRs) with binary transcript calls (BT) resulting into RNAP-guided transcript segments (RTSs). New transcripts (highlighted in red) were discovered on opposite strands **(a, b)**, as well as in intergenic regions **(c, d)**.



Supplementary Figure 5 | Flowcharts of the molecular biology tool box for the elucidation of the organizational components. Various genome-scale methods were deployed and developed to determine the meta-structure. Methods are depicted here include **(a)** transcription profiling, **(b)** transcription start site (TSS) profiling, **(c)** chromatin immunoprecipitation coupled to microarrays (ChIP-chip), and **(d)** proteomics.

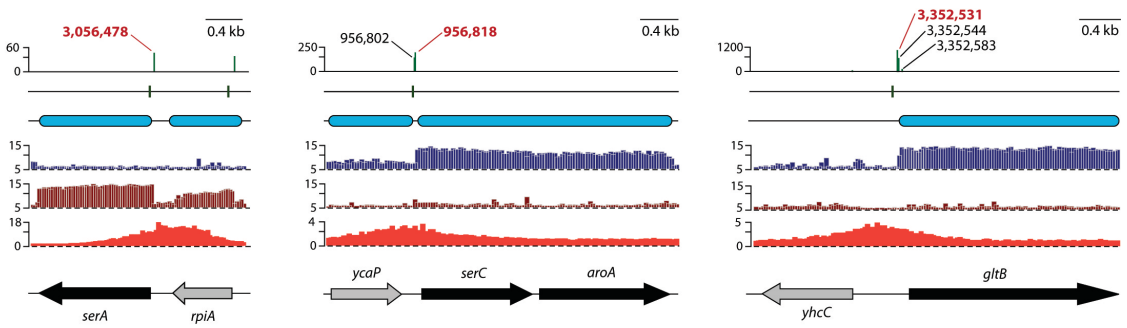


Supplementary Figure 6 | Overlapping pORFs. (a) Frequency of peptide detection in the region where overlapped pORFs were found, **(b)** Examination of translation directionality of the overlapped pORFs based on the mRNA transcript profiles. The red arrows indicate false positives that were detected as pORFs.



Supplementary Figure 7 | Number of unique peptides from pORFs with accurate and inaccurate boundaries. Among 803 pORFs mapped to the validated ORFs (from EcoGene), a total of 507 pORFs showed accurate translation start/stop positions (filled circle). pORFs with non-matching translation start positions (296 pORFs) exhibited poor peptide coverage (open circle). Due to this coverage limitation, additional methods (e.g., proteomics with N-terminal modification) have to be applied to obtain a more comprehensive and accurate ORF map at a genome-scale.

a

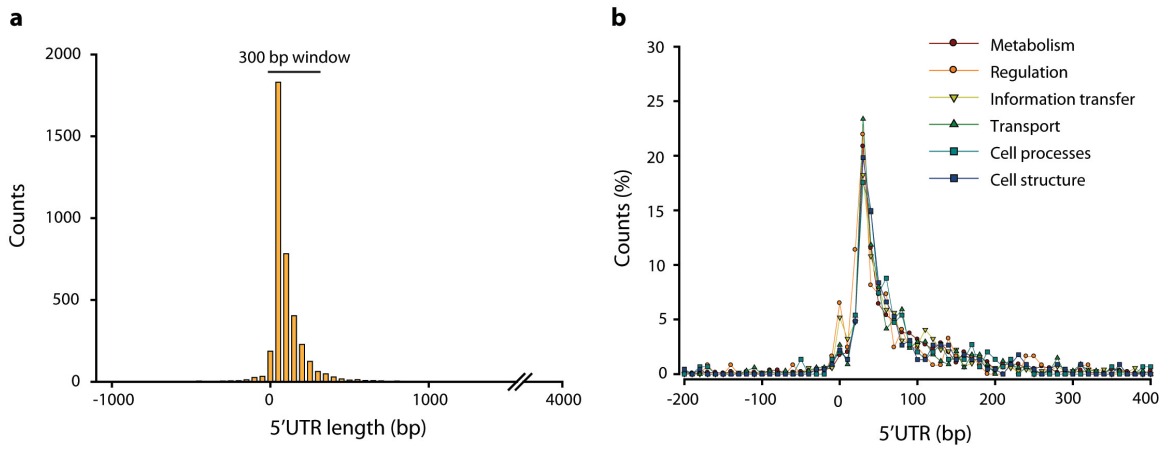


b

TU	Strand	Known TSS	Detection	Reads	Regulation	Lrp ChIP-chip
ilvIH	+	85,394		0		
ilvIH	+	85,420		0		
ilvIH	+	85,534		0		
ilvIH	+	85,597	85,597	10	+	2.167
serC-aroA	+		956,802	170		
serC-aroA	+	956,818	956,818	235	+	1.348
dadAX	+	1,236,732	1,236,732	4	-	2.628
dadAX	+	1,236,748	1,236,748	76	-	2.628
dadAX	+	1,236,761	1,236,761	11	-	2.628
stpA	-		2,796,550	230		
stpA	-	2,796,558	2,796,558	315	+	4.504
stpA	-	2,796,578	2,796,578	37		
stpA	-	2,796,600	2,796,600	6		
gcvTHP	-		3,048,789	62		
gcvTHP	-	3,048,794	3,048,794	283	+	3.495
serA	-	3,056,478	3,056,478	57	+	3.511
serA	-	3,056,571		0	-	3.511
gltBDF	+	3,352,531	3,352,531	1,100	+	1.838
gltBDF	+		3,352,544	700		
gltBDF	+		3,352,583	87		
livKHMGF	-		3,595,627	181		
livKHMGF	-		3,595,638	1,051		
livKHMGF	-	3,595,753	3,595,753	1,230		
livKHMGF	-	3,595,778	3,595,778	214	-	3.396
ilvL	+	3,948,241	3,948,241	5		
ilvL	+	3,948,313	3,948,313	1,007	+	1.710
lysU	-	4,352,820	4,352,820	160		
lysU	-	4,352,828		0	-	4.419
fimAICDFGH	+	4,540,717		0	+	3.376
fimAICDFGH	+		4,541,107	23		
osmY	+	4,609,176	4,609,176	215	-	2.797
osmY	+		4,609,257	1,213		
osmY	+		4,609,269	223		
osmY	+		4,609,356	708		
osmY	+		4,609,391	172		
livJ	-		3,597,715	103		
livJ	-	3,597,785	3,597,785	724	-	3.635

Supplementary Figure 8 | Use of alternative TSSs. **(a)** The *serA* gene, *serC-aroA* operon, and *gltBDF* operon have multiple experimentally verified TSSs. We detected the dominant TSS (3,056,478) for the *serA* promoter, which is highly activated by the transcription factor Lrp. Another experimentally confirmed TSS (3,056,571) is likely to be utilized less under this growth condition. The transcription factor Lrp also activates one experimentally verified TSS (956,818) of the *serC* promoter, which was detected as a

dominant TSS in this study. In addition, we found another TSS (956,802) at the *serC* promoter. The other previously confirmed TSS (3,352,531) at the *gltB* promoter was detected as a dominant TSS with Lrp-binding signal. **(b)** List of TSSs regulated by the transcription factor Lrp. We observed the alternative TSSs at the various promoter regions regulated by Lrp.



Supplementary Figure 9 | 5'UTR length of various functional categories. (a) Distribution of 5'UTR shows a median length maximun of ~36 bp, **(b)** comparison of 5'UTR length (in base pairs) showed no difference between different functional categories.