BMC Microbiology

Research article

Open Access

# Identification of new genes in *Sinorhizobium meliloti* using the Genome Sequencer FLX system

Chunhong Mao*, Clive Evans, Roderick V Jensen and Bruno WS Sobral

Address: Virginia Bioinformatics Institute, Virginia Polytechnic Institute and State University, Blacksburg, VA 24061, USA

Email: Chunhong Mao* - chmao@vt.edu; Clive Evans - cevans@vbi.vt.edu; Roderick V Jensen - rvjensen@vt.edu; Bruno WS Sobral - sobral@vbi.vt.edu

* Corresponding author

Published: 2 May 2008

*BMC Microbiology* 2008, **8**:72 doi:10.1186/1471-2180-8-72

Received: 9 January 2008
Accepted: 2 May 2008

This article is available from: http://www.biomedcentral.com/1471-2180/8/72

## Abstract

**Background:** *Sinorhizobium meliloti* is an agriculturally important model symbiont. There is an ongoing need to update and improve its genome annotation. In this study, we used a high-throughput pyrosequencing approach to sequence the transcriptome of *S. meliloti*, and search for new bacterial genes missed in the previous genome annotation. This is the first report of sequencing a bacterial transcriptome using the pyrosequencing technology.

**Results:** Our pilot sequencing run generated 19,005 reads with an average length of 136 nucleotides per read. From these data, we identified 20 new genes. These new gene transcripts were confirmed by RT-PCR and their possible functions were analyzed.

**Conclusion:** Our results indicate that high-throughput sequence analysis of bacterial transcriptomes is feasible and next-generation sequencing technologies will greatly facilitate the discovery of new genes and improve genome annotation.

## Background

*Sinorhizobium meliloti* is a micro-symbiont associated with legume plants. This soil bacterium inhabits nodules on the roots of host legume plants, where it reduces atmospheric nitrogen to organic nitrogenous compounds that can be utilized by its hosts. Because of its agricultural and ecological importance, *S. meliloti* has been extensively studied as a model symbiont. The *S. meliloti* 1021 genome sequence and the initial annotation of the genome were completed in 2001 [1-4]. The *S.meliloti* genome comprises three replicons, the 3.65 Mb chromosome, the 1.35 Mb megaplasmid pSymA, and the1.68 Mb megaplasmid pSymB [4]. According to RefSeq [5], the *S. meliloti* 1021 genome has 6205 predicted protein-encoding genes. Among these, more than one-third were annotated as "hypothetical" or "unknown". Many research papers have been published on *S. meliloti* since its genome sequence was completed. Also, more genomes of closely related species such as *Brucella* spp., *Rhodopseudomonas palustris*, and *S. medicae* WSM419 have been sequenced. Comparative genomics including newly sequenced genomes provides new information about the genome of *S. meliloti*. There is an ongoing need to update and improve its genome annotation. So far, there are no systematic efforts of direct sequencing of its entire transcriptome. Microarray data are available, but most microarray designs are based on annotated genes [6,7]. High-density whole-genome tiling arrays are not yet available.

The goal of this study was to develop a high-throughput experimental approach to search for new genes of *S. meliloti* missed in the previous genome annotation [1-4].

We used pyrosequencing [8] to sequence the transcriptome of *S.meliloti*. The GS FLX system from Roche and 454 Life Sciences can generate more than 100 million bases per sequencing run with an average yield of greater than 400,000 reads of average length of 250 bases. This platform provides a broad range of applications including whole genome sequencing [9-11], transcriptome and gene regulation studies [12-15], metagenomics analysis [16] and amplicon sequencing [17,18]. Although pyrosequencing has been used to sequence microbial genomes, relatively few applications of transcriptome analysis have been reported. Here, we present the first report of sequencing a bacterial transcriptome using the GS FLX platform as an experimental approach for gene discovery.

## Results

### *Gene prediction*

We used an automated gene annotation pipeline provided by PATRIC [19] to predict genes in *S. meliloti*. This pipeline uses a combination of gene prediction programs, Glimmer [20], GeneMark [21,22], TICO [23] and RBSfinder [24] to predict genes and compares with genes in RefSeq. A total of 512 new protein-coding genes (with length >90 nt) in the intergenic regions of the genome were predicted through this automated pipeline (Additional file 1). The number of predicted genes in different length ranges is shown in Figure 1. Most of the predicted new genes are relatively small (length <400 nt). The average length is about 200 nt. These genes were BLASTed against the NCBI non-redundant (NR) protein database [25,26]. The result showed that 159 candidates had BLAST hits in the NR database with E-values less than 0.01, whereas the remaining 353 of the candidates had no significant hits. Small gene size and lack of BLAST hits may be the reasons that the predicted new gene candidates were missed in the original genome annotation process.

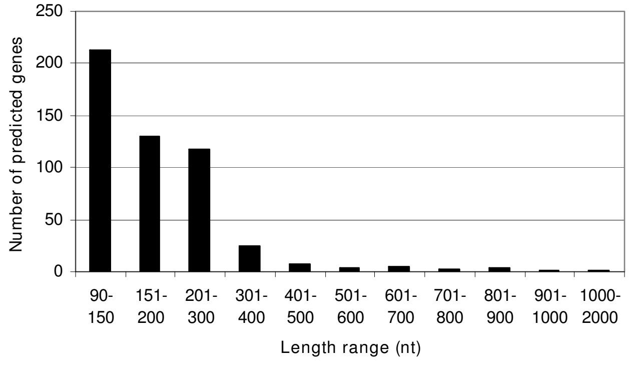

**Figure 1**
**Length distribution of predicted gene candidates.**

### *Sequence analysis*

Total RNAs were extracted from *S. meliloti* 1021 cells grown to mid-exponential phase in the TY medium and treated with DNase I to remove genomic DNA (Methods). The 16s and 23s rRNAs were depleted and the RNA samples were then amplified to produce cDNA fragments of average length about 150 nt (Methods). With two test cDNA samples loaded on to 4 lanes per sample of a 16 lane sequencing plate, the titration run generated a total of 19,005 high quality reads with average length of 136 nt (Table 1). Although our rRNA removal step indicated that more than 90% rRNAs were depleted as judged by Agilent 2100 Bioanalyzer, approximately 90% of the reads still aligned to the rRNA operons (Figure 2). This may be due to relative low mRNA population in the *S. meliloti* cells. Out of 17092 reads aligned to the rRNA operons, 3 reads matched 5s rRNA, 2860 reads matched 16s rRNA, 13983 reads matched 23s rRNA, and the remaining 246 reads aligned to the integenic regions between the rRNA genes in the rRNA operons. For the 1854 non-rRNA sequences, 1774 matched to 737 of the 6271 RefSeq genes (proteins and RNAs) and 59 matched to 32 of 512 new protein-coding genes predicted through our gene prediction pipeline. The remaining 21 sequences mainly matched sequences either immediately before or after a coding region, presumably 5' UTR or 3' UTR.

### *Validation of new genes using RT-PCR*

Twenty new gene candidates with multiple GS FLX sequence hits or long sequence hits (>80 nt) were chosen for further verification using RT-PCR analysis of the original total RNA samples. These new gene candidates were predicted by PATRIC pipeline as protein-coding genes. Figure 3 shows an example of a new gene candidate (VBISMc1000) in the *Sinorhizobium* Genome Browser, which was built using the GBrowse software [27]. All 20 genes were detected in the RT-PCR experiment and the PCR products were sequenced and confirmed (Figure 4). The negative controls indicated that there was no genomic DNA contamination in the RNA samples tested.

To test whether the new genes are co-transcribed with their upstream or downstream flanking genes (if any), a set of primer pairs were designed (Figure 5, Additional file 2) to detect RT-PCR product of the transcripts with the

**Table 1: GS FLX sequencing results**

| | Sample 1 | Sample 2 | Total |
|---|---|---|---|
| # Sequence reads | 8694 | 10311 | 19005 |
| Average sequence length | 139 | 133 | 136 |
| # Sequences aligned to genes | 1165 | 689 | 1854 |
| # Sequences in rRNA operons | 7513 | 9579 | 17092 |
| # Sequences not aligned to the genome (e<0.01) | 16 | 43 | 59 |

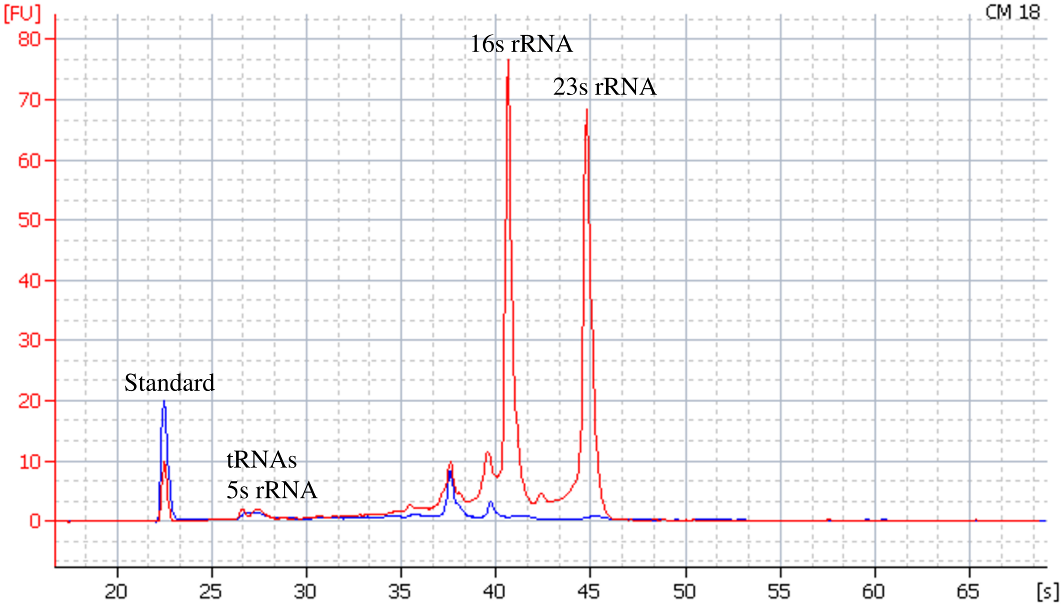

**Figure 2**
**Removal of 16s and 23s rRNAs using MICROB *Express*™ kit from Ambion**. Total RNA samples before (in red) and after (in blue) rRNA depletion were analyzed on the Agilent 2100 Bioanalyzer.

flanking genes. The results are summarized in Table 2. Columns "co-transcribed with upstream gene" and "co-transcribed with downstream gene" indicated whether the RT-PCR products of the transcripts, which span from the upstream flanking gene to the new gene or from the new gene to downstream flanking gene, were detected. Ten of the predicted genes were not detected to be co-transcribed with either upstream or downstream flanking genes (Table 2).

### *Functional annotation of the new genes*

The sequences of putative new genes were searched against the NR database from NCBI and SwissProt from EBI [28] using BLASTX [25] and Smith-Waterman programs ([29]; Table 2). Both programs produced very sim-

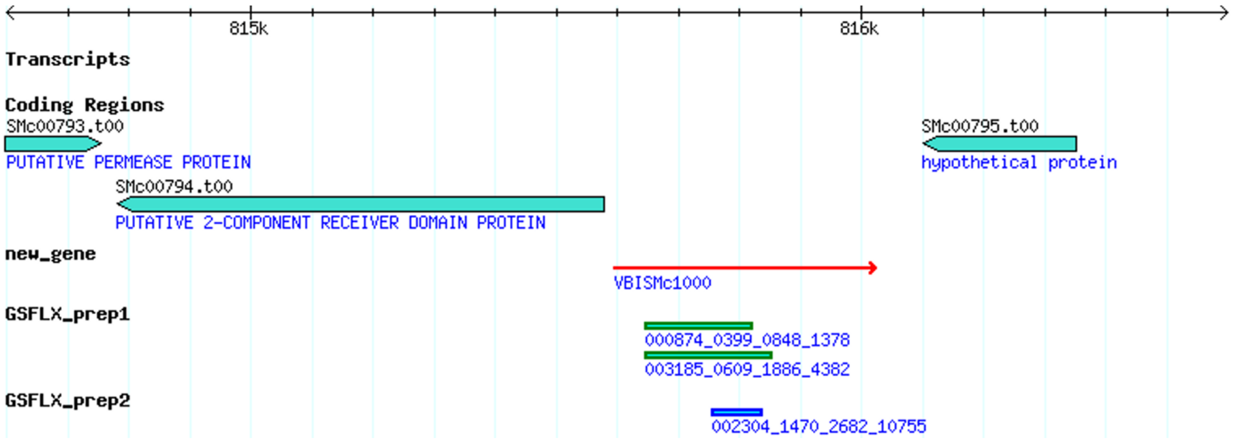

**Figure 3**
**Genome view of a new gene (in red)**. GLX sequences (in green from prep 1 and in blue from prep 2) are aligned to VBISMc1000.

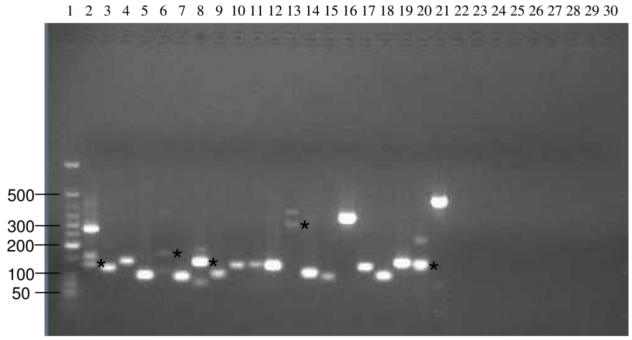

**Figure 4**

**RT-PCR of 20 gene candidates**. Lane 1: low molecular weight DNA ladder from New England Biolabs. Size range: 25 bp to 766 bp. Lane 2-21: 20 gene candidates, VBISMa0080, VBISMa0492, VBISMa1337, VBISMb0078, VBISMb0839, VBISMc0095, VBISMc0802, VBISMc1000, VBISMc1221, VBISMc1492, VBISMc1793, VBISMc2171, VBISMc2174, VBISMc2596, VBISMc2940, VBISMc2955, VBISMc3188, VBISMc3282, VBISMc4046 and VBISMc4289, respectively. For majority RT-PCR reactions, each produced one corresponding PCR product (lane 3-5, 7, 9-12, 14-19 and 21). These PCR products were directly sequenced and their sequences matched to the corresponding gene candidates. Multiple PCR products were found in lane 2, 6, 8, 13 and 20. The bands with the correct PCR product sizes are labeled with *. These PCR products were used to do a second round of PCR to produce enough DNA for sequencing. The sequencing results confirmed that they matched to the corresponding gene candidates. The most abundant PCR product in lane 2 was sequenced and determined to be a part of 23s rRNA sequence. Lane 22-29: negative controls using the RNA sample that was not reverse transcribed and primer pairs of new genes to show no genomic DNA contamination. In each lane of 22 to 29, combined primer pairs of two or three genes were used. Lane 30: no template control. Primer pairs of cm0012a, cm012b and cm016a, cm016b were used.

ilar results: 10 of the 20 new genes had no significant hits, and the other 10 had either a full length or partial match with proteins in the NR database (Table 2). The genes with no significant hits were relatively short with lengths ranging from 120 to 366 nt.

Four predicted genes showed significant matches with genes with known or putative functions (Table 2). VBISMc2940 had 89% similarity to a conserved hypothetical signal peptide protein from *S. medicae* WSM419. VBISMb0839 had 69% similary to a two component transcriptional regulator, LuxR family from *S. medicae* WSM419. VBISMa1337 partially matched to a putative dioxygenase with 25% similarity. VBISMc3282 matched to nodulation protein NolR with 59% similarity. The NolR protein is a transcriptional regulator for common nodulation genes as well as the three nodD copies present in *S. meliloti*. Previous studies have shown that nolR gene in *S. meliloti* strain 1021 has a single insertion in the C-terminal coding sequence which abolishes the DNA-binding ability of the NolR protein [30,31]. Thus, Rm1021 has no NolR activity. NolR$^-$ strains nodulate host plants less efficiently than NolR$^+$ strains. In the RefSeq database, VBISMc3282 was not previously annotated as a gene although it is a mutant form of the *nolR* gene. Here, we demonstrate and confirm that this mutant gene is expressed.

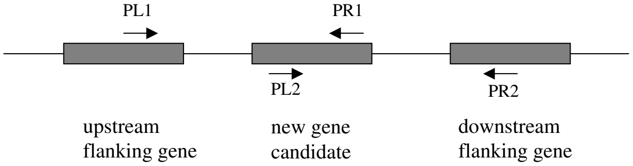

**Figure 5**

**Primer design for testing co-transcription using RT-PCR**. PL1 and PR1 are primer set for testing transcript from upstream flanking gene to new gene and PL2 and PR2 are primer set for testing transcript from new gene to downstream flanking gene.

No neighboring genes were detected to be co-transcribed with VBISMc2940, VBISMb0839 or VBISMa1337, while VBISMc3282 was co-transcribed with the downstream gene SMc01535, a hypothetical protein. The other six genes with BLAST hits matched only hypothetical proteins (Table 2).

### *Gene expression levels*

Because the RNA amplification step was linear (Methods), we expect that the cDNA samples we prepared represent relative mRNA levels in the cell. Thus, for a full sequencing run, with high coverage, the number of sequences would be a good indication of gene expression levels. Due to the low coverage of our pilot experiment, we cannot yet estimate gene expression levels based on number of sequences for each gene. However, we expect that most of the genes that showed five or more matches to our transcriptome sequences should be highly expressed in the cell population (Additional file 3). The known genes with high sequence copy number are consistent with our knowledge about the high expression level of those genes under the same growth condition (our unpublished microarray data).

## Discussion

Our study demonstrated that there are many genes missed in the initial genome annotation and it is useful to have large-scale transcriptome analysis to reveal these genes

**Table 2: Summary of new genes**

| Gene | Replicon* | Predicted start | Predicted end | Strand | Length | Co-transcribed with upstream gene | Co-transcribed with downstream gene | Target description for predicted genes | E-value | Percent similarity |
|---|---|---|---|---|---|---|---|---|---|---|
| VBISMa0080 | A | 65258 | 65494 | - | 237 | SMa0121 | - | - | - | - |
| VBISMa0492 | A | 410418 | 410660 | + | 243 | - | - | - | - | - |
| VBISMa1337 | A | 1191271 | 1191525 | - | 255 | - | - | putative dioxygenase, slightly similar to catechol 1,2-dioxygenase protein [*S. meliloti* 1021] | 2e-4 | 25 |
| VBISMb0078 | B | 66440 | 66619 | + | 180 | - | SMb20056 | hypothetical protein BMEII0534 [*B. melitensis* 16M] | 4e-3 | 30 |
| VBISMb0839 | B | 679069 | 679800 | + | 732 | - | - | two component transcriptional regulator, LuxR family [*S. medicae* WSM419] | 2e-68 | 69 |
| VBISMc0095 | C | 83906 | 84025 | + | 120 | tRNA-Ala | 23s rRNA | - | - | - |
| VBISMc0802 | C | 662922 | 663287 | + | 366 | - | - | - | - | - |
| VBISMc1000 | C | 815598 | 816023 | + | 426 | - | - | hypothetical protein Smed_0338 [*S. medicae* WSM419] | 6e-20 | 36 |
| VBISMc1221 | C | 997383 | 997565 | + | 183 | ctaE | SMc00014 | hypothetical protein Smed_0524 [*S. medicae* WSM419] | 1e-15 | 63 |
| VBISMc1492 | C | 1208848 | 1209090 | - | 243 | - | - | - | - | - |
| VBISMc1793 | C | 1444596 | 1444832 | - | 237 | - | rne | - | - | - |
| VBISMc2171 | C | 1725819 | 1726607 | - | 789 | - | SMc01204 | hypothetical protein Smed_1270 [*S. medicae* WSM419] | 8e-124 | 89 |
| VBISMc2174 | C | 1728081 | 1728221 | - | 141 | SMc01200 | SMc01202 | - | - | - |
| VBISMc2596 | C | 2060940 | 2061143 | + | 204 | - | csp4 | - | - | - |
| VBISMc2940 | C | 2312866 | 2313225 | - | 360 | - | - | conserved hypothetical signal peptide protein [*S. medicae* WSM419] | 2e-54 | 89 |
| VBISMc2955 | C | 2321074 | 2321289 | + | 216 | - | - | - | - | - |
| VBISMc3188 | C | 2520862 | 2520993 | - | 132 | - | - | - | - | - |
| VBISMc3282 | C | 2595016 | 2595537 | + | 522 | - | SMc01535 | Nodulation protein nolR | 9e-50 | 59 |
| VBISMc4046 | C | 3233965 | 3234153 | + | 189 | SMc03108 | - | hypothetical protein pRL110117 [*R. leguminosarum* bv. viciae 3841] | 1e-5 | 56 |
| VBISMc4289 | C | 3417257 | 3419215 | - | 1959 | - | - | hypothetical protein Cvib_0070 [*P. vibrioformis* DSM 265] | 4e-35 | 47 |

*: replicon A: pSymA; B: pSymB; C: Chromosome.

and validate their status. Our results showed that sequencing bacterial transcriptomes using the GS FLX system is feasible and it helps to discover new genes and improve the genome annotation. A full GS FLX sequencing run can produce an average yield of more than 400,000 reads which is 20-fold greater than the yield from our titration run for this study. Even with 90% rRNA population in the sample, there will still be more than 40,000 reads that are non-rRNA transcripts. This provides an average 6X coverage of non-rRNA genes. Our pilot experiment with only 1854 reads already identified 20 new genes. With a full sequencing run, which produces more than 20-fold reads than the titration run, we expect to discover many more new gene transcripts that have been previously missed. However, a full sequencing run with 6X coverage of non-rRNA genes will still not be sufficient to discover all possible new genes expressed, especially for ones with low expression levels, and considering that conditions under which genes are expressed may not be known or studied by any particular set of experiments. According to the previous microarray studies, about 70-80% annotated genes are expressed under the same growth conditions as used in this study ([6] and our unpublished data). Nevertheless, our study suggests two ways to improve the results: the first is to more effectively remove rRNA or use a normalized cDNA library; the second is to employ "deep" sequencing techniques, either by performing multiple GS FLX runs, or by using Illumina [32] or ABI [33] methods which produce millions of reads, but of smaller average length.

## Conclusion

Our study indicated that there are still many genes missed in the initial genome annotation of *S. meliloti*. High-throughput sequence analysis of bacterial transcriptomes is feasible for the identification of new genes. Next-generation sequencing technologies will greatly facilitate the gene discovery process and improve genome annotation.

## Methods

### *Cell culture and RNA isolation*

*Sinorhizobium meliloti* strain1021 was grown at 30°C in TY medium [34] to mid-exponential phase ($OD_{600}$ = 0.6). Cell growth was stopped by adding 1/9th volume of stop solution (5% buffer equilibrated phenol pH 7.4 in ethanol) and placed on ice. Cells were collected by centrifugation in a microcentrifuge at maximum speed for 3 minutes. The cell pellets were stored in -80°C. Total RNA was isolated by using Qiagen RNeasy bacterial RNA purification kit (Qiagen, Valencia, CA). The total RNA was treated with DNase I on mini-RNeasy column before eluted with RNase free water. For RT-PCR experiments, an additional DNase I treatment was done after RNA was eluted from the RNeasy mini column to ensure that there was no genomic DNA contamination. 20 μl of total RNAs eluted from the RNeasy mini-column were treated with 5 μl DNase I (Qiagen) in 10 μl RDD buffer, 1 μl RNase inhibitor (Invitrogen, Carlsbad, CA) and 64 μl RNase free water (Qiagen) at 25°C for 30 minutes. The RNAs were then extracted with phenol/chloroform and precipitated with ethanol using standard protocols. 16s and 23s rRNAs were then depleted using the MICROBExpress™ Bacterial mRNA Enrichment Kit (Ambion, Austin, TX). Total RNAs and rRNA depleted RNAs were quantified and analyzed on the Agilent 2100 Bioanalyzer. 7 μg of total RNA per reaction was used. After 16s and 23s rRNAs were depleted, about 0.5 μg (7%) RNAs was recovered. The total RNA samples had RNA integrity number (RIN value) of 8.0 or better. As shown in Figure 2, more than 90% 16s and 23s rRNAs were depleted. We analyzed more than 10 independent rRNA-depleted preparations on the Agilent 2100 Bioanalyzer. The results were consistent and showed that the 16s and 23s rRNA peaks were greatly reduced in these preparations but not completely removed (Figure 2). In addition, two small peaks immediately before 16s and 23s rRNAs could not be removed by the Ambion MicrobExpress kit. The two peaks were consistently present in all of our RNA preparations.

Two RNA samples were prepared for RNA amplification. Sample 1 was 16s and 23s rRNA depleted RNA sample as described above. Sample 2 was the 16s and 23s rRNA depleted RNA sample ligated to a 3' RNA adptor (5'-$PO_4$-UUCGCUGUUC UUAGCGGCCG CAUGCUC-idT-3'; idT: 3' inverted deoxythymidine) (Dharmacon Research, Lafayette, CO) and a 5' RNA adptor (5'-OH-AUGUGCGCGA CUUCCUGUAG ACGGAACGCU AGAAGAAA-OH-3') (Dharmacon Research). 3' and 5' adaptor ligations were done as described in Argaman *et al.* 2001 [35].

### *RNA amplification and cDNA preparation*

To obtain enough cDNA for sequencing, the 16s and 23s rRNA depleted RNAs (sample 1 and 2) were amplified using Nugen WT-Ovation Pico RNA amplification system [36]. 5 ng of starting RNA was used. The SPIA™ amplified single strand cDNA (2.5 ug) was then taken through a second strand cDNA synthesis, using the following conditions: 5X 2nd strand reaction mix 30 μl (Invitrogen), dNTP, 10 mM 3 μl (Invitrogen), *E. coli* DNA ligase 1 μl (Invitrogen), *E. coli* DNA polymerase I 4 μl (Invitrogen), RNase H 1 μl (Invitrogen), RNase-free water 91 μl (Ambion). The reaction mix was incubated at 16°C for 2 hours. The cDNA was then purified using the Qiagen PCR clean up kit resulting 4 ug of cDNA quantified by using the Nanodrop spectrophotometer. 1 ug of cDNA of each sample was size selected (>100 bp) using Roche's GS FLX library Preparation Guide recommendations (no nebulization was necessary due to the size range of the cDNA GS FLX library Preparation Guide), and a single stranded library was created.

### *GS FLX sequencing and data filtering*

The DNA sequencing libraries for the two samples were combined with the sequencing beads in 4 different concentrations to determine the optimal conditions for emPCR amplification. All 8 preparations were sequenced in 8 lanes of a GS FLX sequencing plate using the standard Roche/454 protocols. Sequencing data was obtained after a 7 hour run on the GS FLX. The 54,162 raw reads from GS FLX sequencing run that passed the sample key code filter (initial bases TCAG) were further filtered by the 454 software to eliminate 10,521 mixed reads (with two or more different DNA strands/bead), 15,604 excessively short reads (less than about 50 bp), and 9,032 interrupted reads ("dots"). 35% of the raw reads passed all filters in this titration run to provide the 19,005 high quality reads used in this study.

### *Sequence analysis and mapping to* S. meliloti *genome*

GS FLX sequences passed filtering criteria were BLASTN-aligned to *S. meliloti* genome. Sequences with matching to rRNA operons were filtered. The remaining sequences were BLASTed against RefSeq genes and our predicted new genes from our gene annotation pipeline.

### *RT-PCR*

5 μg DNase I treated total RNA was reverse transcribed using superscript II with 4 pmoles of equally mixed gene-specific primers for each candidate gene selected (cm012b-cm0031b, Table S1). Primers were designed using Primer3 [37]. For PCR, each 40 μl reaction includes 0.5 μl of 40 μl reverse transcription reaction, 20 μl of 2X GoTaq Green Master Mix (Promaga, Madison, WI) and 0.5 μM of primer pair of each gene (IDT, Coralville, IA). PCR conditions were 95°C 2 min, 30 cycles of 95°C 45 s,

52°C 45 s, 72°C 60 s, and a final cycle of 72°C for 10 min. PCR products were examined by electrophoresis in a 2.5% agarose/TAE/EtBr gel. Sequencing was performed using the BigDye Terminator Cycle Sequencing Kit (Applied Biosystems) and analyzed on an Applied Biosystems model 3730 automated capillary DNA sequencer.

## Authors' contributions

CM designed and executed the experiments, performed data analysis and drafted the manuscript. CE developed the RNA amplification method and performed RNA amplification and cDNA sample preparation for GS FLX sequencing. RVJ helped with data analysis. CE and RVJ wrote portions of the Methods. BWSS and RVJ critically edited and revised the manuscript. BWSS provided funding, coordination and oversight of the project. All authors read and approved the final manuscript.

## Additional material

### Additional file 1

*Supplementary Table S1. New protein-coding genes predicted by PATRIC that are located in the intergenic regions*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2180-8-72-S1.xls]

### Additional file 2

*Supplementary Table S2. Primers*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2180-8-72-S2.xls]

### Additional file 3

*Supplementary Table S3. Genes with high copy number of GS FLX sequences*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2180-8-72-S3.xls]

## Acknowledgements

We thank Timothy Driscoll for helpful discussions, Jessica Kraszewski for assistance in the GS FLX runs and Chunxia Wang for providing microarray data. This work was funded by the Commonwealth Research Initiative (CRI) from the Commonwealth of Virginia and by the Virginia Bioinformatics Institute.

## References

1. Barnett MJ, Fisher RF, Jones T, Komp C, Abola AP, Barloy-Hubler F, Bowser L, Capela D, Galibert F, Gouzy J, Gurjal M, Hong A, Huizar L, Hyman RW, Kahn D, Kahn ML, Kalman S, Keating DH, Palm C, Peck MC, Surzycki R, Wells DH, Yeh KC, Davis RW, Federspiel NA, Long SR: **Nucleotide sequence and predicted functions of the entire Sinorhizobium meliloti pSymA megaplasmid.** *Proc Natl Acad Sci USA* 2001, **98(17):**9883-9888.
2. Capela D, Barloy-Hubler F, Gouzy J, Bothe G, Ampe F, Batut J, Boistard P, Becker A, Boutry M, Cadieu E, Dreano S, Gloux S, Godrie T, Goffeau A, Kahn D, Kiss E, Lelaure V, Masuy D, Pohl T, Portetelle D, Puhler A, Purnelle B, Ramsperger U, Renard C, Thebault P, Vandenbol M, Weidner S, Galibert F: **Analysis of the chromosome sequence of the legume symbiont Sinorhizobium meliloti strain 1021.** *Proc Natl Acad Sci USA* 2001, **98(17):**9877-9882.
3. Finan TM, Weidner S, Wong K, Buhrmester J, Chain P, Vorholter FJ, Hernandez-Lucas I, Becker A, Cowie A, Gouzy J, Golding B, Puhler A: **The complete sequence of the 1,683-kb pSymB megaplasmid from the N2-fixing endosymbiont Sinorhizobium meliloti.** *Proc Natl Acad Sci USA* 2001, **98(17):**9889-9894.
4. Galibert F, Finan TM, Long SR, Puhler A, Abola P, Ampe F, Barloy-Hubler F, Barnett MJ, Becker A, Boistard P, Bothe G, Boutry M, Bowser L, Buhrmester J, Cadieu E, Capela D, Chain P, Cowie A, Davis RW, Dreano S, Federspiel NA, Fisher RF, Gloux S, Godrie T, Goffeau A, Golding B, Gouzy J, Gurjal M, Batut J: **The composite genome of the legume symbiont Sinorhizobium meliloti.** *Science* 2001, **293(5530):**668-672.
5. Pruitt KD, Tatusova T, Maglott DR: **NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins.** *Nucleic Acids Res* 2007, **35(Database issue):**D61-5.
6. Barnett MJ, Toman CJ, Fisher RF, Long SR: **A dual-genome Symbiosis Chip for coordinate study of signal exchange and development in a prokaryote-host interaction.** *Proc Natl Acad Sci USA* 2004, **101(47):**16636-16641.
7. Ruberg S, Tian ZX, Krol E, Linke B, Meyer F, Wang Y, Puhler A, Weidner S, Becker A: **Construction and validation of a Sinorhizobium meliloti whole genome DNA microarray: genome-wide profiling of osmoadaptive gene expression.** *J Biotechnol* 2003, **106(2-3):**255-268.
8. Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, Berka J, Braverman MS, Chen YJ, Chen Z, Dewell SB, Du L, Fierro JM, Gomes XV, Godwin BC, Begley RF, Rothberg JM: **Genome sequencing in microfabricated high-density picolitre reactors.** *Nature* 2005, **437(7057):**376-380.
9. Green RE, Krause J, Ptak SE, Briggs AW, Ronan MT, Simons JF, Du L, Egholm M, Rothberg JM, Paunovic M, Paabo S: **Analysis of one million base pairs of Neanderthal DNA.** *Nature* 2006, **444(7117):**330-336.
10. Pearson BM, Gaskin DJ, Segers RP, Wells JM, Nuijten PJ, van Vliet AH: **The complete genome sequence of Campylobacter jejuni strain 81116 (NCTC11828).** *J Bacteriol* 2007, **189(22):**8402-8403.
11. Poinar HN, Schwarz C, Qi J, Shapiro B, Macphee RD, Buigues B, Tikhonov A, Huson DH, Tomsho LP, Auch A, Rampp M, Miller W, Schuster SC: **Metagenomics to paleogenomics: large-scale sequencing of mammoth DNA.** *Science* 2006, **311(5759):**392-394.
12. Bainbridge MN, Warren RL, Hirst M, Romanuik T, Zeng T, Go A, Delaney A, Griffith M, Hickenbotham M, Magrini V, Mardis ER, Sadar MD, Siddiqui AS, Marra MA, Jones SJ: **Analysis of the prostate cancer cell line LNCaP transcriptome using a sequencing-by-synthesis approach.** *BMC Genomics* 2006, **7:**246.
13. Berezikov E, Thuemmler F, van Laake LW, Kondova I, Bontrop R, Cuppen E, Plasterk RH: **Diversity of microRNAs in human and chimpanzee brain.** *Nat Genet* 2006, **38(12):**1375-1377.
14. Emrich SJ, Barbazuk WB, Li L, Schnable PS: **Gene discovery and annotation using LCM-454 transcriptome sequencing.** *Genome Res* 2007, **17(1):**69-73.
15. Gowda M, Li H, Alessi J, Chen F, Pratt R, Wang GL: **Robust analysis of 5'-transcript ends (5'-RATE): a novel technique for transcriptome analysis and genome annotation.** *Nucleic Acids Res* 2006, **34(19):**e126.
16. Krause L, Diaz NN, Bartels D, Edwards RA, Puhler A, Rohwer F, Meyer F, Stoye J: **Finding novel genes in bacterial communities isolated from the environment.** *Bioinformatics* 2006, **22(14):**e281-9.
17. Sogin ML, Morrison HG, Huber JA, Mark Welch D, Huse SM, Neal PR, Arrieta JM, Herndl GJ: **Microbial diversity in the deep sea and the underexplored "rare biosphere".** *Proc Natl Acad Sci USA* 2006, **103(32):**12115-12120.
18. Taylor KH, Kramer RS, Davis JW, Guo J, Duff DJ, Xu D, Caldwell CW, Shi H: **Ultradeep bisulfite sequencing analysis of DNA methylation patterns in multiple gene promoters by 454 sequencing.** *Cancer Res* 2007, **67(18):**8511-8518.
19. Snyder EE, Kampanya N, Lu J, Nordberg EK, Karur HR, Shukla M, Soneja J, Tian Y, Xue T, Yoo H, Zhang F, Dharmanolla C, Dongre NV, Gillespie JJ, Hamelius J, Hance M, Huntington KI, Jukneliene D, Setubal

JC, Sobral BW: **PATRIC: the VBI PathoSystems Resource Integration Center.** *Nucleic Acids Res* 2007, **35(Database issue):**D401-6.

20. Delcher AL, Harmon D, Kasif S, White O, Salzberg SL: **Improved microbial gene identification with GLIMMER.** *Nucleic Acids Res* 1999, **27(23):**4636-4641.
21. Borodovsky M, McIninch J: **Recognition of genes in DNA sequence with ambiguities.** *Biosystems* 1993, **30(1-3):**161-171.
22. Lukashin AV, Borodovsky M: **GeneMark.hmm: new solutions for gene finding.** *Nucleic Acids Res* 1998, **26(4):**1107-1115.
23. Tech M, Pfeifer N, Morgenstern B, Meinicke P: **TICO: a tool for improving predictions of prokaryotic translation initiation sites.** *Bioinformatics* 2005, **21(17):**3568-3569.
24. Suzek BE, Ermolaeva MD, Schreiber M, Salzberg SL: **A probabilistic method for identifying start codons in bacterial genomes.** *Bioinformatics* 2001, **17(12):**1123-1130.
25. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.** *Nucleic Acids Res* 1997, **25(17):**3389-3402.
26. **NCBI** [http://www.ncbi.nlm.nih.gov]
27. Stein LD, Mungall C, Shu S, Caudy M, Mangone M, Day A, Nickerson E, Stajich JE, Harris TW, Arva A, Lewis S: **The generic genome browser: a building block for a model organism system database.** *Genome Res* 2002, **12(10):**1599-1610.
28. **SwissProt** [http://www.ebi.ac.uk/swissprot]
29. Smith TF, Waterman MS: **Identification of common molecular subsequences.** *J Mol Biol* 1981, **147(1):**195-197.
30. Cren M, Kondorosi A, Kondorosi E: **An insertional point mutation inactivates NolR repressor in Rhizobium meliloti 1021.** *J Bacteriol* 1994, **176(2):**518-519.
31. Wais RJ, Wells DH, Long SR: **Analysis of differences between Sinorhizobium meliloti 1021 and 2011 strains using the host calcium spiking response.** *Mol Plant Microbe Interact* 2002, **15(12):**1245-1252.
32. **Illumina** [http://www.illumina.com]
33. **ABI** [http://www.appliedbiosystems.com]
34. Beringer JE: **R factor transfer in Rhizobium leguminosarum.** *J Gen Microbiol* 1974, **84(1):**188-198.
35. Argaman L, Hershberg R, Vogel J, Bejerano G, Wagner EG, Margalit H, Altuvia S: **Novel small RNA-encoding genes in the intergenic regions of Escherichia coli.** *Curr Biol* 2001, **11(12):**941-950.
36. **Nugen** [http://www.nugeninc.com]
37. Rozen S, Skaletsky H: **Primer3 on the WWW for general users and for biologist programmers.** *Methods Mol Biol* 2000, **132:**365-386.