Published in final edited form as:

Curr Protoc Mol Biol. 2018 October ; 124(1): e69. doi:10.1002/cpmb.69.

Whole Genome Next Generation Sequencing Mutation Identification in *Pseudomonas aeruginosa*

Murat Cetinbas^{1,2}, Shen Yu^{1,2}, and Ruslan I. Sadreyev^{1,3,*}

¹Department of Molecular Biology, Massachusetts General Hospital, Boston, MA 02114

²Department of Genetics, Harvard Medical School, Boston, MA 02115

³Department of Pathology, Harvard Medical School, Boston, MA 12115

Abstract

Identification of spontaneous or chemically-induced bacterial mutations is a powerful tool for investigation of molecular mechanisms, including the mechanism of action of novel antibiotics. However, a major bottleneck to this approach has been the identification of the causative mutation underlying a phenotype of interest. Until recently, this has required time-consuming genetic analysis. However, the advent of relatively inexpensive and rapid next-generation sequencing (NGS) technologies have revolutionized the correlation of bacterial phenotypes and genotypes. In this Unit we describe a simple bioinformatics pipeline to identify differences between sequenced bacterial genomes. We also describe the procedures involved in growing, extracting and purifying DNA, and preparation of sequencing libraries for one bacterial species, *Pseudomonas aeruginosa*. Similar protocols will be applicable to other bacterial species.

Keywords

Next-generation sequencing; bacterial mutations; ethyl methanesulfonate mutagenesis; spontaneous mutation detection

STRATEGIC PLANNING

This unit describes a bioinformatics pipeline to identify differences in the NGS data of a reference bacterial genome and a “mutated” genome derived from the reference genome. These differences can then be used to in turn identify mutations underlying a phenotype of interest. The simplest application of this technology is for the characterization of spontaneous bacterial mutants that have been identified on the basis of a particular genetic screen or selection. In this case, if a relatively small number of genes can be mutated to cause a particular phenotype, *e.g.*, resistance to a particular antibiotic, and if the rate of mutation to a resistant phenotype of each of the relevant genes is approximately the same, it is only necessary to sequence the genomes of relatively few independent mutants. For example, in the particular case on which the protocol described in this Unit is based, 23 independent *Pseudomonas aeruginosa* mutants that are resistant to a new investigational

Corresponding author: Ruslan I. Sadreyev, sadreyev@molbio.mgh.harvard.edu.

antibiotic were sequenced, and 21 of the 23 mutants contained a mutation in a particular *P. aeruginosa* gene, thereby identifying the mechanism underlying resistance to the antibiotic in question. A similar analysis would be involved if a phenotypic screen required the use of a mutagen such as ethyl methanesulfonate (EMS) to increase the mutation to rate. In this case, however, because, a large number of EMS-induced background mutations are expected to be present in the genomes of each bacterial mutant identified in the screen that are unrelated to the phenotype of interest, it may be necessary to sequence more independent mutants to identify the particular mutations that cause the phenotype of interest. In either of these two cases, however, it is advisable to use an independent genetic and/or biochemical test to confirm that mutations identified in a particular gene are the cause of the relevant phenotype. For example, the mutants could be complemented with the wild-type gene or mutations in the identified gene can be generated using a variety of techniques.

Importantly, because of the relatively small size of bacterial genomes, the sequencing samples can be multiplexed to a large degree. The number of samples per sequencing run depends on the genome size and required sequencing depth. Table 1 shows estimates, by bacterial species, of the number of samples that can be multiplexed in a typical run of an Illumina MiSeq instrument that produces approximately 30 million reads per run, assuming various sequencing depths between 10x and 100x.

Because each bacterial species has particular growth requirements and optimal methods for DNA extraction, the “wet lab” aspects of this protocol will vary from species to species. The specific protocol described below was used for the preparation of NGS sequencing libraries for *P. aeruginosa*.

BASIC PROTOCOL 1. Construction of NGS sequencing libraries corresponding to bacterial mutants

Basic Protocol 1 was designed for *P. aeruginosa*. The DNA isolation part of the protocol may have to modified depending on the bacterial species. However, once isolated DNA has been purified, the “wet-lab” protocol described here can be used for DNA from any bacterial species as can the bioinformatics pipeline described in Basic Protocol 2, which is designed to identify differences between DNA isolated from mutants and a reference genome.

Materials

- Isopropanol
- Ethanol
- Bacteria Growth Media such as the Luria-Bertani (LB) medium
- Bacto Agar (BD)
- Appropriate Antibiotics for Selection of Bacterial Mutants
- Phosphate Buffered Saline (PBS, Gibco)
- Tris-EDTA (TE) Buffer (Sigma Aldrich)

NEBNext Ultra II DNA Library Prep Kit (NEB E7645)
NEBNext Multiplex oligos for Illumina (NEB E7335, E7500)
Covaris S2 Sonicator
microTUBE AFA Fiber Crimp-Cap 6x16mm (Covaris)
Covaris microTUBE holder S 5001144
AMPure XP Beads (Agencourt)
Magnetic stand
55°C water bath
37°C incubator or water bath
NanoDrop
Qubit Fluorometer (Life Technologies)
Qubit dsDNA HS Assay kit (Life Technologies)
Benchtop centrifuge
Thermal cycler
Heat block

Isolation of mutant DNA samples

- 1 Streak individual putative mutants as well as the wild-type strain from which the mutants were derived on LB agar plates to obtain isolated colonies and incubate the plates overnight at 37°C.

For example, in the experiment on which this protocol is based, 23 putative *P. aeruginosa* antibiotic-resistant mutants were streaked on LB agar (without antibiotics) and the next day individual colonies were inoculated into 5 ml of LB broth and grown overnight at 37°C with vigorous aeration. 1 ml of each culture was pelleted by centrifugation and washed twice in PBS. The pelleted cells were resuspended in PBS and then serially diluted in PBS to a titer of approximately 1×10^9 per ml (expected mutational frequency $\sim 1 \times 10^7$) and plated with inhibitory concentrations of the investigational antibiotic. The agar plates were incubated at 37°C for ~20 hours until colonies appeared. Two colonies from each culture were picked and re-streaked on fresh antibiotic plates to verify resistance to the antibiotic. One resistant colony from each original culture, as well as a colony of wild-type *P. aeruginosa*, was inoculated into 5 ml LB broth for genomic DNA isolation as described in the next step.

- 2 Inoculate each mutant (from an isolated colony) into 5 ml LB broth in a test tube and aerate overnight vigorously at 37°C.
- 3 Isolate genomic DNA according to the Qiagen DNeasy Blood & Tissue Kit protocol manual.

For *P. aeruginosa*, $\sim 2 \times 10^9$ cells were pelleted as the maximum starting material for DNA preparation. The lysate would otherwise be too viscous, thereby reducing the yield. The kit on average generates 200 μl of genomic DNA at a concentration 50–100 ng/ μl , corresponding to 10–20 μg of DNA for each mutant. Genomic DNA from wild-type *P. aeruginosa* as well as 23 randomly selected mutants were used to construct a multiplexed library for NGS.

Shearing of genomic DNA—Steps 4–9 are carried out to fragment the genomic DNA to the desired size, depending on the sequencing method of choice. For paired-end 150 bp runs, the target insert size should be around 300 bp. The steps below are based on a specific instrument (Covaris S2 sonicator). Please refer to the manual of your instrument to set the correct parameters to obtain desired fragment size.

Note: While longer read lengths have the advantage of facilitating mapping to the genome, the NGS sequencing reagents are more expensive.

- 4 Fill the tank of a Covaris S2 sonicator with deionized water and turn on the sonicator and water cooler. Ensure the water cooler is set to 4°C. Degas for at least 30 min before beginning sonication.
- 5 Dilute 1.5 μg of genomic DNA to a final volume of 55 μl using TE.
- 6 Transfer the 55 μl DNA sample to a Covaris 6x16 mm microTUBE with AFA fibre and crimp cap. Remove any bubbles by pipetting up and down.
- 7 Mount the sample tube into a microtube holder S 500114, and place the holder in the water tank.
- 8 Sonicate the sample using the following settings:
 - Intensity: 5
 - Duty Cycle: 10%
 - Cycles/burst: 200
 - Time (sec): 120
 - Number of cycles: 2.
 - Power mode: Frequency sweeping.
 - Degassing mode: Continuous.
- 9 Remove the holder and retrieve the tube. Transfer the sonicated DNA solution to a fresh 1.5 ml Eppendorf tube. Store the sonicated DNA at -20°C .

Library preparation and validation—The following protocol is for use with the NEBNext Ultra II DNA library prep kit for Illumina. This kit contains end repair reagents and adapters for Illumina sequencing and is a one-kit solution for preparing DNA for NGS library construction directly from fragmented gDNA. The oligos for adaptor ligations are included in the kit but the oligos for PCR enrichment of ligated fragments need to be purchased from NEB or from a separate oligo supplier.

The NEBNext Ultra II DNA library prep kit protocol has detailed instructions on multiple steps of the preparation of the sequencing library, including end preparation, adaptor ligation, size selection, PCR enrichment and clean up. The user starts from fragmented gDNA and winds up with individual tubes of bead-purified samples that are ready to be pooled and QCed for NGS. Steps 10–36 below briefly describe the steps based on the NEBNext Ultra II DNA library prep kit to provide readers an understanding of the process. The vendor-supplied protocol should be followed to ensure successful construction of the NGS library. The workflow of the NEB kit, which explains the concepts visually, can be found at this link: <https://www.neb.com/applications/library-preparation-for-next-generation-sequencing/illumina-library-preparation/nebnext-ultra-ii-dna-library-prep/nebnext-ultra-ii-dna-library-prep>

End Preparation—Steps 10–11 prepare the fragmented input DNA for subsequent steps by end repair, 5′-phosphorylation, and dA-tailing.

- 10 To 50 µl of sheared DNA add the following and mix thoroughly by pipetting:
 - 7 µl End prep reaction buffer
 - 3 µl End prep enzyme mix
- 11 Incubate at 20°C for 30 minutes, followed by 65°C for 30 minutes, and hold at 4°C.

Ligate adaptors—Steps 12–15 ligate a universal oligo to the 5′-end of each fragment and a different universal oligo to the 3′-end of each fragment. This allows the ligated fragments to be PCR amplified and bar-coded for multiplexing.

- 12 To the 60 µl of end prepared DNA fragments from step 11, add the following and mix by pipetting:
 - 30 µl Ligation Master Mix
 - 1 µl Ligation Enhancer
 - 2.5 µl Adaptor for Illumina
- 13 Incubate at 20°C for 15 min.
- 14 Add 3 µl of USER (Uracil-Specific Excision Reagent) enzyme and mix by pipetting up and down.
- 15 Incubate at 37°C for 15 min.

Size selection using AMPure XP beads—Using AMPure XP beads, steps 16–26 purify the adapter-ligated DNA fragments and fragments of the desired size are enriched.

- 16 According to the vendors recommendations based on the desired insert size, add the appropriate amount (30 µl for an expected insert size of 370 bp) of resuspended AMPure XP beads to the ligated DNA fragments (96.5 µl). Mix well by vortexing or by pipetting up and down at least 10 times.
- 17 Incubate for 5 min at room temperature.

- 18 Place the tubes on a magnetic stand. Allow the sample to stand until the solution is clear (at least 5 min). Carefully transfer the supernatant to a fresh tube, avoiding transfer of any beads. Discard the beads (do not discard supernatant).
- 19 Add 15 μ l (or appropriate vendor recommended amount based on insert size) of resuspended AMPure XP beads to the supernatant. Mix well by vortexing or by pipetting up and down at least 10 times.
- 20 Incubate for 5 min at room temperature.
- 21 Place the tubes on a magnetic stand. Allow the sample to stand until the solution is clear (at least 5 min). Carefully remove the supernatant, avoiding removal of any beads. Discard the supernatant (do not discard beads).
- 22 Keeping the tube on the magnetic stand, add 200 μ l freshly prepared 80% ethanol to the beads, incubate for 30 seconds at room temperature and then carefully remove and discard the supernatant.
- 23 Repeat step 22 once.
- 24 Keeping the tube on the magnetic stand, air dry the bead pellet for 5 min.
- 25 To elute the DNA fragments, add 17 μ l 0.1X TE or sterile water to the beads. Mix well on a vortex mixer or by pipetting up and down at least 10 times. Return the sample to the magnetic stand and allow the sample to stand until the solution is clear (at least 1 min).
- 26 Carefully transfer 15 μ l of the supernatant to a fresh PCR tube.

PCR enrichment of adaptor ligated DNA fragments—Steps 27–28 amplify the DNA samples via PCR. Using multiplexing primers, DNA fragments from each individual sample are also tagged with a barcode that allows the different sample to be deconvoluted later to allow simultaneous sequencing of multiplexed genomes.

- 27 To 15 μ l of size-selected ligated DNA fragments from step 26 add the following:
 - 5 μ l Universal PCR primer (10 μ M)
 - 5 μ l Index PCR primer (10 μ M)
 - 25 μ l NEBNext Ultra II Q5 PCR master mix.

A different index primer should be used for each sample for multiplexing. NEB has four sets of multiplexing primers, allowing multiplexing up to 48 samples. For higher numbers, a 96-sample set is also available. Based on our estimates, one MiSeq run can accommodate 15–30 *P. aeruginosa* genomes (6 Mb each), and our experiment verified that one MiSeq run can sufficiently identify mutations in 23 mutant *P. aeruginosa* genomes and one wild-type genome.

- 28 Amplify using the following PCR program:

- 1 cycle: 30 sec 98°C
- 8 cycles: 10 sec 98°C
- 75 sec 65°C

1 cycle: 5 min 65°C

Purify amplified fragments with AMPure XP beads—Steps 29–36 recover DNA samples after PCR amplification. The AMPure XP beads allow recovery at a very high yield.

- 29 Add 45 µl (0.9X) resuspended AMPure XP beads to the completed PCR reaction. Mix well by with a vortex mixer or by pipetting up and down at least 10 times.
- 30 Incubate for 5 min at room temperature.
- 31 Place the tubes on a magnetic stand. Allow the sample to stand until the beads have separated from the solution, and the solution is clear (at least 5 min). Carefully remove the supernatant, avoiding removal of any beads. Discard the supernatant. Do not discard beads!
- 32 Keeping the tube on the magnetic stand, add 200 µl freshly prepared 80% ethanol to the beads, incubate for 30 seconds at and then carefully remove and discard the supernatant.
- 33 Repeat step 32 once.
- 34 Keeping the tube on the magnetic stand, air dry the bead pellet for 5 min.
- 35 To elute the DNA fragments, add 33 µl of 0.1X TE to the beads. Mix well on a vortex mixer or by pipetting up and down at least 10 times. Return the sample to the magnetic stand and allow the sample to stand until the solution is clear (at least 1 min).
- 36 Carefully transfer 30 µl of the supernatant to a fresh tube.

Library validation and quantification—In this protocol libraries are validated and quantified by gel electrophoresis and Qubit assay before pooling. More accurate estimation of the size and concentration of each library can be performed using the Agilent Bioanalyzer and a quantitative PCR assay, but is not essential at this step.

- 37 Analyze 2 µl of the purified library by electrophoresis on a 2.5% agarose gel. The library should run as a tight smear with most fragments 350–400 bp in size. Estimate the average size of each library by comparison to an appropriate DNA ladder.
- 38 Quantify the DNA in 1 µl of the purified library using the Qubit dsDNA HS assay kit.

Pooling libraries

- 39 Using the estimated average library size, calculate the approximate molar mass of each library with the following formula:

$$\text{Library molar mass g/mol} = \text{Library average size (bp)} \times 660 (\text{g bp}^{-1} \text{mol}^{-1})$$

- 40 Calculate the molar concentration of each sample using the following formula:

$$\text{Library concentration (nM)} = \text{Library concentration (pg/ml)} / \text{Library molar mass g/mol}$$

- 41** Pool libraries for multiplexed sequencing. Mix an appropriate volume of each sample and dilute with TE to generate a solution containing an equimolar concentration of each library, a total DNA concentration of 20 nM, in a total volume of 20–100 µl.

The number of libraries to be pooled depends on the sequencing technology being used and the desired depth of sequence coverage. Take care to ensure that each library in the pool has been generated using a different index primer during PCR enrichment to allow demultiplexing.

Accurate quantification of the pooled sample with Agilent Bioanalyzer and quantitative PCR assays is recommended before sequencing. Check whether these are performed as part of the sequencing service you are using.

- 42** Submit pooled sample for high throughput sequencing.

BASIC PROTOCOL 2. Analysis of next-generation sequencing data

The analysis reports all protein coding missense variants between input sequencing samples and a reference genome. To perform this analysis and discover mutational variants in bacterial genomes, we constructed a streamlined version of a computational pipeline based on a previously published pipeline, SnpMap (Lehrbach et al. 2017) and its online version CloudMap (Afgan et al. 2016; Minevich et al. 2012). This streamlined version follows exactly the same steps, but it is shortened for the particular application of comparing bacterial genomes. As an example, we show an application to the *P. aeruginosa* genome (see the test data). However the same pipeline can easily be applied to other bacterial species.

In order to install and run this pipeline, the user needs to have an access and basic familiarity with Unix/Linux command-line interface, and to download and install publicly available computational tools listed below.

Necessary Resources

BWA: <http://bio-bwa.sourceforge.net> (read mapper)

Picard: <http://broadinstitute.github.io/picard> (mapped read processing and manipulation)

GATK: <https://www.broadinstitute.org/gatk> (variant caller)

SNPeff: <http://snpeff.sourceforge.net> (variant processing and annotation)

SnpSift: <http://snpeff.sourceforge.net/SnpSift.html> (variant annotation)

SnpMap_bacteria: https://github.com/MolBioBioinformatics/SnpMap_bacteria (pipeline wrapper)

Step by step Implementation

1. As a preliminary step, download and install the required software components listed above if they are not already installed in your system.
2. Download the FNA file for the reference genome of interest, for example: http://pseudomonas.com/downloads/pseudomonas/pgd_r_17_2/Pseudomonas_aeruginosa_UCBPPA14_109/Pseudomonas_aeruginosa_UCBPP-PA14_109.fna
3. By using the index utility of BWA (see manual at <http://bio-bwa.sourceforge.net/bwa.shtml>), create an alignment index for the reference genome.
4. *(Optional)* Test the pipeline installation by downloading the test dataset (https://github.com/MolBioBioinformatics/SnpMap_bacteria) and running the pipeline on the small test input FASTQ files sample_R1.fastq and sample_R2.fastq using this command:

```
erl SnpMap_short.pl sample paired
    /pathtoGenome/Pseudomonas_aeruginosa_ucbpp_pa14.fasta
```

where *pathtoGenome* is the path to the directory with the reference genome FASTA sequence file, for example, /usr/data/genomes/. Upon successful execution of the pipeline, the user can verify that the produced output files are the same as the expected output files provided in the test dataset.

5. Make sure that the names of the input FASTQ files of interest follow this naming convention: SampleName_R1.fastq and, in the case of paired-end reads, SampleName_R2.fastq, where “SampleName” is any text string identifying this sequencing sample.

In the case of one FASTQ file with single-end reads, run this single-line command:

```
perl SnpMap.pl SampleName single-end /pathtoGenome/genome.fasta,
```

where *SampleName* is the name of the sample matching the file name SampleName_R1.fastq, and *pathtoGenome/genome.fasta* is the path to the reference genome FASTA sequence file, for example /usr/data/genomes/*Pseudomonas_aeruginosa_ucbpp_pa14.fasta*.

In the case of two FASTQ file with paired-end reads, run this single-line command:

```
perl SnpMap.pl SampleName paired-end pathtoGenome/genome.fasta,
```

where *SampleName* is the name of the sample matching the file names *SampleName_R1.fastq* and *SampleName_R2.fastq*, and *pathToGenome/genome.fasta* is the path to the reference genome FASTA sequence file, for example */usr/data/genomes/Pseudomonas_aeruginosa_ucbpbp_pa14.fasta*.

This one-line Perl command will run the variant discovery analysis in 5 steps using publicly available tools described above:

- a. Reads are mapped to the reference genome by BWA aligner.
- b. Picard is used to add group information, sort and index aligned reads.
- c. GATK is used to define intervals for local realignment and perform local realignment of reads around indels.
- d. Picard is used to mark duplicate reads.
- e. Genome variants are identified using GATK UnifiedGenotyper.
- f. Finally, the discovered variants are annotated using snpEff. At the end of this step a table of candidate SNPs will be produced in the VCF format.

As the output, this command will produce the VCF file *SampleName.final.vcf* with detected variants for sample *SampleName*. For multiple samples, run the above command for each sample including the control sample from original non-mutagenized strain.

6. Filter out the background strain variants occurring in the control sample using the R script *filter_variants.R* provided as a part of the pipeline (https://github.com/MolBioBioinformatics/SnpMap_bacteria). Run command:

```
Rscript filter_variants.R sample.final.vcf control.final.vcf
```

where *sample.final.vcf* and *control.final.vcf* are the VCF files with variants detected in mutant samples and control sample, respectively; for example:

```
Rscript filter_variants.R experiment.vcf control.vcf
```

This step produces the VCF file *final_variants.vcf* with genomic variants identified in mutants but not in control.

7. As an alternative to the direct inspection of the VCF file produced at the previous step, use SnpSift to extract the most informative basic attributes of the variants for an easier inspection:

```
java -jar SnpSift.jar extractFields final_variants.vcf -s "," -
e "." CHROM POS REF ALT "ANN[1].EFFECT" "ANN[1].IMPACT"
"ANN[1].GENE" "ANN[1].FEATURE" "ANN[1].BIOTYPE" >final_variants.txt
```

This will create a tab-separated table with columns corresponding to basic attributes of the candidate variants. Column names are described in Table 2.

8. Finally, it is essential to carefully examine each candidate variant in the context of the particular screen and to perform experimental validation of these variants.

COMMENTARY

Background information

Forward genetic analysis in a variety of bacterial species and especially in *Escherichia coli* has provided important insights into many fundamental biological processes. Many of these insights have been gained by the molecular cloning of genes corresponding to mutations that cause a particular phenotype of interest. Until next generation sequencing technologies made full-genome sequencing rapid and affordable, identification of mutant alleles required extensive and time-consuming genetic analysis, for example, by complementing the mutant phenotype using DNA fragments cloned into a suitable plasmid vector. Although this is relatively straight-forward for bacterial species such as *E. coli* for which many recombinant DNA techniques are available, it is tedious or not feasible for many bacterial species. The advent of next generation sequencing technology, however, has allowed identification of candidate causative mutations to be identified in a single experiment that can be performed much more quickly than using conventional genetic approaches. What in the past took at best a few months of effort, a bacterial geneticist can now identify 20 or more different mutations per lane using the Mi-Seq NGS platform within weeks of performing the genetic screen from which mutants corresponding to the mutations emerged.

Critical Parameters and Troubleshooting

Independent mutants—It is critical to only sequence mutants that arise independently. Otherwise it is likely that you will wind up sequencing “sibs”. If sibs are sequenced, many of the sequenced mutants will typically contain the exact same mutations when compared to the reference genome. To avoid sequencing sibs, we typically generate a set of independent cultures derived from single colonies on non-selective media (conditions) corresponding to the number of mutants that will be sequenced. A single mutant from each of these cultures is the verified and sequenced.

Robustness of the phenotypic assay used to identify bacterial mutants—An important factor in the success of this protocol is robust phenotypic identification of the bacterial mutants. Spontaneous mutations in a particular gene with a scorable phenotype typically arise at a frequency of 10^{-6} to 10^{-8} . If the phenotype of interest appears at a higher frequency, it is possible that several target genes can be mutated to give the same phenotype or the assay to identify the mutants is leaky. The higher the frequency at which mutants arise, the more independent clones will need to be sequenced to insure that all of the relevant genes contributing to the phenotype of interest have been identified.

For mutations that occur at a lower frequency, it may become difficult to isolate spontaneous mutants without mutagenesis. For example, we found that when plating *P. aeruginosa* at a titer higher than $\sim 1 \times 10^8$ CFUs per plate in order to identify antibiotic-resistant mutants, a

thin biofilm would form on the surface of the agar leading to dramatic changes in tolerance to antibiotics, thereby making it impossible to isolate resistant mutants. Considering the fact that it is necessary to isolate and sequence multiple mutants, selecting mutants that arise at a frequency of less than 10^{-8} can become prohibitively expensive, especially when using expensive compounds as selective agents.

Knowing when a screen/selection approaches saturation—It can generally be assumed that a genetic screen/selection has reached saturation when multiple phenotypically similar mutant strains contain independent lesions in the same gene. On the other hand, this may not be the case if different target genes exhibit significantly different rates at which a particular mutant phenotype occurs. For example, loss of function mutations may occur at substantially higher rates than gain of function mutations. Thus, if loss of function mutations in gene X and gain of function mutations in gene Y give the same phenotype, the gain of function mutations may be missed if only a limited number of mutants are sequenced.

EMS mutagenesis—Although typical EMS mutagenesis protocols will cause hundreds of mutations per bacterial genome that are not relevant to the phenotype of interest, it should be relatively straightforward to identify causative mutations of interest as long as a sufficient number of mutant clones are sequenced.

Library quality and sequencing coverage—Constructing libraries with a high proportion of fragments that align to the bacterial genome of interest is required to make efficient use of high throughput sequencing approaches. This is usually not an issue with bacterial strains that can be readily cultured *in vitro*. The genomic DNA inserts in the library must be at least twice as long as the read length (assuming paired end reads, see below) to ensure that sequencing capacity is not wasted by sequencing part of each insert twice. If library preparation generates inserts of smaller than the anticipated size, it may be necessary to optimize DNA shearing and/or the size selection steps.

The detection of small-scale genomic variants (SNPs, short insertions, deletions, etc.) can be performed at relatively low sequencing coverage, which allows for larger numbers of multiplexed samples per sequencing run. The average sequencing depth can be estimated as $N_{\text{reads}} \times L_{\text{reads}} / L_{\text{genome}}$, where N_{reads} is the number of sequencing reads in the sample, L_{reads} is the length of a read, or of a read pair in the case of paired-end sequencing (e.g. $L_{\text{reads}} = 300$ for paired-end 150 bp reads), and L_{genome} is the length of the genome. For detecting small-scale genomic variants, a basic recommendation is to achieve 10–30x sequencing coverage. Higher coverage increases the sensitivity of variant detection and is relatively affordable in the case of small bacterial genomes.

Subtraction of background mutations—Due to genetic drift most bacterial strains used in phenotypic screens and selections will contain a large number of polymorphisms when compared to a published reference genome. These variants will be detected when reads from the sample are aligned to the reference genome. A large number of background mutations hampers analysis of candidate causative alleles. To avoid this problem, it is extremely important to include the parental strain used in the mutagenesis screen in the sequencing analysis to identify the variants present in this strain before mutagenesis.

Confirmation of candidate causative alleles—Although the list of candidate lesions identified by this protocol is often short, further analysis may be necessary to definitively correlate the identified gene(s) with a particular phenotype. For many bacterial species, this typically involves the construction of an in-frame deletion mutant in the case of loss of function mutations or complementation of the mutant gene with the wild-type.

Trouble Shooting

The protocol relies on commercial kits and is very robust. However, one potential problem that we encountered is worth noting: Users of NEB kit (E7645) have to choose the correct primer set (*e.g.*, E7335) and mix the primers correctly to successfully amplify the adaptor-ligated fragments (Step 27 in this protocol and Section 4.1 in the NEB online protocol for E7645). The primers are not named in the conventional manner such as a 5' - / 3' - pair or forward / reverse pair. Instead, each pair must be combined with an equal amount of the Universal Primer and one of the Index Primers. In addition, ligation adaptors and USER enzymes are provided in the E7335 kit (and/or other primer sets) but not the E7645 kit. Caution is advised when readers want to mix-and-match reagents from different vendors to ensure acquisition of the complete set of reagents.

Anticipated results

This protocol should generate a list of genes that are disrupted in multiple mutant strains which most likely correspond to the mutant phenotype.

Time considerations

This experimental procedure spans roughly 3 weeks. 1–2 weeks are required for DNA extraction and library preparation. Sequencing usually only takes 1 week, depending on sequencer availability. The computational analysis takes 1–2 days.

Acknowledgments

This work was supported by NIH P30 DK40561. [*Copy Editor: We added this acknowledgment based on the author's submission to S1M. Please ask Author to edit as needed.]

Literature Cited

- Lehrbach NJ, Ji F, Sadreyev S. Next-generation sequencing for identification of EMS-induced mutations in *Caenorhabditis elegans*. *Curr Protoc Mol Biol*. 117(7):29.1–7.29.12. DOI: 10.1002/cpmb.27
- Afgan E, van den Baker D, Beek M, Blankenberg D, Bouvier D, Eeck M, Chilton J, Clements D, Coraor N, Eberhard C, et al. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2016 update. *Nucleic Acids Res*. 2016 gkw343.
- Minevich G, Park DS, Blankenberg D, Poole RJ, Hobert O. CloudMap: A Cloud-Based Pipeline for Analysis of Mutant Genome Sequences. *Genetics*. 2012; 192:1249–1269. [PubMed: 23051646]

Table 1

Approximate numbers of multiplexed samples per MiSeq run for various bacterial genomes.

Species	Genome Size (Kbp)	Approx. number of samples per run for various levels of genome sequencing coverage (depth)			
		10x	25x	50x	100x
<i>Agrobacterium thizogenes</i>	7265	120	48	24	12
<i>Pseudomonas aeruginosa</i>	5900	150	60	30	15
<i>Salmonella enterica serovar Pullorum</i>	4930	180	72	36	18
<i>Escherichia coli</i> K12	4670	190	76	38	19
<i>Enterococcus faecalis</i> H4	3250	275	110	55	28
<i>Staphylococcus aureus</i>	2900	310	124	62	31
<i>Lactobacillus acidophilus</i>	1846	485	194	97	49
<i>Streptococcus thermophilus</i>	1826	490	196	98	49
<i>Helicobacter pylori</i>	1760	510	204	102	51
<i>Mycoplasma genitalium</i>	573	1570	628	314	157

For each bacterial species, genome size is indicated, followed by the approximate numbers of sequencing samples that can be included in a single Illumina MiSeq run for a required sequencing depth (25x, 50x, or 100x). These estimates are based on the assumption of an Illumina MiSeq run producing approximately 30 million paired-end 150 bp reads.

Table 2

Format of the candidate SNP table

Column	Contents
1	Chromosome ID
2	Chromosome position
3	Reference allele
4	Alternate allele
5	Effect
6	Impact
7	Annotation
8	Gene Name
9	Gene Feature
10	Biotype